UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Marília
Faculdade de Filosofia e Ciências
Programa de Pós-Graduação em Ciência da Informação

JANUÁRIO ALBINO NHACUONGUE

O campo da Ciência da Informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno

APOIO

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

CAPES
Coordenação de Aprefeiçoamento de Pessoal de Nível Superior

Marília – SP
2015

JANUÁRIO ALBINO NHACUONGUE

O campo da Ciência da Informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação como um dos requisitos para a obtenção do título de Doutor em Ciência da Informação – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista (UNESP), Campus de Marília.

Área de concentração: Informação, Tecnologia e Conhecimento

Linha de Pesquisa: Informação e Tecnologia

Orientador: Prof. Dr. Edberto Ferneda

Marília - SP
2015

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação - Faculdade de Filosofia e Ciências (FFC/UNESP – Marília)

Nhacuongue, Januário Albino.
N576c O campo da Ciência da Informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno / Januário Albino Nhacuongue. – Marília, 2015.
194 f. ; 30 cm.

Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2015.
Bibliografia: f. 189-194
Orientador: Edberto Ferneda.

1. Ciência da informação. 2. Pós-modernismo. 3. Teoria da informação. 4. Recuperação da informação. 5. Mineração de dados (Computação). 6. Tecnologia da informação. 7. Conhecimento e aprendizagem. I. Título.

CDD 005.73

Januário Albino Nhacuongue

O CAMPO DA CIÊNCIA DA INFORMAÇÃO: CONTRIBUIÇÕES, DESAFIOS E PERSPECTIVAS DA MINERAÇÃO DE DADOS PARA O CONHECIMENTO PÓS-MODERNO

Banca examinadora:

Prof. Dr. Edberto Ferneda (Orientador)
Departamento de Ciência da Informação
Faculdade de Filosofia e Ciências (FFC)
Universidade Estadual Paulista (UNESP) – Campus de Marília

Profª. Dra. Zaira Regina Zefalon (Membro externo)
Departamento de Ciência da Informação (DCI)
Universidade Federal de São Carlos (UFSCar)

Prof. Dr. Guilherme Ataíde Dias (Membro externo)
Centro de Ciências Sociais Aplicadas (CCSA)
Universidade Federal da Paraíba (UFPB)

Profª. Dra. Maria José Vicentini Jorente (Membro interno)
Departamento de Ciência da Informação
Faculdade de Filosofia e Ciências
Universidade Estadual Paulista (UNESP) – Campus de Marília

Prof. Dr. Elvis Fusco (Membro externo)
Departamento da Ciência da Computação
Centro Universitário Eurípedes de Marília (UNIVEM)

Local: Universidade Estadual Paulista – Faculdade de Filosofia e Ciências – Campus de Marília

Data: 17 de Abril de 2015

Marília - SP
2015

O atrativo do conhecimento seria pequeno se no caminho que a ele conduz não houvesse que vencer tanto pudor

Friedrich Nietzsche

Dedico este trabalho às minhas estimadas filhas
Cindy e Janyra
e à minha querida mãe
Ana

AGRADECIMENTOS

À Deus, a fonte inesgotável de todo o poder ilimitado.

Aos meus pais, Albino Lisboa e Ana Gilberto Novele, pela abnegada e incessante dedicação na construção da minha personalidade. Um reconhecimento especial vai à minha mãe, a mulher audaz que lutou afincadamente pela alfabetização dos seus filhos.

Às minhas filhas, Cindy e Janyra, das quais peço perdão e espero que no futuro compreendam o motivo da minha ausência nas respectivas fases de crescimento.

Aos meus irmãos que sempre me apoiaram neste longo e sinuoso caminho acadêmico.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) que financiaram esta pesquisa.

Ao Ministério do Interior de Moçambique, especialmente ao Alberto Ricardo Mondlane, ex-ministro do interior e atual governador da Província de Niassa, pela longa experiência de aprendizado, confiança e incentivo no aprimoramento do conhecimento científico.

Ao Programa de Pós-Graduação em Ciência da Informação da UNESP de Marília que, de forma sublime, aceitou o meu projeto e permitiu a minha contribuição científica em várias atividades de pesquisa dentro e fora da faculdade.

Ao professor Edberto Ferneda, pelo qual tenho profunda admiração pelo caráter, dedicação e simplicidade, demonstradas em todas as etapas da orientação.

Aos professores do Programa de Pós-Graduação em Ciência da Informação da UNESP de Marília, pelas lições obtidas nos processos formais e informais de ensino e aprendizagem.

Às professoras Zaira Regina Zefalon, Maria José Vicentini Jorente e aos restantes membros da banca examinadora, pelas observações sem as quais não seria possível atingir a proporção atual desta pesquisa.

Às professoras Silvana A. B. G. Vidotti, Plácida L. V. A. da Costa Santos e Mariângela S. L. Fujita, pelas quais tenho profunda admiração e total gratidão.

À Secretaria da Pós-Graduação, aos funcionários do Departamento de Ciência da Informação, ao pessoal da Biblioteca e outros profissionais da instituição que me apoiaram com zelo em diversas inquietudes da demanda acadêmica.

Aos amigos e colegas com os quais convivi, aprendi e dividi os vários momentos da academia e da vida social.

Muito obrigado!

LISTA DE ILUSTRAÇÕES

Figuras

Figura 1: Linha de tempo da Web e Internet 44
Figura 2: História da Web Social (1969 – 2012).....52
Figura 3: Nível de conectividade nas redes sociais em Janeiro de 2014, com base em usuários ativos 58
Figura 4: Modelo de coleta de metadados. 103
Figura 5: Exemplo do uso do Padrão Dublin Core 114
Figura 6: Processo de representação e recuperação da informação. 121
Figura 7: Problema da busca por palavras-chave 142
Figura 8: Interface interativa do WEKA. 166
Figura 9: Relações assimétricas entre os principais atores da política de ensino superior no Brasil 169
Figura 10: Mineração de dados com Graph API Explorer 175

Tabelas

Tabela 1: Diferenças esquemáticas entre modernismo e pós-modernismo.....37
Tabela 2: Abordagem básica para a representação da informação 91
Tabela 3: Algumas regras gerais que ajudam na definição do significado dos números usados como tags de campo (MARC 21).....95
Tabela 4: Tipologias e funcionalidades de metadados.....97
Tabela 5: Elementos do padrão Dublin Core, Versão 1.1: Descrição de Referência 113
Tabela 6: Diferentes Tipos de Atributos. 160
Tabela 7: Teorias adotadas e técnicas utilizadas em sete análises estruturais 172
Tabela 8: Permissões de token de acesso e os dados a serem minerados.....178

LISTA DE ABREVIATURAS E SIGLAS

AACR2	Anglo-American Cataloguing Rules
API	Application Programming Interface
ARPANET	Advanced Research Projects Agency Commissions
CERN	Conseil Européen pour la Recherche Nucléaire
CDU	Classificação Decimal Universal
CI	Ciência da Informação
DBD	Divisão de Bibliotecas e Documentação
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
FQL	Facebook Query Language
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
ID	Identifier
KWIC	Keyword in Context
LCSH	Library of Congress Subject Headings
MARC21	Machine Readable Cataloging
MD	Mineração de Dados
NISO	National Information Standards Organization
OAI	Open Archives Initiative
SDK	Software Development Kit
SGML	Standard Generalized Mark-up Language
SRRI	Sistemas de representação e recuperação da Informação
SRI	Sistemas de recuperação da Informação
RI	Recuperação da Informação
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

NHACUONGUE, Januário Albino. **O campo da Ciência da Informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno**. Marília: UNESP, 2015. 194f. Tese (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação da Faculdade de Filosofia e Ciências, UNESP – Marília.

RESUMO

O trabalho faz uma abordagem sobre a gênese do campo da Ciência da Informação (CI) e analisa as principais contribuições e desafios impostos pela tecnologia, no que tange à representação e recuperação da informação. O objeto da pesquisa é a Ciência da Informação e o contexto, por um lado, resulta da revolução das ciências, na dicotomia entre a busca pela essência e o foco nos problemas humanos, em concomitância com a relação entre a ciência e a tecnologia. Por outro, do aumento dos recursos informacionais digitais e da complexidade, tanto dos ambientes de produção, comunicação e uso da informação, como dos modelos de representação. Para tal, usou o método qualitativo de caráter descritivo, cujos procedimentos técnicos foram centrados na pesquisa bibliográfica e documental de materiais relativos às variáveis. A partir do delineamento sobre a origem e desdobramentos da CI enleados à tecnologia e do respectivo objeto (informação), identificou como problema da pesquisa, a intangibilidade de algumas informações da Web Social, no ponto de vista do acesso. Assim, a pesquisa partiu da seguinte pergunta de partida: é possível utilizar-se da Mineração de Dados (MD) como uma forma de garantir a recuperação da informação intangível na Web Social? Por conseguinte, a pesquisa identificou como objetivo geral: propor a mineração de dados como solução para a recuperação da informação intangível em ambientes da Web Social. Assim, o trabalho chegou às seguintes conclusões: com base na noção de campo proposta por Pierre Bordieu, a CI é um campo científico e a sua gênese está aliada aos problemas informacionais humanos e à tecnologia. A maioria das suas abordagens é anterior à explosão informacional no período Pós-Guerra e foram incorporadas a partir de relações interdisciplinares, principalmente, com a Biblioteconomia, Arquivologia, Documentação, Museologia e Ciência da Computação. Porém, quanto à institucionalização como campo científico, a CI consolidou-se no período Pós-Guerra, com as manifestações da pós-modernidade, algumas das quais incidem sobre a indústria cultural e diversificação dos meios de produção e consumo, culminando com a explosão informacional e fragmentação da informação. Neste contexto, ao mesmo tempo em que a CI alcança um patamar alto na construção do conhecimento através dos Sistemas de Representação e Recuperação da Informação (SRRI), a tecnologia impõe novos desafios para a recuperação. Daí que, no âmbito interdisciplinar característico da área, a tese propõe a mineração de dados, não só para a descoberta de conhecimento em grandes volumes de dados, como também para o acesso e tratamento de informações em ambientes da Web Social, com opções de visualização e detalhes de granularidade. Deste modo, a pesquisa pretende alargar a contribuição do campo da CI, congregando perspectivas de análise de diferentes áreas sobre o conhecimento na Web Social. O trabalho teve limitações no teste de algumas aplicações sobre a mineração e, por isso, recomenda mais pesquisas sobre o tema.

Palavras-chave: Ciência da Informação. Conhecimento. Pós-modernidade. Representação e recuperação da informação. Mineração de Dados. Tecnologia da informação.

NHACUONGUE, Januário Albino. **The field of Information Science: contributions, challenges and perspectives of data mining for the post-modern knowledge**. Marília: UNESP, 2015. 194f. Project (Doctorate Degree in Information Science) - Graduate Program in Information Science, College of Philosophy and Sciences, UNESP - Marília.

ABSTRACT

The work is a discussion of the genesis of the field of Information Science (IS) and analyzes the main contributions and challenges posed by technology, regarding the information representation and retrieval. The object of research is the Information Science and the context on the one hand, results of the revolution of the sciences, in the dichotomy between the search for the essence and the focus on human problems, in tandem with the relationship between science and technology. On the other, the increase in digital information resources and the complexity of both production environments, communication and use of information, such as the representation models. To do this, it used the qualitative method of descriptive character, whose technical procedures were focused on bibliographical and documentary research materials related to variables. From the design of the origin and developments of IS ensnared technology and its object (information), identified as the research problem, the intangibility of some information from the Social Web, the point of view of access. Thus, the research came from the following starting question: is it possible to use the Data Mining (DM) as a way to ensure the retrieval of intangible information in the Social Web? Therefore, the survey identified the general objective: propose the data mining as a solution for the retrieval of intangible information in the Social Web environments. Thus, the work reached the following conclusions: based on the notion of field proposed by Pierre Bourdieu, Information Science is a scientific field and its genesis is allied to human problems and informational technology. Most of their approaches is prior to the informational explosion in the postwar period and were incorporated from interdisciplinary relations, especially with the Library, Archival, Documentation, Museology and Computer Science. However, as the institutionalization as a scientific field, IS consolidated in the post-war period, with the manifestations of postmodernity, some of which focus on the cultural industry and diversification of the means of production and consumption, culminating with the information explosion and information fragmentation. In this context, while the IS achieves a high level in the knowledge construction through the Information Representation and Retrieval Systems (IRRS), the technology imposes new challenges for retrieval. Hence, in the characteristic interdisciplinary scope of the area, the thesis proposes data mining, not only for knowledge discovery in large volumes of data, but also to access and process information in the Social Web environments, with viewing options and granularity details. Thus, the research intends to expand the field IS contribution, bringing analytical perspectives of different areas of knowledge in the Social Web. The study had limitations in testing some applications on mining and therefore recommends more research on the topic.

Keywords: Information Science. Knowledge. Postmodernity. Information representation and retrieval. Data Mining. Information Technology.

SUMÁRIO

1. Introdução 14
1.1 Contextualização da pesquisa 15
1.2 Problema da pesquisa 19
1.3 Hipótese 22
1.4 Justificativa 22
1.5 Objetivos 24
1.6 Metodologia 24
1.7 Estrutura do trabalho 25
2. Pós-modernismo e Perspectiva histórica da Web 27
2.1 Visão sobre o capítulo 28
2.2 Pós-modernismo 28
2.3 World Wide Web 38
2.4 História da Web Social 50
3. Ciência da Informação 60
3.1 Visão sobre o Capítulo 61
3.2 Ciência e Tecnologia 61
3.2.1 O fim da Guerra e os espaço da CI 63
3.2.2 O paradigma moderno 66
3.3 A Ciência da Informação como campo científico 67
3.4 Gênese da Ciência da Informação 70
3.4.1 Período pré – Segunda Guerra Mundial 71
3.4.2 Período pós – Segunda Guerra Mundial 72
3.5 O caráter interdisplinar da Ciência da Informação 77
3.6 O Objeto de estudo da Ciência da Informação 79
4. Representação da Informação 84
4.1 Visão sobre o Capítulo 85
4.2 Conceito da representação da informação 85
4.3 Métodos da representação da informação 88
4.4 Metadados 96
4.4.1 Funções dos metadados 98
4.4.2 Processo de coleta de metadados 102
4.4.3 Esquemas de metadados 104

4.5 Vocabulários Controlados 106
4.6 Representação da informação em ambientes digitais 110
4.6.1 Dublin Core (DC) 110
4.6.2 Níveis de interoperabilidade do padrão DC 111
4.6.3 Elementos do padrão DC 112
5. Recuperação da Informação 116
5.1 Visão sobre o Capítulo 117
5.2 Conceito da recuperação da informação 117
5.3 Sistemas de Representação e Recuperação da Informação 120
5.4 História dos Sistemas de Representação e Recuperação da Informação 123
5.5 Pesquisadores pioneiros e suas contribuições na área dos SRRI 127
5.5.1 Mortimer Taube (1910–1965) 127
5.5.2 Hans Peter Luhn (1896–1964) 129
5.5.3 Calvin Northrup Mooers (1919–1994) 131
5.5.4 Gerard Salton (1927 - 1995) 132
5.5.5 Karen Spärck Jones (1935 - 2007) 133
5.6 Recuperação da informação na Web 134
5.6.1 Classificação de relevância usando termos (TF-IDF) e semelhança 135
5.6.2 Relevância usando *hiperlinks* e classificação por popularidade 137
5.6.3 PageRank e outras medidas de popularidade 138
5.6.4 Sinônimos, homônimos e ontologias 140
5.7 Findability 141
5.8 A ambiguidade na Recuperação da Informação 144
5.9 Efeitos da RI: o conhecimento mediado pela tecnologia 146
6. Mineração de Dados da Web Social 153
6.1 Visão sobre o Capítulo 154
6.2 Mineração de dados 154
6.2.1 Considerações sobre os dados 159
6.2.2 Mineração de dados e a Ciência da Informação 162
6.3 Recursos de Mineração de Dados 164
6.4 Análise estrutural das redes sociais 168
6.5 Mineração de dados da Web social (Facebook) 172
7. Considerações finais 181

7.1 Visão sobre o Capítulo 182
7.2 Considerações 182
7.3 Limitações e recomendações 186
Referências 188

1

Introdução

1.1 Contextualização da pesquisa

O progresso técnico e social do conhecimento pode ser contextualizado em analogia com a própria história da humanidade. As observações sobre o universo, a busca do entendimento sobre a essência das coisas e o consequente processo de comunicação subsidiaram experiências em cada estágio das vivências humanas. A comunicação, por sua vez, sempre motivou o armazenamento de informações para a transmissão a outros sujeitos ou gerações futuras.

A necessidade da preservação da informação sempre existiu desde a fase da pré-história, com o surgimento do homem há cerca de 3,5 milhões de anos. Contudo, foi na antiguidade em que teve o seu maior advento, com o surgimento da escrita por volta de 4000 A.C. e a emergência das ciências. A ciência, enquanto conjunto de conhecimentos empíricos, teóricos e práticos adquiridos mediante a observação e a experiência sobre a natureza, constitui o subsídio humano para a construção social. Por outras palavras, enquanto a escrita introduziu novas convenções para simplificar a comunicação, a ciência padronizou e institucionalizou o processo de raciocínio lógico. Um dos padrões da ciência, em especial à ciência moderna, foi o método científico, por meio de regras específicas de produção do conhecimento científico, tanto a partir de observações sobre experiências, como de correlações sobre conhecimentos já existentes.

A ciência moderna, cuja gênese se atribui a Galileu com o método científico e autonomia da pesquisa científica, ganhou maior proporção com o advento da tecnologia, com a invenção da tipografia através da prensa de Gutenberg, em 1436, seguida das inovações do inglês Henry Mill, do italiano Pellegrino Turri, entre outros sucessores. A escrita, por um lado, proporcionou a exteriorização do conhecimento para que fosse compartilhado e agregado aos diferentes contextos geográficos, temporais, políticos, sociais, culturais que circunscrevem a identidade de cada ser cognoscente. Por outro lado, sistematizou o processo de comunicação humana, através de códigos, linguagens e interpretações sintáticas e semânticas que permitem induções ou deduções nas quais o ciclo informacional se encontra enraizado.

Ao lado da escrita, a restruturação dos campos científicos também proporcionou novos cenários de abordagens temáticas sobre o conhecimento. A Revolução Científica ocorreu no período entre os séculos XVI e XVIII e foi

caracterizada principalmente pelo rompimento do vínculo entrelaçado entre a Filosofia e a Ciência no geral, passando esta a consagrar-se como conhecimento estruturado e com viés prático, isto é, integrada aos problemas humanos. No final do século XIX ocorreu outra revolução que quebrou os padrões da ciência moderna, com a emergência de novos campos científicos com o foco na universalidade da dimensão complexa do homem. Tal complexidade das ciências emergentes foi herdada da cultura contemporânea da própria sociedade, através da pós-modernidade que quebrou as grandes narrativas, pelos questionamentos de verdade, razão, universalidade e identidade, bem como pela sofisticação das tecnologias de informação e comunicação.

Francis Bacon foi um dos revolucionários com a sua filosofia do conhecimento como poder, que se contrapunha à filosofia Aristotélica enraizada na essência ou natureza das coisas. Para Bacon, além do silogismo existem outras formas de raciocício lógico, como a experiência e todo o esforço da busca pelo entendimento só se justifica se tiver fins práticos. É neste novo paradigma filosófico do uso que nasceu a Ciência da Informação (CI).

A CI é um dos campos emergentes que se consagrou com a pós-modernidade, na segunda metade do século XX, principalmente com o movimento acelerado das tecnologias de comunicação. O seu desafio foi notabilizado pela busca de soluções tecnológicas de caráter interdisciplinar que permeassem a produção, a organização, o armazenamento, a disseminação e o uso da informação, cada vez mais excessiva, numa autêntica revolução da estrutura do acesso à informação e da construção do conhecimento na época. Por isso, a Ciência da Informação situa-se no paradigma emergente da universalidade e complexidade que caracterizam qualquer ciência pós-moderna, isto é, ela não é concebida como ciência constante, na medida em que o seu objeto traduz a essência da complexidade e diversificação dos problemas informacionais humanos e tecnológicos, em cada estágio da própria sociedade.

Na relação com outros campos científicos em prol do conhecimento, o campo da CI desdobra-se em diversas linhas de atuação que, segundo Le Coadic (2004), atuam sobre o ciclo informacional (construção, uso e comunicação). Neste processo, a recuperação da informação, no sentido amplo e com enfoque social, ganha destaque pelo fato de configurar-se como o fim no rol da maioria das abordagens teórico-conceituais, processuais e metodológicas envolvidas no tratamento,

representação, organização e gestão da informação. É por meio destas abordagens que se materializa o objeto da CI.

O objeto de estudo da CI, segundo Buckland (1991), é a informação como conhecimento, como processo e como coisa. Todavia, para garantir a manipulação, organização, disseminação e recuperação, o campo incide especificamente sobre a informação como coisa ou informação registrada em algum suporte, na medida em que é a única passível de representação pelos sistemas de recuperação da informação. Esta característica da área levanta outros problemas da complexidade, resultantes da dinâmica atual das estruturas de disseminação e compartilhamento de informações. Entre as referidas estruturas, destaca-se a Web social que reveste uma característica antagônica, pois, enquanto por um lado é fechado quanto ao processo de produção e difusão de informações, por outro, é aberto em relação ao modo como essas informações fluem dentro do ambiente. Na Web social, as informações revestem as características de espontaneidade, versatilidade e inconsistência. Além disso, o acesso a elas é condicionado por laços de afinidade ou vínculos entre os sujeitos envolvidos no processo da comunicação. Mesmo sendo de caráter tecnológico, esta situação afeta a CI na medida em que está entrelaçado ao objeto informação no geral.

Saracevic (1992) traçou o perfil da CI a partir da sua natureza interdisciplinar e principalmente com base na sua estreita ligação com a tecnologia da informação e na sua peculiaridade do comprometimento pela responsabilidade social de transmissão de conhecimento, num aspecto que transcende a tecnologia. Esta linha de pensamento coaduna com o entendimento de Le Coadic (2004), ao afirmar que, preocupada em esclarecer o problema social da informação, a CI situa-se no campo das ciências sociais que permitem uma compreensão social e cultural, com recurso à interdisciplinaridade. Assim, mesmo que as informações do Facebook ou Twitter, na sua maioria, sejam singelas, existe a necessidade da sua preservação e recuperação para o usuário, pois a Ciência da Informação é eminentemente social e a sua contribuição é voltada para a sociedade em todas as suas dimensões. A sociedade visada pelos olhares e fazeres da CI é simultanêamente fator e produto do pós-modernismo.

Algumas manifestações do pós-modernismo, de acordo com Harvey (2004), são a efemeridade, a fragmentação, a descontinuidade e a fugidez. Sendo fruto da pós-modernidade, a Web herdou estas características que, na sua maioria, incidem

sobre a informação. Por isso algumas informações da Web são efêmeras, fugidiças ou traduzem alguma sensação de imaterialidade que se reflete na sua intangibilidade. Além disso, a complexidade tecnológica às vezes compromete o seu acesso pelos tradicionais métodos e processos de representação e recuperação em uso na CI. Por isso, sendo a Ciência da Informação uma ciência complexa e dinâmica, torna-se premente a busca de soluções de índole interdisciplinar para se adequar a este novo cenário informacional.

No âmbito da relação interdisciplinar entre a CI e outras áreas de conhecimento, a Mineração de Dados é um processo advindo da Ciência da Computação que permite tanto a descoberta de informações úteis em grandes depósitos de dados, como o acesso e o tratamento de informações difusas e fragmentadas[1]. Assim, a presente pesquisa propõe a mineração de dados na CI, de modo a expandir o universo da recuperação da informação e proporcionar melhor retorno à sociedade, por meio do acesso às informações fragmentadas para indivíduos ou grupos.

A MD envolve diversas tecnologias ou métodos de processamento da informação e a escolha de cada um deles depende da complexidade das tarefas de mineração, dos objetivos a alcançar com a mineração, do grau de especialização dos profissionais no assunto e dos custos envolvidos. No caso específico desta pesquisa, a mineração de dados foi abordada no sentido genérico, sem a profundidade técnica que requer maior infraestrutura para o uso de algorítmos e métodos estatísticos de classificação ou agrupamento de informações.

Em suma, a pesquisa aborda a Ciência da informação como campo científico, mostrando a sua gênese, características, contribuições na recuperação da informação, desafios impostos pela transcedência da tecnologia e perspectivas de caráter científico e social, como a mineração de dados para ampliar o universo da construção do conhecimento. Para alcançar este objetivo, evidenciou-se um olhar qualitativo aplicado de índole descritiva, da relação dos conceitos e variáveis baseados no levantamento do referencial teórico bibliográfico e documental.

O outro aspecto abordado na pesquisa refere-se à complexidade da própria tecnologia. As tecnologias de informação e comunicação, através das quais a CI

[1] Informações quebradas, tanto pelo imperativo subjetivo e ideológico que norteia os modelos de representação individual, como pela complexidade tecnológica que assiste os processos de produção, tratamento, disseminação e uso. São informações contextuais, quebradas pelo modelo do hipertexto e pelas diversas plataformas tecnológicas de representação, tornando difícil a sua coesão.

vem contribuindo para o acesso e o uso da informação para a transformação social, revestem as características de versatilidade e complexidade. A tecnologia é, por si própria, fator de inclusão e exclusão. O aumento das ferramentas da Web que permitem o compartilhamento de informações traduz a inclusão, na medida em que mais usuários interferem sobre esse conhecimento, agregando-lhe valor. Contudo, também traduz a exclusão porque muitas pessoas ainda vivem desprovidas de condições que lhes permitam a tal interferência nos moldes propostos por essas tecnologias ou pela sociedade no geral. Além disso, a complexidade da tecnologia traduz um fator crítico em relação à sua implementação e eficácia no ambiente organizacional. Estes aspectos também se situam na esfera dos problemas atuais da CI, pois esta se desenvolve sob perspectiva tecnológica.

1.2 Problema da pesquisa

A CI contribui significativamente para a transferência da informação, através de processos e métodos que permitem a disseminação e o compartilhamento de informações para a construção de conhecimento, por meio de estruturas cognitivas dos indivíduos situados em todas as áreas do saber. Como campo científico, a CI tem uma natureza interdisciplinar e estreita ligação com as Tecnologias de Informação e Comunicação (TIC).

O objeto de estudo da CI, conforme a tripartição de Buckland (1991) é a informação como processo, como conhecimento e como coisa, mas a informação como coisa ou informação registrada em algum suporte constitui o epicentro da área, pelo fato de apresentar a peculiaridade de ser a representação tangível de códigos, sinais, dados, textos, filmes, etc., usados pelos sistemas de informação, bibliotecas, arquivos. Por isso que como campo científico, conforme Saracevic (1996), a CI centra-se tanto na natureza, manifestações e efeitos da informação e conhecimento, como nos processos da comunicação e uso da informação. Porém, a abordagem destes aspectos na atual fase da explosão informacional traz problemas de forma e contexto em relação à eficácia e eficiência dos processos e métodos apropriados pela área no âmbito de um dos seus objetivos – a recuperação da informação. Por exemplo, a Web “pulveriza” a informação pela onipresença e versatilidade, não só da própria informação, como também dos ambientes informacionais. Deste modo, a informação na Web torna-se ubíqua, mas

simultaneamente fragmentada, desvanecida e, muitas vezes, percebida como imaterial ou intangível.

A Web é a maior coleção de documentos do mundo, por isso, atualmente a recuperação da informação subjaz de técnicas que incidem sobre ela. Contudo, em alguns ambientes da Web as informações fluem constantemente de forma não estruturada[2] e são propensas a mudanças instantâneas. Particularmente, as informações da Web social são caracterizadas pela espontaneidade, versatilidade e inconsistência. Ademais, encontram-se preservadas em estruturas condicionadas que, às vezes, comprometem a localização, o acesso e a manipulação por qualquer usuário e pelos atuais sistemas de representação e recuperação da informação.

Com base nas referidas características da informação em ambientes digitais, o trabalho partiu da análise de dois níveis do conceito de "tangibilidade ou intangibilidade", na proposição feita por Buckland (1991). O primeiro nível referente à materialidade ou imaterialidade da informação, ou seja, à exteriorização da informação em algum suporte e o segundo relativo ao acesso. É sobre o acesso que recai a pesquisa, isto é, sobre a intangibilidade de recursos, mesmo que de certa forma estejam materializados em algum suporte. Por outras palavras, a pesquisa considera que dentro da CI, para que a informação seja considerada tangível, não basta que ela esteja exteriorizada em algum suporte, pois existem outros fatores que contribuem para a tangibilidade e que vão além da exteriorização. Um dos fatores que contribuem para a intangibilidade da informação digital é a complexidade tecnológica que afeta as estruturas de produção, tratamento, organização, disseminação e uso. Outros fatores estão relacionados às variações linguísticas, às ideologias conjunturais, à cultura e às terminologias de áreas científicas. Por isso, enquanto a tangibilidade no nível da exteriorização é absoluta, no nível do acesso e representação ela se torna relativa. Isto é, no nível do acesso varia de acordo com a situação particular de cada sujeito. Daí que a intangibilidade do suporte no qual algumas informações da Web se encontram registradas, compromete a identificação, a localização e, por conseguinte, o acesso, a representação, a recuperação, o uso e a manipulação pelos sistemas adotados na Ciência da Informação.

[2] O termo não estruturado refere-se à ausência de padrões específicos para a estrutura do texto e da linguagem.

O problema da intangibilidade da informação conduziu à seguinte pergunta de partida: é possível utilizar-se da mineração de dados como uma forma de garantir a recuperação da informação intangível na Web Social? Este problema é complexo e exige uma reflexão profunda na área por meio de relações interdisciplinares, neste caso, da relação interdisciplinar com a Ciência da Computação.

No caso da Web social, os usuários cada vez mais recorrem a ela para compartilhar informações nos Blogs, no Twitter, no Facebook, no Linkedin, no Instagram, entre outros ambientes. Para elucidar, segundo dados do Statista Inc.[3], no primeiro trimestre de 2015, o Facebook tinha 1.440 milhões de usuários ativos mensais, dos quais 1.250 milhões usavam dispositivos móveis. Isto significa que cerca de 1,4 bilhões de usuários desta rede social postaram informações de difícil recuperação, já que além da autenticação por senhas, também dependem de vínculos entre os respectivos sujeitos da comunicação. Mesmo que se tratem de informações que na sua maioria não tenham o cunho científico, existem comunidades científicas que usam as redes sociais para disseminar informações importantes para a sociedade. Além disso, a CI é genuinamente social, dinâmica e os seus intentos se desdobram em estratégias que visam à busca de soluções para problemas dos indivíduos, grupos e organizações, independentemente dos respectivos graus de estruturação e importância social.

A presente abordagem mostra que, ao lado das técnicas e processos tradicionais de representação e recuperação da informação, torna-se premente fazer abordagens de caráter investigativo que possam culminar em soluções inovadoras face às características de versatilidade e complexidade da própria tecnologia. Uma das soluções enaltecidas pelo presente estudo consiste em acoplar a mineração de dados aos sistemas de representação e recuperação da informação, de modo a recuperar a informação intangível e agregá-la com maior valor no processo de construção do conhecimento humano.

[3] Statista Inc. é uma das principais empresas de estatísticas sobre a internet que fornece ferramentas para a pesquisa de dados quantitativos, estatísticos e informações relacionadas. Os dados providenciados destinam-se tanto para o uso acadêmico, como para o empresarial.

1.3 Hipótese

Face ao problema anteriormente identificado, a pesquisa partiu da seguinte hipótese da pesquisa: a mineração de dados permite a recuperação de informações intangíveis em ambientes da Web Social.

De referir que a complexidade tecnológica e conceitual do ciclo informacional torna a informação, fragmentada, efêmera, fugida e inconsistente, traduzindo uma sensação de imaterialidade que se reflete na intangibilidade para os processos e métodos adotados na CI, comprometendo deste modo, o acesso, o uso e a produção de novas informações. Assim, o estudo defende a seguinte tese: ao lado dos sistemas de representação e recuperação da informação utilizados na Ciência da Informação, a mineração de dados vai dar condições para a tangibilidade, e ampliar a recuperação de informações em ambientes da Web.

As abordagens teórico-conceituais e práticas sobre a MD constituem a essência da resposta para o problema da pesquisa, na medida em que permitem a identificação e a descrição de processos que garantem a recuperação da informação para a satisfação da multiplicidade das necessidades informacionais do usuário final.

1.4 Justificativa

Os processos e métodos inerentes à CI desempenham um papel preponderante na difusão da informação para a sociedade, permitindo o acesso e uso para a resolução de diversos problemas ou necessidades, isto é, para o conhecimento em ação. Esta ação intervencionista é enaltecida por Saracevic (1996), no que concerne tanto à influência da área no modo como a informação é manipulada na sociedade e pela tecnologia, como na compreensão e zelo dos problemas, processos e estruturas associados ao conhecimento, à informação e ao comportamento humano frente à informação, principalmente pelo enfoque nas práticas profissionais.

O conhecimento pós-moderno está enraizado na tecnologia e esta, por sua vez, incide sobre a informação digital ou sobre ambientes informacionais digitais. A tendência crescente da informação digital é demonstrada por Gill (2008) que considera a Web como a maior coleção de documentos do mundo. Por exemplo: pesquisas de julho de 2007 efetuadas pela Netcraft sobre o protocolo da transmissão de dados HTTP, indicavam 125.626.329 pedidos de nomes de

servidores de sites, um aumento em 162780% de pedidos, comparativamente ao mês de janeiro de 1996 em que a mesma pesquisa indicava apenas 77.128 pedidos. A pesquisa apenas se refere à "Web visível", pois na "Web invisível" existe uma quantidade maior de conteúdos da Web servidos a partir de bancos de dados em resposta a uma entrada de usuário ou que requerem algum tipo de autenticação de usuário ou login, tornando-se não localizáveis para os motores de busca. Muitos conteúdos da "Web invisível" situam-se fora da esfera de domínio público, tanto por se tratarem de páginas privadas, como pela exigibilidade de taxas ou assinaturas pagas para acessá-los.

Algumas redes sociais configuram-se na Web invisível, pois no geral, os recursos de informação disponibilizados nesses espaços dependem da autenticação de usuários e do vínculo com os produtores ou disseminadores da informação, embora existam alguns aplicativos usados por alguns profissionais para o acesso às informações nesses ambientes.

O crescimento da informação digital também é secundado por Stewart (2008), ao considerar que anualmente o mundo produz cerca de cinco exabytes (10^{18} bytes) de informação nova, sendo 92% armazenada em meios digitais e diariamente cerca de 31 bilhões de e-mails são enviados. Todavia, apenas 15% do total de informações são organizadas e estruturadas em esquemas definidos pelos sistemas de informação e podem ser facilmente localizadas, manipuladas e recuperadas. Os restantes 85% são constituídos por conteúdos não estruturados, tornando difícil a sua recuperação. Além disso, mesmo com a indexação automática, fontes de informação eletrônicas, uso de palavras-chave ou assuntos, os motores de busca geralmente apresentam listas de páginas que contêm apenas os termos usados na busca, sejam ou não usados na forma e no contexto pretendido pelo usuário.

A riqueza do conhecimento humano, proporcionado pela quantidade de informações é limitada pelo fenômeno de "infoglut" que, segundo o Business Dictionary (2012)[4], significa excesso de informação ou massas de informação em contínuo crescimento, tão mal catalogadas ou organizadas (ou não organizadas em todo) que é quase impossível navegar por elas para pesquisar ou tirar qualquer conclusão ou significado. Esta situação levanta a questão sobre o modo pelo qual

[4] "Infoglut." Business Dictionary, WebFinance, Inc., 2012.

construímos o nosso conhecimento, isto é, até que ponto o nosso pensamento ou conhecimento é influenciado pelos motores da busca na Web?

A imprecisão e ambiguidade que ainda caracterizam os sistemas de recuperação da informação prova a necessidade do aprofundamento de pesquisas na CI. A mineração de dados vai auxiliar a recuperação de informações que atualmente não se configuram nas técnicas e nos processos dos sistemas de representação e recuperação da informação. Deste modo, a CI pode expandir o seu comprometimento social de mediação, proporcionando recursos para o conhecimento em ação.

1.5 Objetivos

O objetivo geral da pesquisa é: propor a mineração de dados como solução para a recuperação da informação intangível em ambientes da Web Social. Deste modo, objetivou-se especificamente:

- Contextualizar a gênese e os problemas que a CI se predispôs a resolver através do objeto “informação”;
- Analisar os sistemas de representação e recuperação da informação dos pontos de vista histórico, conceitual e funcional;
- Apresentar uma perspectiva histórica da Web para estudar a origem e os desdobramentos da sua complexidade;
- Delinear a fragmentação da informação nos ambientes informacionais digitais colaborativos.

1.6 Metodologia

No trabalho fez-se o levantamento do referencial teórico sobre a CI, enquanto campo científico e os problemas que se predispõe a resolver através dos seus métodos científicos, para demonstrar a importância dos sistemas de representação e recuperação da informação e da mineração de dados, como mecanismos direcionais ao processo de construção do conhecimento na sociedade. Para tal, a pesquisa centrou-se numa abordagem qualitativa que segundo Lakatos (2001), visa interpretar e compreender os diferentes fenômenos, dados e ação dentro da organização em estudo, sua relação com a teoria e contexto, de modo a dar um enfoque descritivo do significado fenomenológico na procedência.

Do ponto de vista da sua natureza, Gil (1994) aponta que a pesquisa pode ser básica ou aplicada. A básica incide sobre conhecimentos novos, sem foco na aplicação prática; enquanto que a aplicada visa conhecimentos dirigidos à satisfação de necessidades específicas. Deste modo, a presente pesquisa é aplicada, pois visa à satisfação de problemas sobre a recuperação da informação em ambientes da Web.

No concernente ao procedimento técnico, fez-se a revisão bibliográfica e a análise documental sobre materiais relativos às principais variáveis do problema.

O objeto de estudo é a Ciência da Informação, no seu aspecto de complexidade propiciado pela pós-modernidade, no enfoque tecnológico e os respectivos desdobramentos enleados à característica da informação em ambientes informacionais digitais colaborativos.

1.7 Estrutura do trabalho

Além das páginas preliminares, o trabalho encontra-se organizado em sete capítulos. O primeiro Capítulo refere-se à introdução, no qual se faz a contextualização do tema da pesquisa a partir da gênese da CI, mostrando-se o enlace com a tecnologia que condicionou o seu desenvolvimento em relação ao maior contributo científico e social da recuperação da informação. Também se levanta a problemática enleada às especificidades tecnológicas que assistem ao objeto de estudo da área da CI, em diferentes estágios do seu desenvolvimento, que culminam em limitações de caráter técnico e metodológico no que tange ao tratamento de informações para a disseminação, acesso e uso. Assim, traça-se a hipótese sobre a MD de modo a permitir a recuperação de informações intangíveis em ambientes digitais da Web Social. De igual modo, mostra-se a pertinência da abordagem deste assunto, através da justificativa baseada na tendência crescente da informação digital e das ferramentas tecnológicas de compartilhamento, majoritariamente destituídas de esquemas formais de organização, o que compromete a sua recuperação para o usuário. Finalmente, levantam-se os objetivos da pesquisa e a metodologia usada para sua satisfação e respostas à questão principal da pesquisa.

O segundo Capítulo aborda os aspectos complexos que configuram os desafios da CI, maximizados pela pós-modernidade enquanto corrente que traduz as

mudanças de pensamento científico, em certa medida influenciado pelo advento da tecnologia. Esta abordagem se baseia nos vários questionamentos que incidem sobre a fase da pós-modernidade, mormente os aspectos que vão desde a crítica à tecnologia até o critério de verdade do conhecimento. Também se mostra a influencia da pós-modernidade na concepção da Web.

O terceiro Capítulo é relativo à CI. Neste capítulo aborda-se a CI como campo científico de conhecimento a partir da noção de campo proposta por Pierre Bordieu, a sua gênese e peculiaridades que estiveram na sua origem, o seu caráter interdisciplinar, o seu objeto de estudo (informação) e sua relação com o conhecimento.

Os Capítulos quatro e cinco debruçam sobre algumas contribuições da CI no conhecimento humano, enaltecendo os aspectos da representação e recuperação da informação, que se figuram no epicentro das suas necessidades enquanto campo científico inovador no conhecimento humano. Assim, abordam-se os sistemas de representação e recuperação da informação numa perspectiva histórica, destacando alguns pesquisadores ou pesquisas que contribuíram significativamente para a revolução da área, bem como alguns métodos e processos de representação e recuperação da informação.

O sexto Capítulo analisa as perspectivas da CI face às limitações impostas pelo período atual da versatilidade e complexidade da tecnologia, aliadas ao crescente volume de informações disponibilizadas na Web. Neste contexto, analisa-se o processo de implementação da mineração de dados, principalmente a mineração da Web Social como alternativa para estender a área da recuperação da informação e, por conseguinte, reforçar a dimensão social e humana na construção do conhecimento.

Finalmente, o sétimo Capítulo resume as considerações finais destacadas durante a abordagem dos diversos aspectos do tema, colocando a necessidade de mais aprofundamentos teóricos sobre o assunto para o enriquecimento do campo da Ciência da Informação.

2

Pós-modernismo e Perspectiva histórica da Web

2.1 Visão sobre o capítulo

Na presente seção analisa-se a gênese e as manifestações da pós-modernidade, de modo a elucidar o contexto da idealização da Web e a complexidade enleada ao atual processo de construção de conhecimento.

Num panorama de caráter histórico, o capítulo também debruça sobre a World Wide Web que atualmente se configura no epicentro de todas as dimensões do universo informacional, para mostrar os principais eventos e ações que contribuíram para a sua arquitetura. Com estes objetivos pretende-se identificar os espaços e o papel da CI na mediação, para potencializar o acesso e uso da informação em embientes colaborativos, com o enfoque social. De igual modo, objetiva-se traçar vínculos que demonstrem o enfoque da CI na Web e na conturbada complexidade humana de representação e comunicação.

2.2 Pós-modernismo

Pós-modernismo ou pós-modernidade são expressões que denotam uma ambiguidade em relação à sua conceituação e origem. Se, para alguns autores, a era pós Segunda Guerra Mundial, que vai desde 1945, já seria marcada pelas manifestações do pós-modernismo; para outros, essa fase ainda seria uma extensão do Modernismo.

A melhor diferenciação[5] entre pós-modernismo e pós-modernidade pode ser encontrada em Eagleton (1996, p.3): “a palavra pós-modernismo refere-se em geral a uma forma de cultura contemporânea, enquanto o termo pós-modernidade alude a um período histórico específico”. Para o autor, no pós-modernismo predomina uma corrente de pensamento que propõe a queda de paradigmas do iluminismo, pelos questionamentos dos critérios de verdade e das noções de razão, identidade e objetividade, progresso ou emancipação universal, sistemas únicos, bem como pela perda das grandes narrativas ou fundamentos definitivos de explicação. Em contrapartida, segundo Eagleton (1996, p.3), “vê o mundo como contingente, gratuito, diverso, instável e imprevisível”, dominado por um conjunto de culturas ou interpretações desunificadas que geram certo grau de ceticismo em relação à

[5] Mesmo considerando a diferença entre pós-modernismo e pós-modernidade, pela relação entre os dois termos, Eagleton (1996) apenas assume o termo pós-modernismo para tecer críticas sob o aspecto político e teórico, numa visão socialista e marxista.

objetividade da verdade, da história e das normas. Um dos reflexos do pensamento pós-modernista verifica-se nos constantes questionamentos ou críticas à ideologia emancipadora da cultura ocidental e aos fundamentos ortodoxos do cristianismo.

A queda de paradigmas do iluminismo, segundo Eagleton (1996), foi fortemente marcada pela mudança da forma do capitalismo prevalecente no ocidente que passou a ser transitório e descentralizado, essencialmente marcada pela tecnologia, consumismo e pela indústria cultural. Na nova forma de capitalismo, os meios de produção foram dominados pelas indústrias de serviços, finanças e informação e a política de classes foi dominada pelas minorias identitárias. Assim, pós-modernismo é este novo estilo de cultura cujas manifestações vão desde a arte superficial, descentrada, infundada, auto-reflexiva, divertida, eclética e pluralista, que mistura a elite e o popular, até a experiência cotidiana.

A gênese do pós-modernismo é associada ao fracasso da ideologia assente na autonomia da razão e consolidada pela revolução industrial e pelo capitalismo. Este movimento de ruptura da radicalidade moderna caracterizou-se por uma transversalidade que atravessou todas as dimensões sociais, desde os pensamentos filosóficos até à cultura ocidental e religião. Eagleton (1996) cita alguns casos de quebra de paradigmas, protagonizados principalmente por minorias desacreditadas pelo sistema dominante, como é o caso do Congresso Nacional Africano[6]. Tais minorias eram principalmente constituídas por jovens que denunciavam a opressão através de manifestações populares de massa e de crítica ao poder, por exemplo, na linguagem e na sexualidade, cultivando o lema da "liberdade ou libertação".

O culto da liberdade que centraliza a tônica do discurso autonomizante e anárquico foi extensivo ao campo do conhecimento científico. O pensamento enraizado na busca pelo entendimento do universo, reservado às elites filosóficas, foi contrastado pela proliferação de diversos cursos de formação técnica e profissional nas universidades e em outras instituições acadêmicas, concebidas como alternativas para a autoafirmação na sociedade.

No contexto da revolução tecnológica e explosão informacional nasceu a World Wide Web. A legitimidade da Web justifica-se pela capacidade de hibridização

[6] Movimento fundado em 1912, na África do Sul para a proteção dos direitos da população negra do país, contra a o regime de segregação racial conhecido por apartheid e que teve Nelson Mandela como principal ativista político.

de linguagens nas quais os recursos informacionais se apresentam. Porém, não se pode ignorar a sua ideologia revolucionária sobre os critérios de verdade, razão e método na produção de conhecimento. As manifestações atuais no compartilhamento e produção de informações na Web social traduzem a ideologia da busca pela igualdade e liberdade de pensamento. As redes sociais são espaços de compartilhamento de informações e construção de conhecimento, desprovidos da tradicional estrutura de regência de conteúdo que dava primazia apenas aos grandes e renomados cientistas. Contudo, há aqui um dilema de forma e essência que importa referenciar: se o conhecimento que se constrói refuta a noção de verdade e lógica, então ele se torna afora dos modelos de representação e, por isso, incapaz de ser compreendido. Este espécime se circunscreve nas próprias manifestações do pós-modernismo, pois a maioria está assente em premissas contraditórias em si mesmas.

Harvey (2004) critica a contrariedade pós-modernista a partir do espaço, pois este, enquanto torna libertador a identidade humana, também propicia uma vulnerabilidade à psicose e ao pesadelo totalitário. Neste contexto, as tecnologias de informação e comunicação ganham destaque naquilo que o autor denomina plasticidade da personalidade humana, por meio da diversidade dos modelos de produção e representação do conhecimento, de tal forma que seja altamente permeabilizadora, no sentido positivo, ou conflitante, no sentido negativo. Além disso, também condicionam a “destruição criativa”, representando verdades eternas por meio da destruição das verdades já existentes. Para Jameson (2004), são esses condicionalismos que propiciam a proliferação de teorias do fragmentário, que acabam duplicando a alienação e reificação do presente. É sobre a maioria destas questões que se deduz a crise do pensamento pós-moderno.

Na perspectiva de Eagleton (1996), a crise no pensamento resulta na crise do próprio sistema em que o pensamento esteve assente. No entendimento de Jameson (2004, p.6), centra-se “em apreender e expor o conteúdo total, procurando dar conta das contradições do presente e evitar, ao máximo, as armadilhas da ideologia”. Por isso que a busca incessante pela verdade tornou-se uma atividade paranormal de contemplação do subversivo, no qual o entendimento do irracional consubstancia uma alternativa. Ao desprender-se das totalidades, o pensamento pós-modernista torna-se incoerente, inconsistente e tende a desvanescer circunstancialmente.

Na pós-modernidade, a razão e a verdade não são absolutas, mas sim contextuais e legitimadas pela diversificação dos modelos de apropriação e representação de cada ser cognoscente. Neste cenário, tudo vale pelo que é levado a ser e não pelo que é. Enquanto o homem constrói o conhecimento, o conhecimento manipula a realidade no qual o homem se constrói. Isso acontece numa esfera de dominação e completude, de tal forma que se dispensam quaisquer tentativas de crítica ou autocrítica. Esta é uma das características pós-modernistas que nasceu na modernidade, como subjetivismo radical e individualismo desenfreado. Harvey (2004) justifica que as duas características nasceram com a exploração da estética como domínio cognitivo no século XVIII, que suscitava múltiplas facetas de interpretação à imensa variedade de artefatos culturais produzidos sob diferentes condições sociais. Assim, a interação consubstanciava uma atividade eminentimente subjetiva de criação de uma verdade, mediante o juízo estético do artista individual. Esse subjetivismo também assistia ao processo de produção da arte e mostrava como a nossa realidade poderia ser construída e reconstruída através da estética, por tanto, uma representação da realidade passível de ambiguidades, contrastes e mudanças. Por isso que o subjetivismo e o individualismo traduzem parte da essência do pós-modernismo.

Para Santos (2004), o pós-modernismo é o nome aplicado às mudanças ocorridas nas ciências, nas artes e nas sociedades avançadas desde 1950, quando, por convenção, se encerra o modernismo (1900-1950). O autor acrescenta que esta fase nasce com a arquitetura e a computação e a sua ação foi notabilizada na filosofia nos anos 70, com a crítica da cultura ocidental. Na atualidade, as suas manifestações incidem sobre a cultura, através da moda, do cinema, da música e da tecnologia. Harvey (2004) considera que a tônica dessas mudanças foi a partir de 1970, principalmente com a remodelação da noção de tempo e espaço centrado. Assim, Harvey cita o livro *“soft city”* de Jonathan Raban, publicado em 1974 que trouxe à tona muitas das características pós-modernas, espelhadas na vida urbana de Londres, como: individualismo subjetivo baseado no acúmulo do capital ou posse, aparência, racionalidade conflitante e esquema de organização social de múltiplas possibilidades.

Harvey (2004, p.19) fundamenta que o pós-moderno é uma reação à monotonia da visão de mundo do modernismo universal:

> geralmente percebido como positivista, tecnocêntrico e racionalista, o modernismo universal tem sido identificado com a crença no progresso linear, nas verdades absolutas, no planejamento racional de ordens sociais ideiais, e com a padronização do conhecimento e da produção. O pós-moderno, em contraste, privelegia a heterogeneidade e a diferença como forças libertadoras na redifinição do discurso cultural. A fragmentação, a indeterminação e a intensa desconfiança de todos os discursos universais ou totalizantes são o marco do pensamento pós-moderno.

O pós-modernismo refuta as metanarrativas ou interpretações que visem universalizar quaisquer ideias. O problema é que, conforme Santos (2004) acrescenta, o pós-modernismo atua no plano cotidiano, através da tecnologia de massa e individual que visa à proliferação de informações, estímulo à diversão e ao consumo pelo prazer, culminando com ideologias niilistas, vazias ou desprovidas de valores e de sentido para a maioria dos sistemas consagrados na sociedade. Esta situação, segundo Harvey (2004), é determinada pelo pós-modernismo, através de padrões de debate que definem o modo do discurso e estebelecem parâmetros para a crítica cultural, política e intelectual.

O presenteísmo e o simulacro, cultivados através do entretenimento, subsidiam a busca incessante pela perfeição ou identidade na estética ou até no conhecimento que acabam por produzir seres extravagantes. Este fato é comprovado pela história atual do compartilhamento constante de informações e respectiva emissão de juízos de valor sobre quaisquer tipos de informações, num cenário que desfalece a figura do especialista.

O entretenimento proporcionado pela imagem, por sua vez, é uma simulação perfeita que, nos termos de Quéau (1999), modifica a relação do sujeito com o real, eliminando os limites de atuação entre o real e o virtual. Esta situação suscita novas formas de interação e apropriação na relação entre sujeitos e destes com o objeto. Assim, a tecnologia impera sobre os sentidos, provocando novas sensações e, por conseguinte, novos conceitos e novas formas de representação.

Santos (2004) aponta para uma digitalização do social que caracteriza o pós-modernismo pela funcionalidade imposta pela velocidade e instantaneidade. Por outras palavras, tudo é feito para servir o momento, mas um momento que só existe no ego de cada sujeito, pela ausência de um referencial material ou existencial. Nota-se uma ausência da substância no sujeito devido à extrema diferenciação na moda, numa total cultura daquilo que o autor denomina personalização pela aparência e narcisismo que leva à extravagância. Neste cenário, a televisão e a

indústria de publicidade encontram o espaço hiper-real, espetacular, que excita e alegra. Nada mais vale pelo que é, mas sim pelo que é levado a ser; por exemplo, o valor e a necessidade de um produto são determinados pela capacidade de levar à felicidade, como se esta consubstanciasse um valor metricamente calculável.

No conhecimento, o pós-modernismo desfalece o critério da verdade para a validação do pensamento. A realidade é criada com base em conceitos individuais internalizados de acordo com a formação e cultura. A verdade perde os caráteres de objetividade e universalidade e ganha as características de contextualidade e relatividade. O problema da comunicação, da semântica e da representação está relacionado ao que Lyotard (1988, p.16) denomina por “jogos de linguagem”, por meio dos quais os sujeitos criam relações com objetos e com outros sujeitos. Isso justifica as constantes invenções das palavras ou dos discursos nos quais se encontra embasada a diversidade dos modelos de representação de cada indivíduo.

A Web está assente na maioria das manifestações pós-modernistas, fortemente influenciadas pelo advento da tecnologia. A instantaneidade, a ubiquidade, o presenteísmo e outras características que circundam este espaço informacional modificam o próprio sujeito ou usuário, através de pensamentos circunstâncializados, busca incessante pela autoafirmação que, às vezes, culmina com comportamentos niilistas e narcisistas. Um dos desafios da CI no que tange ao processo de construção do conhecimento está assente neste paradoxo, numa tentativa de padronização das estruturas de produção, representação e disseminação da informação, de modo que seja compreendida em diferentes contextos de interação e apropriação. Trata-se do paradoxo, pois enquanto a CI luta pela padronização dos processos envolvidos no ciclo informacional para garantir a uniformidade na representação e acesso, os pressupostos tecnológicos provocam questionamentos e revoluções às estruturas e padrões.

2.2.1 Do modernismo ao pós-modernismo

A maioria das características descritas sobre a pós-modernidade é oriunda do modernismo e só ganhou proporções retumbantes com a tecnologia e com o capitalismo transcendente. A modernidade, segundo Baudelaire (1863 apud HARVEY, 2004, p.21), “é o transitório, o fugidio, o contigente; é uma metade da arte, sendo outra metade o eterno e o imutável”, ou seja, o modernismo reveste uma característica dual de pouca duração e constante desvanecimento, ao mesmo tempo

eterno e imutável. Tanto na cultura como no conhecimento, o modernismo foi caracterizado pela fragmentação, efemeridade e mudanças caóticas que hoje assolam o pós-modernismo.

Algumas das manifestações modernistas incidiram sobre a ciência, com a emergência de novos campos científicos, disciplinas ou linhas que passaram a congregar novos métodos de interação com objetos, originando novas teorias, correntes de pensamento e críticas às narrativas já consagradas. Esta liberdade de criação e subjetivismo nos quais a Web se corporifica suscita vários questionamentos sobre o alcance equitativo e eficácia na sociedade. Conforme Harvey (2004), a transitoriedade e a fragmentação provocam descontinuidades ou rupturas com qualquer condição histórica precedente. Por isso que atualmente há muita informação difícil de ser compreendida e, consequentemente, pouca crítica devido à ausência de referenciais teóricos-conceituais-históricos. A fragmentação também conduz à incoerência e à inconsistência, na medida em que nenhuma verdade é absoluta, isto é, todas as premissas estão propensas a mudanças circunstanciais no tempo e no espaço.

O pensamento iluminista que advogava a igualdade, liberdade, fé na inteligência humana e razão universal foi contrastado pelas constantes indagações aos conceitos de universalidade e igualdade. Um exemplo elucidativo pode ser observado a partir das críticas à cultura ocidental, mormente na política, economia e religião. Como Harvey (2004) fundamenta, as duas grandes Guerras mundiais foram motivadas pela revolta à lógica emancipadora subjacente da racionalidade iluminista de dominação e opressão, maquilhada pelo senso comum sobre a natureza ou ciência. O problema é que o iluminismo determinava os modos de produção da ciência, estabelecendo seres com razão superior e guardiões do conhecimento, juízes éticos ou verdadeiros cientistas, a quem cabia especificar as condições para o exercício da razão. O iluminismo ilustrava uma concepção de sabedoria elitista, em relação a qual o pós-modernismo desafia pela atual e prevalecente ideologia do culto de individulidade e liberdade.

Lyotard (1988) chama atenção para o antagonismo pós-modernista, questionando o conceito de legitimação. O problema é que, enquanto o pós-modernismo refuta a ideologia sobre emancipação universal ou sociedade sem classes que legitimavam o modernismo, associa a informação e o conhecimento aos

meios de produção, isto é, aceita a informação como mercadoria e lucro e, por isso, como novo instrumento de dominação e controle capitalista.

É sobre o modernismo influenciado pelo iluminismo que se constrói a maioria das oposições pós-modernas. Segundo Harvey (2004, p.35), tal acontece porque

> o iluminismo considerava axiomática a existência de uma única resposta possível a qualquer pergunta [...], o mundo poderia ser controlado e organizado de modo racional se ao menos se pudesse apreendê-lo e representá-lo de maneira correta. Mas isso presumia a existência de um único modo correto de representação que, caso pudesse ser descoberto [...], forneceria os meios para os fins iluministas.

A ideologia iluminista teve o colapso depois de 1848, com o foco em outros sistemas de representação. Harvey (2004) cita o exemplo da mudança na linguística, proposta pela teoria estruturalista da linguagem de Saussure que preconizava que o sentido das palavras é determinado antes pela sua relação com outras palavras do que pela sua referência a objetos. Outras mudanças foram influenciadas pelo movimento socialista de luta de classes face às disparidades baseadas no acúmulo do capital que contestavam qualquer forma de hierarquia, consumismo e burguesia. Alguns destes movimentos provocaram uma anarquia, desordem e desrespeito que o modernismo viria a enfrentar através de novas formas de compreensão e exploração de múltiplas perspectivas. Assim, nasceu o perspectivismo e o relativismo modernista que procurava representar a complexidade envolvida. Este cenário foi herdado pelo pós-modernismo e atualmente caracteriza boa parte das relações entre os sujeitos e destes com o objeto.

Conforme Harvey (2004) aponta, mesmo com as mudanças supracitadas, o iluminismo voltou a ganhar a hegemonia com o modernismo universal no período pós-guerra, principalmente com a ideologia do progresso e emancipação humana, atrelada ao capitalismo corporativo e aos centros do poder. Esta fase do iluminismo foi contrastada pela despolitização e retorno ao culto do individualismo exacerbado na arte e no conhecimento, centrados no liberalismo.

Os movimentos sociais antimodernistas deram lugar ao pós-modernismo. Por isso que o pós-modernismo consagrou-se como movimento de transformação cultural emergente nas sociedades ocidentais, centrado não na mudança total de paradigmas no contexto cultural, social e econônico, mas nas políticas e formações discursivas. Esta transformação, segundo Harvey (2004), teve início em 1968 com o

movimento de resistência ao modernismo universal, aliado ao capitalismo liberal e ao imperialismo. A grande mudança ideológica deste movimento foi o foco nas pessoas e não nos ideais abstratos teóricos e doutrinários do Homem, em diversos contextos sociais. Foi neste período em que Engelbart (2003) vislumbrava novas abordagens sobre a inovação descontínua e computação interativa, de modo a tornar colaborativa, a construção do conhecimento. É na mudança de pensamento sobre o processo de acesso e uso da informação e da tecnologia para aumentar a capacidade humana que nasceu a Web. De salientar que esta mudança também foi fortemente impulsionada pelo capitalismo ou por interesses corporativos.

A mudança pós-modernista, alega McHale (1987 apud HARVEY, 2004), foi uma passagem de um dominante "epistemológico" a um "ontológico", ou seja, de um modelo de relação complexa com objetos segundo a perspectiva individual, para uma coexistência, colisão e interpenetração de realidades diferentes. Deste modo, a liberdade do pensamento passou a permitir inferências sobre outros modelos de análise e instrumentos de recolha e processamento de dados. Não se pretende com isto exaltar soberbamente as manifestações pós-modernistas, na medida em que também refletem incoerências de qualquer movimento social. A melhor sustentação sob este olhar é encontrada em Harvey (2004), ao afirmar que quando os sentimentos modernistas foram solapados, desconstruídos, separados ou ultrapassados, passou a reinar pouca certeza quanto à coerência e significado do pensamento pós-modernista que torna difícil avaliar, interpretar e explicar a maioria das mudanças. O problema é que o pós-modernismo é um movimento cultural de difícil explicação e a maioria das suas manifestações tem natureza ou está ligada ao modernismo. Daí que Harvey (2004) até aventa a hipótese de o pós-modernismo ser apenas nova fase do modernismo em que ideias latentes como fragmentação, efemeridade, descontinuidade e mudanças caóticas no pensamento[7], tornaram-se explícitas e dominantes. Esta afirmação é desenvolvida no âmbito do olhar crítico face à representação das diferenças entre o modernismo e pós-modernismo, ilustradas na tabela 1.

[7] Ao contrário do modernismo, o pós-modernismo faz correlações de fatos para produzir ideias sem se opor a outras ideias ou definir conceitos que traduzam a universalidade contextual. Para o pós-modernismo, há situações, fatos e verdades que podem acontecer de forma caótica ou que não caberiam em premissas ou afirmações assumptivas universais.

Tabela 1: *Diferenças esquemáticas entre modernismo e pós-modernismo*

Modernismo	**Pós-modernismo**
Romantismo/simbolismo	Parafísica/dadaísmo
Forma (conjuntiva, fechada)	Antiforma (disjuntiva, aberta)
Propósito	Jogo
Projeto	Acaso
Hierarquia	Anarquia
Domínio/logos	Exaustão/silêncio
Objeto de arte/obra acabada	Processo/*performance/happening*
Distância	Participação
Criação/totalização/síntese	Descriação/desconstrução/antítese
Presença	Ausência
Centração	Dispersão
Gênero/fronteira	Texto/intertexto
Semântica	Retórica
Paradigma	Sintagma
Hipotaxe	Parataxe
Metáfora	Metonímia
Seleção	Combinação
Raiz/profundidade	Rizoma/superfície
Interpretação/leitura	Contra a interpretação/desleitura
Significado	Significante
Lisible (legível)	*Scriptible* (escrevível)
Narrativa/*grande histoire*	Antinarrativa/*petite histoire*
Código mestre	Idioleto
Sintoma	Desejo
Tipo	Mutante
Genital/fálico	Polimorfo/andrógino
Paranóia	Esquizofrenia
Origem/causa	Diferença-diferença/vestígio
Deus pai	Espírito santo
Metafísica	Ironia
Determinação	Indeterminação
Transcendência	Imanência

Fonte: *Hassan (1985, apud HARVEY, 2004, p.48).*

Sobre a tabela 1, Harvey (2004) alerta sobre o perigo da simplicidade expressada na diferenciação de relações complexas que até envolvem vários campos científicos, como Linguística, Antropologia, Filosofia, Retórica, Ciência Política e Teologia. Para explicar a situação, o autor cita a abordagem de Lyotard sobre a linguagem, ou seja, sobre o número indeterminado de jogos de linguagem sobre o qual recaem as intersecções de cada sujeito na relação com outros sujeitos ou com o objeto. Tais intersecções dependem da situação particular de cada sujeito (tempo, espaço, estrutura mental, estratégia, etc.). Por isso que o conhecimento é a principal força de produção e o principal problema proporcionado pela atual explosão informacional consiste na definição desse poder disperso em vários elementos narrativos numa heterogeneidade de jogos de linguagem. Este é um dos problemas da complexidade humana que afeta os SRRI e a Ciência da Informação, no geral. A a descrição, a indexação, a busca e o uso da informação são atividades intrínsecas por excelência e nenhum esquema de representação humana pode traduzir a sua plenitude.

A diversidade dos modelos de representação, interação e apropriação produz o que Harvey (2004, p.51) denomina “determinismos locais”, isto é, indivíduos ou grupos com certo tipo de conhecimento e inseridos num contexto particular que, em função dessas condições, são levados a controlar e a considerar o que é válido

como conhecimento. Por outras palavras, o pós-modernismo cultiva a compreensão da diferença, da alteridade e da liberdade, isto é, da heteropia[8]. No entendimento de Jameson (2004), estas diversidades impõem a necessidade de investigação, compreensão e enfretamento para o olhar crítico cultural no estágio atual da história.

A maioria das características da pós-modernidade anteriormente referidas refletem-se sobre os atuais ambientes de produção e disseminação da informação. A Web situa-se no epicentro desses ambientes e, por isso, reveste a problemática da complexidade cuja Ciência da Informação e outros campos científicos têm de lidar.

2.3 World Wide Web

Berners-Lee (1996) considera a World Wide Web como o universo interativo de informações compartilhadas, tanto para a comunicação humana, como para a comunicação com a máquina. Neste contexto, a revolução da Web foi propiciada pelo interesse comercial entre pequenos grupos de usuários que atualmente enriquecem o processo de busca e compartilhamento de informações através de debates, pesquisas, iniciativas e padrões que permitem o livre acesso e a interoperabilidade.

Segundo Berners-Lee (1996, p.1), a Web é "o universo da informação acessível através da rede global". Por outras palavras, é o espaço de interação entre usuários situados em diferentes contextos sociais, espaciais, temporais, em prol da informação nos mais variados tipos de linguagem (texto, imagem, vídeo e som), isto é, da hipermídia.

A Web é o atual espaço informacional que norteia os negócios entre indivíduos, grupos e organizações, quer no âmbito pessoal, quer no contexto institucional. É o espaço no qual se encontra enraizado o processo de construção de conhecimento por relações intersubjetivas. É a maior conquista secular para a humanidade, no universo da informação e, constitui o auge de várias idealizações do passado sobre a estrutura concomitante de preservação e acesso da informação e interação em prol do conhecimento.

[8] Termo usado por Focault para designar a coexistência, num espaço impossível, de um grande número de mundos possíveis fragmentários ou justaposição ou superposição de espaços incomensuráveis (HARVEY, 2004, p. 52).

A história da Web tem uma natureza dicotômica, semelhante à história da Ciência da Informação. Enquanto a maioria dos pesquisadores, mormente norte-americanos, apontam o período Pós-Guerra e o artigo de Vannevar Bush como principais marcos, outros pesquisadores como o bibliotecário australiano Warden Boyd Rayward consideram que a ideia sobre o hipertexto, que norteou o desenvolvimento da Web, foi concebida anteriormente.

Rayward (1994) aponta que foi Paul Otlet quem desenvolveu sistemas complexos de organização para integrar dados bibliográficos, imagens e textos que mais tarde impulsionaram e foram adaptadas para o hipertexto. Mesmo sem usar a terminologia hipertexto, Otlet pode ser considerado o precursor da Web, na medida em que em 1893 já levantava a necessidade de um sistema internacional de tratamento de informações que envolvia o uso de catálogos para novas formas de publicação, a gestão de bibliotecas, arquivos e museus como agências inter-relacionadas de desenvolvimento colaborativo, a criação de uma enciclopédia universal e da rede universal para todo o conhecimento humano. A ideia da Web colaborativa que prevalece atualmente já se vislumbrava na concepção de Otlet sobre a Bibliografia, pois para ele a importância do livro se resumia na identificação e organização do respectivo conteúdo, fontes e conclusões, no âmbito do conhecimento colaborativo.

A fragmentação do hipertexto em nós para ser recuperado por vínculos associativos também pode ser atribuída a Otlet, a partir das quatro categorias de informação (fatos, interpretação de fatos, estatísticas e fontes) com base nas quais cada artigo ou capítulo deveria ser analisado, para ser representado em cartões que formariam o catálogo. A partir da Classificação Decimal Universal (CDU) e com base no Instituto Internacional de Bibliografia, Otlet e La Fontaine criaram o primeiro banco de dados de cartões que além de arquivos textuais, também classificaram imagens, estabelecendo a primeira enciclopédia que pretendia reunir todo o conhecimento humano e os padrões de publicação. A Conferência das Associações Internacionais organizada por Otlet e La Fontaine em 1910 culminou no palácio mundial, que envolvia coleções do museu, da biblioteca, do catálogo bibliográfico e arquivos documentais universais. Para Otlet, essas coleções representavam todo o conhecimento humano (livros, documentos, catálogos e objetos científicos), classificado de acordo com os padrões e podem ser considerados como parte dos subsídios que nortearam a construção da Web (RAYWARD, 1994, p.239).

O sistema complexo de indexação da informação desenvolvido por Otlet pode ser considerado como uma versão do Google atual. Aliás, pode se afirmar que o sistema de busca por assuntos ou termos na Web foi subsidiado pelo sistema internacional de busca por assuntos ou pelo número CDU, propostos por Otlet e La Fontaine. O sistema de arquivos limitado pela tecnologia da época permitia pesquisas a partir de números de cartões de índice e extrair informações sobre o recurso informacional representado, num mecanismo de fragmentos que caracteriza o hipertexto.

Castells (2003, p.13) entatiza que o projeto da rede de computadores – Internet, teve o auge em 1990, com o desenvolvimento da World Wide Web pelo inglês Berners-Lee, a partir da síntese do online system de Engelbart; do Xanadu, hipertexto aberto e auto-evolutivo de interligação de informações desenvolvido por Ted Nelson, bem como do sistema hypercard de interconexão de informações de Bill Atkinson. Este projeto foi posteriormente influenciado, por um lado, pela Microsoft, com a introdução do *software Windows 95* e do navegador *Internet Explorer* em 1995 e, por outro, pelos movimentos estudantis de produção de softwares abertos, de redes de computadores e de distribuição de protocolos de comunicação.

Vannevar Bush (1945) é considerado um dos maiores precursores da Web, pois através do seu artigo *"As we may think"*, levantava o problema sobre o aumento da especialização científica necessária para o progresso que se refletia no defasamento dos métodos de transmissão e revisão dos resultados das pesquisas. Como solução, propôs o *"memex"*, um dispositivo de armazenamento de microfilmes e comunicação instantânea para ampliar a memória e garantir um novo tipo de recuperação da informação. Em última análise, o *memex* pode ser considerado como Web, não só pelo armazenamento de informações em forma de microfilme ou texto contínuo e integrado entre diferentes mídias, como também pela comunicação entre usuários e recuperação da informação por associação ou "trilhas" que vinculam os documentos. Para Rayward (1994), estas ideias conduziram ao desenvolvimento de uma nova forma de enciclopédia, o hipertexto, para conectar vários textos em diferentes fronteiras do documento, de modo a suportar sistemas online que permitam o acesso e o desenvolvimento colaborativo, independentemente da distância entre os usuários.

Independentemente de qualquer posicionamento, a Web foi proporcionada pelo aumento do volume de informações, pelo advento das tecnologias de produção

e disseminação da informação, aliadas à indústria cultural, pelo capitalismo transitório e descentralizado, cujo auge foi sintetizado pela pós-modernidade. Por isso, a Web nasceu e desenvolveu-se na complexidade que envolve várias teorias e questionamentos que se prendem com os seguintes aspectos: democratização do acesso à informação, inclusão e exclusão, efetividade da tecnologia dada a sua versatilidade e limitação, comunicação subjetiva, poder de transformação social pela Web, alcance da sociedade da informação, especificidades antropológicas da Web colaborativa, fragmentação e descontinuidades. Estas e outras inquietações subsumem-se nas estratégias pelas quais se constrói o conhecimento humano, tendente a viabilizar a complexidade dos problemas envolvidos no mesmo contexto.

Para Berners-Lee (1996), a Web foi criada com base na ideia sobre o hipertexto que, por sua vez, foi sintetizada a partir das propostas de Vannevar Bush sobre o *Memex*, do sistema online ou *NLS,* de Douglas Engelbart, e das proposições sobre o termo hipertexto, na cunhagem de Ted Nelson. Assim, segundo Berners-Lee (1996), a Web foi desenvolvida em 1989 com observância dos princípios de *design* de *software* para a rede e dos seguintes critérios baseados em sistemas pessoais que já funcionavam na altura:

- Um sistema de informação deve ser capaz de registrar associações[9] aleatórias entre quaisquer objetos arbitrários, ao contrário da maioria dos sistemas de banco de dados;
- A ligação entre dois sistemas envolvendo diferentes usuários deve ser gradativa, não sendo necessário o uso de operações não escaláveis, como *links* de fusão de bancos de dados;
- Qualquer tentativa de restringir os usuários como um todo para o uso de linguagens particulares ou sistemas operacionais foi sempre condenado ao fracasso;
- A informação deve estar disponível em todas as plataformas, incluindo as gerações futuras;
- Não padronização dos dados para restringir o modelo mental dos usuários;

[9] Atualmente essas associações são feitas a partir de *links* de informações no âmbito das estratégias de busca do usuário. O registro dessas associações é de suma importância, na medida em que traduz o respectivo conhecimento e, por isso, podem ser usadas para aprimorar recuperações futuras das mesmas informações ou de outras a elas associadas, bem como para sugerir outros usuários para o mesmo tipo de associação.

- O profissional responsável deve entrar ou corrigir as informações organizacionais representadas com precisão no sistema da forma mais simples possível.

Conforme Berners-Lee (1996) acrescenta, a Web foi desenvolvida de modo a permitir a interação intuitiva entre a pessoa e o hipertexto para que a máquina representasse, a partir da informação legível, o estado dos nossos pensamentos, interações e padrões. Deste modo, o aprendizado da máquina tornou-se uma ferramenta de gestão muito poderosa em grandes organizações. De fato, a padronização dos processos de representação da informação e a interoperabilidade entre usuários humanos e não-humanos, permite que atualmente as máquinas produzam o conhecimento por associações que ultrapassam a capacidade humana de reflexão, devido ao volume excessivo de informações.

Os identificadores universais de recursos constituem outro ponto crucial na arquitetura da Web. O hipertexto funciona com base em *links* que estabelecem ligações consistentes entre documentos ou recursos e os identificadores como *Hypertext Transfer Protocol (HTTP)* – Protocolo de Transferência de Hipertexto indica o espaço para os restantes pontos de cadeia, isto é, o protocolo de comunicação para a conexão, comunicação e transferência de dados entre sistemas. Cada recurso de informação é identificado através do respectivo *Universal Resource Identifier (URI)* – Identificador Universal de Recurso e localizado diretamente por meio do respectivo *Universal Resource Locator (URL)* – Localizador Universal de Recurso. Enquanto a URI indica o nome, a URL é relativo ao endereço, mas na Web tal diferenciação é insignificante, pois normalmente se usam as URLs como endereços. Os identificadores universais de recursos, por um lado, são opacos, na medida em que o software não tem permissões para tirar quaisquer conclusões sobre o objeto referenciado. Por outro, identificam os recursos de forma genérica[10], sem especificar as características peculiares do recurso referido (BERNERS-LEE, 1996, p.4). O problema é que este nível da estrutura da Web não

[10] Um URI, por exemplo, pode identificar um livro que está disponível em vários idiomas e em vários formatos de dados. Outro URI poderia identificar o mesmo livro em um idioma específico, e outro URI poderia identificar uma edição específica do livro em um determinado idioma (Berners-Lee, 1996, p.4).

envolve a descrição temática e conceitual, nas quais poderiam se esboçar as diferentes manifestações do recurso ou obra.

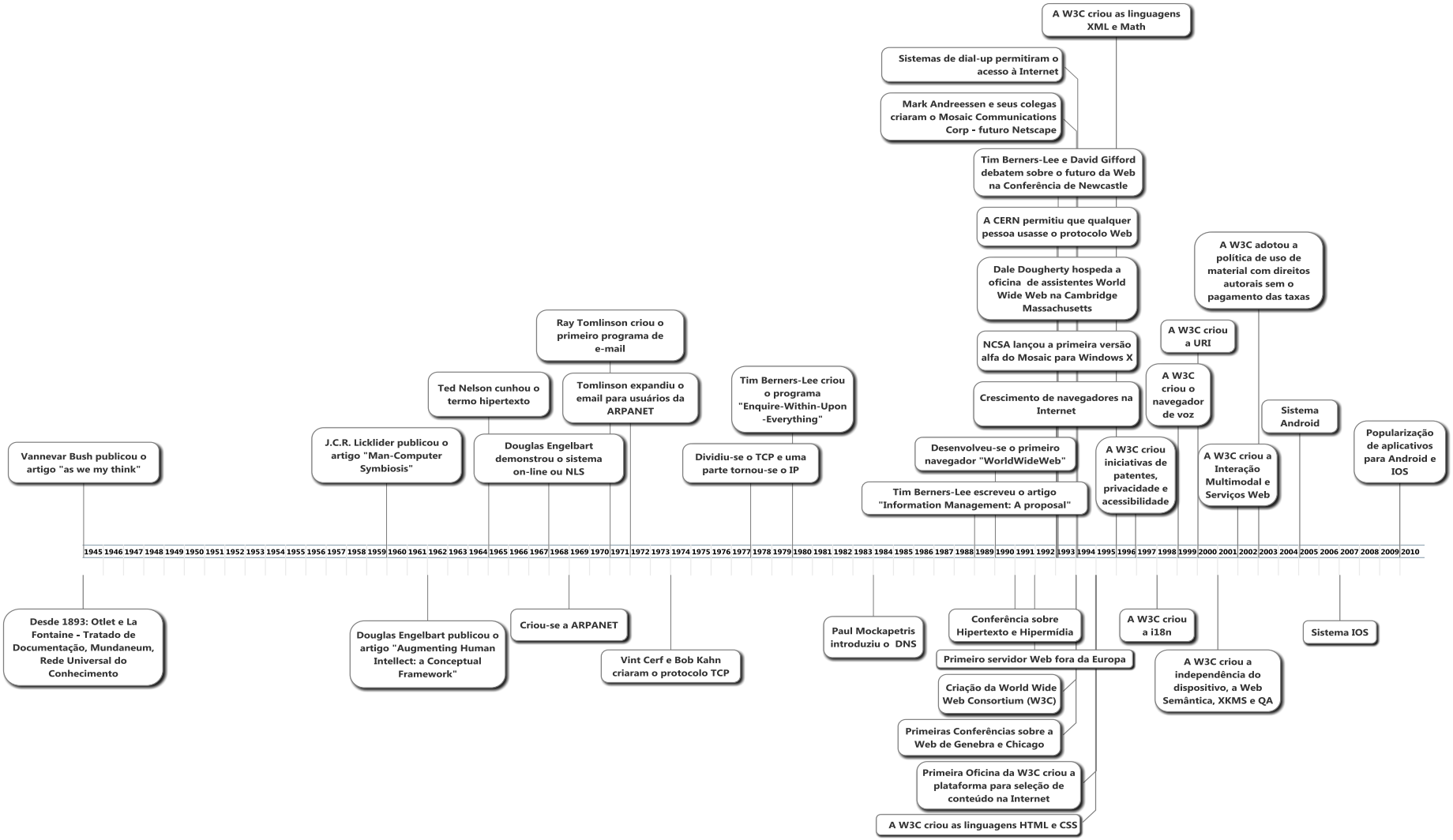
Para o intercâmbio do hipertexto, de acordo com Berners-Lee (1996), adotou-se a *Hypertext Markup Language (HTML)* – Linguagem de Marcação de Hipertexto como um formato de dados para serem transmitidos através da escrita, à semelhança dos sistemas baseados em *Standard Generalized Markup Language (SGML)*.

Como se referiu anteriormente, antes dos precursores norte-americanos, Paul Otlet já havia idealizado o modelo de hipertexto, tanto por meio de fragmentos que representam as diversas facetas do documento, como através do sistema de indexação e busca por assuntos ou números de classificação. Em 1934 também desenvolveu a ideia da rede de conhecimento ou multimídia, combinando cartões de índice e telefone para tornar acessível o conhecimento da época e melhorar a colaboração entre os pesquisadores. Porém, muitos desconsideram o contributo de Otlet e afirmam que Vannevar Bush foi o maior propulsor das ideias sobre o hipertexto e Internet.

Conforme a figura 1, a história da Web pode ser sintetizada a partir dos seguintes marcos:

- Em 1945, Vannevar Bush propôs no seu artigo *"as we my think"* a criação de um dispositivo mecânico fotoelétrico para a expansão de memória, através de *links* entre documentos em microfichas. Para Gromov (1995), antes do período pós-guerra, o primeiro passo na história da Internet foi a instalação do cabo para realizar a comunicação através do oceano Atlântico, em 1858 e o lançamento do satélite Sputnik pela União Soviética em 1957 que mobilizou pesquisas em redes de computadores e tecnologia de comunicações, através da Agência de Projetos de Pesquisa Avançada (ARPA);

***Figura 1**: Linha de tempo da Web e Internet*



***Fonte:** Criada pelo autor com base no gráfico do World Wide Web Consortium (2004).*

- Em 1960, J.C.R. Licklider publicou o artigo "*Man-Computer Symbiosis"* e em 1968, em co-autoria com Robert Taylor, publicou o artigo "*The Computer as a Communications Device*", descrevendo a sua visão sobre a computação, a razão para a construção da Internet e o futuro sobre o uso da rede. Licklider era Diretor do *Information Processing Techniques Office (IPTO)*, uma divisão do Pentagon´s Advanced Research Projects Agency (ARPA), na qual usou os fundos do governo para financiar pesquisas que levaram ao desenvolvimento da Internet, do mouse, do hipertexto, do projeto MAC[11] e de computadores pessoais que posteriormente originaram as primeiras comunidades online;
- Em 1962, Douglas Engelbart publicou o artigo *"Augmenting Human Intellect: a Conceptual Framework"*. No artigo, Engelbart enalteceu os ganhos alcançados pelos longos anos de pesquisa em computação e propôs novos desafios sobre a computação interativa, numa inovação descontínua, para solucionar os problemas humanos. Sobre este aspecto, Gromov (1995) acrescenta que a Internet foi desenvolvida como ferramenta para criar a massa crítica dos recursos intelectuais, pela possibilidade de comunicação e troca de informações à distância;
- Em 1968, Engelbart demonstrou o funcionamento do sistema online ou *NLS*, desenvolvido no seu próprio laboratório de pesquisa entre 1960 e 1970, que facilitava a criação de bibliotecas digitais, o armazenamento e recuperação de documentos eletrônicos utilizando hipertexto, a interação através do mouse, a criação de interfaces gráficas e execução em janelas para a leitura de emails, bem como variedade de opções de processamento de texto[12];
- Em 1969 foram criadas as comissões da *Advanced Research Projects Agency commissions (ARPANET)* e os primeiros nós foram conectados. Um dos nós foi o laboratório de Engelbart, permitindo o

[11] Tornou-se famoso por pesquisas inovadoras em Sistemas Operacionais, Inteligência Artificial e da Teoria da Computação.

[12] As pesquisas de Engelbart, desde a NACA Ames Laboratory – o precursor da NASA e o Instituto de Pesquisa de Stanford, locais onde inicialmente trabalhou até o seu laboratório pessoal denominado Augmentation Research Center, visavam o uso da tecnologia para resolver os problemas da complexidade humana, ou seja, para aumentar o intelecto humano.

uso do seu sistema online para recursos da ARPANET com base no modelo do hipertexto já em funcionamento;

- Em 1971, Ray Tomlinson criou o primeiro programa de e-mail para enviar mensagens através de uma rede distribuída e em 1972 expandiu-o para usuários da ARPANET, usando o símbolo "@" como parte do endereço;
- Em 1974, Vint Cerf e Bob Kahn publicaram um protocolo para a interconexão de pacote de rede, que especificava em detalhes o projeto da criação do Protocolo de Controle de Transmissão – *Transmission Control Protocol (TCP)*. Em 1978 dividiu-se o TCP e uma parte tornou-se o Protocolo de Internet – *Internet Protocol (IP)* para a identificação de dispositivos e redes;
- Em 1984, Paul Mockapetris introduziu o Sistema de Nomes de Domínios – *Domain Name System (DNS)* para o gerenciamento de nomes de domínios em endereços da rede IP;
- Em março de 1989, Tim Berners-Lee escreveu o artigo *"Information Managment: a proposal"*, uma proposta sobre a gestão de informações gerais sobre os aceleradores e experimentos na *Conseil Européen pour la Recherche Nucléaire (CERN)*, atualmente denominada *Organisation Européenne pour la Recherche Nucléaire*, que discutia os problemas de perda de informações sobre sistemas complexos e propunha uma solução baseada em um sistema de hipertexto distribuído. A segunda versão do artigo foi publicada em maio do mesmo ano;
- No final de 1990 foi desenvolvido o primeiro navegador *"WorldWideWeb"* e o navegador em modo de linha, editor e servidor, culminando em dezembro de 1990 com a primeira comunicação entre cliente e servidor Web através da Internet;
- Em dezembro de 1991, Berners-Lee demonstrou o seu papel na Web por meio de pôster apresentado na Conferência sobre Hipertexto e Hipermídia, em Texas, nos EUA, que reuniu estudiosos, pesquisadores e profissionais de diversas áreas para analisar a forma, o papel e o impacto do hipertexto e hipermídia, bem como o poder transformador

da hipermídia e sua capacidade de alterar a nossa forma de leitura, escrita, discussão, trabalho, troca de informações e entretenimento;

- Em dezembro de 1992 foi implementado o primeiro servidor Web fora da Europa, criado pela Universidade de Stanford e a partir de 1993 verificou-se o crescimento de navegadores, como Midas, Erwise, Viola e Samba;
- Em março de 1993, o *National Center for Supercomputing Applications (NCSA)* lançou a primeira versão alfa do Mosaic para Windows X;
- Em abril de 1993, a CERN permitiu que qualquer pessoa usasse o protocolo Web e o código *royalty-free*. Em junho do mesmo ano, Dale Dougherty da O'Reilly hospedou a oficina[13] de assistentes World Wide Web na Cambridge Massachusetts, EUA;
- Em 1994, Mark Andreessen e seus colegas deixaram o Centro Nacional de Aplicações de Supercomputação para formar o *Mosaic Communications Corp*, que mais tarde se tornou Netscape. No mesmo período, Sistemas de *dial-up* tradicionais (CompuServe, AOL, Prodigy) começaram a permitir o acesso à Internet;
- Em 1994 foi criada a World Wide Web Consortium (W3C) com o lema "perceber, liderar a Web e potencializar a sua evolução". Segundo, Berners-Lee (1996), o consórcio da Web visava unificá-la ou evitar a fragmentação dos padrões da arquitetura que acabariam por destruir o universo das informações, em consolidação;
- Em 1995, a W3C desenvolveu a linguagem HTML e a linguagem de folhas de estilo – *Cascading Style Sheets (CSS)* para a construção de páginas Web. Enquanto a HTML proporciona a estrutura de página Web para vários dispositivos, o CSS permite a apresentação independente através de cores, leiaute e fontes, separando a estrutura do conteúdo;
- Em 1996, a W3C desenvolveu a *Extensible Markup Language (XML)* para tornar o conteúdo da Internet legível para humanos e

[13] Segundo a W3C, o principal objetivo da oficina era discutir as prioridades de desenvolvimento, mormente desenvolvimentos em HTTP e HTML, bem como sobre recursos que estavam sendo adicionados à biblioteca de código de domínio público, e os servidores de base (por exemplo, autenticação e contabilidade, mecanismos de folhas de estilo, objetos interativos, necessidades de software de fornecedores de serviços Internet, entre outros).

computadores, isto é, para ler documentos XML e fornecer o respectivo conteúdo e estrutura. Igualmente, desenvolveu a *Mathematical Markup Language – MathML* como uma aplicação XML para descrever a notação matemática e capturar a estrutura e o conteúdo.

- Em 1997, a W3C publicou as primeiras recomendações W3C para HTML – HTML 3.2 e lançou o *International Program Office (IPO)* para a iniciativa de acesso à Web. A Web acessível é aquela que permite o uso por pessoas portadoras de deficiências, através de diferentes combinações de sentidos e capacidades físicas, bem como de diferentes plataformas de hardware e software, mídia, culturas e países. Também desenvolveu o *Document Object Model*, uma interface de plataforma e linguagem neutra para permitir que programas e *scripts* acessem e atualizem dinamicamente o conteúdo, estrutura e estilo de documentos. De igual modo, introduziu a Política de Patentes – *Patent Policy*, no processo de produção de padrões Web; a privacidade – *privacy* através da tecnologia *Do Not Track (DNT)* que permite que os usuários expressem as suas preferências em torno do rastreamento online enquanto navegam pela Web; a *Synchronized Multimedia Integration Language (SMIL)* que permite apresentações audiovisuais interativos, integrando *streaming* de áudio e vídeo com imagens, textos ou qualquer outro tipo de mídia;
- Em 1998, a W3C criou o Conselho Consultivo e a internacionalização – *Internationalization (i18n)*, para garantir o desenvolvimento de conteúdos, aplicações e especificações que atendam às necessidades de pessoas diferentes (cultura, linguagens, sistemas de escrita);
- Em 1999, a W3C criou o navegador de voz para permitir que os desenvolvedores criem aplicações de voz habilitada, baseadas em tecnologias Web;
- Em julho de 2001, a W3C criou a Independência do dispositivo, de modo a permitir o acesso da Web a partir de qualquer aparelho (telefones, leitores de livros eletrônicos, televisores, leitores de música, aparelhos domésticos, etc.). Também criou a Web semântica que, ao contrário da Web dos documentos, é uma Web de dados, na medida em que garante interações confiáveis na rede, através de dados

vinculados. Igualmente adotou a *XML Key Management (XKMS)* que especifica protocolos para distribuir e registrar chaves públicas, para a obtenção das principais informações (valores, certificados, dados de confiança ou gestão) de um serviço Web, bem como a Garantia de Qualidade – *Quality Assurance (QA)* que visa preservar a qualidade das páginas da Web (padrões da W3C, estrutura HTML, CSS, links, interfaces amigáveis e a navegabilidade por meio de vários dispositivos);

- Em fevereiro de 2003, a W3C adotou a política de uso de material com direitos autorais, sem a necessidade de pagar as respectivas taxas de licença. Também adotou a forma de grupos de trabalhos compostos por membros da W3C e por especialistas convidados para abordar a complexidade e dinamismo da Web e outras questões;
- A história da Web também foi influenciada pelo desenvolvimento dos sistemas android, IOS e outros que permitem o acesso e compartilhamento de informações por meio de dispositivos móveis, como tablets, smartfones, smarttvs e bluray players.

Alguns especialistas criticam o modo como a Web está sendo implementada, fundamentando sobre o papel da informação e a complexidade da tecnologia. Mesmo com algumas vicissitudes e divergências, a Web teve um desenvolvimento explosivo. De acordo com a W3C (2004), enquanto em 1991 havia apenas um servidor Web, em 2004 já havia mais de 46 milhões, ou seja, o uso da Internet para a comunicação, desenvolvimento de negócios, ensino, pesquisa, entre outras ações, teve um crescimento exponencial. A popularização da Internet foi tão alta que em 2012 havia cerca de 2,4 bilhões de usuários online e 6 bilhões de subscrições de serviços online através de dispositivos móveis.

Uma das premissas que estiveram no desenvolvimento da Web e que ainda constitui um dos seus maiores desafios é a interação humana. Mesmo que as máquinas produzam correlações interessantes na produção do conhecimento, ainda é da alçada humana, a capacidade de raciocínio e interpretação sobre a qualidade e a relevância da informação. Este aspecto traduz uma das facetas da complexidade herdada da pós-modernidade e fruto da dimensão humana. O pensamento, a apropriação, a representação e a interação são fatores que se circunscrevem na maior problemática da comunicação humana, na medida em que envolvem aspectos

endógenos e exógenos do ser (educação, cultura, crenças, costumes, sistemas políticos, etc.).

Além dos fatores descritos, Berners-Lee (1996) aponta três desafios da Web para o futuro: a melhoria da infraestrutura, o aumento da web como meio de comunicação e interação entre as pessoas e o uso de dados compreensíveis por máquinas para permitir a análise e a resolução de problemas dos humanos. Ora, estes desafios se relacionam com a abordagem de Engelbart sobre o uso dos computadores para o aumento do intelecto humano. De referir que, de nenhum modo o uso de computadores significa substituir a cognoscência humana de relações intersubjetivas por socialização na educação, cultura e interação.

2.4 História da Web Social

A Web social pode ser conceituada como o período histórico da Web, caracterizada essencialmente pela interação contínua entre os usuários, no compartilhamento de informações e pelo provimento de serviços online (política, educação, saúde, comércio eletrônico, entretenimento, etc.).

A Web social caracteriza-se como uma das fases cronológicas da própria Web. Numa entrevista dirigida pelo programa Milênio da GloboNews (2009), pertencente à Rede Globo de Televisão, o jornalista norte-americano Andrew Keen considerava que, na história da Internet, a Web 1.0 predominou no período entre 1990 e 1999 e foi caracterizado pela disponibilização de conteúdos na rede por empresas tradicionais de comunicação. Esta Web consistia apenas no provimento da informação digitalizada e, por isso, estava destituída da transformação cultural que caracteriza a Web 2.0, por meio de publicações e compartilhamento de conteúdos por usuários não especialistas ou formalmente credenciados em determinados assuntos.

Melo Júnior (2007) também considera que a Web 1.0 era caracterizada por páginas estáticas e isso só começou a mudar em 1996 com a implementação do Javascript e tecnologias como *PHP, ASP, JSP*, incrementando os negócios na Web, como o *e-commerce*. Alguns recursos como *Flash* também incrementaram a dinâmica das páginas e o interesse dos usuários em notícias, *chats*, compras, etc.

O termo Web 2.0 foi cunhado por Tim O´Reilly em 2003 e, segundo Melo Júnior (2007), esta fase da Web não teve mudanças significativas influenciadas por

recursos tecnológicos. O que mudou foi o foco da Web para a interação e o compartilhamento de informações. As próprias páginas passaram a permitir maior interação, através da Interface de Programação de Aplicações ou Aplicativos, vulgo API´s, voltados para o uso de serviços e não para softwares.

A Web 2.0, por sua vez, proporcionou a emergência massiva de redes de usuários, movidos pelo interesse em determinados assuntos ou peculiaridades no compartilhamento de determinados recursos informacionais. Essas redes deram origem ao termo Web social.

Na linha de recuperação da informação, a Web Social provoca mudanças significativas e constantes nos processos e métodos envolvidos na CI, pelo fato de envolver novas terminologias, linguagens, tecnologias e estratégias de busca e uso da informação. Um exemplo elucidativo se verifica no uso das folksonomias e adequação dos sistemas aos modelos mentais de representação, busca e processamento pelos usuários finais.

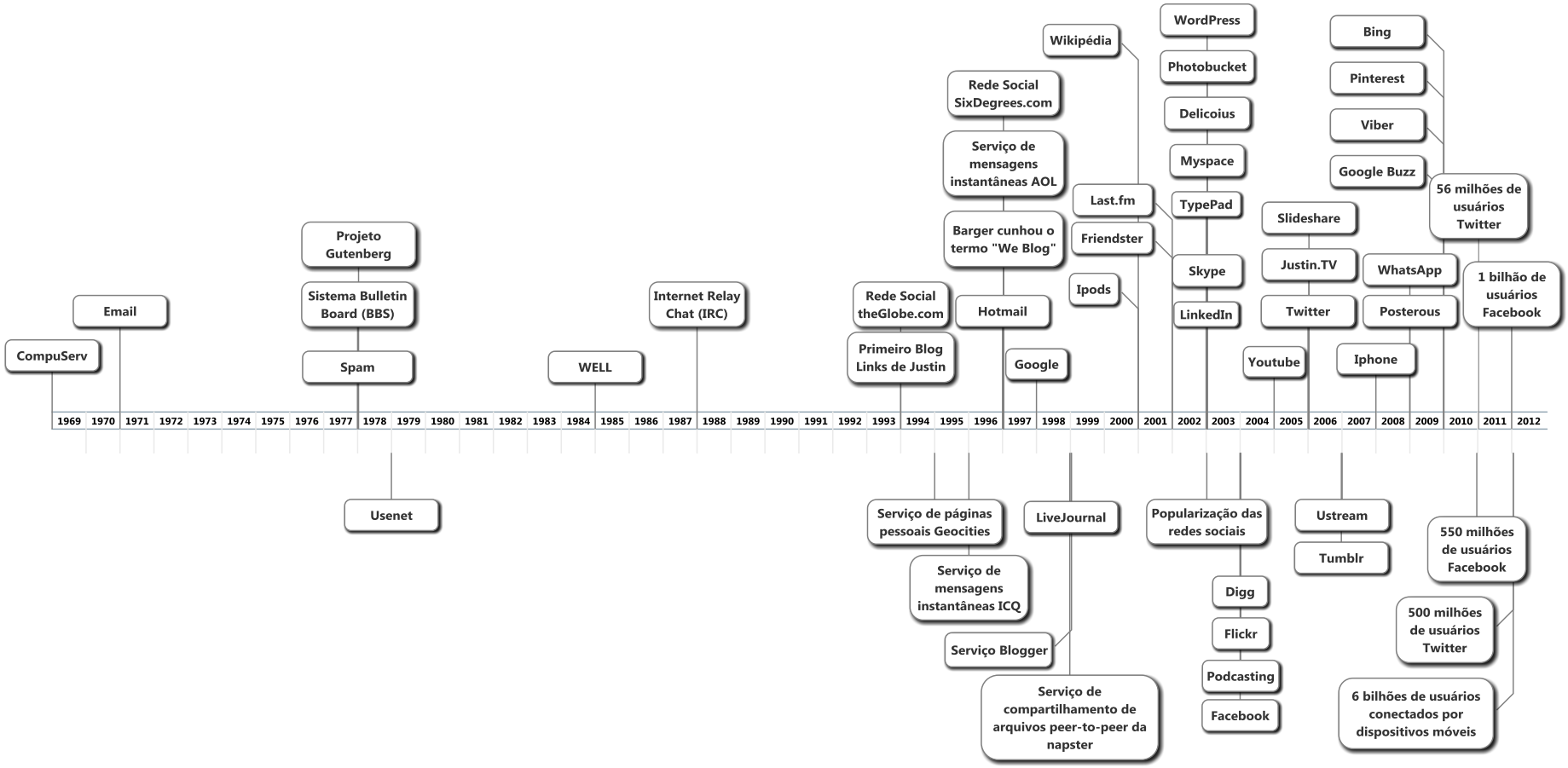
Na figura 2 encontram-se sintetizados alguns marcos sobre a história da Web Social, alguns dos quais perfazem a própria história da Internet.

Segundo a MediaBistro[14] (2013), o compartilhamento de informações através da rede teve início em 1971, com a criação do serviço de email que permitiu as primeiras discussões na Web.

O projeto Gutenberg consiste na primeira biblioteca digital fundada por Michael Hart em 1971, centrada no processo da democratização da informação através da digitalização, arquivamento e livre distribuição de obras literárias. Na prossecução do lema “quebrar as barreiras da ignorância e literacia”, até hoje os voluntários do projeto se engajam na disponibilização de todo o material bibliográfico possível, em formato aberto para que seja acessível por qualquer computador ou usuário. Do mesmo modo, se preocupam com a preservação dos registros bibliográficos por tempo indeterminado.

[14] MediaBistro é um site fundado em 1993 por Laurel Touby como local de encontro para profissionais de jornalismo, editoração e outras indústrias relacionadas à mídia em Nova York. Atualmente possui ofertas de emprego, cursos de formação, eventos, fóruns e publica várias notícias, em particular sobre as mídias sociais.

***Figura 2**: História da Web Social (1969 – 2012)*



***Fonte:** Criada pelo autor com base na MediaBistro (2013).*

Além do projeto Gutenberg, o lançamento da WELL em 1985, como Whole Earth 'Lectronic Link por Stewart Brand e Larry Brilliant, consagrou-se como a primeira comunidade online ou "comunidade virtual"[15] de compartilhamento de informações que inspirou várias comunidades de debate atualmente existentes na Web.

Outro marco da Web Social foi a criação em 1988 do protocolo de comunicação por bate-papo – *Internet Relay Chat (IRC)*, de acordo com Stenberg (2011), por Jarkko Oikarinen, na Universidade de Oulu, na Finlândia. Em julho de 1990, o protocolo tinha em média 12 usuários em 38 servidores da Internet e em 1993, durante a Guerra do Golfo, foi usado para a troca de informações em tempo real nas Universidades do Oriente Médio. A sua decadência deveu-se à concorrência de outros serviços que agradavam mais os usuários, por um lado, o programa de mensagens instantâneas ICQ, que foi desenvolvido pela empresa israelense Mirabilis e depois adquirida pela empresa America Online (AOL). Atualmente o ICQ pertence ao Mail.Ru Group e está disponível na versão 8 para bate-papo no Windows e em dispositivos móveis; por outro, o MSN Messenger, criado pela Microsoft em 1999 que inovou com a conversa por áudio e vídeo.

Segundo a página pessoal Justin Links (2014), o jornalista e empresário americano Justin Hall usa a Web para publicar as suas notas pessoais, desde janeiro de 1994. Por esse motivo, em dezembro de 2004, o New York Times Magazine considerou-o como fundador dos blogs pessoais e o ano de 1994, por sua vez, a criação do primeiro blog.

Em 1997, o norte-americano Jorn Barger, cunhou o termo "Weblog" para caracterizar as suas publicações na Web. O termo foi posteriormente encurtado para "Blog" por Peter Merholz, em 1999. Atualmente os blogs integram as redes sociais, no compartilhamento de informações e construção do conhecimento por indivíduos, grupos e organizações. No mesmo ano de 1997, Andrew Weinreich criou o site "sixdegrees.com" que se consagrou como primeiro modelo da rede social atual. O site consistia em relacionamentos entre usuários organizados em degraus e o envio das mensagens entre eles observava esse critério de estratificação.

Em 1998 criou-se a Google, uma empresa de serviços online e softwares que inovou o processo de busca de informações na Web. Em 1999, a Google criou o

[15] Termo cunhado por Howard Rheingold.

“Blogger” que revolucionou a publicação e o compartilhamento de informações, tanto para pessoas especialistas, como para não especialistas.

Em 1999, segundo o site MediaBistro, criou-se o LiveJournal como uma comunidade virtual de interação e compartilhamento de informações, através de blog, jornal ou diário. O LiveJournal prevalece nos dias atuais e qualquer usuário pode criar uma conta, compartilhar informações textuais, pessoais, tags, classificar o conteúdo publicado, especificar se deseja comentários ou tornar público o conteúdo publicado, entre outras opções. Outra rede social de compartilhamento de conteúdos online e de mídia (vídeos, fotos, mensagens e comentários) é Friendster, fundada em 2002 pelo canadense Jonathan Abrams. Atualmente, Friendster é um site de jogos sociais sediado em Kuala Lumpur, na Malásia.

Outro marco fundamental na história da Web social foi a criação em 1999, da tecnologia *peer-to-peer*, que passou a permitir a troca de arquivos de mídia, tais como livros, música, filmes e jogos entre usuários conectados na rede. Esta tecnologia permitiu a criação em 2002 da Last.fm, um serviço de descoberta de músicas e recomendações baseadas nos perfis dos usuários, bem como de ipods em 2001, pela empresa norte-americana Apple, para a reprodução de músicas.

A criação da Wikipédia, em janeiro de 2001, por Jimmy Wales e Larry Sanger, pode ser considerada como um dos principais acontecimentos no panorama da Web social. Segundo o próprio site, a Wikipédia é uma enciclopédia virtual de conteúdo e acesso livres, apoiado e organizado pela Fundação sem fins lucrativos Wikimedia. Atualmente, tem cerca de dezoito bilhões de visualizações por usuários que podem editar o conteúdo e, por isso, tornou-se a maior e mais popular obra de referência geral na Web. Historicamente falando, se recuarmos para a fusão das coleções de bibliotecas das sociedades científicas ao Escritório Internacional de Bibliografia de Otlet, em 1907 que, segundo Rayward (1994), visava formar a biblioteca coletiva da comunidade científica que mais tarde se tornaria a Biblioteca Internacional, é pacífico considerar Paul Otlet como um dos pesquisadores cujas ideias inspiraram a idealização da Wikipédia.

O termo Web Social ganhou popularidade em 2003, com o uso massivo das redes sociais e pela possibilidade de os usuários desenvolverem vocabulários compartilhados, como tags e folksonomias, para a classificação do conteúdo. No mesmo ano, lançou-se o Linkedin, fundado em dezembro de 2002, por Reid Hoffman e outros membros da PayPal, como uma rede social de relacionamentos

para negócios. Alex Welch e Darren Cristal também fundaram o Photobucket para a hospedagem de imagens e vídeos, tanto para fins pessoais, como para negócios, pela comunidade dedicada à preservação e compartilhamento do ciclo de vida de fotografias e vídeos.

Em 2003, Joshua Schachter criou o Delicious[16], uma rede social voltada para o armazenamento, compartilhamento e descoberta de favoritos da Web, através da classificação social dos usuários, baseada em termos livres ou folksonomia. Chris DeWolfe e Tom Anderson, por sua vez, criaram o Myspace, uma rede social de relacionamentos com ênfase na música. Igualmente, Ryan Boren e Matthew Mullenweg criaram o aplicativo WordPress para o gerenciamento de conteúdo na Web e que inovou a publicação de informações com ferramentas de criação avançadas de blogs. De igual modo, criou-se o Typepad para hospedar blogs de empresas de mídia, como BBC e Sky News. Janus Friis e Niklas Zennström, em cooperação com Ahti Heinla, Priit Kasesalu e Jaan Tallinn criaram o Skype, uma rede social de relacionamentos através de mensagens instantâneas, chamadas de voz e vídeo que, desde o ano de 2011, pertence à Microsoft.

Em 2004, Kevin Rose, Owen Byrne, Ron Gorodetzky e Jay Adelson criaram o Digg, um site que agrega informações por meio de links de outros sites e que os usuários podem avaliá-las. Stewart Butterfield e Caterina Fake também criaram o Flickr que foi lançado pela empresa Ludicorp, para hospedar imagens e vídeos que podem ser compartilhados pelos próprios usuários e pelos profissionais que trabalham com fotografias ou blogs. Igualmente, criou-se o Podcasting[17] para a publicação de arquivos de mídia digital, principalmente no formato de áudio, por meio de feed RSS.

Em 2004, Orkut Büyükkökten, engenheiro turco e funcionário da empresa Google criou a rede social Orkut, para o relacionamento entre usuários, através de bate-papo e compartilhamento de arquivos de mídia. O Orkut foi a primeira rede social mais usada no Brasil e só perdeu a preferência dos usuários para o Facebook e Twitter. Por isso, em Setembro de 2014 a Google anunciou o seu encerramento.

[16] Uma das importâncias desta rede social consiste no envolvimento do próprio usuário na classificação dos conteúdos, por meio de termos livremente escolhidos e no agrupamento de links por temas que podem ser compartilhados para proporcionar uma navegação por categorias de assunto.

[17] O Podcasting assemelha-se à rádio sob demanda, mas a sua principal diferença está na variedade de opções, algumas das quais assentam sobre o poder de decisão do usuário sobre o tempo e lugar para o acesso a informações, que podem ser músicas, notícias ou comentários sobre espetáculos, conteúdos ou histórias educativas.

No mesmo período, Mark Zuckerberg e seus colegas da Universidade de Harvard Eduardo Saverin, Andrew McCollum, Dustin Moskovitz e Chris Hughes criaram o Facebook, uma rede social para o compartilhamento de informações pessoais, de grupos e de empresas ou organizações. Atualmente o Facebook é a rede social mais usada em muitos países como: EUA, Inglaterra, Portugal, Brasil, Índia, Turquia, França, Itália.

Em 2005, Chad Hurley, Steve Chen e Jawed Karim, antigos funcionários da PayPal, criaram o Youtube que desde 2006 pertence à empresa Google. O Youtube é atualmente a maior rede social de compartilhamento de arquivos de vídeos e em 2007 foi lançada a versão da interface em português no Brasil.

Em 2006, Jack Dorsey, Evan Williams, Biz Stone e Noah Glass criaram o Twitter, uma rede social de compartilhamentos de “tweets” ou mensagens de até 140 caracteres, que podem ser agrupadas por tema ou tipo, antecedida da hashtag “#”. O símbolo “@”, seguido do nome do usuário, é aplicado para respostas a outros usuários e as mensagens podem ser retuitadas entre os usuários. Em outubro do mesmo ano também se criou o *SlideShare*, uma rede social de compartilhamento de arquivos privados e públicos em *PowerPoint, PDF, Keynote,* apresentações *OpenDocument* e em formato de vídeo. Além do acesso, os usuários podem votar, comentar e compartilhar o conteúdo publicado. Atualmente é considerado um dos principais serviços de fomento à educação.

Em 2007, Justin Kan, Emmett Shear, Michael Seibel e Kyle Vogt criaram a Justin.TV para a hospedagem e compartilhamento de vídeos ao vivo entre os usuários. Desde 10 de fevereiro de 2014, o serviço foi rebatizado por Twitch Interactive Inc. e atualmente inclui jogos online.

Em 2007, John Ham, Brad Hunstable e Dr. Gyula Feher criaram o Ustream para permitir que os seus amigos militares no Iraque pudessem se comunicar com as suas famílias. Atualmente a empresa oferece serviços de transmissão ao vivo de vídeos na política, no entretenimento para emissoras de televisão e rádio, na tecnologia para empresas; conta com parceiros como Samsung, Logitech, CBS News, NASA, Dell, Sony, entre outros. No mesmo ano, David Karp criou o Tumblr, uma rede social em forma de blog que permite a publicação de arquivos multimídia (textos, imagens, vídeos, links, citações, áudio) e o acompanhamento dos usuários entre si e das respectivas publicações. Desde 2013, o Tumblr pertence à empresa Yahoo e atualmente hospeda mais de 200 milhões de blogs.

Em 2009, criou-se o Posterous, uma plataforma de blogs que foi encerrado em 2013; a maioria dos seus funcionários foram recrutados pelo Twitter e o termo “unfriend” foi adicionado ao dicionário americano Oxford, para significar o ato de remover alguém da lista de amigos ou contatos numa rede social. No mesmo período, Brian Acton e Jan Koum, ambos ex-funcionários da empresa Yahoo, criaram o WhatsApp Messenger, um serviço de troca de mensagens instantâneas e compartilhamento de arquivos de imagens, áudio, vídeo e localização entre usuários, por meio de Smartphones. Atualmente o WhatsApp pertence ao Facebook e conta com cerca de 600 milhões de usuários ativos.

Em 2010, Kevin Systrom e Mike Krieger criaram a rede social Instagram que permite aos usuários filtrar fotografias e vídeos para compartilhá-los no próprio Instagram ou em outras redes sociais, como o Facebook, Twitter, Flickr e Tumblr. O Instagram foi adquirido pelo Facebook em abril de 2012 e está disponível para dispositivos móveis na App Store, Google Play e Windows Phone Store. No mesmo ano, também se criou o motor de busca Bing, com enfoque nas redes sociais e Ben Silbermann, Paul Sciarra e Evan Sharp criaram o Pinterest, uma plataforma de personalização de mídia no qual os usuários podem carregar, salvar, classificar e gerenciar imagens e vídeos, bem como compartilhá-los. Em 2011 as mídias sociais ficaram disponíveis e acessíveis para qualquer dispositivo (computador, tablet, SmartPhone, Smart TV, Bluray Player, etc.) em qualquer lugar.

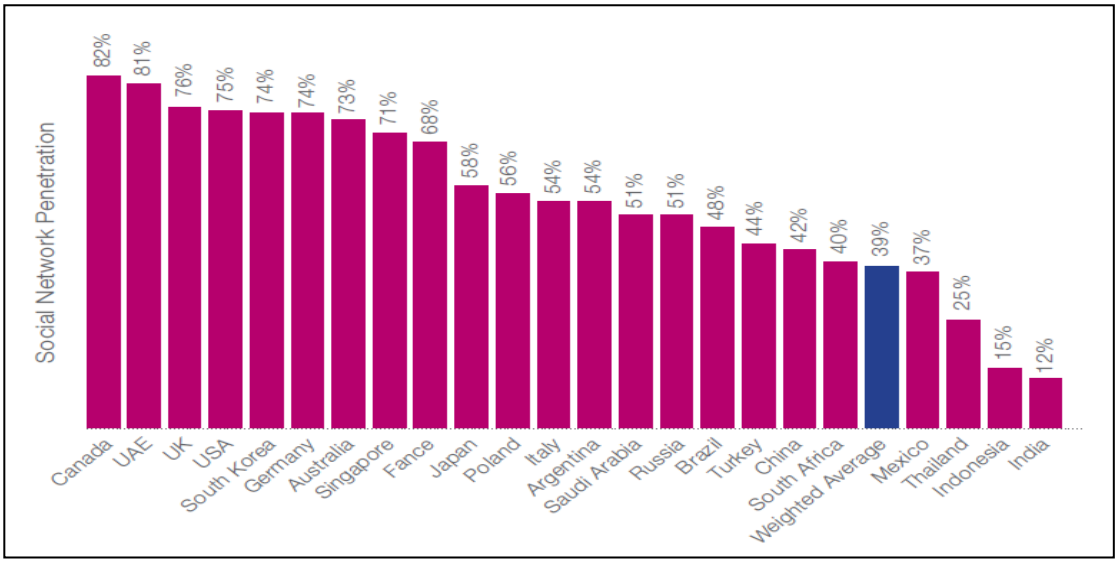
A maioria das abordagens atuais sobre as redes sociais deve incidir sobre a Web na medida em que dúvidas não se levantam quanto à consagração desta como o ambiente da produção, difusão, acesso e uso da informação. Esta tendência cada vez crescente também é favorecida pelas comunidades online engajadas no compartilhamento de informações em diferentes áreas do saber.

O crescimento da Web e das redes sociais na Internet é comprovado pelo relatório da Broadband Commission for Digital Development[18] (2014), publicado em setembro, que indicava que até o final de 2014, cerca de 2,9 bilhões de pessoas correspondentes a 40% da população mundial estariam online e metade da população mundial estaria conectada pela rede global até 2017. No mesmo cenário,

[18] Criada em Maio de 2010, pela iniciativa da União Internacional de Telecomunicações (ITU) e da Organização das Nações Unidas para a Educação, Ciência e Cultura (UNESCO), para intensificar os esforços da ONU no cumprimento dos Objetivos de Desenvolvimento do Milênio, através da implantação de conexões de banda larga de alta capacidade e velocidade para a Internet, como parte essencial da infraestrutura moderna, com benefícios econômicos e sociais consideráveis nos países em todos os estágios de desenvolvimento.

o relatório indicava que 2,6 bilhões de smartphones estavam online e 5,6 bilhões estariam até 2019. Indubitavelmente, a Web produz mudanças incrementais nos indivíduos, grupos e organizações, principalmente na educação que foi a área que teve maior crescimento nos serviços móveis, no período entre 2005 e 2013. Todavia, sendo a Web uma tecnologia, não se pode exacerbar acerca destes fatos, pois, segundo o mesmo relatório, muitos usuários são de países desenvolvidos, como Canadá, Inglaterra, EUA, Coreia do Sul e Alemanha.

Figura 3: *Nível de conectividade nas redes sociais em Janeiro de 2014, com base em usuários ativos*



Fonte: *Relatório da Broadband Commission (2014, p.14).*

Conforme a figura 3, os países com economias emergentes como Rússia, Brasil, China, África do Sul e Índia, a conectividade situa-se em 51%, 48%, 42%, 40% e 12%, respectivamente. Nos países subdesenvolvidos, majoritariamente africanos, o uso da internet não atinge 10% dos usuários. Assim, é caso para questionar quem são os membros da propalada “Sociedade da Informação” e qual é a sua real contribuição sobre o conhecimento no todo.

A tecnologia inclui, por meio de espaços de difusão e compartilhamento de informações, mas ao mesmo tempo exclui, pois muitas pessoas ainda carecem dessa tecnologia e, por isso, não podem contribuir com total propriedade nesses ambientes.

A Web foi desenvolvida como qualquer outra tecnologia, isto é, para ajudar os humanos, neste caso, auxiliando na produção, armazenamento, disseminação, acesso e uso da informação. Algumas questões complexas da Web foram herdadas

da pós-modernidade e fazem parte dos principais desafios da CI, enquanto campo do conhecimento. O problema é que a Web permite que qualquer usuário assuma o papel de especialista no ciclo informacional. Assim, como garantir a produção, representação e recuperação de informações consistentes e seguras, face ao subjetivismo e liberdade que caracterizam a pós-modernidade? A resposta aos questionamentos propiciados pelos avanços tecnológicos envolve todas as áreas do conhecimento, na medida em que traduzem a essência humana. Por isso muitas abordagens ainda se mostram necessárias e a total transformação social pela informação e tecnologia, por sua vez, ainda se mostra como um desafio a alcançar.

A inclusão das questões sobre a pós-modernidade na Ciência da Informação é importante para o entendimento conceitual e contextual dos aspectos que justificam as práticas dentro do campo. Por exemplo, permitem entender que a representação da informação não é mera atividade repetitiva de descrição, mas sim processo complexo que visa à completude informacional face às vicissitudes linguísticas e culturais, de modo abrangente e catalisador para o processo de produção do conhecimento.

No Capítulo seguinte, analisam-se continuamente alguns aspectos da pós-modernidade que estiveram na origem da CI como campo científico e na complexidade envolvida no seu objeto.

3

Ciência da Informação

3.1 Visão sobre o Capítulo

A história da produção e difusão do conhecimento é antiga e pode ser ilustrada em paralelo com o próprio processo de evolução do homem. A CI figura-se no rol das ciências emergentes do Século XX, cuja origem está associada ao fenômeno tecnológico e ao paradoxo subjetivista de apropriação e representação de objetos.

O Capítulo sobre a CI foi desenvolvido no âmbito metodológico das principais variáveis da pesquisa. A sua abordagem resulta da necessidade transcendental de situá-la no cenário das ciências tradicionais, modernas e pós-modernas, através das quais o universo do conhecimento humano se permeabiliza. Por isso, antes de elucidar as principais características pelas quais a CI contribui para o conhecimento, tornou-se imprescindível primeiramente discuti-la como campo científico. Para tal, foi necessário, por um lado, teorizar as noções de campo científico, e por outro, dimensionar a CI no rol das características de um campo científico, como: gênese, aspectos que influenciaram a sua origem e nortearam o seu desenvolvimento, objeto de estudo, especificidades que consagraram a sua inovação para cada estágio da sociedade e a sua relação com o conhecimento.

3.2 Ciência e Tecnologia

A relação entre a Ciência e Tecnologia é antiga e pode ser contemporizada em diferentes estágios da historiografia da sociedade. Donald Stokes (2005) analisa essa relação em concomitância com a relação entre a pesquisa básica e a pesquisa aplicada. Assim, enquanto a pesquisa básica procura ampliar a compreensão dos fenômenos de um campo da ciência e é caracterizada pela originalidade, liberdade dos pesquisadores, avaliação dos resultados publicados e distancia temporal entre a descoberta e utilização prática, a pesquisa aplicada, por sua vez, está voltada para a necessidade ou aplicação por parte de um indivíduo, grupo ou sociedade. Em suma, a pesquisa básica visa ao entendimento e a aplicada, a utilização. Por analogia, a pesquisa básica seria reservada à Ciência no todo, enquanto que a aplicada estaria a serviço da Tecnologia.

De acordo com Stokes (2005), as questões sobre a dicotomia entre a pesquisa básica e a aplicada remontam à antiguidade clássica, com os filósofos gregos do Século VI e V a.C. e prevaleceram na civilização grega pela

desvinculação da investigação da prática, através da atribuição das atividades práticas às pessoas de menor posição social. Aristóteles e Platão são exemplos de apologistas que exaltavam a busca pelo entendimento como tarefa filosófica da ciência e não do uso. Posteriormente, com a emergência da ciência grega na Europa, esta manteve a mesma ideologia da superioridade da ciência pura, destituída de fins práticos. Apenas no século XIII, no final da Idade Média e do início da Idade Moderna, a componente utilitária começou a ganhar espaço na Europa, difundida por clérigos que ocupavam posições relevantes nas igrejas e que enfatizavam o trabalho manual. Francis Bacon pode ser considerado o maior precursor da quebra do paradigma grego, através da sua filosofia do conhecimento como poder e da técnica como sinônimo do conhecimento. Contudo, nos anos subsequentes verificou-se um distanciamento entre a ciência e a tecnologia na Europa, fruto da diferença de posição social e situação econômica entre os mentores da ciência e da tecnologia.

A separação entre a ciência pura e a aplicada também foi fortemente cultivada pelo iluminismo na idade moderna. Para Harvey (2004) o iluminismo determinava os modos de produção da ciência, especificando as condições para o exercício da razão. Por isso que o pensamento pós-modernista destaca-se pela crítica ao iluminismo, com ideologias sobre a individualidade e liberdade. Tais ideias pós-modernistas estiveram na origem de campos científicos como CI, voltada para a peculiaridade na qual se circunscreve a individualidade de cada usuário.

A ideologia acerca da superioridade da ciência pura sobre os usos práticos revelou-se uma utopia no campo do conhecimento humano e do progresso técnico. Este fato, segundo Stokes (2005) foi comprovado pelo próprio processo da Revolução Industrial, cujos progressos foram da camada laboral de menor posição social e com pouco conhecimento sobre a ciência. Por isso, nas fases posteriores à Revolução Industrial, verificou-se novamente a quebra do paradigma grego enraizado nos pensamentos filosóficos e voltou-se à fusão entre a ciência e a tecnologia, na qual o século XIX foi testemunha, com a aplicação de métodos científicos aos processos industriais tecnológicos de produção. As pesquisas visando à busca do entendimento nas universidades passaram a ser norteadas por problemas de caráter prático, portanto, um modelo de construção de conhecimento que prevalece atualmente. Porém, na Europa, concretamente na Alemanha, institucionalizou-se a separação entre a pesquisa pura e a pesquisa aplicada, ao se

criar escolas técnicas para capacitar profissionais para a indústria, obedecendo à estrutura hierárquica de comando por meio de institutos de pesquisa e universidades.

Conforme Stokes (2005) acrescenta, o progresso alemão estimulou os Estados Unidos a enviarem os seus quadros para a formação em institutos e universidades alemãs. Na medida em que esses quadros eram empregues nas universidades, esta simbiose alterou a visão do progresso norte americano centrado na fusão entre a ciência e tecnologia e criou a ideia dessas universidades como centros de pesquisa da ciência pura. Este entendimento viria a ser contrariado pela indústria química e elétrica, nas quais a fusão entre a ciência e a tecnologia se mostrava cada vez mais necessária, através da ação concomitante de tecnólogos empíricos e engenheiros com formação prática.

3.2.1 O fim da Guerra e os espaço da CI

Com o fim da Segunda Guerra Mundial que se aproximava, havia uma preocupação dos cientistas norte-americanos sobre o futuro da ciência que, na sua maioria, estava aliada a pesquisas de natureza bélica. A solução para esta inquietação foi feita através do plano para a manutenção do investimento do governo federal, proposto por Vannevar Bush ao presidente norte-americano Roosevelt. Tal plano visava à criação de uma componente intelectual científica e ostensiva para a defesa do país. Bush propôs a criação do *National Defense Research Committee – NDRC* que viria a ser o *Office of Scientific Research and Development – OSRD* da presidência dos EUA. Conforme Stokes (2005, p.81), o referido Escritório de Pesquisa e Desenvolvimento Científico “alcançou sucesso notável em orientar toda a capacidade da ciência e da engenharia industrial dos Estados Unidos para fins militares”. Parte do seu sucesso deve-se ao vinculo que se estabelecia com a elite científica das universidades e da indústria.

Face às criticas que o projeto do Escritório de Pesquisa e Desenvolvimento Científico recebia sobre o colapso da economia norte-americana, aliado ao controle do mercado por meio de patentes sobre o resultado das pesquisas, Bush sentia o medo da retirada dos fundos de pesquisas essencialmente bélicos na época de paz que se aproximava. Como solução e em resposta ao pedido do presidente Roosevelt, Bush enviou o relatório *Science, The Endless Frontier*, ao presidente

Truman, em julho de 1945. Com o referido relatório, pretendia restaurar a autonomia da ciência, restringindo a ação do governo sobre as pesquisas, ao mesmo tempo em que garantia o apoio total daquele. Para tal, aceitou a criação do *National Research Foundation* que viria a ser substituído pelo *National Science Foundation*, para acolher a pesquisa básica. Para alcançar os seus objetivos, Bush estabeleceu dois princípios: o primeiro de que a pesquisa básica não visa fins práticos e o segundo de que a pesquisa básica é precursora do progresso tecnológico. Assim, seria possível garantir o financiamento incondicional em pesquisa básica com a promessa do retorno em inovação tecnológica (STOKES, 2005).

Vannevar Bush é considerado o propulsor da CI, através do artigo publicado em 1945. Este assunto é detalhado no ponto sobre a gênese da CI. A ação do Bush ao publicar o referido artigo não consistia apenas na intenção de criar a solução tecnológica para o problema da explosão informacional. Mas, despertar tal necessidade, enquanto responsável pela pesquisa básica, de modo a atrair investimentos que teriam retorno em produtos tecnológicos. A mobilização subsequente dos cientistas por ele coordenados para discutir as suas propostas prova a intenção da separação entre a ciência e tecnologia. Além disso, a emergência da CI também pode ser analisada do ponto de vista estratégico de poder e controle no período de paz que se aproximava. Ora, por muito tempo o poder americano manifestou-se por pesquisas militares ostensivas, testemunhadas pelas duas bombas atômicas. Com o fim da guerra, também existia uma preocupação de domínio sobre as demais potências, através do uso do conhecimento como poder. Assim, havia a necessidade de criar uma ciência capaz de condensar a informação explosiva que as tecnologias de informação e comunicação potencializavam, para decisões estratégicas e fomento nos meios de produção. Por isso que o período pós-guerra foi marcado pelo crescimento da economia norte-americana, propulsionada pelos modernos processos de produção.

Mesmo com a transformação em lei, a proposta de Bush foi severamente criticada em relação a muitos aspectos, alguns dos quais relacionados com a sua autonomia, patentes resultantes das pesquisas universitárias, autonomia das Ciências Sociais que estavam agrupadas a outras ciências como forma de lhes reduzir o mérito da pesquisa básica. Após cinco anos, o projeto Bush teve o colapso, principalmente pela ausência de colaboradores no Congresso e na presidência. Porém, mais tarde verificou-se novamente o financiamento do governo em pesquisa

básica em resposta ao lançamento de Sputnik, em 1957, pela União Soviética. Conforme Stokes (2005) mesmo com o colapso do projeto Bush, a sua ideologia acerca da relação ciência-tecnologia prevaleceu nas diversas atividades de fomento à pesquisa no período pós Segunda Guerra, considerando a pesquisa básica como a única propulsora da inovação tecnológica, ou seja, desconsiderando utilizações tecnológicas que inspiraram pesquisas básicas. Tal percepção é contrariada pelo paradigma moderno que prevalece na sociedade industrial que desacredita a competitividade na economia mundial baseada apenas em fortes investimentos na ciência pura, movidos pela simples curiosidade.

A quebra do paradigma pós-guerra sobre a relação Ciência-Tecnologia não significa nenhuma reformulação conceitual da unicidade da pesquisa. Pelo contrário, impõe uma dicotomia simbiótica tanto na origem, como no desenvolvimento da universalidade do conhecimento humano. A melhor explicação sobre estes aspectos talvez seja a dada no relatório anual da *National Science Foundation*, em 1951, por Conant que foi colega de Bush, presidente de Harvard e membro do *National Science Board:*

> ninguém pode traçar uma linha divisória precisa entre a pesquisa básica e a pesquisa aplicada [...], arrisco-me a sugerir que faríamos bem em descartar inteiramente as expressões “pesquisa aplicada” e “pesquisa fundamental”. Em seu lugar eu colocaria as palavras “pesquisa programática” e “pesquisa descompromissada”, pois há uma distinção suficientemente clara entre uma pesquisa dirigida a uma meta específica e uma exploração descompromissada de uma ampla área da ignorância humana (CONANT 1951, apud STOKES 2005)

Com esta afirmação fica claro que existem dois tipos de pesquisa que se distinguem pelo seu objetivo. Por um lado, a básica ou descompromissada que é movida pela curiosidade da busca pelo entendimento de um fenômeno, ação ou fato natural ou social e, por outro, a aplicada ou programática que procura solucionar um problema específico. O mais importante é que qualquer tipo de pesquisa é fundamental para o universo do conhecimento humano e, por isso, qualquer política de separação ou categorização nas instituições de produção, acesso e fomento de pesquisa torna-se redutível e desnecessária, principalmente pelo entrelaçamento entre os dois tipos de pesquisa no âmbito das necessidades humanas.

3.2.2 O paradigma moderno

A dinâmica atual impera uma relação de complementaridade e não de dependência sobre o campo das pesquisas básicas e aplicadas ou da Ciência e tecnologia. Esta é a ideia secundada por Stokes (2005, p.39), ao considerar que:

> se a pesquisa básica pode ser diretamente influenciada por objetivos aplicados, então a ciência básica não pode mais ser vista apenas como uma remota geradora de descobertas científicas, movida a curiosidade, descobertas a serem posteriormente convertidas em novos produtos e processos pela pesquisa aplicada e pelo desenvolvimento, nos estágios subsequentes da transferência tecnológica.

Com esta afirmação, Stokes (2005) renuncia o modelo linear da relação Ciência-Tecnologia e a visão simplista que coloca a tecnologia como embrião da Ciência, mostrando que muitas práticas são frutos de inovações tecnológicas destituídas da ciência pura. Aliás, muitas vezes, são soluções inovadoras de caráter tecnológico que motivam a busca por um entendimento complexo de certas realidades sociais ou científicas. O próprio status atual da ciência se deve à inovação tecnológica. Para o autor, foi Vannevar Bush quem cultivou a ideia de que a ciência é a principal fonte da inovação tecnológica e que prevalece atualmente em alguns modelos de análise da relação Ciência-Tecnologia através do processo da transferência de tecnologia: Pesquisa básica – Pesquisa aplicada – Desenvolvimento – Produção e Operações.

De referir que para o processo de construção do conhecimento, as fronteiras entre a Ciência e Tecnologia se manifestam de forma intermitente e difusa. Equivale isto dizer que a busca pelo entendimento é tão necessária quanto a procura de soluções para satisfazer as necessidades particulares. Este fato é comprovado pela coexistência pacífica de diversas linhas de pesquisa dentro da CI, algumas relacionadas a questões epistemológicas sobre a natureza das coisas, outras voltadas para a tecnologia, visando à aplicação.

Segundo Stokes (2005), a relação entre a Ciência e Tecnologia geralmente ocorre numa simbiose latente entre os métodos da pesquisa básica e da pesquisa aplicada e que, não raras vezes, resulta em conflito na Ciência. É porque quando se trata de abordagens de caráter científico, a busca pelo entendimento pode culminar com soluções que visam resolver problemas específicos, tal como aconteceu com os estudos de Pasteur que, procurando entendimentos sobre os processos de doenças

e de microbiológicos, desembocaram em soluções para prevenir a deterioração na produção de vinagre, cerveja, vinho e leite, entre outras aplicações.

A aparente categorização e simbiose latente entre a pesquisa básica e a aplicada são paradoxalmente aceites pelas ciências modernas e pós-modernas. À semelhança da Inteligência Artificial e da Ciência da Computação, entre outros campos, a CI encontra a sua validade nessa característica que traduz as suas linhas de pesquisa ou das suas disciplinas. Existe uma tentativa de separação dos métodos e dos processos da pesquisa básica e aplicada que culmina numa abordagem difusa do campo no todo, pelo respectivo caráter social e aplicado. Para satisfazer às necessidades dos usuários, existe uma busca pelo entendimento do contexto do próprio usuário, da informação, dos fluxos informacionais e dos processos de comunicação. Talvez esta questão não suscite muitas reflexões na área da CI pelo caráter de insignificância, já que a sua gênese está aliada e comprometida com a relação Ciência-Tecnologia como um microorganismo no qual a busca incessante pelo entendimento só se justifica no uso.

3.3 A Ciência da Informação como campo científico

A noção de campo científico está enraizada nas características das ciências tradicionais, como Matemática, Física, Biologia, Geografia, Química e Filosofia, na busca pelo entendimento do universo e solução de problemas da humanidade. Tais ciências desenvolveram-se através de métodos que visavam respostas a um conjunto de inquietações, corporificadas por meio de objetos de estudo. Esta divisão do conhecimento humano é por si contraditória, pois parte da premissa de que a especificidade das partes coaduna na compreensão geral de todo o universo. Porém, sabe-se que o diálogo entre diferentes campos sobre o mesmo assunto é tão complicado e complexo, tão quanto complexo é o diálogo entre diferentes disciplinas do mesmo campo científico.

A CI, à semelhança de outros campos científicos emergentes, teve uma gênese conturbada, fundamentalmente assente no descrédito tanto da sua estrutura como campo científico, como da sua intencionalidade traduzida através do seu objeto de estudo e dos seus métodos de pesquisa. Para demonstrar o caráter científico da CI como campo, talvez seja importante e oportuno recorrer a Pierre

Bordieu, não apenas por critérios de afinidade, mas principalmente pela abordagem que fez sobre o campo numa perspectiva social.

Na Conferência e debate, organizados pelo grupo *Sciences en Questions*, em 1997, em Paris, Bordieu (2004) fez uma reflexão sobre a lógica do mundo científico. Para o autor, o campo no sentido científico é um microcosmo ou espaço de atuação entre agentes (indivíduos e instituições) que produzem, reproduzem e difundem a ciência. Este espaço mesmo que sofra imposições externas, possui leis próprias, através das quais resiste a essas pressões externas e que lhe conferem o respectivo grau de autonomia. Equivale por outras palavras dizer que o indicador de autonomia de um campo é a sua capacidade de quebrar as resistências exercidas pelo exterior ou torná-las imperceptíveis. Por isso, quando os problemas políticos são notórios no seio das disciplinas de um campo isso se torna um indício da fraca autonomia desse campo no todo.

Uma das características do campo é a existência de relações de forças consubstanciadas através das relações objetivas entre os seus agentes. No caso do campo científico, a pesquisa, produção literária e o lugar da difusão dessa produção são consequências dos interesses e da posição tomada pelos seus agentes[19], ou seja, os agentes "fazem os fatos científicos", e consequentemente, o campo científico. Conforme Bordieu (2004) acrescenta, essa imposição é indireta e depende do capital científico de alguns agentes, em detrimento da maioria. Deste modo, um desafio ou contraste à ordem de um campo pode acarretar o descrédito dentro da estrutura desse campo.

Com base nas características do campo referenciadas, afirma-se que a CI é um campo científico, pois é um espaço de produção e disseminação do conhecimento científico. Tal como os outros campos, possui forças externas a partir das quais as suas instituições e agentes se estabelecem e se desenvolvem com maior ou menor influência dentro da estrutura.

Saracevic (1996) aponta que qualquer campo científico é definido a partir dos problemas que se predispõe a resolver e no caso da CI, existe um vínculo permanente entrelaçado entre a pesquisa empírica e a prática profissional. Esta característica justifica-se na origem do campo, enquanto proposta inovadora na

[19] A estrutura das relações objetivas entre os agentes do campo determina a posição ocupada por esses agentes dentro do respectivo campo e a sua vigorosidade condiciona o objeto de pesquisa nesse campo (BORDIEU, 2004).

ciência de uma área técnica específica em resposta ao problema do excesso da informação.

Para Bordieu (2004, p.26), cada campo tem a sua forma específica de capital, o capital simbólico que se constitui a partir de “atos de conhecimento e reconhecimento”, através de citações, prêmios, etc. Tal reconhecimento se traduz numa espécie de estratégia de jogo pela parte do agente que consegue se antecipar às pesquisas sobre temas ou assuntos que terão maior interesse e desdobramento no futuro.

Sobre as propriedades específicas dos campos científicos, Bordieu (2004, p.30) aponta que “quanto mais os campos científicos são autônomos, mais eles escapam às leis sociais externas”. Por outras palavras, campos autônomos apenas sofrem críticas e contrastes através de argumentos e demonstrações; enquanto que os campos sem normas e estruturas rígidas são facilmente manipulados por meio de intervenções não científicas (construções sociais).

Bordieu (2004) destaca duas espécies de capital científico, por um lado o temporal ou político, institucional e institucionalizado, ligado à ocupação de posições estratégicas nas instituições científicas e aos meios de produção e reprodução; por outro, o poder específico ou de prestígio pessoal resultante do reconhecimento do agente pelas suas qualidades científicas. Enquanto o capital institucional se fortalece através de estratégias políticas de participação em reuniões, eventos, bancas, comissões, entre outras, ele se torna legítimo e inquestionável; em contraste, o capital pessoal depende da contribuição reconhecida do agente em prol da ciência, por meio de invenções ou pesquisas e publicações em órgãos de renome, e por isso, se torna passível de contraste e crítica[20]. O autor acrescenta que, mesmo diferentes, os dois capitais podem coexistir na mesma estrutura do campo e podem ser transponíveis. Deste modo, conforme Bordieu (2004, p.43), o campo científico é um “espaço de pontos de vista”, pois pode ser observado em função da posição, às vezes imperceptível, ocupada dentro do campo e das contribuições dos polos da estrutura autonomizante. E uma constatação pessoal e interessada pode ser validada como universal em função da posição tomada pelos sujeitos.

[20] Além das características citadas, o capital institucionalizado depende apenas dos processos burocráticos de nomeação ou eleição do agente; enquanto que o pessoal depende do carisma ou aceitação do agente (BORDIEU, 2004).

Sobre a autonomia com base no qual o campo resiste às influências políticas ou sociais externas, é importante realçar que a CI é uma ciência nova e em fase de estruturação. Como tal, ainda sofre influências externas porque o capital científico institucional pesa mais sobre os poucos pesquisadores renomados da área. Ademais, especificamente no Brasil, há uma tendência generalizada de influência e de controle da academia através das instituições de fomento de pesquisa, como Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), que além de classificar os programas de pós-graduação, avaliam o desempenho anual de cada programa, por meio de vários requisitos, dentre eles a publicação nos periódicos por elas também classificados. Alia-se a este fato a editoria da maior parte desses periódicos por pesquisadores consagrados na área que, tanto de forma implícita ou explícita, contribuem para a temática das publicações pelo critério de aceitabilidade ao escopo do periódico.

O problema de autonomia e das forças existentes no campo é resumido por Saracevic (1996) como sendo o problema da ecologia informacional. A ecologia enquanto espaço de coexistência entre produtores, editoras, revistas, instituições de fomento de pesquisas, etc., é caracterizada por conflitos de interesse e que indubitavelmente afetam o modo de produção do conhecimento ou a forma pela qual determinado campo científico se desenvolve. A CI, por sua vez, reflete esta problemática que caracteriza todos os campos científicos ou a sua maioria. Conforme se referiu anteriormente, a liberdade científica do campo em relação à política só se consegue pelo fortalecimento da ação do respectivo corpo de pesquisadores, através de atos de conhecimento e reconhecimento. Tal reconhecimento leva tempo para se concretizar e é passível de contrastes, à semelhança da própria ideologia ou temática advogada pelo autor durante a construção do seu capital.

3.4 Gênese da Ciência da Informação

A gênese da CI ainda é objeto de controvérsias. Contudo, a maioria dos pesquisadores a situa em dois períodos: um anterior ao final da Segunda Guerra Mundial e outro posterior a guerra.

3.4.1 Período pré – Segunda Guerra Mundial

O professor Aldo Barreto é um dos poucos autores que vislumbram os traços da CI antes do período consagrado pela explosão informacional no final da Segunda Guerra Mundial. Na sua historiografia, Barreto (2008, p.2) considera que a gênese da CI subjaz da inovação tecnológica e que "o fluxo de informação e sua distribuição ampliada e equitativa" sempre foi uma das aspirações do homem, desde a fase das pinturas rupestres, passando pela invenção da escrita e "divinização"[21] da informação, até a Internet. Neste contexto, embora de forma indireta, percebe-se que este entendimento do autor está relacionado ao papel da CI sobre a busca de soluções sobre a produção, a organização e a disseminação da informação, visando uma acessibilidade e usabilidade ótimas. Daí, a ligação da CI com o papel das primeiras redes de saber universal, que tinham o mesmo propósito. Eis alguns exemplos das redes de saber citados pelo autor: a Academia de Lince, de 1603; a Royal Society, de 1660, e reconhecida em 1662; as Academias de Ciências de Londres, de París e de Berlim, criadas em 1665, 1666 e 1700, respectivamente, e alguns jornais científicos.

Além das academias e das enciclopédias, estas enquanto redes de conhecimento distributivo, é na rede universal de conhecimento de Otlet e La Fontaine que se vislumbra o marco importante da CI:

> os determinantes colocados anteriormente permitem refletir com mais liberdade a questão da ciência da informação em um desenrolar histórico descritivo, que tem somente a validade no contexto do desenvolvimento histórico da informação e conhecimento. Permitem ainda verificar que o ideal do acesso ao conhecimento livre e para todos não surgiu com a Internet (BARRETO, 2008, p.3).

De fato, muitas questões que norteiam as diretrizes de atuação da área foram incorporadas ou retomadas a partir dos estudos de Paul Otlet. Esta percepção é compartilhada por Rayward (1994), ao considerar que a maioria das abordagens sobre a CI negligenciam os trabalhos de Otlet. Para o autor, foi Otlet que desenvolveu o sistema complexo de organização para integrar bases de dados bibliográficos, imagéticos e textuais que atualmente funcionam como hipertexto.

Rayward (1994) é pragmático ao assumir o Tratado de Documentação de Otlet de 1934, como a primeira obra da CI, pelo fato de ser a primeira abordagem

[21] Segundo Barreto (2008), a informação foi mantida restrita por diversas épocas, tanto pela relutância de questionamentos e interpretação das escrituras sagradas, como pela exiguidade dos usuários com habilidades para lidar com as informações.

sobre a documentação como sistema integrado de organização da informação, através de cartões padronizados que representassem as facetas do documento e que pudessem ser armazenadas no banco de dados e recuperadas por um sistema de busca multifacetado.

Para Barreto (2008), Vannevar Bush “pode ser considerado o pioneiro da ciência da informação e 1945 sua data fundadora pela publicação do seu artigo”, no qual, pela limitação humana de lidar com o volume excessivo de informação, propôs o memex para armazenar e recuperar a informação, através da associação de palavras. Conforme o autor acrescenta, as ideias de Bush foram discutidas em Londres, na Conferência da Royal Society de 1948, culminando com a criação da CI como um novo campo científico para lidar com a nova situação.

A gênese da CI também pode ser associada à *Conferences on Training Science Information Specialists*, realizadas em 1961 e 1962, nos EUA, envolvendo docentes e bibliotecários da Universidade de Geórgia, com o intuito de treinar especialistas de informação no contexto norte-americano. Pelo fato de a referida conferência ter sido realizada apenas numa visão formativa no contexto norte-americano, este marco é contestado como de índole global para a CI como campo científico (BARRETO, 2008, p.4).

3.4.2 Período pós – Segunda Guerra Mundial

A maioria dos autores aponta a explosão informacional, essencialmente marcado pela quantidade de relatórios que documentavam a Segunda Guerra Mundial e as conferências para debater a solução tecnológica proposta por Vannevar Bush como o marco da CI.

A CI, segundo Capurro e Hjorland (2003), nasceu nos 50, aliada ao desenvolvimento e disseminação de uso de computadores, no final da Segunda Guerra Mundial. Esta periodicidade também é defendida por Wersig (1993).

Na Conferência Internacional sobre conceções de Biblioteconomia e Ciência da Informação: perspectivas históricas, teóricas e empíricas, realizada na Universidade de Tampere, na Finlândia, Wersig (1993) apresentou uma palestra, abordando a Ciência da Informação a partir do respectivo objeto, o estudo do conhecimento pós-moderno. Neste sentido, a percepção do autor coincide com a de Barreto (2008), sobre a gênese da CI aliada à inovação tecnológica. Contudo,

contrariamente a Barreto, o autor considera que o início da CI foi marcado pela explosão informacional no fim da Segunda Guerra.

Para Wersig (1993), ao contrário das ciências clássicas cuja gênese está enraizada na busca do entendimento completo de como o mundo funciona, a CI nasce da necessidade de desenvolver estratégias para resolver problemas causados por ciências e tecnologias do clássico. Esta ação intervencionista da CI impõe mudanças significativas no conhecimento, através de mudanças de paradigmas, por exemplo, da visão técnico-sistema para visão usuário/humano.

É na referida mudança paradigmática que se traduz a real necessidade e contribuição de uma ciência emergente como a CI, pois o foco em aspectos que transcendem a tecnologia, como usuário final e as suas especificidades, tanto na origem, manifestações, efeitos e comunicação, como na busca e uso da informação, suscita um interesse trans-inter-multidisciplinar por parte de outras áreas do conhecimento. Os cientistas da computação, os engenheiros, os *designers*, entre outros, passam a desenvolver produtos sob o olhar da CI, no que tange aos requisitos de usabilidade, acessibilidade, inclusão, interfaces amigáveis, etc. A própria filosofia às vezes se depara com o dilema da necessidade de reformulação de alguns conceitos sobre o conhecimento e a moral, com a prática de compartilhamento no uso e na produção da informação e pelas variações linguisticas que impõem novos e diversificados modos de interação e representação.

Ao contextualizar a origem da CI a partir do problema da informação, Wersig (1993) considera como marco inicial da área o período pós-guerra. Conforme o autor aponta, a CI nasceu no século XX com a documentação que procurava solucionar o problema do “dilúvio da literatura”, passando pelo foco na recuperação da informação e culminou com aspectos complexos da tecnologia, tornando-se CI. Para o autor, a CI nasce na mudança do papel do conhecimento para indivíduos, organizações e culturas, com as manifestações do pós-modernismo.

A mudança do papel do conhecimento foi influenciada, em grande parte, pelo advento da tecnologia. Como Wersig (1993) aponta, um dos efeitos desta mudança foi a despersonalização do conhecimento, através da tecnologia da comunicação, onde o conhecimento pessoal e transmitido pela oralidade, com a invenção da impressão, passou a ser traduzido em suportes e disseminado pela multiplicidade de usuários. Neste período, a documentação teve um papel importante na criação de catálogos alfabéticos.

Um dos problemas apontados por Wersig (1993) e que ainda são objetos de análise na CI se prende com a fonte e a apropriação da informação. Com a tecnologia, a fonte de informação torna-se cada vez menos evidente. Esta situação é comum na Web e acaba por ofuscar os ganhos da referida despersonalização do conhecimento, já que informações importantes são preteridas pela falta da clareza da respectiva fonte ou informações não confiáveis são usadas para assuntos complexos e sensíveis. De igual modo, na mediação da informação, algumas organizações se apropriam dessa informação como se delas se tratasse, prejudicando a visibilidade dos reais autores, o acesso e uso para outras finalidades.

O segundo efeito da mudança do papel do conhecimento apontado por Wersig (1993) está relacionado à credibilidade do conhecimento, através da tecnologia que permite a observação. Se, no passado, o conhecimento produzido poderia ser testado por muitos pesquisadores, atualmente, com a sofisticação das teorias, metodologias e técnicas de coleta e processamento de dados, o conhecimento produzido fica restrito a certos grupos ou comunidades.

> Se mais conhecimento se torna despersonalizado, por outro lado, mais conhecimento tem de ser acreditado […]. A situação se tornará mais complicada com as novas tecnologias […]. Por isso, cada vez mais temos que ter cuidado com os dados de observação em dois aspectos: em primeiro lugar temos que aceitar a tecnologia que originou os dados, e depois levar em conta o que poderia ter acontecido com os dados brutos em processo de transformação. Para aceitar o conhecimento temos que ser muito críticos em relação às tecnologias de coleta e manipulação (WERSIG, 1993, p.232).

Por isso, a CI procura soluções para esse problema através da política de acesso a dados abertos. Segundo o OpenGovData (2007), para que o dado seja aberto, deve reunir oito princípios, a saber: ser completo, primário, atual, acessível, compreensível por máquinas, não discriminatório, não proprietário e livre de licença. Assim, pesquisas afins podem ser desencadeadas, envolvendo outras ferramentas de manipulação. Além de garantir a prova das conclusões tiradas pelo autor dos dados, o processo permite a reutilização dos dados para outras finalidades. Por isso, torna-se imprescindível a descrição detalhada de todo o processo envolvido na sua obtenção.

O terceiro efeito da mudança do papel do conhecimento segundo por Wersig (1993) é a fragmentação do conhecimento por meio da tecnologia de apresentação. O autor aponta três razões dessa fragmentação: volume excessivo; autonomização das áreas de conhecimento, através de padrões próprios de cada campo ou

disciplina e sistemas finais, enquanto ideologias de pensamentos sobre o mundo. Estes fatores culminam com a “diversificação de tecnologias de apresentação de conhecimento”, levando as pessoas a um conhecimento fragmentado de acordo com as suas crenças, culturas e áreas de formação ou atuação.

O quarto efeito da mudança do conhecimento segundo Wersig (1993) é a racionalização do conhecimento, através da tecnologia da informação. Uma das razões para essa racionalização se prende com o aumento de tecnologias para a disseminação e acesso à informação e das potencialidades de uso do conhecimento para a solução dos problemas da sociedade. O pensamento de Engelbart (2003) está enraizado neste efeito da mudança, isto é, da transformação social pelo uso da tecnologia.

Engelbart (2003) analisa a complexidade da tecnologia e dos problemas humanos, propondo o uso de computadores para aumentar o intelecto humano. Desde 1960, abordava questões sobre a inovação descontínua que visava mudanças no modo como se usava os computadores, para a colaboração online ou computação interativa, pois para o autor, a produção de conhecimento não é uma atividade individual, mas de grupo. A maioria das suas pesquisas que na época faziam pouco sentido para a humanidade, pelo estágio em que a computação se encontrava, atualmente se encontram em prática nos nossos processos de produção e uso da informação. O sistema online desenvolvido entre 1960 e 1970 no seu próprio laboratório *“Augmentation Research Center”* é prova das suas aspirações. Este sistema facilitou a criação de bibliotecas digitais, o armazenamento e recuperação de documentos eletrônicos utilizando hipertexto, a criação de novas interfaces gráficas que permitiram a leitura de e-mails, novas opções de processamento de texto e vídeo teleconferência. Muitos conceitos do sistema online foram aplicados na ARPANET e continuam em uso na Web.

Wersig (1993) também apela à necessidade de criação de uma base teórica capaz de atender a área da CI. Neste sentido, propõe um modelo básico de redefinição ou ampliação de conceitos amplos para os propósitos internos da CI, a reformulação de conceitos que são tratados em outras áreas com o viés da CI e o entrelaçamento dos modelos com os conceitos reformulados.

Na mesma conferência de Tampere, na Finlândia, Saracevic (1996) traçou o perfil da CI a partir de três características: a interdisciplinaridade na natureza da CI, a ligação da CI à tecnologia e o papel ativo da CI na evolução da sociedade da

informação. Neste sentido, o pensamento do autor vai de acordo com a corrente que atribui a gênese da CI ao movimento científico que procurava solucionar o problema informacional no período pós-guerra. O artigo *"As we my think"*, publicado por Vannevar Bush em 1945 caracterizou, segundo o autor, o embrião da CI, pela proposta tecnológica de associação automatizada de ideias na recuperação para colmatar o problema de excesso da informação e pelos diversos esforços conjugados posteriormente em prol da importância estratégica da informação para indivíduos, grupos e organizações.

A ideia da recuperação da informação que segundo Saracevic (1996) caracterizou a essência da CI na sua origem, desenvolveu-se sob três vertentes: descrição intelectual da informação, especificação intelectual da busca e especificidade sobre sistemas e técnicas a empregar. Neste sentido, o foco da CI na recuperação introduziu mudanças, como: de cartões perfurados aos CD-ROMs, de sistemas simples aos interativos, de bases documentais às bases do conhecimento, de textos escritos à multimídia, entre outras. De igual modo, pesquisas empíricas foram desencadeadas sobre a natureza da informação, uso e usuários, interação homem – máquina, relevância, avaliação dos sistemas de recuperação, entre outras.

As mudanças na recuperação e as pesquisas sobre a informação culminariam na indústria da informação que, embora não singularmente responsável, contribui significativamente para a atual fraca crítica da ciência, não só pela apropriação e controle, como também pela banalização da informação. Na atual Web, muitas informações consistentes não são recuperadas devido a políticas centralizadoras dos locais de publicação e difusão, e em contrapartida informações fúteis são frequentemente propagadas para manter o usuário às grandes mídias.

A CI como campo científico emergiu nos anos 50 e os seus problemas estão sintetizados tanto na natureza, manifestações e efeitos da informação e conhecimento, como nos processos da comunicação e uso da informação:

> A CI é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais (SARACEVIC, 1996, p.7)

Esta conceituação traduz uma essência do conceito de Borko (1968) e que revolucionou a área e acrescenta aos problemas da CI, a necessidade da informação, os usos e os contextos social e institucional.

3.5 O caráter interdisplinar da Ciência da Informação

A relação interdisplinar da CI com outras áreas se justifica pela própria natureza do campo e pelos problemas que a área se predispôs a resolver. Como afirma Wersig (1993), a CI surgiu para resolver os problemas de outros campos; além do problema “informação” ter origem em outras áreas, também é por elas tratado, por isso, algumas soluções utilizadas atualmente foram incorporadas a partir dessas áreas. Por outras palavras, a natureza, manifestação e efeitos da informação, bem como o processo da comunicação são assuntos que além de envolverem a complexidade humana, não caberiam na análise de um único campo científico.

Saracevic (1996) também analisa a interdisciplinaridade da área, a partir da relação com a Biblioteconomia, com a Ciência cognitiva, com a Ciência da Computação e com a Comunicação. Em relação à Biblioteconomia, de acordo com Shera (1972, apud SARACEVIC, 1996), as bibliotecas contribuem para o sistema total de comunicação na sociedade. Por isso, a CI e a Biblioteconomia compartilham o papel social e o comprometimento com os problemas da efetiva utilização dos recursos bibliográficos. Importa realçar que essa interdisplinaridade não retira, de algum modo, a peculiaridade que caracteriza cada área no seu papel social de conhecimento e comunicação humana.

A interdisplinaridade entre a CI e a Biblioteconomia é frequentemente notória no aspecto técnico de preservação e comunicação da informação. Se atualmente a CI logra sucessos na recuperação interativa da informação, atendo a especificidade dos usos e usuários, tal sucesso esta a mercê dos contributos da Biblioteconomia numa camada inferior e, às vezes imperceptível, principalmente através da representação descritiva e temática dos recursos informacionais. Em cada contexto, a qualidade de um sistema de recuperação da informação depende dos processos e técnicas usadas pela Biblioteconomia no tratamento da informação, desde a catalogação, metadados, indexação, classificação, entre outros. Atualmente, os laços entre as duas áreas são estreitos e fortificados no leque de ações que visam à

preservação, integridade e portabilidade face à obsolescência das plataformas tecnológicas. Também se estendem para a interoperabilidade dos sistemas para agregação de um valor maior à informação.

Com a Ciência da Computação, Seracevic (1996) aponta que a relação se baseia na automação, através da aplicação dos computadores e da computação na recuperação da informação. Enquanto a Ciência da Computação trata os algoritmos que transformam os códigos binários em informação, a CI trata essa informação e a sua comunicação para os humanos. Conforme o autor acrescenta, muitos pesquisadores da Ciência da Computação contribuíram para o desenvolvimento da CI, alguns dos quais com pesquisas que inicialmente não tinham ligação com a CI na sua plena natureza, como: sistemas inteligentes, hipertexto, interação homem – computador, entre outras.

A CI também tem uma forte relação interdisciplinar com a Ciência Cognitiva que, segundo Saracevic (1996), parte da premissa da inteligência como produto dos processos cognitivos, do cérebro e da mente e suas manifestações. A CI e a Ciência Cognitiva mantêm estritas ligações através da Inteligência Artificial.

Atualmente na CI, a recuperação da informação por meio de sistemas inteligentes ou que simulam o modo operacional da mente humana subjaz do contributo da Inteligência Artificial. As redes neurais são sistemas que permitem o aprendizado da máquina, simulando o processo de associação neural do cérebro humano para potencializar futuras recuperações da informação. Essa contribuição da Ciência Cognitiva é perceptível nos motores de busca atuais, como o Google, que permitem a associação dos termos de busca para localizar documentos similares para um usuário ou a associação dos termos entre diversos usuários.

A relação entre a CI e a Comunicação segundo Saracevic (1996) se baseia na informação como fenômeno e comunicação como processo. De fato, as duas áreas até podem ser confundidas principalmente no que tange à inquietude da necessidade de transformação social através da informação e dos modelos nos quais ocorre tal transformação. Porém, trata-se de dois campos distintos e que caminham juntos em prol do conhecimento coletivizado para a ação.

A interdisciplinaridade não é uma característica singular da CI. Atualmente, mesmo os campos tradicionais recorrem a ela para entender e explicar fenômenos emergentes, alguns dos quais propiciados pelas manifestações pós-modernas. Ela se justifica e se torna necessária sempre que se pretende atingir o objetivo das

ciências, na construção e difusão do conhecimento para a sociedade. Por exemplo, o Direito recorre à Antropologia, Sociologia e Moral para entender e reformular o conceito de casamento no direito civil, que atualmente inclui a união entre pessoas do mesmo sexo.

3.6 O Objeto de estudo da Ciência da Informação

O objeto da CI (informação) ainda suscita algumas contestações que refletem parte da problemática do campo. Segundo Le Coadic (2004), o problema é que a sua incorporação dentro do campo está relacionada a outros conceitos, como conhecimento e comunicação. Estes conceitos também são objetos de outros campos científicos. Um exemplo elucidativo é o da “teoria da informação” na Matemática e da “teoria de código genético” na Biologia. Contudo, o objeto informação na CI está relacionado à cognição e à comunicação humanas e transcendem qualquer objeto de outros campos pelo caráter universal e social. Para entender melhor o objeto da Ciência da Informação, parte-se do artigo de Buckland.

Buckland (1991) estabelece uma tripartição do objeto da CI baseada em três momentos da informação: informação como processo, informação como conhecimento e informação como coisa.

A informação como processo corresponde ao ato de informar, comunicar ou falar algo, isto é, ao contato intrínseco entre os sujeitos ou entre o sujeito e o objeto. Rayward (1994) aponta que o sistema de hipertexto, com base no qual a Web se desdobra, consiste em fragmentos de textos que constituem os nós e nos vínculos associativos entre diferentes nós. Assim, a informação como processo equivale às associações que o usuário faz na busca de informações para satisfazer as suas necessidades, através de *links* que conectam diversos nós ou facetas do mesmo documento, ou ainda nós entre diferentes documentos. Essas conexões do usuário perfazem algum tipo de informação, na medida em que são semânticas e representam os seus conceitos, as suas ideias, as suas estratégias, a sua cognição e as suas características. Por isso que a leitura do hipertexto não é sequencial ou linear, depende da estrutura mental de cada usuário.

A informação como conhecimento resulta da apreensão ou percepção de algo durante os processos anteriores, ou seja, a transformação que ocorre no sujeito na relação com outro sujeito ou com objeto. A informação como coisa é a exteriorização

do que foi percebido ou apreendido em objetos informativos que contêm a qualidade de conhecimento comunicado, comunicação, informação ou algo informativo. A informação como coisa apresenta a peculiaridade de ser a representação tangível de códigos, sinais, dados, textos, filmes, etc., usados pelos sistemas de informação, bibliotecas, museus, com a finalidade de informar o usuário (processo) para que possa adquirir um conhecimento (informação como conhecimento). Equivale isto dizer que a manipulação, a operacionalização, o armazenamento, a representação e a recuperação só são possíveis com a informação física ou como coisa (BUCKLAND, 1991).

A peculiaridade da informação registrada dentro da CI coaduna com a percepção de Le Coadic (2004, p.4), ao afirmar que “a informação é um conhecimento inscrito (registrado) em forma escrita (impressa ou digital), oral ou audiovisual, em um suporte”. Neste sentido, o documento é o objeto que contém a informação.

A CI teve uma origem conturbada, pelo fato de alguns não a aceitarem como campo científico devido à falta de clareza em relação aos seus métodos e objeto de pesquisa. Wersig (1993) considera que a CI é uma ciência pós-moderna e alguns não a aceitavam como campo científico porque não tinha único objeto e método exclusivo. O objeto do campo, informação, não era aceito porque, além da falta da clareza sobre o seu conceito, parecia assunto tratado por outras áreas. Para o autor, como ciência pós-moderna, a CI não precisa de um método específico porque lida com uma diversidade de problemas, causados pelas ciências clássicas e pela tecnologia, nos quais não caberia apenas um único método. Deste modo, a CI não deve ter uma estrutura à semelhança dos campos tradicionais. Para lidar com os seus propósitos, o autor propõe o desenvolvimento dos seguintes métodos para a CI:

- Análise da comunicação em contextos organizacionais;
- Análise de estruturas de conhecimento (sistemas baseados em conhecimento);
- Avaliação de tecnologias de informação e comunicação;
- Avaliação do efeito informacional (em particular visual) e das estruturas de apresentações do conhecimento.

Estes métodos podem ser aliados a metodologias de abordagens, como estudos de caso e pesquisa social qualitativa.

Segundo Wersig (1993, p.233), informação é conhecimento em ação, isto é, “esse conhecimento tem de ser transformado em algo que suporta uma ação específica dentro de uma situação específica”. Para o autor, é com base neste princípio que a CI se desenvolveu, reforçando os efeitos problemáticos de tecnologias sobre o uso do conhecimento, ao criar no passado sistemas que garantem a recuperação fracionada da informação. Atualmente, a CI se empenha em soluções para lidar com a despersonalização do conhecimento e sua fragmentação, bem como no desenvolvimento de outras formas adequadas de racionalização que estejam abertas a todos os tipos de conhecimento.

A informação gera o conhecimento, por isso é procedente a afirmação de que outras áreas científicas também trabalham com a informação. Contudo, se essas áreas tratam de informação no contexto específico dos seus métodos e do seu objeto, a CI vai muito além disso. O fundamento pode ser encontrado em Kochen (1974, apud SARACEVIC, 1996), que analisa a CI pelo papel centralizador da recuperação, a partir do enfoque do usuário, da informação em si materializada em suporte como documento e dos tópicos da representação. Por outras palavras, além da informação que visa construir conceitos ou conhecimento específico de um domínio, a CI trabalha a informação no todo, através de processos e métodos desde a produção, organização, armazenamento, disseminação até a recuperação e uso, atendendo às características de cada fase e à especificidade do sistema, da técnica, do suporte, da tecnologia e dos intervenientes de cada etapa.

O objeto informação justifica o caráter tecnicista no qual a área da CI teve origem, através da Bibliotecnomia e Documentação. Mesmo considerando as divergências que caracterizaram o percurso das duas áreas, o objetivo comum de tornar acessível a informação que no passado esteve reservada aos eruditos, ao clero e nobreza dos mosteiros e aos filósofos medievais, prevalece e ganha maior repercursão com as tecnologias aliadas ao contexto da comunicação. Por outras palavras, desde a Biblioteca de Alexandria, passando pela imprensa de Gutenberg, pela Documentação de Otlet e pela classificação de Dewey até a Ciência da Informação, a atuação dos profissionais da informação e o emprego das técnicas e processos de representação e recuperação da informação são movidos pela necessidade de tornar a informação registrada, manipulável, acessível para o uso e agregação de valor, de modo a permitir a produção de novos conhecimentos.

Capurro e Hjorland (2003) conceituam a informação a partir das suas características de novidade e relevância, necessárias para que o sujeito a selecione e a interprete, de modo a transformar a sua mente pelo conhecimento comunicado. Estas características subsumem-se à essência da CI, enquanto campo científico voltado para os processos de produção, organização, armazenamento, disseminação, acesso e uso da informação. Porém estes conceitos suscitam divergências conceituais e terminológicas. O problema é que a novidade e relevância estão relacionadas aos jogos linguísticos pelos quais cada usuário constrói as respectivas intersecções. Assim, tornam-se intrínsecos e difíceis de exteriorizar por qualquer modelo de representação humano. Por isso que no âmbito subjetivista da representação e recuperação é difícil determinar o que é informação.

A peculiaridade da CI reside em atos e processos que se antecipam aos usuários, através de delineamento de estratégias e ações que consubstanciem as suas necessidades informacionais. Por outras palavras, a CI projeta nos sistemas de recuperação da informação as necessidades informacionais, as estratégias de busca, o *findability*, antes da própria interação do usuário com o sistema. Se esta ação, por um lado, traduz uma das maiores contribuições da área, por outro, elucida o seu maior desafio. Adivinhar aquilo que o usuário precisa para satisfazer as suas necessidades informacionais, as estratégias que irá usá-las para alcançar a informação necessária e principalmente o julgamento que fará acerca da informação encontrada, são tarefas laboriosas e revestidas de imprecisão e ambivalência. Estes aspectos são discutidos pormenorizadamente no Capítulo quatro.

A Web como tecnologia de disseminação da informação, modifica tanto a noção de tempo e espaço da transferência da informação entre o emissor e receptor que passa a ser real ou online, como o acesso, dando a possibilidade do julgamento da relevância na interação. Também permite a simbiose de várias linguagens em um único recurso informacional (som, texto e imagem) e a interação concomitante entre diversos conteúdos e fontes informacionais (BARRETO, 2008, p. 5).

Barreto (2008) estabelece uma distinção entre informação e conhecimento a partir de dois níveis de fluxos de informação, por um lado os fluxos internos e baseados na produção, visando à organização e controle do sistema de armazenamento e recuperação, por outro, os fluxos externos que consubstanciam a exteriorização do pensamento do sujeito para o receptor que o assimila como conhecimento, com base nas suas condições particulares (cultura, espaço, outras

informações associáveis, etc.). Neste sentido, o autor aponta para três tempos no desenvolvimento da CI e que, de alguma maneira, são estanques:

- Tempo gerência da informação (de 1945 a 1980);
- Tempo relação informação e conhecimento (de 1980 a 1995);
- Tempo do conhecimento interativo (de 1995 a atualidade).

Outro aspecto referenciado por Barreto (2008) na historiografia da CI é o cognitivismo, enquanto estudo do comportamento de assimilação do conhecimento por seres humanos, máquinas e a interação entre os dois. Deste modo, a informação para a CI passou a revestir a característica da intencionalidade de gerar conhecimento no indivíduo.

Na argumentação de Barreto (2008), percebe-se o olhar de Buckland (1991) no que tange à conceituação da informação (como processo, como conhecimento e como coisa). Por isso que o conhecimento, enquanto modificação do estado mental ou intelectual do sujeito, por meio da apropriação da informação, e consequente representação, essa informação depende da materialidade (como coisa), para que seja tangível. Dentro do contexto de relação informação – conhecimento, importa destacar a questão da "ação" como assimilação da informação para aspectos (políticos, econômicos e sociais) peculiares do sujeito no meio em que se encontra, de acordo com as suas especificidades (mentais, culturais, acadêmicas, religiosas, entre outras).

Em síntese, a CI é um campo de conhecimento que se situa no rol das ciências emergentes, frutos da Revolução Científica e Tecnológica. A maioria dos aspectos abordados na área são anteriores ao período pós-guerra, mas este período é considerado como principal marco pela institucionalização da CI como campo científico. A CI tem uma natureza interdisciplinar, tecnológica e social, justificadas pelo objeto de estudo a "informação", visando à recuperação e usabilidade, por isso, está inserida no paradigma da complexidade, como ciência em constante transformação em função das situações concretas de cada estágio da sociedade. O comprometimento com a dimensão social do usuário é frequente e notório através das respectivas linhas de pesquisa, em observância às especificidades e contextos de comunicação e interação, para uma efetiva transformação ou conhecimento em ação. O Capítulo 4 analisa algumas questões relacionadas com a dimensão social do usuário, envolvidas na representação da informação.

4

Representação da Informação

4.1 Visão sobre o Capítulo

A CI, conforme se referiu anteriormente, tem uma natureza interdisciplinar, social e tecnológica, através da qual desempenha o papel imprescindível no conhecimento humano, pelo enfoque na natureza da informação e conhecimento e na comunicação humana. Embora a contribuição da área possa ser analisada por diferentes vieses, a representação e recuperação da informação constitui o epicentro da sua abordagem, pois o cunho científico do campo no conhecimento foi ganho pela análise da natureza, manifestações, efeitos da informação, do conhecimento e da comunicação, visando à recuperação e uso. O Capítulo tem o objetivo de elucidar as diversas contribuições da CI sobre o conhecimento, no que concerne a representação da informação.

4.2 Conceito da representação da informação

Para Rayward (1994), o Tratado de Documentação publicado por Otlet em 1934 foi a primeira discussão sistemática e moderna para a organização da informação na qual a documentação de Otlet corresponde a aquilo que hoje se conhece por representação e recuperação da informação. A base para o problema do Tratado de Documentação era o defasamento das técnicas tradicionais de processamento da informação e Otlet propunha a integração de sistemas mecânicos para o tratamento da informação, como o cartão 3x5, baseado no princípio monográfico, no sistema de gerenciamento de banco de dados e na Classificação Decimal Universal (CDU). Assim, Otlet e La Fontaine criaram os primeiros repositórios de cartões no qual estava representado todo o conhecimento da época e padronizaram o sistema de publicação. Também criaram os primeiros sistemas de busca por assunto ou por número CDU.

A representação é a fase do tratamento da informação que antecede e condiciona a recuperação. Como o próprio termo representação sugere, é um processo em que a sua eficácia, por um lado, está condicionada à especificidade no conteúdo e contexto, e por outro, ao maior nível de detalhes possíveis na descrição, de modo a incluir a maior parte das interpretações possíveis. A representação da informação consiste na extração de alguns elementos do documento, como palavras-chave ou atribuição de termos ao documento, como descritores ou cabeçalhos de assunto, com vista à caracterização e apresentação da sua essência.

Segundo Chu (2007), a representação da informação pode ser feita através da combinação dos seguintes meios: extração de um trecho do documento, indexação, categorização, sumarização e extração de termos do documento. O autor ainda considera que o processamento da informação e a gestão da informação, mesmo com diferentes significados, são muitas vezes considerados sinônimos da representação da informação, pois a ênfase do processamento é a forma como a informação é tratada para fins de recuperação e a gestão lida com a gama de atividades associadas à informação, desde a seleção até a preservação.

Para Marcondes (2001 apud SIQUEIRA, 2003), a representação consiste na caracterização de um objeto, de tal forma que o usuário possa entender essa representação na sua mente mesmo com a ausência do objeto.

Para Siqueira (2003) os humanos usam alfabetos, diagramas, desenhos, fluxogramas, linguagens de programação de computador, interfaces de computador, símbolos, idiomas, notas musicais etc. para exteriorizar as representações de objetos, de modo a facilitar a compreensão, resolução de problemas, cálculos e o crescimento do conhecimento. Ao contrário, a representação interna consiste no uso da mente, redes semânticas, etc. para representar em parte ou no todo o ambiente em que o ser humano se encontra.

Conforme Lesk (1997 apud CHU, 2007), já que não existe um único sistema de representação conhecido pelos usuários e que seja capaz de satisfazer todas as necessidades informacionais dos mesmos, o processo de representação deve ser simples e eficiente. Neste sentido, Peschl (2002 apud SIQUEIRA, 2003) acrescenta que “a forma escolhida de representação tem que ser apropriada à tarefa que o usuário da representação tem que realizar”.

A diversidade dos métodos ou sistemas de representação traduz a essência das manifestações da pós-modernidade. No Capítulo 2 referiu-se que Harvey (2004) aponta sobre este aspecto, em concordância com as abordagens de Lyotard sobre o número indeterminado de jogos de linguagem sobre os quais recaem as intersecções que acabam por produzir determinismos locais e heteropia. O problema central da representação está relacionado à comunicação e cognição humana. Cada sujeito cria os próprios modelos de representação de acordo com a sua formação, cultura e variações linguísticas, na sua maioria, associadas a outros aspectos antropológicos. Por exemplo, a Amazônia para o índio brasileiro não tem o mesmo significado atribuído pelo cidadão europeu. Além dos diferentes modelos humanos, a

representação deve atender aos vários estágios da produção ou transformação. Esses estágios ou transformações são designados por instanciações.

As instanciações (frequentemente designadas como “versões” ou “edições” ou ainda “manifestações” de uma obra) são artefatos de pontos individuais no momento em que uma determinada obra é estabelecida, ou seja, registrada para a preservação ou disseminação. A obra é a ideia primária ou conteúdo ideacional cujas instanciações classificam-se em: derivadas, quando não há alteração da semântica ou do conteúdo ideacional, e mutação, quando há alteração da semântica ou do conteúdo ideacional, ou mesmo de ambos (SMIRAGLIA et al., 2005a).

A instanciação de uma obra existe sempre que o trabalho é realizado no tempo ou se manifesta em forma física. O problema surge quando múltiplas instanciações de uma obra (várias edições, traduções, etc.) existem e devem ser colocadas em um sistema de recuperação para auxiliar na seleção das instanciações para o interesse do usuário. Do mesmo modo, coisas podem ser representadas por metadados que podem existir em múltiplas instanciações, por exemplo, fotografia impressa, negativo, fotografia digital (SMIRAGLIA, 2005b).

Smiraglia et al. (2005a), citam o seguinte exemplo de instanciações, mutações e derivações: um autor cria uma história e escreve-a como um manuscrito (primeira instanciação); tem uma versão editada dela publicada como a primeira edição (segunda instanciação); quando a história se torna um sucesso é traduzida para outro idioma (terceira instanciação, neste momento a mutação); ao vendê-la para ser transformada em um roteiro televisivo ou cinematográfico (quarta instanciação, outra mutação); quando as imagens cinematográficas também se tornam um êxito (quinta instanciação) e são vendidas em DVD´s (sexta instanciação, ou segunda instanciação das imagens cinematográficas); a fama repetida da obra incentiva novas edições do original, tanto em edições de livro, como em edições em brochuras (sétima e oitava instanciações e estas são consideradas derivações).

Para efeitos de recuperação de informação, todas as instanciações, mutações e derivações devem ser representadas para que o usuário do sistema possa recuperá-las isoladamente ou tenha a informação sobre todo o processo evolutivo que caracterizou uma determinada obra recuperada. O artefato é único e imutável, mas a sua representação, tal como a fotografia pode ser contida em *Websites* ou em livros e para cada representação podem existir em muitos *sites* de reprodução que podem ser assuntos para múltiplas instanciações.

Gill (2008) também discute a problemática da distinção entre obra e suporte e considera que o modelo *Functional Requirements for Bibliographic Records (FRBR)* - requisitos funcionais para registros bibliográficos estabelece a distinção em quatro entidades: obras, expressões, manifestações e itens.

Para a representação da informação, destaca-se a abordagem sobre os métodos de representação da informação, metadados e vocabulários controlados.

4.3 Métodos da representação da informação

Dos diversos métodos de representação da informação, a CI destaca-se pelo uso dos seguintes: indexação, categorização, sumarização e catalogação.

A indexação é um método através do qual facetas importantes de um documento original são representadas por meio de termos (palavras ou frases), a partir dos respectivos conceitos abordados ou traduzidos para a linguagem de indexação pré-selecionada. Chu (2007) classifica a indexação em relação ao tipo de representação e ao método de produção.

Quanto ao tipo de representação, a indexação pode ser:

- **Derivativa ou por palavras - chave** – quando os termos da indexação são extraídos a partir de termos do documento original;
- **Atribuída** – quando os termos da indexação são atribuídos ao documento a indexar a partir de um vocabulário controlado. Estes termos podem ser descritores quando constam do próprio vocabulário ou identificadores quando atribuídos pelo indexador para novos conceitos.

Quanto ao método de produção, o autor aponta para a existência de quatro categorias de indexação:

- **Intelectual** – quando efetuada pelos homens;
- **Mecânica** – quando efetuada pela máquina;
- **Automática** – quando os computadores efetuam tanto a parte mecânica, como a intelectual. De referir que não existe indexação que seja totalmente automática porque o processo em si é revestido por certo grau de subjetividade, isto é, os computadores dependem das relações ou associações programadas pelos humanos;

- **Automatizada** – quando os computadores asseguram apenas as operações mecânicas de indexação enquanto que a parte intelectual continua a ser garantida pelos indexadores humanos.

A indexação derivativa se baseia em ficheiros de índice elaborados com o objetivo de armazenar os termos representativos dos documentos, associando os termos aos documentos e os termos à respectiva frequência, posição e peso. Para efeitos de elaboração dos ficheiros podem-se aplicar métodos linguísticos ou estatísticos. Os linguísticos consistem em técnicas de processamento em linguagem natural (análise morfológica do léxico, processamento sintático e semântico)[22]. Os estatísticos se baseiam na frequência de termos. A Wordnet[23] é um exemplo do método linguístico de processamento semântico que usa uma rede léxico-conceptual estruturada em torno de um conjunto de relações das palavras.

O uso conjugado dos métodos linguísticos e estatísticos é importante para a otimização dos ficheiros porque enquanto os primeiros são limitados pelas variações de diferentes línguas, os segundos o são pelas variações léxicas.

Além dos métodos linguísticos e estatísticos, Popovic e Willett (1992) consideram que o uso de *stemming* para reduzir as variáveis morfológicas no processamento em linguagem natural pode proporcionar melhores resultados tanto na indexação como na RI. É porque na maioria das vezes, essas variantes têm interpretações semânticas semelhantes e podem ser consideradas equivalentes para aplicações de RI. Por exemplo, as variáveis informação, informativo, informacional podem ser simplificadas à raiz “informa” para efeitos de recuperação. O problema é que o processo pode aumentar a exaustividade e comprometer a precisão dos documentos relevantes específicos à busca.

Os desafios atuais da representação da informação recaem sobre a indexação automática na tentativa de afastar a ação do indexador humano de processos repetitivos e, por conseguinte, de elevados custos, mediante a adoção de algoritmos de aproximação, como a frequência de palavras-chave. No caso específico de ambientes hiper-estruturados como a Web, os termos de índice são os

[22] O uso de métodos linguísticos é essencial para resolver os problemas de sinonímia e polissemia na RI.

[23] O Centro de Linguística da Universidade de Lisboa desenvolveu a base de dados de conhecimento linguístico do Português WordNet.PT, a partir do modelo EuroWorldNet, que pode ser utilizada em várias áreas da Linguística Computacional e da Engenharia da Linguagem, tais como tradução automática, sistemas de indexação e busca da informação, sistemas periciais, aplicações para o ensino do Português.

próprios *hiperlinks* que também asseguram o processo da busca, por isso os autores de documentos tornam-se indexadores dos mesmos (CHU, 2007).

A indexação é um dos problemas atuais da CI, pois a maioria dos autores que indexam as próprias páginas na Web é desprovida de conhecimentos técnicos sobre o processo. Além disso, ainda não existem vocabulários controlados para todas as áreas de conhecimento e que possam ser empregues na indexação de documentos sobre essas áreas. Consequentemente, o processo de busca de informações ainda é uma tarefa laboriosa e revestida de ambiguidades e imprecisões. O usuário da informação precisa adivinhar os termos usados na indexação do documento que necessita e os resultados são inúmeros documentos que geralmente contêm apenas os termos de busca, independentemente das variações linguísticas ou do contexto pretendido.

A categorização, de acordo com Chu (2007), denota uma representação sucessiva e hierárquica das informações por categorias. Este processo observa regras pré-estabelecidas de classificação, como a Classificação Decimal de Dewey (CDD) e a *Library of Congress Classification (LCC)*, tradicionalmente aplicadas nas bibliotecas e serviços de informação, já que na atual fase da sociedade em rede as informações são listadas como *hiperlinks*. Outro processo atual de categorização consiste em taxonomias ou estruturas hierárquicas de informação.

A categorização conforme Rocha (2004) pode ser automática e esta notabiliza-se através da indexação de documentos a partir de padrões encontrados pelo uso de estatísticas em artigos já indexados.

A sumarização é um método de representação de documentos através da condensação das respectivas informações em textos mais curtos e de fácil compreensão. Este processo de acordo com Chu (2007) envolve os seguintes tipos:

- **Resumos** – quando o conteúdo do documento é apresentado de forma clara e precisa. Os resumos podem ser *informativos* quando se parecem com o original pela fidelidade das informações neles contidos; *indicativos* quando apenas se detêm no assunto principal do original e *críticos* quando representam e analisam a informação original. Porém, os resumos críticos não se enquadram no contexto da representação pelo fato de conterem informações diferentes do original;
- **Sumários** – quando os pontos principais do documento original são organizados de tal forma que se possa compreender a sua essência;

- **Extratos** – quando se extraem partes do documento para representar a totalidade do seu conteúdo.

Além das técnicas referenciadas, Chu (2007) aponta para a existência das citações e *strings*. Segundo Malin (1968 apud CHU, 2007), uma citação implica uma relação entre uma parte ou totalidade do documento citado e a parte ou totalidade do documento citante.

No contexto atual da rede, as citações desempenham um papel preponderante não só na valorização de obras de autores mais citados, como no estabelecimento de relações entre diversos documentos que citaram o mesmo autor para efeitos de recuperação da informação. As *strings* podem ser frases ou conjunto de frases que representam um documento, tanto pela criação manual da sequência de caracteres para resumir o tema de um documento, como pela criação mecânica de entrada de índice com base na sequência dada para o efeito de representação. São exemplos da indexação de *strings* o *PREserved Context Index System (PRECIS)* e o *Nested Phrase Indexing System (NEPHIS)*. Os métodos de representação da informação são resumidamente apresentados na seguinte tabela.

Tabela 2: *Abordagem básica para a representação da informação*

Abordagem / Característica	Indexação		Categorização		Sumarização			Outros	
Tipo de Representação	Derivativa	Atribuição	Classificação	Taxonomia	Resumo	Sumário	Extração	Citação	String
Entidade Representada	Parte		Inteira		Inteira		Parte	Inteira	
Quadro de Representação	Não	Sim	Sim	Não	Não			Não	Talvez
Método de Representação	Automático	Manual e Automático	Manual	Manual e Automático	Manual	Manual e Automático		Automático	Automatizado

Fonte: *Traduzido pelo autor com base em Chu (2007, p. 32).*

A partir da tabela 2 pode-se notar que não existe nenhum método de representação válido para todas as situações, isto é, a escolha do método deve atender aos objetivos a alcançar e aos usuários por satisfazer.

Embora não tenha sido referenciado na tabela 2, a catalogação também é uma forma de representação da informação, na medida em que permite o tratamento, organização e recuperação da informação através da representação documentária descritiva, isto é, através de processos padronizados, descreve os documentos com observância dos princípios de integridade, clareza, precisão, lógica

e consistência, para que as informações neles contidos sejam recuperadas para o uso e re-uso (SIQUEIRA, 2003).

Para garantir a unicidade na identificação, localização e recuperação dos documentos, o *Anglo-American Cataloguing Rules (AACR)* estabelece regras que assistem ao processo de construção de catálogos e outras listas, geralmente aplicadas em bibliotecas. Estas regras abrangem a descrição e o provimento de pontos de acesso para materiais bibliográficos.

A segunda edição das regras, segundo o *Anglo-American Cataloguing Rules* concilia os textos britânicos e norte-americanos da edição de 1967, ampliando o estilo e as grafias. A primeira parte das regras lida com a prestação de informações que descrevem o item que está sendo catalogado; enquanto que a segunda é relativa à determinação e ao estabelecimento das posições (pontos de acesso) em que a informação descritiva está sendo apresentada aos usuários do catálogo, bem como à elaboração de referências a essas posições.

Atualmente, a grande aplicação da catalogação e por sinal o seu maior grande desafio centra-se no provimento de catálogos online para o acesso de documentos em ambientes informacionais digitais, como bibliotecas digitais.

Para permitir o intercâmbio de registros bibliográficos e catalográficos, criou-se o formato *Machine Readable Cataloging – MARC.* Segundo a *Library of Congress* (2009), o MARC surgiu na necessidade de tornar os registros de catalogação legíveis para máquinas, isto é, para permitir que as máquinas lessem e interpretassem os dados bibliográficos, ou a informação do próprio catálogo. Um registro MARC inclui: a descrição do item, a entrada principal e as entradas secundárias, os cabeçalhos de assunto e a classificação ou o número de chamada.

Conforme a *Library of Congress* (2009), para a descrição de itens bibliográficos (título, declaração de responsabilidade, edição, detalhes específicos relevantes, informações sobre a publicação, descrição física, série, notas e números padrão), seguem-se as regras da AACR2. Para a entrada principal e entradas secundárias, também se usam as regras da AACR2. Estas regras determinam os pontos de acesso para o registro e a forma como devem ser tomados. Por exemplo: podem determinar que, pelas características de certo livro, haja a necessidade de a entrada no catálogo ser por mais de um autor ou mais de um título. Para cabeçalhos de assunto, usam-se os cabeçalhos da *Sears List of Subject Headings*, da *Library of Congress Subject Headings (LCSH)*, e outros para especificar o assunto no qual o

item será enquadrado. Por isso que os cabeçalhos de assunto se enquadram no vocabulário controlado, pois garantem a consistência e reduzem ambiguidades sobre os itens no católogo. Um exemplo elucidativo é dos livros sobre gatos e outros sobre felinos, em que ambos devem ser listados no assunto sobre gatos, de modo a reduzir a ambiguidade. Por último, para o número de chamada usa-se a Classificação Decimal de Dewey ou da *Library of Congress*, para organizar os itens de acordo com os respectivos assuntos, ordem alfabética dos autores ou nomes dos autores. Assim, os usuários podem fazer associações entre assuntos ou autores no processo de busca informacional.

O registro MARC é importante para criar os meios através dos quais o computador irá interpretar as informações, através de campos para informações bibliográficas. Esses campos podem ser fixos ou ilimitados, conforme o tipo de informações. Conforme a *Library of Congress* (2009, tradução nossa):

> Se um registro bibliográfico tiver sido marcado corretamente e salvo em um arquivo de dados de computador, programas de computador podem, então, ser escritos para pontuar e formatar as informações para imprimir um conjunto de cartões de catálogo, ou para exibir as informações na tela do computador. Os programas podem ser escritos para pesquisar e recuperar certos tipos de informação dentro de campos específicos, e também para exibir listas de itens que satisfazem os critérios de pesquisa.

O MARC também permite o compartilhamento de recursos bibliográficos e o uso de sistemas de automação de bibliotecas, que geralmente são pagas. Além disso, garante a portabilidade e compatibilidade entre sistemas, sem perda dos dados. Pode ser usado para adicionar, editar ou examinar tanto registros bibliográficos como gravações sonoras, software de computador, mapas e outros itens que não sejam livros, através do mesmo conjunto de etiquetas. Segundo a *Library of Congress* (2009), essas etiquetas são constituídas por: campo, tag, indicador, subcampo, código de subcampo e designador de conteúdo.

O campo é cada parte que compõe um registro (autor, título, etc.), e pode ser subdividido em um ou mais subcampos. Como os nomes dos campos são longos, eles são representados por tags de três dígitos. Por isso, a tag indica o campo e o tipo de dados. Quando a tag é exibida em conjunto com os respectivos indicadores, devem-se considerar os primeiros três dígitos para a tag ou campo. Além dos campos, também existem os indicadores. Os indicadores são as duas posições de caracteres que seguem uma tag, com a excepção dos campos 001 a 009. O uso da primeira, da segunda ou de ambas as posições do indicador varia de acordo com o

campo. Nos casos em que nenhuma posição é utilizada, o indicador se refere como indefinido, através do caracter "#". Os valores de cada indicador podem ser números de 0 a 9 ou letras, mas raramente se usam as letras (*LIBRARY OF CONGRESS*, 2009).

Segundo a *Library of Congress* (2009), o subcampo é formado por cada tipo de dados dentro do campo, e é precedido pelo respectivo código[24]. O código do subcampo é representado por uma letra minúscula, ou, às vezes, por um número, precedido por um delimitador. O delimitador é um caractere usado para separar os subcampos (por exemplo: $). O conjunto de tags, indicadores e códigos dos subcampos forma os designadores de conteúdo, pois rotula e explica o registro bibliográfico.

Em síntese, embora o registro MARC seja complexo, principalmente para usuários não familiarizados com o formato, garante a preservação e compartilhamento dos recursos, e dos respectivos metadados. A tabela 3 ilustra parte da complexidade do registro MARC, sobre algumas das suas regras.

A CI desempenha um papel preponderante na revolução tecnológica como a quarta revolução do ciclo informacional, na concepção de Le Coadic (2004), pela capacidade de síntese de informações necessárias à construção do conhecimento. Esta ação é notabilizada pelo emprego de métodos como a catalogação, para permitir tanto o armazenamento e integridade, como a comunicação, recuperação e intercâmbio dos recursos.

[24] Os campos 001 a 009 não têm subcampos (*LIBRARY OF CONGRESS*, 2009).

Tabela 3: *Algumas regras gerais que ajudam na definição do significado dos números usados como tags de campo (MARC 21).*

Regra	Tag	Tipos de dados	Observações
Tags divididas por centenas	0XX	Informações de controle, números, códigos	Em MARC 21, muitas vezes a notação XX é usada para indicar um grupo de tags relacionadas (Ex: 1XX refere-se a todas as tags 100: 100, 110, 130, e assim por diante). Os campos 9XXs são para usos localmente definidos, tais como números de código de barras locais, ou para anexar outros tipos de informações para registros. Os campos X9Xs ou cada um desses grupos - 09X, 59X, etc. também são reservados para uso local, exceto o 490).
	1XX	Entrada principal	
	2XX	Título, declaração de responsabilidade, edição e informações sobre a publicação	
	3XX	Descrição física, etc.	
	4XX	Declarações de série	
	5XX	Notas	
	6XX	Entradas secundárias de assunto	
	7XX	Outras entradas secundários diferentes do assunto e série	
	8XX	Entradas secundárias se série (outras formas de autoridade)	
Pontos de acesso	1XX	Campos (entradas principais)	Estes são os campos que estão sob controle de autoridade, isto é, que se devem preencher seguindo uma forma reconhecida ou estabelecida (ex: Library of Congress Name Authority file para nomes, Library of Congress Subject Headings ou Sears List of Subject Headings para tópicos ou nomes geográficos). Também é importante o controle local da autoridade, através do uso de cabeçalhos de assunto ou listas de nomes adotados localmente, de modo a garantir a unicidade na descrição.
	4XX	Campos (declarações de série)	
	6XX	Campos (cabeçalhos de assunto)	
	7XX	Campos (entradas secundários diferentes do assunto e série)	
	8XX	Campos (entradas secundárias de séries)	
Conteúdo paralelo	X00	Nomes pessoais	Os campos que requerem o controle de autoridade também usam a tag de construção paralela. Em geral, nos campos 1XX, 4XX, 6XX, 7XX e 8XX, um nome pessoal terá os dois últimos algarismos 00. Ao combinar o "Conteúdo paralelo" com "Tags divididas por centenas", conclui-se que: se o tema do livro (6XX) for uma pessoa (Lincoln, Abraham), o tag será 600; se o assunto do livro for uma empresa (Apple Computer, Inc.), a tag será 610; se o tema do livro for um lugar (Estados Unidos), a tag será 651.
	X10	Nomes institucionais	
	X11	Nomes de reuniões	
	X30	Títulos uniformes	
	X40	Títulos bibliográficos	
	X50	Termos de tópicos	
	X51	Nomes geográficos	

Fonte: *Criada pelo autor com base na Library of Congress (2009, tradução nossa).*

4.4 Metadados

O termo metadados, segundo Smiraglia et al. (2005a) foi cunhado em 1969 por Jack E. Meyers e registrado em 1986 como uma marca dos EUA para a empresa na qual Meyers fundou. De seguida, o termo foi adotado por áreas ligadas à informação, como Ciência da Computação, Estatística, Banco de Dados, etc. De acordo com o autor, metadados são "dados codificados e estruturados que descrevem as características de entidades portadores de informações para auxiliar na identificação, descoberta, avaliação e gestão das entidades descritas". Equivale isto dizer que metadados são dados que descrevem outros dados, aplicados tanto para documentos de qualquer natureza através da catalogação e indexação, como especificamente para recursos eletrônicos ou digitais. Os metadados podem ser gerados manualmente ou através de *softwares* específicos.

De acordo com a *National Information Standards Organization* – NISO (2004)[25], metadados são informações estruturadas que descrevem, explicam, localizam ou que permitem a recuperação, o uso e o gerenciamento de recursos de informação. O termo metadado é usado de forma diferente por diversas comunidades. Algumas usam-no para se referir a máquina de informação compreensível, enquanto outras usam apenas para registros que descrevem recursos eletrônicos. No ambiente de bibliotecas, o termo metadado é comumente usado para qualquer esquema formal de descrição de recursos, aplicando-se a qualquer tipo de objeto (digital ou não-digital).

Segundo a NISO (2004), a tradicional catalogação em bibliotecas é uma forma de metadados. Além dos padrões de metadados MARC 21 e AACR2, outros esquemas foram desenvolvidos para descrever vários tipos de objetos textuais e não textuais, incluindo livros publicados, documentos eletrônicos, objetos de arte, materiais educativos e de formação e conjuntos de dados científicos.

Gill (2008) considera que o termo metadados provêm do grego *"meta"* e do latim *"data"* e significa literalmente dados sobre dados, ou seja, dados que descrevem outros dados. São exemplos os cartões de catálogo de bibliotecas que

[25] A NISO é uma associação sem fins lucrativos, credenciada pela *American National Standards Institute (ANSI)*, que identifica, desenvolve, mantém e publica normas técnicas para gerenciar informações em ambiente digital. Essas normas aplicam-se para tecnologias tradicionais e atuais no âmbito das necessidades informacionais, incluindo a recuperação, mudança de formato, armazenamento, metadados e preservação.

contêm dados sobre o conteúdo e localização de livros. Contudo, neste exemplo o autor chama a atenção sobre a diferença entre informação e seus suportes, isto é, o cartão de catálogo apenas contém a informação sobre livros, mas não reflete a informação dos livros em si.

Gill (2008) também considera que uma outra propriedade de metadados que não é tratada adequadamente pelas definições de padrões é que os metadados são normalmente estruturados para modelo de atributos mais importantes do tipo de objeto que eles descrevem. Por exemplo, cada componente do padrão de metadados para registros bibliográficos *MARC 21* encontra-se delineado por tags de campos que identificam o significado de cada peça de informação (autor, título, assunto, notas, série, etc.).

Lagoze et al. (1996) desenvolveram os tipos de metadados mais extensos, comparativamente às tipologias de Gilliland-Swetland´s (2000), Greenberg´s (2001) e Caplan´s (2003). A tabela a seguir ilustra as tipologias de metadados desenvolvidos por estes autores conforme às respectivas funcionalidades (SMIRAGLIA et al, 2005a, p.21).

Tabela 4: *Tipologias e funcionalidades de metadados.*

Lagoze et al (1996)			**Gilliland-Swetland (2000):** Tipologia dos 5 tipos de metadados	**Greenberg (2001):** Tipologia de 4 tipos de metadados (2 subtipos dos Metadados de Uso	**Caplan (2003):** Tipologia dos 4 tipos de metadados
Tipologia dos 7 tipos de metadados	Funções dos metadados "Este tipo de metadados facilita"	Exemplos de Elemento			
Identificativos/ Metadados Descritivos	Descoberta de recursos/ Recuperação da informação	Criador (autor), título, assunto	Metadados Descritivos	Metadados de Descoberta	Metadados Descritivos
Metadados Administrativos	Gerenciamento de recursos	Preço, condição	Metadados Administrativos e de Preservação	Metadados Administrativos	Metadados Administrativos
Metadados de Indicação de Termos e Condições	Utilização de recursos	Direitos, restrições de reprodução	Metadados Administrativos, Preservativos e de Uso	Metadados de Uso Técnico, Uso Intelectual e Administrativos	Metadados Administrativos e de Articulação
Metadados de Classificação de Conteúdo	Uso de recursos por audiências apropriadas	Audiência	Metadados de Uso	Metadados de Uso Técnico e Uso Intelectual	Metadados Administrativos e de Articulação
Metadados de Proveniência	Autenticação de recursos e outras atividades relacionadas com a proveniência	Criador, fonte	Metadados Administrativos e de Uso	Metadados de Autenticidade e Administrativos	Metadados Administrativos
Metadados de Articulação/ Relacionamento	Articulação dos recursos com os outros relacionados	Relação, fonte	Metadados Administrativos	Metadados de Autenticidade e Administrativos	Metadados de Articulação
Metadados Estruturais	Necessidades de recursos de hardware e software	Taxa de compressão	Metadados Técnicos e de Uso	Metadados de Uso Técnico	Metadados Estruturais

Fonte: *Traduzido pelo autor com base em Smiraglia et al (2005, p. 21).*

Conforme a tabela 4, Lagoze et al. (1996) ampliaram e sintetizaram o universo dos metadados. Para os autores, existem os seguintes tipos de metadados: identificativos ou descritivos, administrativos, de indicação de termos e condições, de classificação de conteúdo, de proveniência, de articulação ou relacionamento e estruturais. Destes, destacam-se três tipos principais:

- **Identificativos ou descritivos** – descrevem recursos de informação para permitir a sua descoberta ou recuperação. Tais recursos podem

ser dados bibliográficos ou recursos da *Web* (imagem, áudio, vídeo). São exemplos: o criador ou autor, o título, o assunto ou resumo, as palavras-chave;

- **Técnicos** – são relativos ao funcionamento, por exemplo: informações sobre *hardware* e *software*, digitalização, autenticidade e segurança;
- **Uso** – indicam as informações sobre o uso dos recursos, por exemplo: exibição,controle de acesso, versões, etc.
- **Preservação** – indicam as informações sobre a conservação e preservação dos recursos, por exemplo, condições físicas do recurso;
- **Administrativos** – fornecem informações para auxiliar o gerenciamento de recursos, como por exemplo, quando e como o mesmo foi criado, tipo de arquivo e outras informações técnicas, e quem tem acesso a ele.

A CI desempenha um papel preponderante tanto na construção de padrões de metadados, como na sua implementação, através da ação dos respectivos profissionais em bibliotecas ou outras unidades de informação e em comunidades de debates sobre ações que visam o seu melhoramento e adequação para cada estágio da sociedade. Os metadados são imprescindíveis para a sobrevivência dos recursos de informação na atual estrutura mutável e inconsistente da Web e permitem a recuperação futura para o uso em diferentes contextos das necessidades dos usuários.

4.4.1 Funções dos metadados

Os metadados, conforme se referiu anteriormente, tem uma diversidade de funções que garantem a qualidade, a localização, o acesso e a preservação da informação.

Segundo a NISO (2004), os metadados podem descrever recursos em qualquer nível de agregação. Podem descrever um conjunto, um único recurso, ou uma componente de um recurso maior (por exemplo, a fotografia de um artigo). Também ajudam os catalogadores a tomarem decisões sobre o registro de um catálogo pelo respectivo conjunto de volumes ou por cada volume específico no conjunto. Os metadados também podem ser usados para a descrição, em qualquer nível do modelo de informações dispostas nos requisitos funcionais para registros bibliográficos da *International Federation of Library Associations and Institutions* -

IFLA: obra, expressão, manifestação ou item. Por exemplo, pode-se usar o registro de metadados para descrever um relatório, uma edição especial desse relatório, ou uma cópia específica de uma das edições desse relatório.

Segundo a NISO (2004), muitas vezes os metadados são incorporados em documentos HTML e nos cabeçalhos de arquivos de imagem. O armazenamento de metadados com o objeto que descreve evita a sua perda, elimina problemas de ligação entre dados e metadados, e garante a atualização conjunta do metadado e do objeto. No entanto, é impossível incorporar metadados, em alguns tipos de objetos. O armazenamento separado de metadados pode simplificar o gerenciamento do próprio metadado e facilitar a busca e recuperação. Neste caso, os metadados são armazenados em um sistema de banco de dados e ligados aos objetos descritos.

A NISO (2004) considera as seguintes funções de metadados: descoberta de informações relevantes, organização de recursos eletrônicos, interoperabilidade e integração de recursos, identificação digital e o arquivamento e a preservação.

A **descoberta ou localização de recursos** permite: que os recursos sejam encontrados por critérios de relevância, a identificação dos recursos, a apresentação por proximidade dos recursos similares, a distinção dos recursos não similares e a apresentação da informação sobre a localização do recurso.

A **organização de recursos eletrônicos** é feita mediante a agregação dos sites ou portais, através da organização de links para recursos com base em público ou tópico. Essas listas podem ser construídas como páginas estáticas, com os nomes e as localizações dos recursos “codificados” no HTML. No entanto, é mais eficiente e cada vez mais comum para construir essas páginas dinamicamente a partir de metadados armazenados em bancos de dados.

A **interoperabilidade**[26]: o emprego dos metadados na descrição dos recursos permite que os mesmos sejam inteligíveis, tanto para seres humanos, como para máquinas. O uso de esquemas definidos de metadados e dos protocolos de transferência compartilhados permite que os recursos que trafegam pela rede sejam localizados e recuperados com facilidade. Por isso, o protocolo Z39.50 é comumente usado para a pesquisa de sistema de cruz, pois mesmo sem compartilhar os

[26] Segundo a NISO, é a capacidade de vários sistemas com diferentes plataformas de hardware e software, estruturas de dados e interfaces para trocar dados com perda mínima de conteúdo e funcionalidade.

metadados, mapeia as suas próprias capacidades de busca para um conjunto comum de atributos de pesquisa. A *Open Archives Initiative - OAI* baseia-se neste princípio, para tornar interoperáveis os metadados produzidos em diferentes contextos, através do respectivo protocolo.

A **identificação digital** pode ser feita através de números padrões que os esquemas de metadados incluem, para identificar o trabalho ou objeto ao qual os metadados se referem. A localização de um objeto digital também pode ser feita através de um nome de arquivo, *URL* ou *Digital Object Identifier – DOI*[27].

Em relação ao **arquivamento e à preservação**, a maioria dos esforços sobre metadados giram entorno da descoberta de recursos criados recentemente. No entanto, há uma preocupação relutante sobre a permanência em estado utilizável dos recursos informacionais no futuro. A informação digital é frágil, ou seja, pode ser corrompida ou alterada, intencionalmente ou não. Pode se tornar inutilizável com a mudança da mídia de armazenamento e das tecnologias de hardware e software. Por isso os metadados são fundamentais para assegurar que os recursos irão sobreviver e continuar a ser acessíveis no futuro. O arquivamento e a preservação exigem elementos especiais para rastrear a linhagem de um objeto digital (de onde veio e como ele mudou ao longo do tempo), de modo detalhar as suas características físicas, e para documentar o seu comportamento, a fim de imitá-lo em tecnologias futuras. Muitas organizações internacionais têm trabalhado na definição de esquemas de metadados para a preservação digital, por exemplo, a Biblioteca Nacional da Austrália, e o projeto britânico *Cedars*. O grupo *PREservation Metadata: Implementation Strategies – PREMIS*, desenvolveu um conjunto de elementos fundamentais e estratégias para a codificação, armazenamento e gerenciamento de metadados de preservação dentro de um sistema de preservação digital.

Indubitavelmente, os metadados desempenham uma função importante na transferência da informação, mormente na fase da recuperação da informação digital prevalecente na atualidade. Todavia, alguns tratadistas ainda se mostram relutantes em relação aos reais e efetivos progressos dos metadados na construção do conhecimento baseado na Web. De acordo com Gill (2008), Cory Doctorow, jornalista, blogueiro e escritor de ficção científica Cory Doctorow enumera os “sete obstáculos intransponíveis entre o mundo tal como a conhecemos e meta-utopia”

[27] Segundo a NISO, o DOI é um identificador persistente e é preferencial porque os locais de objetos muitas vezes mudam, tornando a URL padrão (e, portanto, o registro de metadados) inválido.

para defender a tese de que os metadados criados por seres humanos nunca irão garantir na totalidade a descoberta de recursos na Web.

Os sete argumentos de Doctorow (2001 apud GILL, 2008) podem ser sintetizados nas seguintes características:

- Confiabilidade – os metadados na Web não podem ser totalmente confiáveis, pois existem muitos criadores de conteúdo Web que publicam metadados de forma desleal, a fim de tirar o tráfego adicional a seus sites;
- Complexidade – a representação e a produção de metadados são tarefas complexas e a maioria dos editores de conteúdo da Web não é suficientemente motivada e treinada para catalogar cuidadosamente o conteúdo que publica;
- Concorrência – concorrentes de diversos padrões de metadados nunca vão concordar entre si. Além disso, inadvertidamente alguns criadores de conteúdo Web publicam metadados enganosos;
- Subjetividade – os esquemas de metadados não são neutros, isto é, são subjetivos ou variam de acordo com a interpretação dada pelo produtor e pelos usuários. Mesmo com o uso de determinados padrões, sempre existirão diferentes formas de descrever algo, ou seja, a descrição também é subjetiva.

Embora Gill (2008) considere que os argumentos de Doctorow são norteados por uma crítica excessiva e radical, também concorda com a prática desleal de uso de alguns metadados, como *spam meta-tag* para aumentar a visibilidade das páginas e, por conseguinte, a posição de topo no ranking dos motores de busca. Esta ação é principalmente motivada por interesses comerciais. Por isso, a fase atual da explosão informacional não reduz a ação do campo da CI; pelo contrário, amplia o desafio de capacitação de profissionais como bibliotecários, arquivistas, museólogos, entre outros, para atuarem com competência na descrição de recursos digitais com metadados e esquemas de metadados interoperáveis. Deste modo, muitas informações que são perdidas na Web poderão ser recuperadas e aplicadas em diferentes domínios de conhecimento humano.

A necessidade da ação dos profissionais da informação é extensiva ao próprio processo de criação de metadados, por meio de diretrizes permenorizadas

que reduzam a subjetividade no seu uso. Sobre este assunto, Foulonneau e Riley (2008) observam que, antes da escolha do padrão de metadados para a descrição de recursos ou implementação em um SRRI, é necessário definir e testar as diretrizes sobre o uso para esses padrões. Para assegurar que os metadados criados ou a implementar preencham a sua finalidade, é preciso uma variedade de perspectivas incluindo as de peritos na criação de metadados, necessidades de usuário, armazenamento do conteúdo a longo prazo e disponibilização aos usuários finais, recursos e orçamento, bem assim de questões sobre a gestão do projeto.

Para todos os tipos de metadados em criação, especificamente os descritivos, técnicos e de direito, deve-se especificar na frente e no documento das respectivas diretrizes exatamente o que deve ser descrito e se o material existe em várias versões. Por exemplo, se o material na coleção tiver sido digitalizado a partir de originais analógicos, o criador de metadados deve saber se o registro (ou qualquer elemento dentro do registro) descreve o conteúdo intelectual do recurso, o suporte analógico ou qualquer seu suporte digital. Tal acontece no elemento título para um item bibliográfico, em que o mesmo pode se referir ao título do próprio item em descrição ou ao título sobre o item, por isso, para dissipar quaisquer dúvidas as diretrizes sobre o uso do elemento título deverão ser claras.

Alguns padrões como AACR2 definem a “fonte principal da informação” como aquela a partir da qual os dados para um determinado elemento devem ser tomados. Neste caso, quando a norma não prevê a fonte, as diretrizes podem fornecer alguma orientação sobre onde encontrar informações para uso em um elemento específico. De igual modo, as diretrizes de uso de metadados devem indicar e fornecer listas ou *links* para vocabulários controlados para todos os elementos de metadados para os quais são adequados, abrangendo os casos em que os conceitos necessários não estejam disponíveis no vocabulário controlado e, por tanto, tenham que ser identificados pelo indexador.

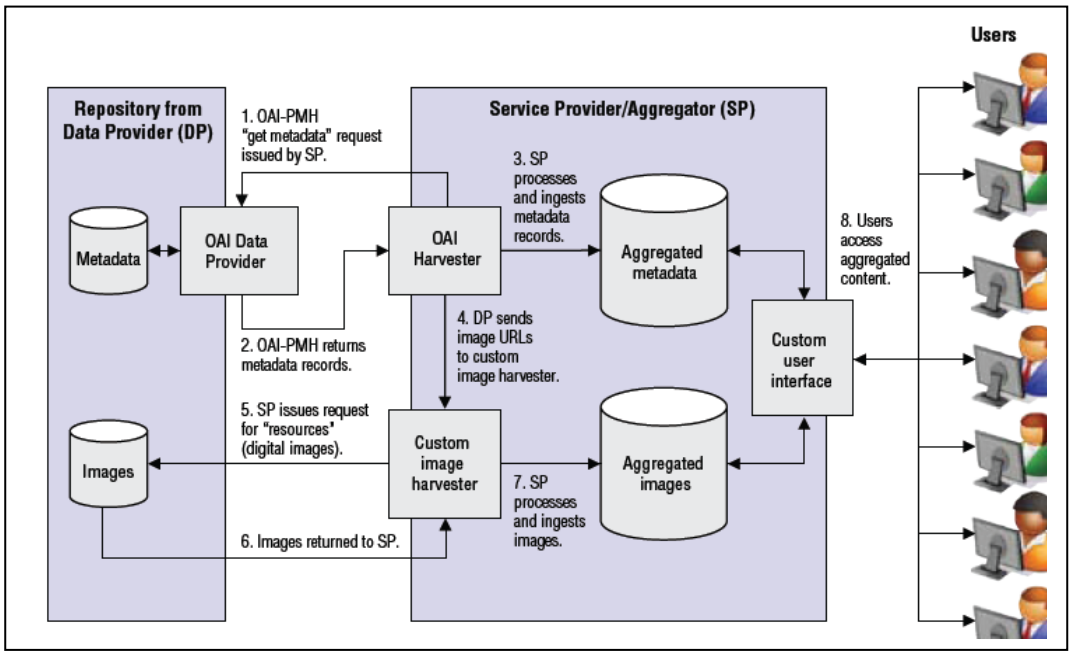
4.4.2 Processo de coleta de metadados

Gill (2008) considera que o protocolo *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* fornece um método alternativo que visa tornar os metadados da “*Web* invisível” mais acessíveis, através da exposição de registros de

metadados na *Web* para que outros sistemas de computador compatíveis com o mesmo protocolo possam acessá-los e recuperá-los.

O protocolo OAI-PMH permite a interoperabilidade entre dois diferentes sistemas de computador, um provedor de dados OAI e um coletor de dados OAI que também pode ser um provedor de serviços OAI. Um provedor de dados OAI é uma fonte de registro de metadados, enquanto que um coletor de dados OAI recupera os registro de metadados a partir de um ou mais provedores OAI. A recuperação do registro de metadados ocorre quando há conformidade dos dois sistemas (provedor e coletor) com protocolos de troca de informações.

Figura 4: *Modelo de coleta de metadados.*



Fonte: *Gill (2008, p. 18).*

Segundo o esquema da figura 4, o modelo de coleta de metadados consiste em três categorias: o repositório do provedor de dados, o serviço de provimento ou agregação e os usuários. O repositório permite o armazenamento de metadados e imagens que serão posteriormente recuperados pelo protocolo OAI-PMH. O serviço de provimento processa e recupera os registros de metadados e os usuários acessam conteúdos agregados[28] por meio da interface do usuário.

O modelo funciona da seguinte maneira: quando o usuário solicita um conteúdo, o protocolo OAI-PMH processa o pedido interligando o coletor, o provedor de dados e o repositório de metadados, baseando-se no pedido do serviço de

[28] Conteúdos agregados, porque contêm a informação em si e os respectivos metadados descritivos.

provimento e retorna os registros de metadados para esse conteúdo. Tratando-se de imagens, a partir do coletor, o provedor de dados envia URLs de imagens para o coletor de imagens e retorna as imagens com base no pedido do serviço de provimento.

O protocolo OAI-PMH pode suportar todo o esquema de metadados expresso na linguagem XML, porém determina como requisito mínimo que todos os provedores de dados OAI sejam capazes de fornecer registros de metadados *Dublin Core XML*. Ao expor o catálogo de metadados como um provedor de dados OAI e registrá-lo com o *Sitemap* do Google, os conteúdos da "*Web* invisível" podem ser rastreados, indexados e disponibilizados pelos motores da Google, tornando-se acessíveis para os usuários finais (GILL, 2008).

De salientar que o modelo de coleta de metadados possui uma interface do usuário tanto para o acesso, como para a entrada de dados no banco de dados, por isso mesmo, os conteúdos dos usuários do sistema podem ser disponibilizados e os seus metadados interoperáveis em contextos similares do mesmo domínio.

4.4.3 Esquemas de metadados

Os esquemas de metadados segundo Smiraglia et al (2005a), são desenvolvidos para a descrição de recursos de modo a facilitar a sua preservação e recuperação. Devido a sua complexidade, os autores alertam para a necessidade de desenvolvimento de esquemas de metadados estabelecendo a diferença entre conteúdo e suporte ou obra, bem como da compreensão dos conceitos a representar, formas como cada termo que representa o conceito central poderá ser derivado e como a descrição de metadados será usada para a recuperação da informação. A negligência da complexidade dos esquemas de metadados leva à criação de ferramentas de recuperação baseadas apenas nos recursos presentes no momento, essencialmente selecionados por acaso, sem o conhecimento das suas características, bem como de outros recursos similares que o sistema poderá recuperá-los.

Segundo a NISO (2004), esquemas de metadados ou esquema são "conjuntos de elementos de metadados projetados para uma finalidade específica, como a descrição de um tipo particular de recurso de informação". A definição ou significado dos próprios elementos é conhecida como a semântica do esquema . Os

valores dados para elementos de metadados são o conteúdo. Esquemas de metadados geralmente especificam nomes de elementos e sua semântica, contudo, também podem especificar regras facultativas sobre a formulação do conteúdo (por exemplo, como identificar o título principal); sobre a representação do conteúdo (por exemplo, regras de capitalização[29]) e sobre os valores permitidos do conteúdo (por exemplo, os termos devem ser utilizados a partir de um vocabulário controlado específico).

De acordo com a NISO (2004), também pode haver regras de sintaxe para a forma como os elementos e o seu conteúdo deve ser codificado. Um esquema de metadados sem regras de sintaxe prescritas é chamado sintaxe independente. Os metadados podem ser codificados em qualquer sintaxe definível. Muitos esquemas de metadados atuais usam *Standard Generalized Mark-up Language – SGML* ou *Extensible Mark-up Language – XML*. A linguagem XML foi desenvolvida pela W3C, como uma forma estendida de HTML que permite conjuntos de tags definidos localmente e fácil troca de informações estruturadas. A linguagem XML visa à facilidade de compartilhamento de informações na Web. A metalinguagem SGML permite definir a linguagem de marcação de um documento.

Segundo Rocha (2004), os esquemas de metadados são empregues na *Web por* comunidades especializadas em diversas áreas do conhecimento para padronizar a descrição dos seus recursos, por exemplo, (*International Conference on Dublin Core and Metadata Applications* de 2001, em que áreas como ambiente, governos, educação, biblioteconomia, etc., criaram esquemas para seus sectores). Para evitar a sobreposição de esquemas definidos para as mesmas finalidades, usam-se Registros de Metadados, Perfis de Aplicação e a iniciativa *Dublin Core*.

Os esquemas de metadados conforme Smiraglia et al (2005a), reúnem as seguintes características:

- Os elementos de metadados são reunidos numa coleção para apoiar uma função, ou uma série de funções (por exemplo, a descoberta de recursos, gestão, utilização, etc.), para um objeto de informação;
- Uma coleção de elementos de metadados forma uma estrutura em que os valores dos dados são adicionados. Os valores dos dados podem

[29] A capitalização consiste em destacar a primeira letra de uma palavra em maiúsculo e as restantes em minúsculo, observando certas instruções. Por exemplo, a capitalização da letra “C” na expressão Ciência da Informação, obedecendo à regra capitalizar título precedido do termo informação.

ser incontroláveis ou controláveis (por exemplo, tomados a partir de uma fonte como LCSH ou uma lista padronizada de valores);

- Uma coleção de elementos de dados, com seus atributos formalizado em uma especificação (ou um dicionário de dados).

Os esquemas de metadados fornecem estruturas de conhecimentos para diferentes áreas, de modo a permitir a descoberta e uso da informação dentro das mesmas.

Muitos esquemas de metadados são desenvolvidos em diferentes ambientes ou domínios do conhecimento. No entanto, os mais comuns são: *Dublin Core - DC, The Text Encoding Initiative - TEI, Metadata Encoding and Transmission Standard – METS, Metadata Object Description Schema - MODS, The Encoded Archival Description - EAD,* entre outros*.*

Dublin Core é usado para a descrição de recursos digitais (este esquema é detalhado no ponto sobre a representação da informação em ambientes digitais). ***The Text Encoding Initiative*** é um projeto internacional para desenvolver diretrizes para a marcação de textos eletrônicos, tais como romances, peças de teatro e poesia, principalmente para apoiar a investigação nas Ciências Humanas. ***Metadata Encoding and Transmission Standard*** foi desenvolvido para descrever objetos de bibliotecas digitais complexos, criando instâncias de documentos XML que expressam a estrutura de objetos digitais de bibliotecas, os metadados descritivos e administrativos associados e os nomes e localizações dos arquivos que compõem o objeto digital. ***Metadata Object Description Schema*** é um esquema de metadados descritivo derivado do MARC 21 que visa à transmissão de dados selecionados a partir de 21 registros MARC existentes ou permitir a criação de descrição registros de recursos originais. ***The Encoded Archival Description*** foi desenvolvido com o intuito de marcar os dados contidos nas ajudas sobre os registros de catálogos. Estas ajudas podem ser pesquisadas e exibidas online, são explicativas e estruturadas de forma hierárquica e geralmente começam com uma descrição da coleção como um todo, indicando os tipos de materiais que contém e a sua importância.

4.5 Vocabulários Controlados

Os vocabulários controlados enquadram-se no âmbito da linguagem para os SRRI. Harpring (2010, p.12, grifo nosso) considera que um vocabulário controlado “é

um conjunto organizado de palavras e frases, inserido em um domínio específico, usado para indexar o conteúdo e/ou recuperá-lo através da navegação e pesquisa". Este conjunto de palavras ou frases tem a finalidade de organizar informações e atribuir terminologias para catalogar e recuperar essas informações.

Os vocabulários controlados são necessários tanto na fase da representação para subsidiar os termos que o indexador precisa atribuir ao documento, como na fase da recuperação, pois os usuários podem utilizar diferentes termos sinônimos ou termos mais genéricos para se referirem ao mesmo conceito. Neste sentido, os vocabulários controlados reúnem condições variantes e sinônimas de conceitos e de ligação de conceitos em uma ordem lógica ou categorias.

Os vocabulários controlados podem estar associados a formatos controlados que estabelecem regras sobre os tipos de dados permitidos e formatação de informação. Estes formatos podem ser agregados ao termo *unicode*, controlar medidas, coordenadas geográficas ou outras informações em campos onde números ou códigos são usados. As restrições podem ser estabelecidas nos campos para regular o número de dígitos permitidos, a expressão de casas decimais e números negativos, etc. Por exemplo, os formatos controlados podem ser usados em datas, como data da descoberta ou criação de uma obra de arte, combinando áreas controladas e um campo de data de exibição (HARPRING, 2010).

Harpring (2010) considera a existência de dez tipos de vocabulários estruturados:

- **Relacionamentos:** pode ser um estado de conexão ou uma associação entre campos ou tabelas em um banco de dados para um vocabulário controlado. Também podem incluir *links* que organizam os termos e fornecem o contexto ou listas de cabeçalhos de assunto, taxonomias, tesauros, etc.;
- **Cabeçalhos de assunto ou títulos:** palavras ou frases que se atribuem aos livros, artigos ou outros documentos para descrever o assunto ou o tema dos textos e agrupá-los com textos de assuntos semelhantes. São exemplos de cabeçalhos de assunto, o *Library of Congress Subject Headings* (*LCSH)*, que fornece uma lista abrangente de termos preferenciais, e o *Medical Subject Headings (MeSH)*, usado na medicina para indexar artigos de jornais e livros;

- **Listas controladas:** simples listas de termos usados para controlar a terminologia. Nestas listas é obrigatório observar as seguintes características: cada termo é único, termos não são sobrepostos em sentido ou termos da mesma classe não podem ter o mesmo sentido, termos são todos os membros da mesma classe, termos são iguais em qualidade ou especificidade e termos são organizados em ordem alfabética ou em outra ordem lógica. As listas controladas têm grande aplicação em projetos específicos de banco de dados e na área de SRRI, permitem que o catalogador ou indexador tenha apenas uma pequena lista para a escolha de termos, garantindo assim maior consistência e reduzindo a probabilidade de erro;
- **Listas de anéis de sinônimos:** conjunto de termos considerados equivalentes para fins de recuperação. Enquanto os relacionamentos de equivalência são feitos entre termos e nomes com mesmos significados, os anéis são feitos para termos com sinônimos diferentes, mas que tenham equivalência, isto é, que podem ser substituídos um pelo outro para efeitos da expansão da consulta na recuperação. Devido à sua complexidade, normalmente os anéis de sinônimos são construídos manualmente e usados por peritos familiarizados com conteúdos específicos de uma determinada área e de sistemas de informação;
- **Arquivos de Autoridades:** conjunto de nomes consagrados ou títulos e referências cruzadas para formas preferidas de variantes ou formas alternativas. Um arquivo de autoridade pode ser qualquer tipo de vocabulário controlado, com exceção da lista de anéis de sinônimos;
- **Taxonomias:** classificação ordenada de termos, aplicável a um domínio específico. Uma taxonomia é organizada em uma estrutura hierárquica, e cada termo pertence a uma ou mais relações pai/filho. São exemplos de taxonomias, a classificação científica dos animais e de plantas;
- **Esquemas de classificação alfanuméricos:** códigos controlados (letras ou números ou letras e números), que representam conceitos ou posições. O sistema de classificação da biblioteca do congresso - *Library of Congress Classification (LCC)* e a Classificação Decimal de

Dewey (CDD) são exemplos de um esquema de classificação alfanumérico;

- **Tesauros:** rede semântica de conceitos únicos, incluindo relacionamentos entre sinônimos, contextos mais amplo e mais restrito (pai/filho) e outros conceitos relacionados. Os tesauros podem conter três tipos de relações: equivalência (sinônimos), hierárquica (inteira/parte, gênero/espécie ou instância) e associativa;
- **Ontologias:** instrumentos para a descrição semântica. As ontologias têm algumas características em comum com taxonomias e tesauros, contudo enquanto as ontologias usam relações semânticas estritas entre os termos e atributos com o objetivo de representação do conhecimento em uma forma legível para a máquina, os tesauros fornecem ferramentas para a catalogação e recuperação. As ontologias são utilizadas na *Web* Semântica, inteligência artificial, engenharia de *softwares* e arquitetura da informação como formas de representação do conhecimento em formato eletrônico sobre um determinado domínio do conhecimento. Silberschatz et al. (2006) definem as ontologias, como estruturas de relacionamento hierárquico entre palavras, geralmente, mediante o uso das expressões “é um” ou “parte de”. Por exemplo: o homem é um animal ou o homem faz parte dos mamíferos. Deste modo, uma busca sobre o termo “homem” também poderia recuperar documentos sobre mamíferos, pois o homem é um mamífero.
- **Folksonomias:** conjunto de conceitos representados por termos e nomes (*tags*) que são compilados através da classificação social. A classificação social é o método através do qual indivíduos e grupos criam, gerenciam e compartilham *tags* (termos, nomes, etc.) para anotar e categorizar recursos digitais em ambientes sociais *online*. As folksonomias não suportam os mecanismos de busca organizada e outros tipos de navegação, mas oferecem pontos de acesso adicionais não incluídos no vocabulário formal.

4.6 Representação da informação em ambientes digitais

O termo digital na área de RRI refere-se aos recursos ou ambientes virtuais, cuja leitura se processa através do código binário (0 e 1). Neste sentido, importa frisar que o digital não significa imaterial porque, mesmo na forma digital, o conteúdo informacional ou ambiente ocupa um determinado espaço e pode ser manipulado, como prova da sua existência (LÉVY, 1999).

Atualmente, alguns ambientes digitais são representados pela Web que, contrariamente ao modelo tradicional de indexação, apresenta uma complexidade pela dificuldade na seleção de documentos em função das áreas de conhecimento, dos objetivos do sistema e dos usuários finais.

Segundo Rocha (2004), devido ao volume da informação na *Web*, os diretórios abertos[30] e a indexação automática surgem como alternativas para o processo de representação da informação (ROCHA, 2004). Gill (2008) e Stewart (2008) apontam o problema da *Web* ser um ambiente que, contrariamente às bibliotecas, é caracterizado pela falta da organização ou de esquemas gerais de estruturação dos conteúdos para permitir a fácil localização, recuperação e manipulação, fato agravado pela indisponibilidade de alguns conteúdos na "*Web* invisível", conforme se referenciou anteriormente.

4.6.1 Dublin Core (DC)

Segundo a NISO (2004), o conjunto de elementos do esquema de metadados DC surgiu de discussões em uma oficina de 1995, patrocinado pela *Online Computer Library Center - OCLC* e pelo *National Center for Supercomputing Applications – NCSA*.

Segundo a DCMI, o termo *"Dublin"* foi cunhado pelo fato da referida oficina ter sido realizada em Dublin, Ohio, nos Estados Unidos da América e *"Core"* porque os seus elementos são amplos e genéricos para a descrição de recursos.

O padrão DC surgiu pelo aumento significativo de recursos eletrônicos e pela incapacidade dos bibliotecários para catalogá-los, definindo um conjunto de elementos que pudessem ser usados por autores para descrever seus próprios

[30] Mecanismos de busca em que especialistas em Ciência da Informação estabelecem categorias, enquanto que indexadores indexam manualmente páginas da *Web* às categorias estabelecidas. O *Open Directory* é um projeto de comunidade de indexadores que visa colmatar o número excessivo de páginas da *Web* (ROCHA, 2004, p. 112).

recursos na Web. Por isso que foi desenvolvido para ser simples e conciso na descrição. Todavia, o esquema tem sido usado com outros tipos de materiais e em aplicações que exigem alguma complexidade. Segundo a NISO (2004), a tal complexidade é justificada por alguns defensores de uma visão estruturalista que, contrapondo-se à visão simplicista, enfatizam a necessidade de distinções semânticas complexas e extensas para comunidades específica.

Atualmente o padrão DC reúne uma comunidade dedicada à promoção e adoção generalizada de padrões de metadados interoperáveis, bem como ao desenvolvimento de vocabulários específicos de metadados para descrever recursos com vista à sua descoberta através de sistemas inteligentes. A referida organização denomina-se *Dublin Core Metadata Initiative (DCMI)* e atua sob o lema “tornando fácil encontrar a informação”.

No geral, o conjunto completo de vocabulários ou termos de metadados de DCMI não só inclui os quinze elementos, como também o conjunto de classes de recursos, esquemas de codificação de vocabulário e de sintaxe. Para garantir a interoperabilidade que é a característica fundamental do padrão *Dublin Core*, recomenda-se o uso de vocabulários controlados registrados pela DCMI, ao contrário, torna-se obrigatório o seu registro como DCMI.

4.6.2 Níveis de interoperabilidade do padrão DC

O desenvolvimento de metadados capazes de serem compartilhados ou re-utilizados por diversas pessoas, organizações, serviços, etc. em diferentes plataformas tecnológicas compatíveis constitui um dos pressupostos fundamentais da DCMI. Neste contexto, a iniciativa destaca os seguintes níveis de interoperabilidade:

- Nível 1 (definições do termo compartilhado): neste nível a interoperabilidade entre os metadados usando aplicativos é baseada na partilha de definições em linguagem natural. Em ambientes de aplicação, como *intranet* ou sistema de biblioteca os usuários escolhem os termos a usar em seus metadados na forma como se encontram definidos;
- Nível 2 (interoperabilidade semântica formal): neste nível a interoperabilidade entre os metadados usando aplicativos é baseada

no modelo compartilhado formal proporcionada pela RDF, que é usado para apoiar o linked data[31];

- Nível 3 (conjunto de descrição da interoperabilidade sintática): neste nível os aplicativos são compatíveis com o modelo de dados vinculados e compartilham uma sintaxe abstrata para registros de metadados comprováveis, o "conjunto de descrição";
- Nível 4 (conjunto de descrição da interoperabilidade do perfil): neste nível os registros são trocados entre metadados usando aplicativos seguintes.

Os níveis 3 e 4 são experimentais comparativamente aos níveis 1 e 2, na medida em que não são bem apoiados com ferramentas de *software*.

4.6.3 Elementos do padrão DC

Na sua origem, o padrão DC era composto por 13 elementos. Na versão 1.1 é composto por quinze elementos que, segundo a DCMI, entraram em normalização desde 1998, tendo evoluído para atuais noções da *Web* Semântica através da inclusão da atribuição de domínios formais, intervalos e definições em linguagem natural. Estes elementos por um lado permitem o compartilhamento por diversas áreas de conhecimento ou instituições na organização e classificação das informações e, por outro, podem ser suportados por vários motores de busca, como *Ultraseek, Swish-E, Microsoft's Index Server, Blue Angel Technologies MetaStar, Verity Search 97 Information Server*.

Rocha (2004) considera que os quinze elementos encontram-se organizados em três grupos, a saber: conteúdo, propriedade intelectual e instância. Neste sentido, descrevem o conteúdo do recurso os seguintes elementos: título, assunto, descrição, linguagem, fonte, relação e abrangência. Por sua vez, são relativos à propriedade intelectual os elementos: criador, publicador, contribuidor e direitos. Por último, a data, o tipo, o formato e o identificador são relativos à instância.

[31] *"Linked Data"* conforme a iniciativa DCMI descreve uma prática mais indicada para expor, compartilhar e se conectar pedaços de dados, informações e conhecimento sobre a *Web* Semântica usando endereços da *Web* e *RDF*.

Tabela 5: *Elementos do padrão Dublin Core, Versão 1.1: Descrição de Referência*

Categoria	Elemento	Nome	Identificador	Definição	Comentário
Conteúdo	Título	Título	Title	O nome dado ao recurso	Tipicamente, um Título será o nome pelo qual o recurso é formalmente conhecido
	Assunto	Assunto e Palavras Chave	Subject	Tópicos do conteúdo do recurso	Tipicamente, um Assunto deverá ser expresso por palavras chave, frases ou códigos de classificação que descrevem o conteúdo do recurso. Como boa prática recomenda-se a seleção de termos de vocabulários controlados ou de sistemas de classificação formais
	Descrição	Descrição	Description	Uma descrição do conteúdo do recurso	Descrições podem incluir, sem estarem limitadas a tal: um resumo, um índice, uma referência a uma representação gráfica do conteúdo ou uma descrição textual
	Linguagem	Língua	Language	A língua do conteúdo intelectual do recurso	Como boa prática recomenda-se para valores do elemento Língua a utilização do RFC 1766 [RFC1766], o qual inclui um código de língua de duas letras (retirado da norma ISO 639 [ISO639]), seguido opcionalmente por um código de duas letras para o país (retirado da norma ISO 3166 [ISO3166]). Por exemplo, 'en' para Inglês, 'fr' Francês, ou 'en-uk' para o Inglês do Reino Unido
	Fonte	Fonte	Source	Uma referência a um recurso de onde o presente recurso possa ter derivado	O presente recurso pode ter derivado do recurso Fonte na sua totalidade ou apenas em parte. Como boa prática recomenda-se a referência ao recurso fonte através de um identificador em conformidade com um sistema de identificação formal
	Relação	Relação	Relation	Uma referência a um recurso relacionado	Como boa prática recomenda-se referir o recurso através de uma cadeia de caracteres ou número em conformidade com um sistema de identificação formal
	Abrangência	Cobertura	Coverage	A extensão ou o alcance do recurso	Cobertura inclui tipicamente uma localização espacial (o nome de um lugar ou coordenadas geográficas), um período no tempo (a sua designação, data ou intervalo de tempo), ou jurisdição (o nome de uma entidade administrativa). Como boa prática recomenda-se a seleção de valores de vocabulários controlados (como por exemplo o "Thesaurus of Geographic Names" [TGN]), devendo ainda ser usados, quando apropriado, preferencialmente nomes de lugares e designações de períodos no tempo, em vez de identificadores numéricos tais como coordenadas ou intervalos de datas
Propriedade Intelectual	Criador	Criador	Creator	A entidade responsável em primeira instância pela existência do recurso	Exemplos de Criador incluem uma pessoa, uma organização ou um serviço. Tipicamente, o nome de um criador deve ser usado para indicar uma entidade
	Publicador	Editor	Publisher	Uma entidade responsável por tornar o recurso acessível	Exemplos de um Editor incluem uma pessoa, uma organização ou um serviço. Tipicamente, o nome de um Editor deve ser usado para indicar a entidade
	Contribuidor	Outro Contribuinte	Contributor	Uma entidade responsável por qualquer contribuição para o conteúdo do recurso	Exemplos de Outro Contribuinte incluem uma pessoa, organização ou serviço. Tipicamente, o nome de um Outro Contribuinte deve ser usado para indicar a entidade
	Direitos	Gestão de Direitos	Rights	Informação de direitos sobre o recurso ou relativos ao mesmo	Tipicamente, este elemento deverá conter uma declaração de gestão de direitos sobre o recurso ou uma referência a um serviço que fornecerá essa informação. Tal poderá compreender informação sobre direitos de propriedade intelectual, direitos de autor ou outros. A ausência deste elemento não permite formular qualquer hipótese válida sobre quaisquer direitos que possam incidir sobre o recurso
Instância	Data	Data	Date	Uma data associada a um evento do ciclo de vida do recurso	Tipicamente, uma Data deve ser associada à criação ou disponibilidade do recurso. Como boa prática recomenda-se para codificação de valores de datas um perfil da norma ISSO 8601 [W3CDTF], segundo o formato AAA-MM-DD
	Tipo	Tipo do Recurso	Type	A natureza ou gênero do conteúdo do recurso	Tipos incluem termos descrevendo categorias genéricas, funções, gêneros, ou níveis de agregação para o conteúdo. Recomenda-se como boa prática a seleção de valores a partir de vocabulários controlados (por exemplo, a lista do documento de trabalho "Dublin Core Types" [DCT1]). Para descrever a manifestação física ou digital do recurso, deve ser usado o elemento Formato
	Formato	Formato	Format	A manifestação física ou digital do recurso	Tipicamente, o Formato deve incluir o tipo de meio do recurso, ou as suas dimensões. Este elemento deve ser usado para determinar as aplicações informáticas ou qualquer tipo de equipamento necessário para reproduzir ou operar com o recurso. Exemplos de dimensões incluem tamanho e duração. Como boa prática recomenda-se a seleção de valores a partir de vocabulários controlados (como por exemplo a lista de "Internet Media Types" [MIME] definindo formatos e meios)
	Identificador	Identificador do Recurso	Identifier	Uma referência não ambígua ao recurso, definida num determinado contexto	Como boa prática recomenda-se a identificação do recurso por meio de uma cadeia de caracteres ou por um número de acordo com um sistema de identificação formal. Exemplos de sistemas de identificação formais incluem o "Uniform Resource Identifier" (URI) (incluindo o "Uniform Resource Locator" (URL)), o "Digital Object Identifier" (DOI) e o "International Standard Book Number" (ISBN)

Fonte: *Criada pelo autor com base na tradução portuguesa da DCMI.*

Conforme a NISO (2004), os qualificadores podem ser utilizados para refinar ou restringir o escopo de um elemento ou para identificar o esquema de codificação usado na representação de um valor de elemento. Por exemplo, o elemento "Data" pode ser usado com o qualificador de refinamento "Criação" para limitar o significado do elemento "Data" com o dia em que o objecto foi criado. A "Data" também pode ser usada com um qualificador de esquema de codificação para identificar o formato em que a data é gravada ou registrada, por exemplo, conforme a norma ISO 8601 para a representação de data e tempo.

Todos os elementos do padrão DC da tabela 5, segundo a NISO (2004), são opcionais, podem ser repetidos e podem ser apresentados em qualquer ordem. Embora a descrição *Dublin Core* recomende o uso de valores controlados para campos onde eles são apropriados (por exemplo, vocabulários controlados para o campo Assunto), isso não é obrigatório. No entanto, os grupos de trabalho foram criados para discutir listas de autoridade, para certos elementos, tais como tipo de recurso. Enquanto o padrão DC deixa regras de conteúdo para a implementação

particular, a DCMI incentiva a adoção de perfis de aplicação (regras específicas de domínio) para domínios específicos, como educação e governo.

A figura 5 ilustra um exemplo do uso do esquema de metadados DC para a descrição do texto. Pode-se constatar que apenas foram utilizados dez elementos do padrão DC, o que comprova a opção facultativa do seu uso. O elemento *Title* (título) é relativo ao conteúdo e indica o nome do recurso *"metadata demystified"*.

Figura 5: *Exemplo do uso do Padrão Dublin Core*

Dublin Core Example

Title="Metadata Demystified"
Creator="Brand, Amy"
Creator="Daly, Frank"
Creator="Meyers, Barbara"
Subject="metadata"
Description="Presents an overview of metadata conventions in publishing."
Publisher="NISO Press"
Publisher="The Sheridan Press"
Date="2003-07"
Type="Text"
Format="application/pdf"
Identifier="http://www.niso.org/standards/resources/Metadata_Demystified.pdf"
Language="en"

Fonte: *National Information Standards Organization – NISO (2004).*

O elemento *Creator* (criador) indica a propriedade intelectual da criação ou entidade responsável em primeira instância pela existência do recurso e, conforme se referiu anteriormente, a sua repetição é permitida; neste caso, para acomodar todos os nomes dos criadores do recurso. O mesmo acontece com o elemento *Publisher* (editor), também relativo à propriedade intelectual, que foi repetido para acolher os responsáveis por tornar o recurso acessível (NISO Press e The Sheridan Press, respectivamente).

O elemento *Subject* (assunto), relativo ao conteúdo do recurso, indica os tópicos tratados no recurso texto. O elemento *Description* (descrição) descreve o conteúdo do recurso, neste caso, descrição textual. O elemento *Date* (data) relativo à instância, indica a data associada à criação ou disponibilidade do recurso. O

elemento *Type* (tipo), também relativo à instância, indica a natureza ou gênero do recurso (texto). O elemento *Format* (formato), igualmente relativo à instância, indica as características físicas ou digitais do recurso (aplicação PDF). O elemento *Identifier* (identificador) indica a referência ao recurso, definida pelo contexto. Por último, o elemento *Language* (linguagem) indica a língua do conteúdo do recurso, neste caso, inglesa.

Conforme a NISO (2004), devido à sua simplicidade, o uso do padrão DC é atualmente extendido para outras comunidades de pesquisadores além de bibliotecários, por exemplo, para curadores de museus, colecionadores de música, entre outras. Há centenas de projetos em todo o mundo que utilizam o DC, ou para catalogar, ou para coletar dados a partir da Internet. Os temas variam de patrimônio cultural e de arte para matemática e física.

A transversalidade do esquema DC por diversas áreas científicas contribui para a sistematização do conhecimento universal, não só pela possibilidade de armazenamento e preservação da informação, como também pela possibilidade de intercâmbio de dados nas relações humano-humano, humano-máquina e máquina-máquina. Este intercâmbio permite que a informação disponibilizada possa ser testada por outros interlocutores do processo da comunicação, agregando mais valor à informação inicial ou empregando outras tecnologias que conduzam à novas inferências.

A representação da informação em ambientes digitais constitui atualmente um dos grandes desafios da área da Ciência da Informação, pelo fato de consistir em processos subjetivos que envolvem variações linguísticas e problemas comunicacionais. Além da representação, a área da recuperação também é caracterizada pelos mesmos problemas que resultam em ambiguidades e imprecisões. O Capítulo seguinte analisa minuciosamente alguns aspectos envolvidos na localização, identificação, acesso e uso da informação.

5

Recuperação da Informação

5.1 Visão sobre o Capítulo

Na seção analise o processo de recuperação, uma das linhas de atuação por meio da qual a CI ganha notoriedade no seio de outros campos científicos, mostrando-se a complementaridade dos processos e métodos de representação descritos no Capítulo anterior. De igual modo, analisa-se a contribuição da CI, sob o ponto de vista de recuperação, no conhecimento humano. Neste contexto, abordam-se alguns aspectos da complexidade oriundos da pós-modernidade e prevalencentes na Web, mostrando-se os desafios do campo, principalmente sobre a ambiguidade e efeitos dos SRI.

5.2 Conceito da recuperação da informação

Ferneda (2003) considera que o termo RI foi acolhido numa concepção dualista na área da CI, em termos de funcionalidade. Deste modo, por um lado significa a seleção de documentos a partir do acervo, em função da procura do usuário; por outro, existe o fornecimento de elementos de informação documentária, com base na procura definida pelo próprio usuário. O autor considera ainda que o termo pode-se referir ao fornecimento de uma resposta elaborada a uma demanda ou até ao processo em si do tratamento da informação, como catalogação, indexação e classificação.

Não raras vezes, os SRI são confundidos com os sistemas de bancos de dados. Porém, os dois são distintos e a sua principal diferença, de acordo com Ferneda (2003), reside na natureza do objeto por eles tratado, ou seja, os SRI lidam com objetos linguísticos, como textos, imagens, áudio; enquanto que os sistemas de banco de dados organizam itens de informação ou dados.

Para Silberschatz et al. (2006), os sistemas de informação e os sistemas de banco de dados congregam um aspecto comum, no que tange ao armazenamento e recuperação de informações, embora o foco dos sistemas de informação seja de consultas baseadas em palavras-chave, relevância dos documentos recuperados em relação à consulta, bem como a classificação e a indexação de documentos, ações baseadas em dados consideravelmente não estruturados. Para os sistemas de banco de dados, as ações concentram-se em dados estruturados e organizados em modelos complexos, como controle de transações. Neste sentido, a RI em geral

consiste em localizar documentos relevantes, com base na entrada do usuário, como palavras-chave ou documento de exemplo.

Para Baeza-Yates e Ribeiro Neto (1999), o usuário de um sistema de recuperação está mais preocupado em conseguir informações sobre um assunto do que com a recuperação de dados que satisfazem uma determinada consulta. Por isso, na recuperação de dados existe uma observância estrita aos erros nos objetos recuperados do que na recuperação da informação, pois, enquanto a recuperação da informação lida com textos em linguagem natural que, às vezes, não estão estruturados ou têm uma semântica ambígua, a recuperação de dados lida com dados estruturados e com semântica bem definida. Segundo os autores, a outra diferença entre os dois sistemas se prende ao conceito da relevância. Enquanto todos os dados recuperados são relevantes à consulta, na recuperação da informação, às vezes, nem todos os documentos recuperados satisfazem a necessidade informacional do usuário. Assim, uma das dificuldades da recuperação da informação consiste na interpretação sintática e semântica de um documento, de modo a determinar a sua relevância em relação à necessidade específica do usuário, naquele contexto da busca.

Recuperar informações consiste em identificar, localizar e acessar objetos de informação relevantes dentro de uma coleção, através de consultas que visam satisfazer determinadas necessidades informacionais do usuário. É um processo de busca seletiva da informação, na medida em que nem sempre a totalidade das informações do corpus satisfaz as condições e especificidades da necessidade do usuário, exteriorizada pela consulta.

De acordo com Chu (2007), a recuperação visa o *information access, information seeking* e *information searching,* isto é, o acesso, a busca e a pesquisa de informações, três conceitos que embora correspondam à recuperação da informação, têm diferentes implicações no processo em si. O acesso corresponde ao aspecto da obtenção da informação, a busca incide sobre o usuário que intervém no processo da recuperação e a pesquisa ao aspecto de como procurar as informações, isto é, sobre as estratégias.

Na atual fase da sociedade em rede, outros conceitos como *data mining*[32] (garimpagem de dados) e *resource discovery* (descoberta de recursos) são usados pelos profissionais da informação para associá-los ao processo de RI, mesmo faltando a sua incorporação no campo (CHU, 2007).

A DCMI considera que na terminologia da *Web*, recurso é qualquer coisa endereçável através de uma *URL*, embora as implementações do padrão em si não sejam necessariamente baseadas na *Web*, isto é, o padrão de metadados *Dublin Core* pode ser igualmente para descrever qualquer tipo de recursos, incluindo várias coleções de documentos e formatos não eletrônicos de mídia, em ambientes como museus ou bibliotecas.

Para efeitos da *Web*, cada página HTML é considerada como um documento que pode ser recuperado na sua totalidade ou em classes específicas através das palavras-chave fornecidas pelos usuários. Estes documentos abrangem não só dados textuais, como também dados de vídeo, áudio e imagem (SILBERSCHATZ et al., 2006).

Silberschatz et al. (2006), apontam a importância da recuperação de texto completo para documentos não estruturados, na medida em que o processo considera todas as palavras do documento como palavras-chave ou termos do documento, já que pode não haver informações sobre palavras que devam ser consideradas como chaves para a busca. Contudo, a recuperação do texto completo é negativada pela maior quantidade de documentos irrelevantes à pesquisa, fato que levou ao desenvolvimento de sistemas de classificação de relevância usando termos[33].

O maior desafio dos SRI, que se estende à CI, consiste na oposição entre a exaustividade e a especificidade. Se por um lado o processo de representação visa à recuperação da maior quantidade possível de documentos sobre certos assuntos, por outro, os documentos recuperados devem corresponder à menor quantidade possível conforme a similaridade com a consulta e, por conseguinte, as necessidades informacionais de cada usuário ou relevância. Por outras palavras, a

[32] Mineração de dados ou descoberta de novas informações em função da complexidade dos padrões ou regras em grandes quantidades de dados, isto é, resulta como alternativa para o usuário final já que muitas informações ficam armazenadas em bancos de dados cujo acesso implica o conhecimento do esquema de banco de dados ou da *Structured Query Language (SQL)* - Linguagem de Consulta Estruturada, de modo a formular a consulta para recuperar os dados através da seleção das colunas e linhas por meio da álgebra relacional (ELMASRI; NAVATHE, 2005, p. 624).

[33] Silberschatz et al. (2006) usam a expressão termos para se referir às palavras-chave, já que para o texto completo, qualquer palavra considera-se como chave.

essência da RI incide sobre a necessidade de obter e ordenar documentos relevantes na coleção. Este desafio suscita novas abordagens para cada estágio de desenvolvimento e sofisticação dos SRRI, principalmente pela proporção do volume de informações e versatilidade da tecnologia de produção, processamento, armazenamento e disseminação.

5.3 Sistemas de Representação e Recuperação da Informação

Os SRRI são objetos de controvérsias na atualidade, oriundas dos diferentes tipos de conceituação, abordagem e percepções sobre a sua origem. Alguns autores denominam-nos sistemas de representação e recuperação da informação, alegadamente porque a representação constitui o fator determinante para a recuperação da informação; para outros, são apenas sistemas de recuperação da informação porque a recuperação envolve a representação. Outros autores ainda consideram que os sistemas de informação, principalmente os transacionais, também se enquadram nos SRRI, na medida em que suportam operações de entrada e alimentação de dados, processamento e armazenamento e geração de documentos e relatórios que serão consolidados em outros sistemas para a satisfação de diversas necessidades informacionais do usuário final, como a tomada de decisões.

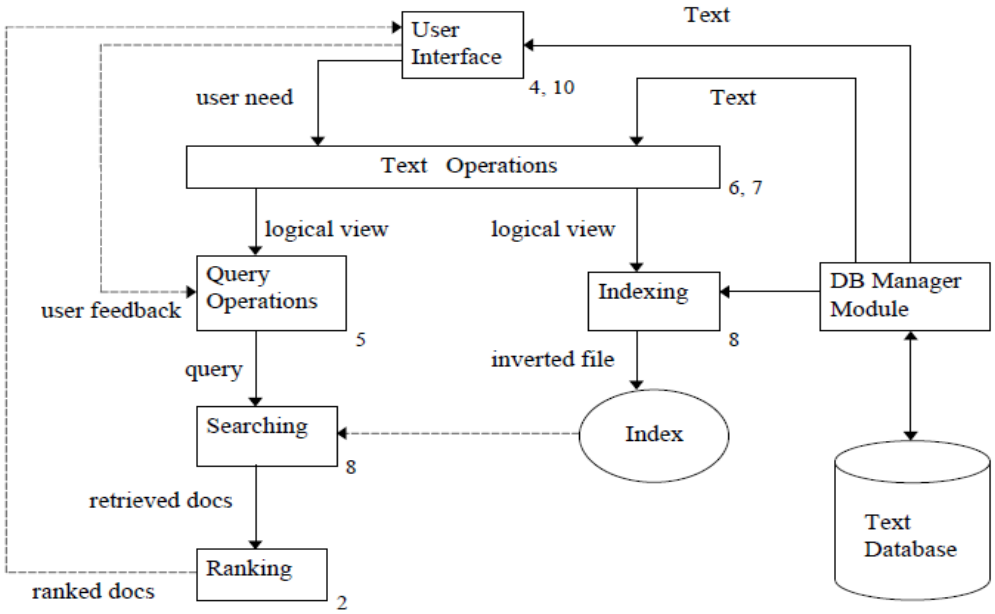
Baeza-Yates e Ribeiro Neto (1999) denominam os sistemas de representação e recuperação da informação apenas por sistemas de recuperação da informação, pois, além da recuperação propriamente dita, também lidam com a representação, a preservação, a organização, e o acesso de itens informacionais.

De acordo com Chu[34] (2007), os SRRI remontam desde a segunda metade do século XIX, com a criação dos esquemas de organização e acesso do conhecimento, como a classificação decimal de Dewey, momento em que foram caracterizados como sistemas de indexação, busca, processamento e gerenciamento da informação. Porém, o seu interesse só despertou atenção para a maioria dos pesquisadores da CI no final da Segunda Guerra Mundial.

O SRRI é composto por três elementos principais: o banco de dados, o mecanismo de busca e a interface. Além destes elementos, a recuperação da informação inclui outros componentes ilustrados na figura 6.

[34] No trabalho adotou-se a terminologia “Sistemas de Representação e Recuperação da Informação”, segundo Chu (2007).

Figura 6: *Processo de representação e recuperação da informação.*



Fonte: *Baeza-Yates e Ribeiro Neto (1999, p.10)*

Conforme Baeza-Yates e Ribeiro Neto (1999), no banco de dados são especificados os documentos que serão utilizados, as operações necessárias e o modelo do texto ou os elementos a serem recuperados.

De acordo com Chu (2007) o banco de dados em representação e organização da informação corresponde a um determinado nível de abstração. Esta abstração, segundo Silberschatz et al. (2006), é desenvolvida pelo projetista do banco de dados em diferentes níveis de complexidade da organização dos dados no sistema, conforme três níveis de visão: *view*, lógico e conceitual.

A visão lógica, segundo Baeza-Yates e Ribeiro Neto (1999), é definida pelas operações do texto. Uma vez definida a visão lógica dos documentos, o gerenciador de banco de dados cria um índice do texto, de modo a permitir uma pesquisa rápida em grandes volumes de informações.

Os termos de índice são criados em observância aos seguintes princípios sobre pesos: um documento representa-se a partir dos termos de índice, e todos os termos têm a mesma importância para representar a sua semântica; cada termo tem um peso que corresponde à sua importância; os termos mais frequentes serão mais representativos. Assim, a criação dos ficheiros de índice tem o objetivo de armazenar os termos representativos dos documentos. O processo da sua criação envolve o processamento do texto nos documentos, a análise lexicográfica, a

ponderação de termos significativos, o *stemming*, e os problemas de sinonímia e polissemia.

Segundo Baeza-Yates e Ribeiro Neto (1999), após a indexação, a recuperação inicia com a especificação da necessidade do usuário, através da consulta ou busca. O sistema processa a busca com base na estrutura do índice, classifica os documentos recuperados segundo a relevância, e retorna o resultado para o usuário. Assim, o usuário pode avaliar os resultados da sua busca e selecionar os documentos do seu interesse para o *feedback* da relevância.

O mecanismo de busca do usuário também é extremamente importante para a recuperação da informação. Segundo Chu (2007), o grau de sofisticação do mecanismo de busca depende dos respectivos algoritmos de busca e dos procedimentos que o sistema de RI agrega. Os procedimentos da busca podem ser básicos ou avançados, sendo os primeiros encontrados na maioria dos sistemas operacionais de RI; enquanto que os segundos resultam de testes e experiências em laboratórios ou modelos. Para os usuários com pouco treinamento ou sem a experiência de busca, recomenda-se o uso de procedimentos básicos por agrupamento de algoritmos, como palavras-chave, lógica booleana, truncagem e busca por proximidade. Já para usuários com formação e experiência, podem-se usar procedimentos avançados adotados em sistemas atuais de RI na *Internet.* Contudo, a organização da informação no banco de dados, aliada às características do sistema de RI adotado e aos objetivos e necessidades dos usuários finais, é que vão determinar o mecanismo de busca ideal para cada situação.

Igualmente, a linguagem é crucial em um sistema de representação e recuperação da informação. Segundo Chu (2007), a informação depende da linguagem, seja escrita ou oral, no processamento, na transferência, ou mesmo na comunicação. A linguagem pode ser dividida em duas categorias: a linguagem natural e o vocabulário controlado. A linguagem natural permite maior grau de especificidade e flexibilidade, na medida em que é normalmente empregue na representação da informação ou na formulação de consultas por usuários sem treinamento, correspondendo à sua linguagem do dia a dia. Já o vocabulário controlado, conforme se refereiu no Capítulo anterior, baseia-se na limitação do vocabulário, da sintaxe, da semântica e da pragmática da linguagem, para adaptá-la a um ambiente ou contexto específico. Por isso, o vocabulário controlado envolve custos elevados pela necessidade tanto da sua arquitetura, como de formação e

treinamento dos usuários finais, porém, reduz a complexidade, a sutileza, e a ambiguidade, que afetam a linguagem natural na representação e recuperação da informação.

Além da linguagem e dos vocabulários controlados, a interface também desempenha um papel preponderante na recuperação. Segundo Shaw (1991 apud CHU, 2007), interface é o que o usuário vê, ouve e toca enquanto interage com um sistema informático. Por isso, na maioria das vezes, ela é usada na representação e recuperação da informação como um elemento de análise do sistema.

Baeza-Yates e Ribeiro Neto (1999) consideram que muitos usuários quase nunca declaram as suas necessidades de informação porque não têm o conhecimento sobre as operações de texto e de consulta, por isso, os sistemas retornam resultados irrelevantes. Para minimizar este aspecto, recomenda-se o uso de interfaces amigáveis. Por exemplo, um sistema amigável deve garantir uma boa interação com o usuário, através de menus de seleção de fácil localização e percepção; de recursos de exibição como *layout* da tela, tipo de letra, cores, organização de texto e imagens, etc.

Atualmente no desenvolvimento de interfaces, principalmente para ambientes digitais, observam-se várias recomendações da W3C que visam fundamentalmente a usabilidade e acessibilidade dos sistemas para todos os usuários, incluindo os portadores de necessidades especiais.

5.4 História dos Sistemas de Representação e Recuperação da Informação

A história dos sistemas de recuperação da informação, segundo Chu (2007), é categorizada em quatro maiores estágios, a saber: processamento elevado, rápido desenvolvimento, desmistificação e rede.

O estágio de processamento elevado (de 1940 a inícios de 1950) foi marcado pela Segunda Guerra Mundial, através do aumento considerável da quantidade de relatórios técnicos e documentos produzidos para anotar as pesquisas e atividades desenvolvidas sobre armamento e estratégias de guerra. Este aumento do volume da informação suscitava o aperfeiçoamento dos aspectos de processamento e gerenciamento, como a seleção, disseminação e preservação da informação. De acordo com Gull (1956 apud CHU, 2007), a quantidade de documentos e relatórios levou ao abandono de algumas técnicas individuais de organização da informação,

como habilidade e memória e levou à conjugação de esforços que culminaram com sistemas específicos de recuperação eficiente da informação, como o manual de índices coordenados introduzido em 1951.

A história dos SRRI antecede a do próprio campo da CI. Este entendimento procede na medida em que alguns processos, técnicas e terminologias atualmente aplicadas nos SRRI são oriundos de outras áreas e foram posteriormente incorporados pelo campo. Ademais, a sua abordagem teve início antes da institucionalização da CI como campo científico e a maioria dos pesquisadores notáveis na área foi formada noutras áreas de conhecimento.

O estágio do rápido desenvolvimento (de 1950 a 1980) foi marcado pelo rápido desenvolvimento e sofisticação dos SRRI, através do envolvimento de computadores entre 1957 e 1959, destacando-se, segundo Salton (1987 apud CHU, 2007), os usos de Hans Peter Luhn, não apenas para a combinação de palavras-chave e classificação de tarefas, como também para a análise do conteúdo.

Chu (2007) aponta que a emergência dos sistemas online, como o DIALOG entre os anos 1960 e 1970 introduziu mudanças significativas na organização e recuperação da informação. Consequentemente, provocaram a migração dos sistemas manuais para a recuperação computarizada da informação. Esta etapa foi marcada por características como emergência de tesauros online, classificação dos emissores ou fontes de documentos, inclusão automática de sinônimos nos formulários de busca, lógica booleana e busca da linguagem natural do texto livre. Neste contexto, alguns sistemas também passaram a incluir programas de coleção automática de dados com vista a monitorar o uso e satisfação dos usuários.

Ao lado dos sistemas online, técnicas automáticas e de automação para a recuperação da informação foram criados e experimentados pela tecnologia da computação, envolvendo pesquisas de muitos profissionais desta área (CHU, 2007).

Considerando o período pós-segunda Guerra Mundial, mormente o artigo de vannevar Bush de 1945 e a Conferência da Royal Society de 1948 como marcos importantes da CI, este foi o período em que a área começou a ganhar a notariedade no campo científico do conhecimento, pois evidenciava enúmeras possibilidades na gestão, organização e comunicação da informação, através dos mais revolucionários processos de transferência da informação. Se, no passado, a interação e apropriação da informação se baseavam no livro ou nos periódicos

científicos, a introdução dos computadores vislumbrava novos contextos que a rede viria a comprovar.

Do ponto de vista científico, este período também foi caracterizado pela institucionalização da RI como disciplina, principalmente com pesquisas de Gerard Salton sobre o modelo de espaço vetorial e conceito de relevância na recuperação da informação.

O estágio de desmistificação (1980 a 1990) foi essencialmente caracterizado pelo uso efetivo dos SRRI para atender o usuário final. Segundo Chu (2007), embora muitos dos sistemas online descritos haviam sido construídos para atender aos usuários finais, com diferentes tipos de necessidades informacionais, pecavam pelo fato de não especificar os mecanismos de busca para os mesmos. Este fato prejudicava a busca, especialmente para usuários sem o adequado treinamento, o que acarretava mais custos pela contratação de assistentes bibliotecários ou outros profissionais de informação que passavam a executar tais tarefas, em benefício dos usuários finais.

O termo usuário final que se confundia com a referida complexidade dos SRRI só começou a ganhar outro significado, tanto com o emprego de computadores pessoais na recuperação da informação, como do *CD-ROM* e de sistemas de catálogos públicos online – *online public access catalog (OPAC)*, nos meados de 1980. No passado, os sistemas online de recuperação da informação eram acessados por vários recursos, como terminais de impressão e tubos de raios catódicos – *cathod Ray tube (CRT)*, por isso, a interação dos sistemas com o usuário que buscava as informações não era atraente ou amigável. Este cenário só mudou com a introdução de computadores pessoais na RI, na medida em que estes passaram a intermediar diretamente a interação entre o usuário e o sistema. De igual modo, a implementação do *CD-ROM* e dos sistemas *OPAC* melhorou a interação com o usuário final, desmitificando os SRI que eram providos online apenas em alguns lugares locais ou terminais. Posteriormente, os SRRI foram progressivamente desenvolvidos para atender os usuários finais, sem a necessidade da presença da intermediação de bibliotecários ou profissionais de informação (CHU, 2007).

O estágio da rede (de 1990 ao presente) é marcado principalmente pelo uso de computadores pessoais e de dispositivos móveis na recuperação ubiquitária da informação, através da rede distribuída.

Segundo Chu (2007), até o ano de 1990 a RI era uma atividade centralizada, pois os respectivos bancos de dados eram fisicamente gerenciados somente em uma única central. Neste sentido, o uso de vários sistemas de RI por diferentes usuários implicava a conexão da central com respectivos sistemas individuais de RI, para que a busca distribuída permitisse o acesso a informações do banco de dados através da rede. Este cenário foi completamente modificado com o advento da *internet* que além de fornecer a infra-estrutura para o acesso geral da rede, também redefiniu o campo da RI, através da aplicação da estatística de palavras-chave, informação de multimídia e outros aspectos. Ainda de acordo com o autor, a recuperação do texto completo tornou-se a norma em vez de exceção na *Internet* e as técnicas de recuperação que anteriormente tinham sido testados em laboratórios ficaram disponíveis nos sistemas de recuperação da *Internet*, como o Alta Vista e Google.

Com a rede, a CI expandiu ainda mais o seu teor científico, possibilitando a disseminação da informação que era concentrada nos institutos de pesquisa, instituições de curadoria, igrejas, entre outros. De igual modo, o acesso à informação permitia a intervenção no conhecimento de usuários situados em diferentes áreas geográficas, contextos políticos e sociais. Assim, a atividade de reflexão e busca pelo entendimento perdia o caráter eminentemente filosófico e ganhava o caráter social, envolvendo mais usuários no ciclo informacional, motivados pela utilidade social.

Algumas questões que se prendem ao acesso à informação podem ser traduzidas pela crise do poder. No passado, a maioria dos sistemas políticos controlava os respectivos cidadãos através de medidas restritivas à informação. Com a *Internet*, muitas informações passaram a ser revestidas pelas características de disponibilidade, acessibilidade, ubiquidade e publicidade. Por isso, os movimentos sociais que surgiram nos anos 30 com o desenvolvimento do capitalismo, foram consolidados pelos ideais de sentimento de injustiça, eficácia de grupo, identidade social e afetividade.

Atualmente, a fase de rede para os SRRI é marcada por abordagens semânticas que procuram agregar maior significado e valor à informação. Neste sentido, a correlação de relações interdisciplinares é massificada pela comunicação e transformação social pela informação. A Filosofia, a Psicologia e a Inteligência Artificial são algumas das áreas que contribuem respectivamente com ontologias,

cognição e comportamento e algoritmos que permitem o aprendizado da máquina, de modo a proporcionar melhores respostas às necessidades informacionais dos usuários.

5.5 Pesquisadores pioneiros e suas contribuições na área dos SRRI

Segundo Chu (2007), muitos pesquisadores contribuíram para o incremento da área dos SRRI, contudo o maior destaque vai para os trabalhos de Mortimer Taube, Hans Peter Luhn, Calvin N. Mooers, Gerard Salton e Karen Spärck Jones.

5.5.1 Mortimer Taube (1910–1965)

De acordo com Chu (2007), Taube trabalhou como bibliotecário nas áreas de circulação, catalogação e aquisição antes de tomar a posição de assistente chefe da referência geral e bibliografia da Biblioteca do Congresso, em 1945. De acordo com Smith (1993 apud CHU, 2007, tradução nossa), em 1952 fundou uma empresa sobre a Documentação, explorando com seus colegas os novos métodos de indexação e recuperação da informação, sob contrato da *U.S. Armed Services Technical Information Agency*. A sua nova abordagem de indexação e busca tornou-se uma indexação coordenada[35].

Conforme Chu (2007), os novos métodos de indexação e recuperação tinham dupla necessidade, por um lado, para contrariar a ação dos sistemas manuais de indexação e recuperação existentes na época, limitada pela imensa quantidade de relatórios técnicos e outras literaturas científicas resultantes das pesquisas durante a Segunda Guerra Mundial; por outro, para acomodar nos dois métodos estabelecidos de representação da informação (o alfabético e o hierárquico) novas disciplinas, novas tecnologias e novas terminologias que envolviam pesquisas e desenvolvimento relacionados à Segunda Guerra Mundial.

A indexação coordenada é baseada na implementação de unitermos e aplicação da lógica booleana na RI. Unitermos são termos particulares selecionados por indexadores para representar as várias facetas de um documento e podem ser considerados como palavras-chave, na medida em que são extraídos de

[35] Segundo Chu (2007, tradução nossa), Taube e Alberto F. Thompson apresentaram um relatório intitulado *"The Coordinate Indexing of Scientific Fields"* – indexação coordenada para campos científicos, que embora nunca tenha sido oficialmente publicado em jornais ou livros, Gull (1987) incluiu posteriormente como apêndice em uma das suas publicações.

documentos originais no esforço para o vocabulário controlado, por exemplo, controle de palavras sinônimas e homógrafas (CHU, 2007).

Segundo Chu (2007), a lógica booleana é uma subdivisão da Filosofia proposta por George Boole[36], em 1849, sobre o fundamento da sua análise sob o aspecto do processo de raciocínio humano e das leis que governam as operações da mente. A lógica baseia-se nos operadores *and, or* e *not* e Taube trouxe este princípio para a área de RI, especificamente na indexação coordenada. De igual modo, introduziu a lógica booleana no ambiente informatizado para a organização e busca de informações através de uso de unitermos ou palavras-chave.

Chu (2007) chama atenção para o fato do termo indexação coordenada não se referir totalmente às características de um SRRI, pois parece apenas um método de indexação. Por isso, e de acordo com a terminologia atual, dever-se-á designar representação e recuperação coordenadas, porque também é usada para a busca. Alguns críticos como Gaivota (1987 apud CHU, 2007), questionam se o método de indexação coordenada se baseia na combinação de palavras ou de conceitos ou ideias.

O método também é questionado pelo fato de não prevenir o "*false drops*"[37] ou ruído, dando a possibilidade de recuperação de todos os documentos que contenham os termos da pesquisa. Por exemplo, a busca de documentos sobre "computador de mesa" retornaria tanto documentos relacionados ao termo computador de mesa, como relacionados à mesa de computador ou até que tivessem os termos computador e mesa. Como se pode depreender, a indexação coordenada também é viciada pela maioria dos problemas que afetam os SRI que dependem da linguagem natural, pela limitação às palavras-chaves sem referência a vocabulários controlados.

De acordo com Chu (2007), mesmo com as desvantagens anteriormente descritas, a contribuição de Taube foi notável na área de SRI, pois o sistema de unitermos vinculados à lógica booleana, em comparação aos catálogos de assunto e classificação, apresenta as seguintes características: baixo custo, tamanho reduzido, análise rápida, mais pontos de acesso por cada unidade catalogada, busca rápida,

[36] Boole considera que o processo de raciocínio ou é de adição de vários conceitos ou classes de objetos para formar mais conceitos complexos, ou de separação de conceitos complexos em individuais ou conceitos simples (CHU, 2007, tradução nossa).

[37] O termo *false drop* refere-se à recuperação de documentos que não são verdadeiramente relevantes para a pesquisa, isto é, só se alcança a revocação perfeita se houver a queda na proporção dos documentos considerados relevantes para o tema da busca.

ritmo lento de crescimento e obsolescência, maior especificidade, universalidade, estrutura lógica, neutralidade, simplicidade e é apropriado para publicações cumulativas.

5.5.2 Hans Peter Luhn (1896–1964)

Segundo Chu (2007), Hans Luhn formou-se como engenheiro em educação e tornou-se o mais famoso inventor da IBM com mais de 80 patentes. As suas pesquisas na área da CI, especificamente na área da RI datam entre 1947 e 1948 quando James Perry e Malcolm Dyson o questionaram sobre a possibilidade de projeção de uma máquina da IBM para pesquisa de estruturas químicas codificadas de acordo com o sistema de notações Dyson[38]. Luhn desenvolveu e testou com eles o pioneiro sistema de busca de informação eletrônica em 1948, chamado *electronic searching selector* – seletor de busca eletrônica e que mais tarde ficou conhecido por *scanner* de Luhn. Em 1953, após várias pesquisas, Luhn publicou o primeiro artigo na área de RI intitulado *"A new method of recording and searching information"*. Doravante, exploraram muitas das importantes aplicações de RI baseadas em computador que atualmente parecem comuns na área.

De acordo com Chu (2007), uma das aplicações desenvolvidas por Luhn é o sistema *Keyword in context (KWIC)*, termo cunhado em 1958 que compreende três elementos essenciais para a representação e recuperação da informação, a saber:

- As palavras-chave, contrariamente às classificações convencionais e cabeçalhos de assunto[39], são empregues para representar e recuperar a pluralidade de facetas do documento;
- Enquanto todas as frases que corporizam o documento formam uma concordância do texto, os títulos ou tópicos das frases formam o contexto para produtos do índice KWIC, ou seja, a partir dos títulos e tópicos extraem-se as palavras-chave;
- A abordagem KWIC também incide sobre a permutação de palavras-chave contidas em títulos ou tópicos. A permutação era muito usada em livros compostos de muitas seções curtas com os seus próprios

[38] Sistema de notação para representar produtos químicos orgânicos, no qual o composto era descrito numa única linha, os símbolos eram usados tanto para os elementos químicos envolvidos, como para os grupos funcionais e sistemas de anéis diferentes (DYSON, 1950).
[39] Alguns códigos antigos incluíam recomendações para a apresentação de cabeçalhos de assunto.

> títulos descritivos, permitindo ao leitor encontrar facilmente um ponto de qualquer palavra de sua posição.

Chu (2007) aponta que Luhn também trouxe contribuições na área da RI, usando métodos estatísticos para desenvolver algoritmos para a produção automática de índices e resumos. A indexação automática se baseia na seleção de palavras essenciais que tenham significados nos documentos, por exemplo, palavras-chave. Assim, palavras que aparecem com maior frequência no documento (ex: artigos, conjunções, preposições) ou que raramente aparecem no documento ou pouco usuais na comunicação ou ainda substantivos ou até termos usados em coleções particulares de documentos podem ser eliminados, adotando uma lista de *stop-words*[40] ou um procedimento estatístico de frequência das palavras.

O uso de *stop-words* para o método de indexação automática pode trazer problemas de precisão e revocação na recuperação dos documentos de áreas específicas, como a medicina na medida em que, por exemplo, a subtração de alguma preposição ou artigo em um laudo médico pode desconstruir o sentido fornecido pelo médico e, por conseguinte, afetar a relevância dos documentos recuperados pelo motor da busca para a necessidade específica da pesquisa.

Já para os resumos automáticos, Chu (2007), considera duas medidas sugeridas para a identificação de palavras significativas e, posteriormente, frases significativas que podem ser mais representativas para um determinado documento. A palavra-chave fornece uma das medidas para a construção de resumos automáticos; enquanto que a outra medida depende da posição relativa dentro de uma frase de palavras significativas. Segundo Luhn (1958 apud CHU 2007), a proximidade de quatro ou cinco palavras insignificantes entre as palavras significantes parece útil para selecionar frases significantes de um documento. A combinação de frequência de palavras-chave e a proximidade de palavras-chave dentro de uma frase é uma metodologia viável para gerar resumos automáticos.

Segundo Chu (2007), Luhn também se destacou no desenvolvimento dos sistemas de disseminação seletiva da informação. A disseminação seletiva da informação é uma aplicação que visa divulgar com eficácia as novas informações científicas para usuários finais ou público-alvo com base em seus perfis. A

[40] Palavras que são filtrados antes ou após o processamento de dados de linguagem natural (texto). A maioria dos motores de busca não consideram algumas palavras ou termos (*stop words*), de modo a economizar o espaço do disco ou acelerar os resultados da busca.

disseminação seletiva da informação tem diversos componentes e etapas de funcionamento, dos quais a criação e manutenção de perfis dos usuários consideram-se tarefas mais importantes e laboriosas. O perfil de interesses dos usuários inclui uma lista de palavras, juntamente com o seu peso ou importância, cada uma indicando o equilíbrio entre as adições e subtrações decorrentes da manutenção do perfil. Neste sentido, o perfil é verificado contra a representação de documentos (ex: abstrações e termos de índices[41]), em um determinado período de tempo (anual, mensal ou semanal).

Embora algumas das contribuições de Luhn tenham sido melhoradas e adaptadas para cada momento da história dos SRI, ele é considerado uma referência obrigatória na área, principalmente pelo desenvolvimento da maioria de aplicações baseadas em computador em uso atualmente.

5.5.3 Calvin Northrup Mooers (1919–1994)

Segundo Chu (2007), Calvin Mooers foi o terceiro pesquisador que contribuiu para a área de RI, principalmente na temática de informação e Ciência da Computação, embora fosse das áreas de Matemática e Física. Em 1950, Mooers cunhou o termo *information retrieval* – recuperação da informação que em seguida foi integrado na área da Ciência da Informação.

O termo recuperação da informação sugere o uso de sistemas para o processo de representação e de recuperação. Assim, a recuperação da informação corresponde ao grau de similaridade entre uma coleção de objetos informacionais relevantes e uma consulta formulada por um usuário, para satisfazer determinada necessidade informacional.

No âmbito da abordagem sobre os SRRI, Mooers criou a seguinte lei: *"an information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it"*, isto é, um sistema de recuperação da informação não tenderá a ser utilizada sempre que se revelar muito complexo e preocupante para o usuário obter informações de que necessita (CHU, 2007).

[41] Descritores que permitem que o acesso seja alcançado através de organização de ferramentas para que o usuário possa procurar sob um ponto de acesso específico ou posição, como autor, título, nome, data.

Segundo Koenig (1987 apud CHU, 2007), a versão da lei de Mooers diz: *"an information system will only be used when it is more trouble not to use it than it is to use it"*, ou seja, um sistema de informação somente será usado se for mais preocupante não usá-lo do que usá-lo. Como se pode depreender, segundo a lei de Mooers, apenas sistemas que refletem as reais necessidades e práticas dos usuários são mais propensos a ser consultados pelos respectivos grupos-alvos.

De acordo com Chu (2007), em 1952 Mooers também desenvolveu o sistema *zatocoding*[42] para armazenar grande número de descritores de documentos em único cartão especial marcado, sobrepondo aleatoriamente códigos descritores de oito dígitos. Este sistema resultaria em apenas um número reduzido, mas tolerável de *false drops* em uma pesquisa bibliográfica, ou seja, em um número reduzido de documentos irrelevantes para a pesquisa, o que nos leva a considerar que, de acordo com a versão da lei de Mooers aliada ao sistema de indexação coordenada de Taube, os usuários poderiam preferi-lo pelo fato de ser o menos desvantajoso entre os dois sistemas.

Mooers também foi responsável pela criação de duas aplicações voltadas para Ciência da Computação: as linguagens *Text Reckoning and Compiling (TRAC)* e *VXM computer*, sendo o primeiro para o tratamento de texto não estruturado em um modo interativo e o segundo, para sistemas de rede multi-computador. Com as suas contribuições na área da recuperação, Mooers teve o seu mérito reconhecido pela *American Society of Information Science*, em 1978, e grande parte das suas ideias encontra-se aplicada na maioria dos atuais sistemas de RI (CHU, 2007).

5.5.4 Gerard Salton (1927 - 1995)

Gerard Salton foi um dos contribuintes mais notáveis no campo de RI. A sua pesquisa incidiu essencialmente sobre o *System for the Manipulation And Retrieval of Texts (SMART)* – sistema para a manipulação e recuperação de textos, conhecido por *"Salton's magical automatic retriever of text"*. Através de pesquisas sobre o modelo de espaço vetorial, a retroalimentação de relevância (ex: usar a relevância dos resultados retornados inicialmente de uma consulta para realizar uma nova

[42] Sistema de seleção baseado em uma mudança automática de cartões perfurados, a *Zatocards*. O sistema liga os códigos e descritores para analisar e recuperar a informação codificada.

consulta), os agrupamentos, a ponderação de termos[43], a recuperação booleana estendida, o valor de descriminação do termo, o dicionário de construção, dependência do termo, a compreensão de texto, e a estruturação e o processamento de texto automático utilizando SMART, Salton mudou fundamentalmente os métodos de processamento de texto completo (CHU, 2007).

Segundo Chu (2007), atualmente muitos sistemas do ramo comercial usam ideias e tecnologias desenvolvidas em SMART, por exemplo, alguns indivíduos autorizados usam a tecnologia na clipagem de notícias ou recorte de jornais.

Salton também introduziu o conceito de relevância na RI, como correspondência ou similaridade contextual entre determinada consulta e informação registrada em documentos. A relevância é uma propriedade da informação ligada à necessidade e contexto do usuário, por isso, traduz a ambiguidade e imprecisão na representação e recuperação.

5.5.5 Karen Spärck Jones (1935 - 2007)

Karen Spärck Jones foi graduada em história e em filosofia e tornou-se a cientista de computação britânica. A sua contribuição para a área de RI, na qual trabalhou por mais de quatro décadas se notabilizou no processamento de linguagem natural (PLN), mas inicialmente trabalhou sobre a abordagem experimental da RI e usou a coleção *Cranfield*[44] para testar o agrupamento de termos para a classificação semântica, que tinha iniciado na sua pesquisa de doutorado. Neste contexto, editou a única monografia alusiva aos métodos experimentais em RI intitulada "*Information Retrieval Experiment"* (CHU, 2007).

Segundo Chu (2007), Jones também se destacou na *Text REtrieval Conference (TREC)*, através de participações como consultora informal para organizadores, membro de comitê do programa e como membro da equipe de pesquisa, tendo contribuído com série de trabalhos de síntese sobre a experiência da TREC. De igual modo, teve destaque na pesquisa sobre métodos estatísticos em RI, refletidos em um dos artigos mais citados da área, publicado em 1972 sob o

[43] Valor atribuído à informação durante a indexação ou recuperação. Ex: No processo de busca, antes da exibição, o computador classifica as referências de acordo com as suas ponderações. Além da ponderação de termos, existe o *query term weighting* - ponderação dos termos da equação de busca e *weighting factor* - fator de ponderação (DICIONÁRIO ELETRÓNICO GRANADA UNIVERSITY 10.5).

[44] Vila situada no noroeste de *Bedfordshire*, na Inglaterra.

tema *Inverse Document Frequency (IDF)* – frequência do documento inverso. Nele, a autora considera que a relevância de um documento não se resume apenas na presença de termos - chaves nele, mas porque esses termos não são frequentes em outros documentos considerados irrelevantes. Ao lado da frequência inversa do documento, o grupo Salton desenvolveu o algoritmo *Term Frequency (TF)* na Universidade de Cornell, formando a combinação *TF.IDF*, um esquema de ponderação mais usado por muitos anos na RI, dado o baixo custo.

De salientar que a *Text Retrieval Conference* é co-patrocinado pelo *National Institute of Standards and Technology (NIST)* e pelo Departamento de Defesa dos EUA e teve início em 1992. Trata-se de uma organização que desempenha um papel preponderante na área dos SRRI, pois visa o apoio na investigação no seio da comunidade de recuperação de informação, fornecendo infra-estrutura necessária para a avaliação em larga escala de metodologias de recuperação de textos.

Sobre o processamento de linguagem natural cujo início data de 1980, o foco de Jones foi a sumarização automática[45], *Question Answering* (QA)[46] – resposta a pergunta e pesquisa em linguagem natural, tendo sido reconhecida pela obra *"Charting a New Course: Natural Language Processing and Information Retrieval"*, publicado em sua honra (TAIT, 2005 apud CHU, 2007).

Conforme acrescenta Chu (2007), Jones também trabalhou sobre a recuperação de documentos falados, adaptando os algoritmos *keywordspotting*[47] que eram usados em técnicas de recuperação de informação baseada em textos para dados de voz, como palestras e noticiários. Deste modo, suas contribuições revitalizaram a área de RI e até influenciaram pesquisas da comunidade de recuperação da informação.

5.6 Recuperação da informação na Web

Os SRI na Web são compostos pelo próprio ambiente da Web, pelos rastreadores ou *crawlers*, pelo repositório, pelo indexador, pelo índice de termos, pelo motor de busca e pelos usuários.

---

[45] Criação de uma versão encurtada de um texto por um programa informático, mantendo os pontos mais importantes do texto original.

[46] Tarefa de responder automaticamente a uma pergunta feita em linguagem natural, através de um programa de computador *(question answering)* que acessa um banco de dados pré-estruturado ou uma coleção de documentos em linguagem natural.

[47] Ramo de reconhecimento de fala, que trata da identificação de palavras-chave em discursos ou expressões orais. Existem *keywordspotting* no discurso livre, no reconhecimento de palavras isoladas, entre outros tipos.

No ambiente da Web, a recuperação da informação tem as seguintes características: volume e dispersão da informação, heterogeneidade da informação, e informação dinâmica e de qualidade variável. Por isso, a recuperação da informação é o processo de determinação da similaridade entre documentos da coleção e a consulta formulada, para ordenar os documentos de acordo com a sua relevância. Essa ordenação depende do modelo que se usa na recuperação. Alguns modelos levam em consideração os seguintes aspectos na ordenação dos documentos na Web: a frequência do termo da consulta em um documento e a frequência do mesmo termo no documento inverso, o PageRank e as variações linguísticas (sinonímia e ontologias).

5.6.1 Classificação de relevância usando termos (TF-IDF) e semelhança

Conforme anteriormente se referiu, o conceito de relevância na RI foi introduzido por Gerard Salton no modelo de espaço vetorial. Na categorização dos modelos de RI estabelecida por Ferneda (2003), o modelo vetorial aparece nos modelos quantitativos, à semelhança do booleano, probabilístico, fuzzy e booleano estendido, na medida em que se baseia em operações matemáticas da estatística. Na categoria dos modelos dinâmicos que se baseiam na interação do usuário com o sistema para a determinação da relevância, encontram-se os sistemas especialistas, redes neurais e algoritmos genéticos.

Ferneda (2003) observa que no modelo vetorial, cada documento é representado como um vetor de termos e cada termo possui um valor que corresponde ao respectivo peso no documento. De igual modo, a expressão da busca é representada por um vetor numérico em que cada elemento corresponde ao peso do termo na busca. Neste sentido, a recuperação é feita através do cálculo do grau de similaridade, tanto entre os documentos em si, como entre um documento e a expressão de busca formulada pelo usuário.

Geralmente, o cálculo dos pesos é feito em função da frequência, ou seja, a classificação de relevância baseada na frequência do termo – *Term Frequency (TF)* parte da contagem do número de ocorrências de um termo *t* em documento *d*, para determinar a relevância desse documento em relação ao termo em questão. Contudo, este método não resolve o problema da relevância, pois conforme Silberschatz et al. (2006)

> o número de ocorrências depende do tamanho do documento, e […] um documento contendo 10 ocorrências de um termo pode não ser 10 vezes mais relevante do que um documento contendo uma ocorrência.

Por isso para classificar a relevância levando em consideração a extensão do documento, aplica-se a seguinte fórmula:

$$TF(d,t) = log\left(1 + \frac{n(d,t)}{n(d)}\right)$$

onde *TF(d,t)* indica a frequência do termo, isto é, relevância do documento *d* em relação ao termo *t*; *n(d)* indica o número de termos no documento e, *n(d,t)* o número de ocorrências do termo *t* no documento *d.* Embora através da fórmula a relevância aumente com mais ocorrências do termo no documento, outros aspectos são levados em consideração para o sistema considerar um documento como o mais relevante, como: ocorrência do termo no título, na lista de autores, no resumo ou na parte inicial do documento (SILBERSCHATZ et al., 2006).

A frequência de documento inversa ou *Inverse Document Frequency (IDF)* é uma medida complementar à frequência do termo, que se aplica para casos em que a consulta tenha vários termos e os respectivos documentos tenham relevâncias diferentes para cada termo, através da fórmula:

$$IDF(t) = \frac{1}{n(t)}$$

onde *n(t)* indica o número de documentos indexados que contêm o termo *t*. Deste modo, a relevância de um documento *d* ao conjunto de termos Q é dada pela fórmula:

$$r(d,Q) = \sum_{t \in Q} TF(d,t) * IDF(t)$$

onde $r(d,Q)$ indica a relevância do documento *d* ao conjunto de termos Q, medida em relação à qual se aplicam outros aspectos, como a proximidade dos termos, isto é, ocorrência de um termo próximo do outro. O sistema organiza os documentos recuperados de acordo com a ordem decrescente da sua relevância (SILBERSCHATZ et al., 2006).

Uma das desvantagens do uso dos termos TF e IDF consiste na determinação da relevância de um documento com base em métodos estatísticos. No Capítulo anterior discutiu-se sobre as especificidades linguísticas, nos quais

maior parte do problema da comunicação humana encontra-se enraizado. Como Harvey (2004), citando a teoria estruturalista da linguagem de Sausurre observou, o sentido das palavras é determinado pela sua relação com outras palavras e não apenas pela sua referência a objetos. Por outras palavras, o significado dado a cada termo de busca depende do contexto intrínseco em que o usuário usa esse termo. Por isso que, tal como na representação, o uso de métodos de processamento em linguagem natural é importante para esmiuçar semanticamente as relações das palavras nos documentos recuperados.

Além da frequência do termo em um documento e da frequência do mesmo termo em documento inverso, a RI também se baseia na semelhança de documentos. Segundo Silberschatz et al. (2006), a recuperação baseada na semelhança do cosseno consiste em o sistema localizar documentos semelhantes com base em termos comuns especificados pelo usuário como termos do documento modelo. Nos casos em que o sistema retorna muitos documentos semelhantes ao documento da consulta, o usuário pode escolher os documentos que considerar relevantes de todo o conjunto recuperado e realizar uma nova consulta, baseando-se na semelhança entre eles e o primeiro documento modelo da consulta, processo chamado *feedback* por relevância.

Através do *feedback* por relevância, o usuário também pode usar um dos documentos recuperados como modelo para uma nova consulta ou até acrescentar outras palavras-chave ao documento da consulta para refinar o processo da busca (SILBERSCHATZ et al., 2006). Uma das desvantagens deste método consiste na demora do processo em si. Além disso, pressupõe-se que o usuário reconheça a sua necessidade informacional e, por isso, conhece o autor, o título ou o assunto do documento. Para usuários que não sabem ao certo o que procuram, o processo torna-se incapaz de satisfazer as suas necessidades de informação.

5.6.2 Relevância usando *hiperlinks* e classificação por popularidade

A relevância por meio de *hiperlinks* que apontam para a página HTML, considerada como um documento para efeitos da recuperação da informação, é um complemento ao processo da busca por relevância de termos TF-IDF, específico para a recuperação na *Web*, na medida em que apresenta grandes coleções de

documentos que quase contém todos os termos especificados numa consulta (SILBERSCHATZ et al., 2006).

A classificação por popularidade ou classificação por prestígio, conforme Silberschatz et al. (2006), consiste em encontrar páginas que são populares e classificá-las antes de outras páginas que contém as palavras-chave especificadas. Equivale isto dizer que em um processo de consulta por um determinado termo, o sistema irá considerar como mais relevantes as páginas mais populares que contenham o termo especificado, embora se considerem outros aspectos, como ocorrência do termo no domínio da página, relevância de termos TF-IDF, etc. Neste sentido, sendo difícil determinar a quantidade de vezes em que os usuários acessam uma determinada página, a popularidade pode ser determinada pela quantidade de arquivos de marcador com *links* à página, como favoritos, bem assim através de *links* para sites relacionados.

Silberschatz et al. (2006), apontam que para efeitos de popularidade, considera-se o *site* todo e não apenas uma página do mesmo, isto é, o domínio ou endereço, mas como alguns *sites* também contêm páginas não relacionadas ao domínio, usa-se o método da “transferência de prestígio” em que uma página passa a ser considerada como popular se tiver um *link* a partir de uma página popular, mesmo que não seja tão conhecida.

Aproveitando-se do método de transferência de prestígio, muitos sites pornográficos são desenvolvidos como coleção de páginas que criam *links* entre si para aumentar a classificação da popularidade e, por conseguinte, serem acessados como páginas relevantes para a consulta.

5.6.3 PageRank e outras medidas de popularidade

PageRank é uma medida de popularidade introduzida pela Google que consiste na técnica de “transferência de prestígio” através do modelo de caminhada aleatória. Este modelo baseia-se na técnica da probabilidade para atribuir a popularidade ou *PageRank* alto às páginas, isto é, parte do princípio de que um usuário pode navegar aleatoriamente em uma página e a partir daí existe a probabilidade de escolher um dos *links* externos da página ou o usuário estará navegando aleatoriamente em algum ponto da página no tempo (SILBERSCHATZ et al., 2006).

A questão principal do *PageRank* é a navegação ou o termo “surfar” que depende das estratégias ou competências de cada usuário no processo de busca da informação. Esta percepção vai de acordo com Page et al. (1998), ao considerarem que a importância de uma página da Web é uma questão subjetiva que depende dos interesses, conhecimentos e atitudes de cada usuário numa heterogeneidade de páginas. Por isso, os autores criaram o *PageRank*, um método que permite classificar as páginas da Web, para medir a importância relativa de cada página. O método baseia-se no número de links que cada página possui. Assim, as páginas com o maior número de *links* são consideradas mais importantes do que aquelas que têm menor número. De igual modo, se uma página tiver um *link* com a outra popular como o Yahoo, será considerada mais importante do que aquelas que apontam para *links* obscuros.

De acordo com Page et al. (1998, p.3), “uma página estará no topo do ranking se a soma dos respectivos *links* for alta. Isso abrange tanto os casos em que a página tenha muitos *links* externos, como os casos em que tenha poucos *links*, mas rankeados como altos”.

A maior desvantagem da medida *PageRank*, segundo Silberschatz et al. (2006), é da recuperação de documentos por popularidade desconsiderando as palavras-chave da consulta. Por isso, adotou-se o uso de palavras-chave no texto de âncora dos *links* para uma página, determinando os tópicos a que a página é relevante, através da *tag* HTML “*a href”*. Por exemplo, para que o site marilia.unesp.br possa ser considerado como relevante para uma busca com o termo unesp, muitos links para marilia.unesp.br deverão ter o termo unesp em seu texto de âncora, através da seguinte *tag*: * unesp *. Contudo, este método também é combinado com os outros anteriormente descritos. Outro método apontado pelo autor é o algoritmo *HITS* que procura calcular a popularidade baseando-se apenas em páginas que contenham as palavras-chave da consulta através de “*hubs* e autoridades”[48], contrariando os que se baseiam em todas as páginas da *Web*. Deste modo, cada página recebe o valor de prestígio de *hub* e de autoridade, respectivamente e segundo Silberschatz et al. (2006), “uma página obtém o prestígio de *hub* mais alto se apontar para muitas

[48] Um *hub* é uma página que armazena os *links* para diferentes páginas relacionadas; enquanto que autoridade é uma página com informações reais sobre um assunto (SILBERSCHATZ et al., 2006, p. 514).

páginas com alto prestígio de autoridade, enquanto uma página recebe prestígio de autoridade mais alto se for apontada por muitas páginas com alto prestígio de *hub*".

5.6.4 Sinônimos, homônimos e ontologias

O uso de sinônimos é um procedimento adotado por alguns sistemas de recuperação da informação que consiste em criar relacionamentos entre palavras, de modo a estender a busca especificada pelo usuário, através do uso dos operadores *"or"* e *"and"*. Porém, conforme considera Silberschatz et al. (2006), este procedimento, na maioria das vezes, resulta em ambiguidades que podem confundir ou até atrapalhar a busca para o usuário, pois o sistema sugere ou adiciona um novo termo consideravelmente sinônimo, sem saber a necessidade ou intenção real do usuário. O mesmo pode acontecer com palavras homônimas que, apesar de terem a mesma grafia e pronúncia, podem ter significados diferentes, por exemplo: o termo "caso" pode-se referir tanto ao substantivo (um caso complicado), como ao verbo (eu caso amanhã). Uma forma de contornar este problema é a adoção de sistemas que admitem consultas baseadas em conceito, ou seja, que analisam a ambiguidade de cada palavra no respectivo documento, de modo que a substituição do termo da pesquisa se faça por outro mais próximo e empregue no mesmo documento.

Embora a consulta baseada em conceitos tenha a desvantagem da sobrecarga no processamento de documentos para retirar a ambiguidade e por isso quase não se usa na recuperação da *Web*, ela permite a recuperação de documentos relacionados através da hierarquia de conceitos, como ontologias (SILBERSCHATZ et al., 2006).

As ontologias vislumbram uma nova etapa no campo da representação e recuperação da informação, não só pela possibilidade de redução de ambiguidades pela definição de terminologias para áreas específicas, como também pelo maior poder de recuperação de documentos através de linguagens legíveis para o computador e para o homem, respectivamente.

5.7 Findability

O termo *findability*, ou encontrabilidade, segundo Morville (2005) significa:

> A qualidade de ser localizável ou navegável;
> O nível no qual um objeto particular é facilmente descoberto ou localizado;
> O nível no qual um sistema ou ambiente suporta a navegação e recuperação.

O *findabilty* é a característica que qualquer sistema tanto de representação, como de recuperação deve ter para garantir que a informação esteja disponível, localizável e acessível para cada necessidade específica do usuário final. Entre os diversos processos de garantia da encontrabilidade, figuram-se os seguintes: construção de mapas de site, adição de caixas de busca, estudos de grupos focais, testes de usabilidade, teleportação[49] e navegação.

Stewart (2008) considera que a navegação reduz a demanda de informações, levando os usuários a seguir por vias conhecidas para a área geral da informação que procuram e, deste modo, reduzir o tamanho da área da busca. Por exemplo, o "você está aqui" é importante para que os usuários se sintam no controle e na direção certa, podendo voltar atrás se errarem e evita o abandono da busca por informações. As taxonomias são outros exemplos recomendados pela autora, na medida em que a sua natureza hierárquica ajuda a educar o usuário, orientando-o através de um assunto. As relações pai/filho inerentes à estrutura da árvore de uma taxonomia são ferramentas poderosas para orientar um usuário a descobrir as associações que não sabia, mostrando como os termos e conceitos são relacionados; logo, podem definir e refinar a sua necessidade de informação.

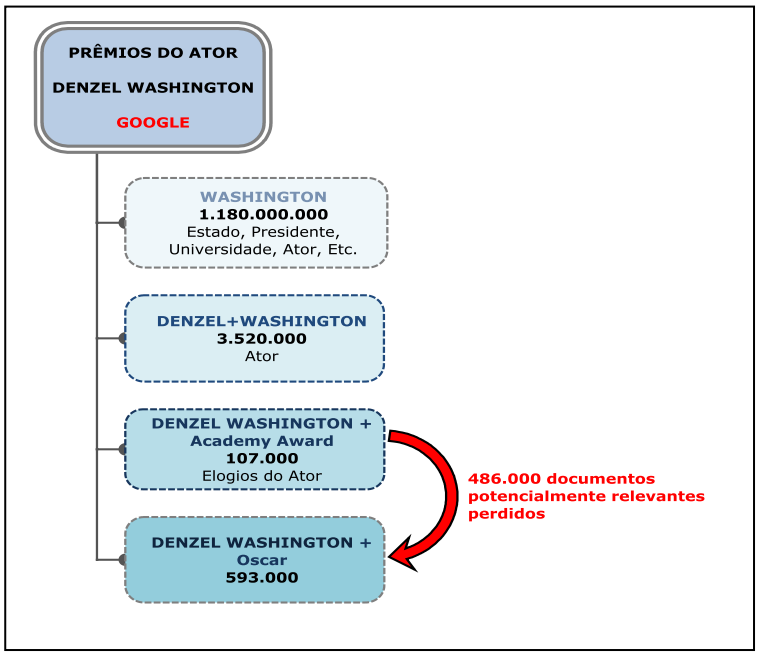
Stewart (2008) desenvolveu uma pesquisa inerente ao problema da busca de informações na *Web* ilustrado pela figura 7. O processo consistia na pesquisa de documentos sobre a premiação de Oscar ao ator Denzel Washington no motor da busca da Google. Numa primeira fase, a autora usou apenas o termo "Washington", tendo recuperado 1.180.000.000 de documentos que além do ator, também se referiam ao presidente dos EUA, instituições e locais que ostentam o termo da pesquisa. Quando acrescentado o termo "Denzel", o número de documentos reduziu para 3.520.000, referindo-se apenas ao ator. Acrescentando o termo oficial da premiação "*academy award*", o número reduziu para 107.000, enquanto que usando o termo mais popular da premiação, isto é, "*oscar*" o número aumentou para

[49] Estratégia de busca de informação que pode ser executada de várias formas e táticas de busca. Além das palavras-chave, pode-se especificar uma URL, abrir um determinado e-mail ou digitando em um caminho de diretório para um determinado documento.

539.000, por tanto, uma diferença de 486.000 documentos relevantes que tinham sido descartados devido à especificação dos termos da busca.

O exemplo apresentado na figura 7 mostra o problema que os sistemas de representação e recuperação da informação por termos ou palavras-chave ainda enfrentam, associados ao problema da indexação efetuada por próprios autores de recursos, majoritariamente desprovidos de conhecimentos técnicos específicos da representação, bem assim da falta de competência dos próprios usuários finais no desenvolvimento das estratégias de busca da informação para as suas necessidades.

Figura 7: *Problema da busca por palavras-chave*



Fonte: *Adaptado pelo autor com base em Stewart (2008, p. 7).*

A figura 7 também ilustra o problema de *precision* e *recall* (precisão e revocação) que afeta a maioria dos sistemas de recuperação da informação. Segundo Stewart (2008, tradução nossa), quanto maior for a precisão, menor será a revocação, pois:

$$\textbf{Precisão} = \frac{Documentos\ relevantes\ recuperados}{Total\ de\ documentos\ recuperados}$$

$$\textbf{Revocação} = \frac{Documentos\ relevantes\ recuperados}{Total\ de\ documentos\ na\ coleção}$$

Por exemplo, se o sistema tiver 100 documentos sobre a Ciência da Informação e o termo da busca for “Ciência da Informação”, uma boa revocação será o total dos documentos, pois a recuperação será na totalidade para o termo especificado. Contudo, a precisão será menor porque provavelmente não serão todos os documentos recuperados pertinentes para o assunto em pesquisa, mesmo que na generalidade tratem sobre o termo especificado na busca (Ciência da Informação).

Silberschatz et al. (2006), também fazem uma abordagem sobre a eficácia e eficiência dos sistemas de RI, baseando-se na precisão e rechamada[50], em função do falso positivo ou falso negativo a que muitos sistemas estão sujeitos. Neste sentido, o falso negativo como não recuperação de documentos relevantes pode ocorrer se os documentos recuperados forem considerados como relevantes e de classificação baixa para o usuário. Já o falso positivo como recuperação de documentos irrelevantes se o sistema classificar com o grau alto documentos irrelevantes, em detrimento dos relevantes para o caso específico da necessidade do usuário.

O grande problema das métricas de precisão e revocação ou rechamada está centrado na dificuldade de saber quando um documento é relevante para um usuário e quantos documentos relevantes para a pesquisa podem ser recuperados do total da coleção.

Silberschatz et al. (2006), também consideram que a recuperação de documentos que contém palavras-chave pode ser otimizado pelo uso de índice invertido que consiste no mapeamento de cada palavra-chave a uma lista de identificadores de documentos que contêm essa palavra-chave. A referida lista, por um lado, pode incluir informações sobre a localização de cada palavra-chave no documento e, por outro, o número de vezes em que a palavra aparece em um documento ou mesmo o número de documentos em que aparece, para determinar a frequência dos documentos para o termo. Contrariamente a Silberschatz et al. (2006), Rocha (2004) considera que esse processo de busca possui resultados imprecisos, tanto pela incapacidade de identificação do significado das palavras do documento, como da busca através de propriedades atribuídas ao documento.

[50] O termo rechamada corresponde ao termo revocação empregue ao longo do projeto.

Conforme se referiu em diversos pontos do Capítulo anterior e deste Capítulo, são imensas as contribuições da CI na produção, gestão, organização, disseminação, acesso e uso da informação. Esta ação do campo permite o desenvolvimento do conhecimento em qualquer outra área. O acesso à informação produzida por diversos campos científicos propicia o uso e a produção de novas informações para a ciência no todo. Porém, mesmo com vários anos de consolidação, a CI ainda denota muitos desafios, principalmente no âmbito da recuperação da informação. Estes desafios ganham maior repercussão científica e social na Web, em que constantemente se produzem volumes excessivos de informações por multiplicidade de usuários, de difícil mapeamento e de diferentes modelos de interação, apropriação e representação.

5.8 A ambiguidade na Recuperação da Informação

Na atualidade, os processos de busca, produção, difusão e uso da informação são condicionados pela tecnologia. O problema é que a tecnologia que assiste o processo da construção do conhecimento propicia mudanças na informação a todo o momento. Além disso, os sistemas de recuperação da informação baseiam-se em aspectos subjetivos para se antecipar às ações dos usuários, tentando adivinhar as suas necessidades informacionais, desenhar as estratégias de busca que serão utilizadas e arrolar as interpretações possíveis para determinar a relevância das informações localizadas. Porém, como se pode identificar e coletivizar ações que, pela própria natureza, são de interpretação individual? Como se pode determinar a relevância de um documento baseada apenas no vínculo com uma palavra-chave? Se a informação é caracterizada pela atualidade e relevância, em que consiste a atualidade para efeitos de *ranking* nos sistemas de busca?

Indubitavelmente, a CI contribui na recuperação e os sistemas de recuperação são extremamente importantes para o acesso e uso da informação. Todavia, mesmo com a adoção de metadados, vocabulários controlados, sistemas inteligentes que permitem o aprendizado da máquina, etc., a área de recuperação da informação ainda se mostra problemática pela imprecisão e ambiguidade. Estes fatores afetam o modo pelo qual se produz a ciência, na medida em que maior parte das interações, associações e construções se baseiam na classificação feita pelos

motores de busca. Este problema não cabe somente aos sistemas ou à CI, na medida em que é um problema da linguagem e da comunicação humana.

Uma das alternativas encontradas pelo campo da CI para se antecipar às ações dos usuários na busca pela informação se prende com os estudos dos usuários. Estudos de usuários[51], segundo Figueiredo (1994, p.7) "são investigações que se fazem para saber o que os indivíduos precisam em matéria de informação, ou então, para saber se as necessidades de informação estão sendo satisfeitas de maneira adequada". Neste contexto, para Costa et al. (2009), existem por um lado estudos enquadrados na abordagem tradicional que são orientados à bibliotecas ou aos centros de informação individual que incidem sobre os serviços prestados ou catálogos, coleções de índice, etc.; por outro, estudos enquadrados na abordagem alternativa que visam investigar um grupo de usuários sobre o seu comportamento ou forma de obtenção e uso da informação.

Partindo da análise do conceito da informação e dos diferentes contextos em que ela se dá, Costa et al. (2009) consideram que o uso da informação subjaz da busca pela satisfação de uma necessidade, que pode ser de resolução de um problema, alcance de um objetivo ou sanação da insuficiência ou inadequação de um conhecimento. Assim, usuários da informação são indivíduos ou grupos que utilizam a informação para satisfazer as suas necessidades.

Existe uma diferença entre necessidade e procura. A procura é consequência de uma necessidade e depende mais do trabalho dos profissionais de informação nas pesquisas sobre os usos para promover os serviços, facilitar o acesso, incluir e treinar os usuários. Já a necessidade depende do aspecto intrínseco do próprio usuário e pode estar relacionada, por exemplo, às atividades de ensino e pesquisa. Conforme Wilson (1981), o comportamento de busca não é uma consequência direta de uma necessidade, pois existem situações em que os usuários não procuram informações devido a barreiras sociais, pessoais, ambientais e, outros não reconhecem a existência da tal necessidade. Esta situação é cada vez mais tenaz na dinâmica da Web atual, em que muitos usuários não sabem o que procuram e as

[51] Figueiredo (1994) considera que os estudos dos usuários datam na segunda metade da década 40, mormente com a Conferência da *Royal Society* de 1948 e posteriormente com a Conferência Internacional de Informação Científica de Washington, em 1958. Porém, para Teruel (2005) as referidas conferências alargaram o seu âmbito, pois os estudos de usuários são oriundos dos trabalhos da comunidade científica sobre pesquisas de hábitos de cientistas na obtenção e uso da informação, para o aprimoramento dos sistemas de informação.

suas decisões são induzidas pela informação encontrada “acidentalmente” durante a navegação ou pelo *ranking* dos sistemas de busca.

O problema da ambiguidade e imprecisão dos SRI prevalece mesmo com diversas inovações que vem influenciando o seu percurso, como: metadados, vocabulários controlados, algoritmos genéticos que permitem o aprendizado da máquina, entre outras. A maioria dos SRI baseia-se nos perfis dos usuários e não nos aspectos intrínsecos que configuram as suas necessidades e estratégias de busca.

Existem diversos métodos de estudos dos usuários. Baptista et al. (2007) dividem os métodos de estudos de usuários em duas fases, a saber: pesquisas quantitativas - década de 1960 a 1980, evidenciadas pelo uso da estatística para a coleta e tratamento de dados, inicialmente com foco apenas na frequência de uso de determinadas fontes, depois com enfoque na transferência, uso e tempo de resposta para o acesso da informação e por último, no planejamento de serviços e sistemas para satisfazer as necessidades dos usuários. A segunda fase foi das pesquisas qualitativas sobre aspectos subjetivos do comportamento dos usuários, destacando-se entre outras, a abordagem de Taylor sobre a busca e transformação de dados em informação útil, através da seleção, análise e julgamento; modelo de Kuhlthau na análise da busca da informação desde o início, seleção, exploração, até a formulação e o *sense making* de Dervin que procura entender o usuário a partir das suas necessidades cognitivas, afetivas, psicológicas e fisiológicas.

Mesmo com o auxílio da Psicologia para a análise de questões sobre o julgamento da relevância, seleção da informação ou sobre as necessidades do fórum individual de cada usuário, ainda não é possível planejar um único sistema capaz de atender as necessidades dos usuários, em todos os contextos informacionais. Cada informação é relevante apenas para o contexto pelo qual cada usuário busca a satisfação da necessidade concreta, mesmo que contenha os termos especificados na busca.

5.9 Efeitos da RI: o conhecimento mediado pela tecnologia

Se no passado, segundo Bordieu (2004), a apropriação da informação consistia em duas correntes, uma semiótica que procurava validar a essência da comunicação à estrutura do texto em si e, outra, marxista que fazia uma interação

entre o texto e o contexto do sujeito para a apropriação, na atualidade essa separação é contrastada pelos espaços concêntricos de difusão e apropriação da informação, marcados pela simultaneidade de probabilidades e hipóteses. A Web social ganha mais proporções pela sua estrutura concomitante e de inúmeras possibilidades, em que a apropriação da informação ocorre num determinado contexto e na interação recíproca desse contexto entre o emissor e o receptor. Por isso, o emissor é ao mesmo tempo o receptor, a mensagem sempre está em transformação dada a interação entre os sujeitos participativos e, por conseguinte, a sua materialização torna-se uma possibilidade. Neste sentido, o conhecimento também é um circuito em transformação.

Uma das questões levantadas por Saracevic (1996) se prende ao modo como o campo da CI se desenvolve, ou seja, se a base dos problemas da CI é construída no enfoque humano ou tecnológico? Conforme o autor aponta, mesmo com as conquistas humanas atuais, a CI ainda denota uma forte pressão tecnológica.

A relação entre a CI e a tecnologia pode ser analisada de diferentes formas. Por um lado, a tecnologia tornou evidente e necessária a intervenção da CI. Exemplos sobre o atual sucesso na recuperação da informação podem ser contextualizados no sentido positivo do enfoque tecnológico pela CI. Por outro, parece que o enfoque tecnológico traduz a essência de dominação, controle e dependência. Em última análise, estes aspectos se circunscrevem nas manifestações da pós-modernidade sobre o consumo pela moda e prazer. Muitas ações na Web são ditadas pelos países desenvolvidos, a serviço de grandes corporações. Muitas pessoas no mundo ainda são analfabetas digitais, e logo se levanta a questão sobre quem a tecnologia está visando no seu desenvolvimento. Por isso, parece que a CI está a cair na crítica levantada por Saracevic (1996), que consiste na alternativa tecnológica ou facilidade de ensinar e ajustar os humanos à máquina do que a máquina aos humanos.

Um dos aspectos negativos na Web, segundo Barreto (2008) está relacionada à quebra da estrutura física dos documentos, através de hiperlinks que apontam para várias páginas, sem nenhuma orientação para o usuário. Por isso mesmo, a CI está empenhada na busca de soluções capazes de reduzir essa ambiguidade que caracteriza a estrutura de navegação da Web atual. A Web semântica, especificamente o *linked data*, aliada aos modelos consistentes de representação da

informação tentam propiciar melhores resultados na busca de informação para os usuários, por meio de relacionamentos.

O outro aspecto negativo que ganha maior proporção com a Web, mesmo sendo oriundo da tecnologia da informação, está no modo de assimilação da informação na atual estrutura aberta, constantemente mutável e de pouco controle por parte dos usuários. Por um lado as pesquisas tendem a ser de maior abrangência temática, mas por outro de menor profundidade, motivadas pelo *"flash"* de informações que aparecem e mudam constantemente nas interfaces dos computadores. Barreto (2008) aponta este aspecto numa abordagem sobre a tendência reducionista da tecnologia, enquanto conjunto de técnicas e ações voltadas para minorias, em analogia com a inovação que atende a pluralidade de interesses comuns.

Além do medo das consequências que a tecnologia provoca na sociedade (segregação pela coletivização), os limites da tecnologia são notórios no distanciamento ou até isolamento das camadas pobres. Para estas, além da falta de recursos tanto financeiros, como cognitivas para interagirem de forma competitiva, a velocidade com que as novas tecnologias se propaga as exclui completamente do cenário. Por isso que dúvidas ainda se levantam em relação ao real alcance da sociedade da informação e do conhecimento. O termo "sociedade" ainda se revela muito abrangente para a atual situação do recurso à tecnologia para a solução de problemas comuns.

A CI deve manter o equilíbrio de abordagem na avaliação da relação homem – tecnologia, para não ser considerada como um campo ideal para a indústria da informação e descompromissada com a sociedade no todo. Saracevic (1996) já questionava a eficiência das aplicações tecnológicas no acesso à informação e na comunicação dos estoques do conhecimento, na medida em que os critérios de eficácia e relevância são estritamente humanos e não tecnológicos. Ademais, para se aferir a eficiência de um sistema dever-se-ia basear na análise de todos os usuários visados, neste caso, a sociedade no todo, o que ainda se mostra impossível, na medida em que muitas pessoas não pertencem a essa "sociedade".

A base da produção do conhecimento atual, e em relação a qual a CI ganha notoriedade, está aliada à tecnologia. Por sua vez, a tecnologia propulsiona as manifestações do pós-modernismo pela instantaneidade, contrária à reflexão. Sobre este aspecto, Debord (2003) faz uma crítica à sociedade de espetáculo, uma

sociedade assente no acúmulo da economia pelos modernos processos de produção, e baseada na representação, uma representação repetitiva e comum em todos os seus aspectos, de tal modo que a tornam real. Porém, uma realidade baseada na referida representação, ou seja, nos espetáculos. Enquanto parte da sociedade, o espetáculo caracteriza toda a ideologia de pensar, agir, ser ou estar nessa sociedade. Esta percepção do autor coaduna com o status das sociedades pós-modernas, orientadas por um discurso mediático e enraizado nas tecnologias de informação e comunicação. No campo do conhecimento, o fator norteado por essas tecnologias desfalece a complexidade enleada ao processo da transferência da informação, da apropriação, da linguagem, entre outros aspectos que caracterizam a diversidade e a especificidade dos espaços informacionais. Sendo a sociedade do espetáculo caracterizada pela representação da realidade em todos os contextos sociais, será o conhecimento uma representação de certa realidade nesta sociedade? O problema da representação é que ela consiste num conjunto de elementos descritivos sobre certa realidade e, por isso, é uma concepção fracionada dessa realidade e de acordo com a visão de quem representa.

Conforme Debord (2003) acrescenta, “o espetáculo não é um conjunto de imagens, mas uma relação social entre pessoas, mediatizada por imagens”. Equivale por outras palavras dizer que o problema não são as tecnologias em si, mas sim na concepção Heideggeriano, segundo Feynman (2001), o discurso tecnocrático que norteia a sociedade, de tal forma que não reste outra solução, senão do uso da tecnologia em todas as dimensões humanas. Por isso que Debord (2003) observa que o espetáculo é o resultado e o modo de produção, ou seja, o modo de vida cristalizado pela publicidade, consumo e entretenimento. O espetáculo enquanto representação da realidade é tão proeminente de tal forma que a própria realidade fica difusa ou perdida na sua exacerbação ou contemplação pelo belo, prazer e omnipresença. Enquanto o espetáculo surge na realidade, a realidade surge no espetáculo. Neste sentido, o conhecimento é uma representação fracionada de uma realidade que circunscreve determinados sujeitos.

Debord (2003) afirma que as diversidades do espetáculo são as aparências organizadas pela própria sociedade como modelo de vida, de tal forma que as críticas ao espetáculo desdobram-se como verdadeiros atentados à vida, pois o espetáculo é o momento histórico ou a formação econômico-social dessa sociedade. Por isso mesmo que o espetáculo se apresenta como algo grandioso, positivo,

indiscutível e inacessível, caracterizado pela aceitação passiva ou inquestionável, dissipado na sua aparência de belo. De fato, atualmente assiste-se uma cultura de mediocridade em relação à essência e ao efetivo alcance da tecnologia. Se por um lado a tecnologia garante grandes proporções nos meios de produção, essa proporção é reduzida pela mediocridade que se assiste no seu uso, aliada à ausência total de críticas em relação a sua forma, processo e alcance. Assiste-se uma iminência de uma sociedade puramente robótica e orientada por um discurso tecnocrático na sua forma de ser e estar.

Um aspecto importante destacado por Debord (2003) prende-se com a finalidade do espetáculo que são os próprios meios, isto é, o espetáculo visa a si mesmo. Como correlato, as coisas são válidas pelo próprio uso como moda. As proporções atuais do Google por um lado, se devem ao seu potencial na recuperação das informações, mas por outro, à sua relação com meios de produção e do poder. As regras e padrões que visam à representação e recuperação da informação, em algumas vezes, são incorporados não apenas pela sua eficiência, mas pelo fato de estarem a serviço de grandes comunidades. E o instrumento de peso ou medida para essa apropriação como Debord (2003) afirmou, é a economia, o símbolo da especialização do poder. Esta situação, segundo Harvey (2004), é uma das manifestações pós-modernistas aliadas aos meios de produção e do poder. Por isso que as economias do primeiro mundo espelham a busca incessante e tentativa de padronização para o terceiro mundo em todas as dimensões sociais.

Os efeitos da apropriação por moda e prazer, conforme Debord (2003), se resumem na degradação do ser na pretensão ter para parecer. Neste sentido, o espetáculo gera um fator hipnótico e encontra na visão o seu maior aliado. Por isso que a indústria da publicidade cresce a cada dia. O valor de um produto é medido pela sua capacidade de “gerar felicidade” no sujeito, como se esta fosse uma métrica calculável por meios puramente técnicos. A necessidade não é mais um ato eminentemente intrínseco de insatisfação pela carência de algum bem, mas um estado deliberadamente produzido e socialmente sonhado, para que esse sonho se torne uma necessidade. Por isso que em termos de preservação de valores, esta sociedade está em contradição consigo mesma. Aliás, os conceitos de valor e moral assumem novos contornos na sociedade.

No âmbito do conhecimento, as academias e institutos de pesquisa se tornaram em verdadeiros campos de disputa entre os pesquisadores. O lema é

publicar para aparecer, pois o potencial de um pesquisador é medido através da sua capacidade de gerar “conhecimento progressivo”, independentemente da sua abrangência, profundidade e consistência. Esta apropriação na abordagem da ciência propicia o que o jornalista e professor Silvio Mieli (2008), na publicação do jornal semanal Brasil de Fato, denominou “síndrome Google de copiar e colar”. Na limitação humana e acadêmica de gerar continuamente temas inovadores, alguns pesquisadores publicam pesquisas devidamente tratados por outros cientistas, alterando-os a forma, idioma ou título. Este comportamento subsume-se na Teoria Darwiniana sobre seleção natural e evolução da espécie. As características que contribuem para a sobrevivência e reprodução tendem a ser mais comuns em relação às prejudiciais, principalmente por meio das respectivas adaptações. Trata-se de uma proposição disjuntiva sobre a dinâmica da sociedade: ou se compadece com o lema da sociedade em que se encontra, ou se extingue.

Debord (2003) considera que muitas vezes o espetáculo é visto sob o ponto de vista dos meios de comunicação, pela sua capacidade de instrumentalização da sociedade. Ora, embora os meios de comunicação não sejam neutros, o problema não está nos meios em si, mas na apropriação destes pelas classes dominantes que detêm o acúmulo do capital. Uma dominação que permeabiliza a divisão e a perda da unidade do mundo, daí a existência do centro e da periferia. Por isso que a ideologia atual sobre o processo de globalização não passa de um novo instrumento de dominação, na medida em que “reúne o separado, mas reúne-o enquanto separado”. É fato para questionar se estamos efetivamente numa sociedade de informação e do conhecimento ou apenas numa sociedade de convergência de mídias?

Os termos “sociedade da informação” e “sociedade do conhecimento” suscitam vários olhares críticos, pois apenas se baseiam nos meios de comunicação, dissociando-se da noção de significado, da capacidade cognitiva, da complexidade, da variação dos modelos de apropriação e representação, entre outros aspectos. Com base na teoria da comunicação, Morin (2004) considera como mitos da comunicação, as afirmações que geram reducionismos, propalando que estamos na sociedade da informação, da comunicação, e do conhecimento. Para Morin (2004), a informação não é o conhecimento; o conhecimento é resultado da organização da informação, e na atualidade há excesso de informação e insuficiência da organização, logo, carência do conhecimento. Quanto mais

desenvolvidos forem os meios que disseminam a informação, menos há compreensão entre os destinatários, porque a compreensão não está ligada à materialidade da comunicação, mas ao político, social, existencial, por isso, é falsa a afirmação de que tudo comunica.

Morin (2004) acrescenta que falar da comunicação é falar da informação, do conhecimento e da compreensão, isto é, da sabedoria, a capacidade de integrar conhecimentos à vida cotidiana. A compreensão é o problema atual da humanidade; ela não está nos meios, mas nos fins, ou seja, apesar de ser auxiliada pela comunicação, depende do aspecto subjetivo profundo do ser, sendo este um problema filosófico. A comunicação só faz sentido quando tomada em conexão com outros fenômenos sócio-culturais e políticos. A maioria destes aspectos está relacionada ao fator material e capitalista do pós-modernismo.

Na sociedade de espetáculo toda a vivência humana é transformada em mercadoria. Uma mercadoria caracterizada pela perda de qualidade por contraparte do quantitativo, devido à abundância. O acúmulo de capital e o consumo são por si sós estratégias de sobrevivência. Por isso que a automação é o setor mais avançado da atual produção industrial e, por conseguinte, o seu modelo.

Muitas questões podem ser levantadas a respeito deste assunto e a sua abordagem ultrapassa os limites da CI. É necessário um pensamento complexo de Morin (2003) que permita uma análise completa e criteriosa sobre os desdobramentos tecnológicos em todos os setores de atividade e de intersecções humanas. Esta análise permitirá maior abrangência e reflexão sobre as contribuições, desafios e perspectivas do campo da Ciência da Informação.

Alguns dos desafios da área, conforme se discute ao longo do trabalho, prendem-se com a complexidade tecnológica e de comunicação humana. Por isso, no Capítulo seguinte analisa-se o processo de mineração de dados como perspectiva imposta pela hibridização de linguagens na Web.

6

Mineração de Dados da Web Social

6.1 Visão sobre o Capítulo

A CI é um campo científico em constante transformação. Esta característica é justificada pela sua gênese, enquanto área estritamente enleada ao paradigma da complexidade, fruto das manifestações da pós-modernidade. Desde a origem até ao estágio atual, a CI foi norteada pela busca de soluções, majoritariamente de índole tecnológica, que visavam à organização, preservação, acesso e o uso da informação excessiva para o conhecimento em ação, em todas as dimensões humanas. Ora, na medida em que o mundo caminha tudo muda: a tecnologia e os processos envolvidos na produção, tratamento e uso da informação; a comunicação; a ideologia assente na construção do conhecimento; a diversificação dos problemas humanos e outros aspectos da complexidade.

O Capítulo sobre a MD enquadra-se no cenário das mudanças de forma, propiciadas pela tecnologia e conceituais, motivadas pelas necessidades situacionais e comportamentais. Por outras palavras, a informação agrega um caráter intransponível para as tradicionais aplicações, técnicas e processos envolvidos nos fluxos para a completude da essência humana que se atinge pelo universo da informação e do conhecimento. Por isso, outras técnicas ou métodos de recuperação e uso da informação se tornam legitimamente justificáveis para o conhecimento no todo. No papel social que transcende a tecnologia, a CI assume a posição de vanguarda na busca e implementação de soluções tendentes à democratização da informação e do conhecimento. A informação é cultura, poder, liberdade, investimento, lucro, cidadania, etc. e todas as intervenções em prol do conhecimento só se podem subsidiar pela informação.

6.2 Mineração de dados

A World Wide Web é o arcabouço do universo do conhecimento humano e a Web Social é o reflexo da ideologia prevalecente sobre a inteligência coletiva. Em vez de ações isoladas, a colaboração na produção e uso de informações que consubstancia a dinâmica da Web Social mostra-se cada vez mais vantajosa, tanto pela despersonalização do conhecimento, como pelos questionamentos e agregações. Com o conhecimento compartilhado, muitas pessoas podem fazer inferências para aceitá-lo, contrastá-lo ou incorporá-lo na situação concreta (cultural, política, econômica e social). Porém, esta propriedade da Web Social é contrastada

pela estrutura de alguns ambientes, de fluxo contínuo de informações, como o Twitter e o Facebook. A recuperação de informações nestes espaços é condicionada, daí a necessidade de adoção de outras soluções, como a mineração de dados.

Além da MD, o conceito de *Big Data* surge como uma das alternativas para o processamento de grandes e complexas coleções de dados, face ao constante crescimento do volume de informações. O processo visa aperfeiçoar a captura, a curadoria, o armazenamento, a busca, o compartilhamento, a transferência, a análise e a visualização dos dados, principalmente para ambientes corporativos. *Big Data* se desenvolve sob o lema de "3V´s", isto é, alta velocidade de coleta e processamento, alto volume de dados e alta variedade de ativos de informação que requerem novas formas de processamento para fundamentar decisões ou permitir a descoberta do conhecimento. Ultimamente, a veracidade e o valor têm sido acrescidos ao processo.

Big Data baseia-se no uso de estatísticas que permitem fazer medições e detectar tendências ou para fazer associações, regressões, etc. que demonstrem relações ou previsões de resultados. Esta tecnologia é de suma importância pela capacidade de processamento de grandes volumes de dados e inferências em curto prazo. Contudo, para a Ciência da Informação o processo pode parecer inviável, não só por suscitar grandes plataformas tecnológicas com *softwares* sofisticados (algoritmos genéticos, processamento em linguagem natural, redes neurais, processamento de sinais, etc.), como também por requerer conhecimentos complexos de quadros especializados. Além disso, o processo tem sido severamente criticado por alguns analistas, pois as inferências estatísticas são feitas com base na diversidade de informações colhidas em diferentes períodos, locais e situações, descontextualizadas da origem e complexidade envolvida na sua coleta. Outra crítica importante se prende à invasão constante da privacidade, na medida em que dados pessoais são coletados, integrados e armazenados em estruturas únicas para apoiar interesses singulares de empresas, instituições, laboratórios, entre outros.

A MD é uma das alternativas propostas por este trabalho, para subsidiar os SRRI, face ao volume excessivo da informação e intangibilidade de algumas informações na Web Social.

A MD surge com o desenvolvimento e sofisticação da área da Computação, mormente no desafio de responder aos grandes volumes de dados armazenados em bancos, através de identificação de padrões e de regras significativas a esses dados. Tal desafio surgiu para colmatar o defasamento das técnicas tradicionais de análise de dados, face às novas características revestidas nos dados, como a sua complexidade, ou seja, dados com diferentes tipos de atributos ou distribuição e armazenamento em fontes diferentes.

Tan et al. (2009) consideram que “a mineração de dados é o processo de descoberta automática de informações úteis em grandes depósitos de dados”. A partir desta conceituação, podem-se destacar quatro elementos essenciais: processo, descoberta automática, informações úteis e grandes depósitos de dados. Como processo, significa que a mineração de dados é um método ou sistema com regras específicas. A descoberta automática implica a obtenção de resultados por meios puramente mecânicos, sendo por isso, em algumas vezes, resultados imprevisíveis à cogitação humana. O elemento informações úteis significa que os resultados obtidos devem ser proveitosos para a tomada de decisões. Por último, o elemento depósito de dados implica, a priori, a existência de um sistema gerenciador de banco de dados para o armazenamento, indexação e processamento de consultas.

A gênese da mineração de dados é associada à simbiose com áreas como: Estatística (amostragem, estimativa e teste de hipóteses), Inteligência Artificial (algoritmos de busca, técnicas de modelagem e teorias de aprendizagem), Ciência da Informação (recuperação da informação), entre outras (TAN et al., 2009).

A mineração de dados pode ser usada para melhorar os sistemas de recuperação da informação e nestes casos, às vezes as duas áreas se confundem. Enquanto a recuperação da informação se baseia em algoritmos e técnicas para organizar, localizar e acessar as informações, a mineração vai além, convertendo dados brutos em informações úteis ou descoberta de conhecimento em banco de dados, por meio do processo chamado *Knowledge Discovery in Databases (KDD)*.

O processo de *Knowledge Discovery in Databases* pressupõe numa primeira fase a entrada de dados que podem ser armazenados em um único repositório ou em diversos locais e em diversos formatos. Em seguida, o pré-processamento para transformar os dados brutos em formato apropriado para a análise por meio das seguintes tarefas: fusão de dados de múltiplas fontes, limpeza de dados para a

remoção de ruídos, observações duplicadas, seleção de registros e características relevantes para a mineração. A fase da mineração em si corresponde à identificação de padrões de dados e a integração com ferramentas de apoio a decisão, através da fase de pós-processamento, de modo a assegurar que apenas resultados válidos e úteis sejam incorporados no sistema de apoio a decisões, com base na visualização de acordo com diversos pontos de vista de analistas (TAN et al., 2009).

Goldschmidt e Passos (2005) afirmam que o processo de KDD envolve o problema, os recursos disponíveis e os resultados. O problema, por sua vez, envolve o conjunto de dados, o especialista do domínio de aplicação e os objetivos da aplicação. Os recursos envolvem os especialistas no processo, os esquipamentos, os ambientes de *software* e os algoritmos de análise. Os resultados incluem o modelo de conhecimento em relação aos objetivos da aplicação e os históricos anteriormente realizados.

Tan et al. (2009) estabelecem uma dupla categorização no que concerne às tarefas da mineração de dados:

- **Tarefas de previsão** – cujo objetivo é de prever o valor de um atributo (variável dependente ou alvo) a partir de valores de outros atributos (variáveis independentes ou explicativas).
- **Tarefas descritivas** – com o objetivo de derivar padrões (correlações, tendências, grupos, trajetórias e anomalias) através dos quais os dados se relacionam.

Um exemplo das tarefas de previsão verifica-se na área da Metereologia. Os técnicos fazem a previsão do tempo com base na combinação dos atributos de outras variáveis explicativas, como clima, temperatura, humidade, massas de ar, etc. Por isso que a mudança de uma variável independente pode afetar a variável dependente, ou seja, a mudança do vento pode alterar toda a previsão do tempo. Já nas tarefas descritivas, um exemplo é da área da saúde, em que se criam padrões com base no relacionamento dos dados. Pessoas com determinadas características genéticas tendem a desenvolver certo tipo de comportamento ou certas anomalias ou enfermidades.

Segundo Tan et al. (2009), na categoria das tarefas de previsão, a **modelagem de previsão** constitui a base para a construção de um modelo da

variável alvo como uma função das variáveis explicativas[52]. Para construir o modelo usa-se, por um lado, a **classificação** para variáveis alvos discretos e, por outro, a **regressão** para variáveis alvos contínuos.

Na categoria das tarefas descritivas, a **associação** é usada para extrair padrões que descrevam características dos dados em um relacionamento. Outra tarefa descritiva é o **agrupamento** cujo foco centra-se na busca de observações semelhantes em um grupo e que tornam o tal grupo diferente dos outros. A detecção de anomalias é outra tarefa descritiva, com base na qual se procuram observações[53] cujas características sejam significativamente diferentes do resto dos dados. Para a detecção de anomalias usam-se algoritmos que devem ter uma alta de detecção e uma baixa taxa de alarme falso, para não considerar como anômalos certos objetos normais. Deste modo, a análise de dados avalia os relacionamentos entre os objetos de dados para determinar o seu agrupamento, classificação ou detecção de anomalias (TAN et al., 2009).

Goldschmidt e Passos (2005) apontam que a associação consiste na busca de itens que, frequentemente, ocorrem de forma simultânea em transações de bancos de dados (ex: associações de compras de produtos em supermercados). A classificação consiste em descobrir funções que mapeiam as classes, para prever o comportamento dentro dessa classe (ex: previsão do comportamento de novos clientes, de acordo com as classes de clientes antigos numa financeira). A clusterização consiste na criação de subconjuntos, cujos elementos possam compartilhar as propriedades que os destinguem dos elementos de outros subconjuntos (ex: clusterização de clientes para obter grupos que compartilhem o mesmo perfil de compra de serviços numa empresa de telecomunicações). A sumarização consiste em procurar, identificar e indicar características comuns entre conjuntos de dados (ex: identificação de perfis de usuários de determinados registros bibliográficos numa biblioteca). A deteção de desvios consiste na identificação de registros que não atendam aos padrões dos dados (ex: senhas de entrada).

Para garantir a execução dos processos de mineração de dados, Goldschmidt e Passos (2005) consideram as seguintes técnicas: técnicas tradicionais, técnicas

---

[52] A classificação e a regressão têm o objetivo de minimizar o erro entre os valores previsto e o real da variável alvo (TAN et al. 2009, p.9).

[53] Estas observações são designadas anomalias ou fatores estranhos (TAN et al. 2009, p.13).

específicas e técnicas híbridas. As técnicas tradicionais incluem as redes neurais, a lógica *Fuzzy*, os algoritmos genéticos, e a estatística. As técnicas específicas à tarefa da mineração podem ser: o algoritmo Apriori, que permite a descoberta da associação, e outros como *DHP, Partition, ParMaxEclat,* etc. As técnicas híbridas podem ser: híbrido sequencial, híbrido auxiliar e híbrido incorporado. De salientar que a escolha de cada técnica depende do problema, dos recursos disponíveis, dos especialistas envolvidos e dos objetivos.

6.2.1 Considerações sobre os dados

Os dados constituem a essência da mineração, por isso o uso adequado de técnicas para a sua coleta, processamento e armazenamento é fundamental para o objetivo final do aprendizado da máquina. Atualmente, os sistemas gerenciadores de banco de dados, enquanto conjuntos formados por dados inter-relacionados e *softwares* que permitem a definição de estruturas para o seu armazenamento, mecanismos para a manipulação e segurança, aliados à perfeita modelagem nos níveis conceitual, lógico e físico, garantem a compacidade, integração, padronização e, por conseguinte, reduzem a ação de limpeza dos dados.

Mesmo com a eficiência dos bancos de dados, etapas de pré-processamento mostram-se indispensáveis para tornar os dados mais apropriados para a mineração. Tal acontece porque os dados podem ser de diversos tipos (qualitativos e quantitativos) e alguns podem ser menos ou mais pertinentes, duplicados ou até menos representativos para a população. Por isso, a sua modificação pode-se mostrar necessária para que sejam adaptados à ferramenta da mineração. Por exemplo, a criminalidade é representada através de números, sendo por isso atributo contínuo e para adaptá-la a algumas técnicas, pode ser representada por outras categorias qualitativas, como: alta, média ou baixa (TAN et al., 2009).

Tan et al. (2009) acrescentam que o conjunto de dados é a coleção dos objetos de dados[54]. Estes objetos de dados são descritos através de atributos[55] cujas propriedades podem ser diferentes dos valores usados para medir o objeto. Esta conceituação foi herdada do modelo relacional, formado por um conjunto de

[54] Os objetos de dados também são conhecidos por registros, ponteiros, vetores, padrões, eventos, casos, exemplos, observações ou entidades (TAN et al., 2009, p.26).
[55] O atributo também pode ser designado por variável, característica, campo, recurso ou dimensão (TAN et al., 2009, p.27).

tabelas que representam a relação entre um conjunto de valores, com as respectivas colunas referentes aos atributos em relação aos quais existe um domínio de cada entidade e linhas que correspondem aos registros ou valores de cada atributo.

Tan et al. (2009) apresentam na tabela 6 os diferentes tipos de atributos.

Tabela 6: *Diferentes Tipos de Atributos.*

Tipo do Atributo		Descrição	Exemplos	Operações
Categorizados (Qualitativos)	Nominal	Os valores de um atributo nominal são apenas nomes diferentes; i.e., valores nominais fornecem apenas informação suficiente para distinguir um objeto de outro. **(=, ≠)**	Códigos postais, números de ID de funcionário, cor dos olhos, sexo	Modo, entropia, correlação de contingência, teste x^2
	Ordinal	Os valores de um atributo ordinal fornecem informação suficiente para ordenar objetos. **(>, <)**	Dureza de minerais {boa, melhor, melhor de todas}, notas, números de ruas	Mediana, porcentagens, testes de execução, testes de assinatura
Numéricos (Quantitativo)	Intervalar	Para atributos intervalares, as diferenças entre os valores são significativas, i.e., existe uma unidade de medida. **(+, -)**	Datas de calendário, temperatura em Celsius ou Fahrenheit	Média, desvio padrão, correlação de Pearson, testes T e F
	Proporcional	Para variáveis proporcionais, tanto as diferenças quanto as proporções são significativas. **(*, /)**	Temperatura em Kelvin, quantidades monetárias, contadores, idades, mas, comprimento, corrente elétrica	Média geométrica, média harmônica, variação porcentual

Fonte: *Tan et al., 2009, p.31.*

Conforme os autores, os atributos nominais e ordinais são categorizados ou qualitativos porque não possuem a maioria das propriedades dos números; enquanto que os intervalares e proporcionais são numéricos ou quantitativos possuem a maioria das propriedades dos números. Cada tipo de atributo possui todas as propriedades e operações dos tipos de atributos acima dele, embora cada operação estatística seja apropriada para certo tipo de atributo.

Os atributos também podem ser classificados em função do número de valores que eles podem receber:

- **Discretos** – quando possuem um conjunto de valores finito ou contavelmente infinito, ou seja, quando são representados através de números inteiros, por exemplo, códigos postais, binários sim/não ou verdadeiro/falso;
- **Contínuos** - quando possuem valores que são números do tipo real, ou seja, quando medidos com precisão limitada, por exemplo, temperatura, altura, peso (TAN et al., 2009).

Tan et al. (2009) destacam três características aplicáveis a maioria de conjuntos de dados e que condicionam as técnicas a usar na mineração:

- **Dimensão** – número de atributos que os objetos possuem. Para dados de alta dimensionalidade pode ser feito o pré-processamento para reduzir a dimensionalidade;
- **Dispersão** – atributos cujos valores diferentes de zero são importantes armazenar e manipular;
- **Resolução** – o nível de resolução determina o padrão dos dados (superfície da terra irregular para uma resolução de metros, mas plana para a resolução de quilômetros).

De igual modo, consideram três categorias de conjuntos de dados: dados de registros, dados baseados em grafos e dados ordenados.

O sucesso da tarefa de mineração de dados é influenciado pela qualidade dos dados envolvidos no processo. Contudo, a maioria dos dados não possui a qualidade adequada pelo fato de terem sido coletados para finalidades diferentes. Para colmatar esta situação, a tarefa de mineração consiste numa primeira fase, segundo Tan et al. (2009), na limpeza de dados, através da detecção e correção de problemas de qualidade de dados e no uso de algoritmos que possam tolerar essa baixa qualidade de dados.

A qualidade dos dados pode estar relacionada tanto a questões de medição e coleta, como de aplicação. Quanto à medição e coleta, o problema pode estar associado a erros humanos, limitações dos dispositivos de medição ou falhas na coleta de dados. Um erro de medição ou coleta pode ser cometido quando o valor registrado se difere do valor real do objeto ou quando se omite ou inclui inapropriadamente algum objeto de dados. Para dados temporais ou espaciais, por exemplo, registros de eletrocardiograma ruídos provocados por distorções podem comprometer a qualidade dos dados (TAN et al., 2009).

A qualidade dos dados, segundo Tan et al. (2009), também pode ser comprometida por um dos seguintes fatores:

- **Externos** – quando os objetos de dados apresentam características diferentes da maioria no conjunto de dados ou alguns atributos que sejam incomuns aos valores típicos para esse atributo[56];

[56] Contrariamente aos ruídos, os externos podem ser interessantes para algumas situações, como detecção de fraudes, cujo objetivo é identificar objetos ou situações incomuns ao padrão pré-estabelecido (TAN et al., 2009).

- **Valores faltando** – quando o objeto não tenha um ou mais valores de atributo, tanto pela falta de coleta, como pelo fato de alguns atributos não serem aplicáveis a todos os objetos. Para estes casos, a solução pode ser: eliminar objetos ou atributos com valores em falta, eliminar objetos em falta ou ignorar os valores em falta durante a análise;
- **Valores inconsistentes** – quando os valores de um objeto de dado são discrepantes entre si ou em relação aos seus atributos (por exemplo, para o objeto de dados delegacia, o valor do atributo código pode não ser referente ao valor do atributo nome correspondente);
- **Dados duplicados** – quando haja dois objetos de dados que representem um único objeto.

6.2.2 Mineração de dados e a Ciência da Informação

A mineração de dados e a Ciência da Informação são duas áreas que convergem para o mesmo objeto, a informação. Se por um lado a mineração visa à descoberta de conhecimento a partir de conjuntos de dados que, a priori, podem não fazer muito sentido para o usuário, por outro lado, a gênese da Ciência da Informação, segundo Borko (1968), está associada a uma perspectiva simbiótica, preocupada tanto com a origem e organização da informação, como no uso de técnicas ou sistemas para a representação, preservação e recuperação dessa informação. Com base nesta referência, pode-se entender que as duas áreas se baseiam em processos técnicos, contudo a informação para a CI, com base na conceituação de Capurro e Hjorland (2003), está imbricada à visão objetiva da teoria da informação e da cibernética, na qual se consideram os fenômenos de relevância e interpretação, principalmente pela significação e relevância contextualizada na cultura, na socialização, na formação, e na interação de cada indivíduo.

O fulcro da diferença entre informação e conhecimento está na origem do termo “mineração de dados” e nas clivagens terminológicas entre os campos científicos. A MD tem origem na Ciência da Computação e, pela complexidade tecnológica e explosão informacional, vem sendo adotado por outras áreas que também trabalham com a informação. Por isso, enquanto na Ciência da Computação a MD visa à descoberta do conhecimento em grandes volumes de dados, na Ciência da Informação o processo visa à informação. Por outras palavras,

mesmo que a máquina permita fazer correlações sobre os dados, a priori desconhecidas pelo ser humano, é este quem atribui os significados relacionados às suas necessidades informacionais. Por isso, as previsões, as associações ou os agrupamentos da mineração de dados terão significados diferentes, em função do contexto, da interpretação e da relevância atribuída por cada usuário. Assim, na CI a mineração de dados permite descobrir e recuperar informações que podem ser convertidas em conhecimento.

A web 2.0 testemunha a visão interpretativa baseada no contexto social e cultural, na medida em que são usuários de diferentes estratos sociais, culturas, espaços geográficos que usam, produzem e compartilham a informação. Por isso que a área da mineração de dados que foi oriunda de unidades de informação isoladas, atualmente ganha mais repercussões pelo foco na web social.

Do ponto de vista da Ciência da Informação, a mineração de dados constitui uma mais valia, não apenas por propiciar uma diversidade de visualizações da informação que, por conseguinte, pode proporcionar uma diversidade de decisões para o usuário, como também pela possibilidade de agregação de diferentes valores à informação, pelas técnicas de mineração de dados, como associação, regressão, classificação, agrupamento em clusters, etc. Por exemplo, a partir da técnica de mineração de dados um profissional de informação (bibliotecário) pode chegar à conclusão de que usuários de um determinado serviço de informação (biblioteca) recorrem predominantemente à mesma fonte (livro), em certo período do dia, da semana ou do mês. Deste modo, pode potencializar o acesso nesses períodos, adquirindo mais exemplares do livro ou reproduzindo cópias e disponibilizando-as nas primeiras estantes. De igual modo, os serviços de referência podem concluir, com base em um banco de dados dos trabalhos acadêmicos de uma universidade, que usuários sobre um determinado tema, utilizaram o mesmo grupo de autores para fundamentar os seus trabalhos e a partir daí, para trabalhos futuros sobre o mesmo tema, pode-se recomendar esses autores.

A mineração de dados fornece uma potencialidade de aplicação em diferentes áreas de conhecimento. Do mesmo modo que se criam associações ou agrupamentos na área do comércio para potencializar as vendas, mais pesquisas sobre a matéria na área da Ciência da Informação poderão culminar com a criação de perfis de usuários para áreas específicas na Web, tanto para nortear a associação ou agrupamento das fontes ou informações consultadas, como para

orientar outros usuários que buscam as mesmas fontes, informações ou assuntos a eles relacionados. Sobre esta ideia, alguns especialistas da área da Ciência da Informação mostram-se relutantes, alegando que tal condição poderia estratificar demasiadamente ou controlar a busca por informações que é de *"per si"*, uma tarefa individual e livre. Contudo, uma boa aplicação da mineração de dados poderia ser mais vantajosa que prejudicial. Ao seu lado, o uso de vocabulários controlados, enquanto linguagens de representação e recuperação, é indispensável para organizar informações e atribuir terminologias para diferentes áreas de conhecimento. Aliás, a mediação da informação é uma atividade presente em toda a cadeia informacional e por sinal, garante a eficiência e eficácia na representação e recuperação da informação.

Sobre a mediação da informação, Almeida Jr. (2008, apud FADEL et al., 2010), considera-a uma atividade que consiste na interferência do profissional da informação, seja direta ou indireta, consciente ou inconsciente, singular ou plural, individual ou coletiva, visando notabilizar a apropriação da informação no âmbito da satisfação de uma necessidade do usuário. Neste contexto, pode ser explícita quando ocorre em ambientes de relações formais entre o usuário e o equipamento informacional, sendo por isso facilmente reconhecida e implícita, quando decorre dos objetivos que norteiam a ação do profissional no geral.

6.3 Recursos de Mineração de Dados

Os recursos de MD são distintos em função da especificidade e complexidade das áreas de conhecimento que se predispõem a cobrir. As tarefas de mineração podem ser aplicadas em Banco de Dados, na Web, em textos, entre outras áreas.

O Weka é um software aberto que reúne algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados que podem ser testados pelos usuários na mineração de diversos dados. Witten et al. (2011) apontam que a ferramenta fornece o suporte para todo o processo de mineração de dados experimental, incluindo a preparação dos dados de entrada, avaliação de sistemas estatísticos de aprendizagem, visualização dos dados de entrada e o resultado da aprendizagem.

Weka significa *Waikato Environment for Knowledge Analysis* - ambiente Waikato para a análise do conhecimento e foi desenvolvido na Universidade de

Waikato, na Nova Zelândia. O sistema é escrito em Java e distribuído sob os termos da *GNU General Public License* - Licença Pública Geral. O Weka fornece uma estrutura de algoritmos aplicáveis à mineração de diversos conjuntos de dados para o aprendizado de máquina, por meio dos seguintes métodos: regressão, classificação, agrupamento, regra de mineração por associação e seleção de atributo. De igual modo, fornece ferramentas para o pré-processamento[57] e visualização dos dados (WITTEN et al., 2011).

Os algoritmos do Weka tomam como base de entrada as tabelas relacionais no formato ARFF que podem ser gerados a partir de um ficheiro do próprio Weka ou com base em uma consulta de banco de dados. Witten et al., (2011) destacam três aplicações fundamentais do Weka que podem ser selecionadas na sua interface *Explorer*:

- A partir de um conjunto de dados, pode-se aplicar um método de aprendizagem e analisar a sua saída para obter mais detalhes sobre os dados;
- A partir de um conjunto de dados, podem-se aplicar modelos de aprendizado para gerar previsões sobre novas ocorrências;
- Podem-se aplicar vários modelos de aprendizado e comparar o seu desempenho, a fim de escolher o mais adequado para a predição ou previsão.

Além da interface *Explorer* do Weka que geralmente se aplica para operações com conjuntos de dados menores, também existe a interface de algoritmos incrementais *Knowledge Flow* para grandes conjuntos de dados. Esta interface permite a especificação da fonte de dados, ferramentas de pré-processamento, algoritmos de aprendizagem, métodos de avaliação e módulos de visualização. A terceira interface *Experimenter* busca uma melhor adequação dos métodos e valores de parâmetros para cada problema, isto é, permite automatizar o processo, tornando mais fácil para executar classificadores e filtros com parâmetros diferentes em um corpus de conjuntos de dados, a fim de coletar estatísticas de desempenho e realizar testes de significância. A interface do experimentador é recomendada para usuários no nível avançado (WITTEN et al., 2011).

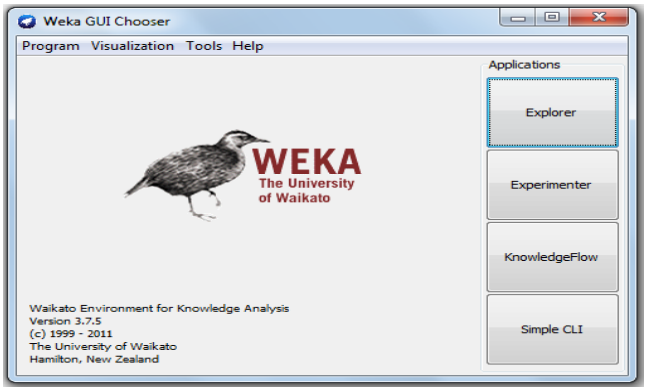
[57] Também chamados filtros, podem ser selecionados em cada menu para adequá-los a cada necessidade de mineração (WITTEN et al., 2011, p. 404).

Como se pode depreender, são diversas potencialidades que a ferramenta pode proporcionar. Na atual fase em que o mundo produz yottabytes (10^{24} bytes) de informação por dia, a mineração de dados surge como alternativa para agregar o conhecimento detalhado sobre os dados, além de diversas opções da sua visualização. No caso concreto do Weka, a maior dificuldade não reside na sua manipulação, mas sim na compreensão do conjunto de dados, do processo que se aplica e da interpretação do significado do resultado na saída dos dados. Por isso que a própria área da mineração de dados surge como simbiose de diversas áreas de conhecimento, neste caso, a estatística torna-se fundamental para o entendimento das previsões, testes da margem de erro ou desvio padrão e das correlações.

O Weka é uma ferramenta recomendada para dados de menor granularidade e obtidos a partir de banco de dados. Para textos ou documentos da Web, o uso de outras ferramentas de mineração pode ser menos complexo e produtivo, em função dos resultados proporcionados e da respectiva interpretação.

O Weka é um *software* de livre acesso e está disponível no seguinte endereço: *www.cs.waikato.ac.nz/ml/weka*. Após a instalação, pode ser inicializado com ou sem o console e a sua interface interativa *Explorer* é apresentado conforme a figura 8.

Figura 8: *Interface interativa do WEKA.*



Fonte: *Elaborado pelo autor com base no software Weka 3.7.5*

Para usar o Weka, inicialmente precisa-se preparar os dados, em seguida, carregar o Explorer, selecionar o método para a construção da árvore de decisão, construir a árvore e interpretar os resultados de saída. O conjunto de dados pode ser construído no próprio programa ou obtido a partir de um banco de dados. A obtenção a partir de um banco de dados pode consistir em uma planilha

apresentado sob a forma de lista de registros com virgulas entre itens, no formato CSV. Neste caso, carrega-se o arquivo em um editor de texto, adiciona-se o nome do conjunto de dados usando a tag @ *relation*, as informações de atributo usando @ *attribute*, e uma linha de dados usando @ *data* e salva-se o arquivo como texto. Contudo, o Explorer pode ler os arquivos no formato CSV, não havendo a necessidade de seguir os procedimentos descritos (WITTEN et al., 2011).

A maior parte das aplicações atuais desenvolvidas com a programação está na Web ou visa a Web. Python é uma linguagem de programação orientada a objetos, usada no desenvolvimento de aplicações para diversos tipos de arquiteturas: celulares, computadores, jogos multimídia, gestão empresarial, mineração de dados, entre outras. O interpretador Python pode ser estendido com novas funções e tipos de dados implementados em C ou C++ ou pode ser usado como uma linguagem de extensão para aplicações personalizadas.

A linguagem Python foi criada nos inícios de 1990, por Guido van Rossum em Stichting Mathematisch Centrum – um Instituto de Pesquisas para a Matemática e Ciência da Computação, na Holanda. Em 2001, criou-se a Fundação do Software Python – Python Software Foundation (PSF), como uma organização sem fins lucrativos, específica à propriedade intelectual relacionada com Python, por isso, todas as versões estão em forma de código aberto.

No Brasil, a comunidade Python Brasil reúne grupos de usuários interessados em difundir e divulgar a linguagem Python, consolidando todo o material editado sobre esta linguagem na língua portuguesa. Atualmente, a Comunidade está empenhada na tradução da documentação da linguagem Python, através do Projeto de Documentação Python-Br.

Segundo Russel (2011, p.22), a instalação do Python para usuários do Windows pode ser feita a partir do endereço (http://www.activestate.com/activepython) que além de adicionar o Python ao seu path no prompt de comando do Windows ou terminal, também adiciona o easy_install, um gerenciador de pacotes que permite instalar pacotes Python sem a necessidade de código-fonte. Também se recomenda que se instale o pacote NetworkX, para a criação e análise de grafos.

6.4 Análise estrutural das redes sociais

A rede social, segundo Emirbayer e Goodwin (1994), é um grafo que mapeia certa realidade através de nós e arestas. Esses nós representam as entidades (indivíduos, classes de indivíduos ou atores) e as arestas, os relacionamentos entre entidades. Assim, as relações na rede consistem no compartilhamento de um ou mais atributos ou ainda fontes de dados.

Mesmo não sendo parte da mineração de dados, a análise estrutural de uma rede social é fundamental para compreender não só a estrutura, como também a dinâmica de funcionamento e as posições relativas dos membros ou atores. Estes aspectos permitem a obtenção de uma riqueza de detalhes para o uso estratégico da rede, principalmente no ambiente da educação, política ou comércio. Algumas redes sociais são complexas e os seus detalhes só podem ser pormenorizadamente analisados com recurso à mineração de dados.

Lemieux e Ouimet (2004) descrevem o processo da análise estrutural das redes sociais, a partir das relações entre os respectivos atores. Essas relações podem ser amigáveis ou positivas e hostis ou negativas. De realçar que mesmo que determinada rede social seja caracterizada por relações positivas, parte dessas relações entre alguns dos seus atores pode ser negativa. Isso acontece quando uma parte dos atores apenas se comunica entre si por intermédio de um terceiro ator, que ocupa uma posição vantajosa na rede[58].

A maioria das relações nas redes sociais é amigável, na medida em que se baseia em laços de afinidade pessoal, temático ou ideológico. Contudo, também existem relações hostis enraizadas em divergências que suscitam algum tipo de interesse no debate sobre determinados assuntos. Neste caso, segundo Lemieux e Ouimet (2004), existe o princípio de grupabilidade que permite a formação de blocos para relações internas positivas e relações externas negativas. Além disso, a relação pode mudar em função dos interesses específicos dos atores, em cada momento ou situação.

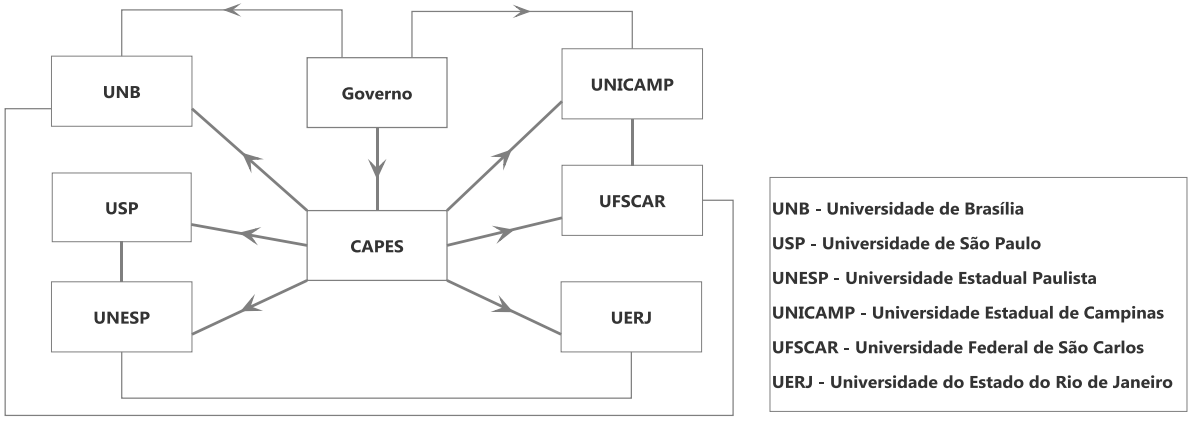
Uma das vantagens da análise estrutural consiste na explicação do processo de produção das políticas públicas com base nas relações entre atores individuais ou coletivos, pois, as políticas públicas constituem interesses e estratégias dos atores no contexto das regras de jogo ou de entendimentos institucionais. Por

[58] Esta característica recebe a denominação de “buracos estruturais”, segundo Burt (1992, apud LEMIEUX e OUIMET, 2004).

exemplo, no âmbito da construção do conhecimento, a análise estrutural permite o enfoque sobre as formas estáveis ou evolutivas que as relações adquirem no contexto interno e externo. Estas formas condicionam o pensamento coletivo, a ideologia ou a crítica adjacente ao sentimento coletivo, dada a influência ou posição estratégica ocupada pelo respectivo mentor.

A maior parte das influências, segundo Lemieux e Ouimet (2004, p.17), ocorre nas “relações orientadas”, isto é, aquelas em que há transmissão de um ator para outro, da informação, bens, serviços e controle. Em “relações não orientadas”, geralmente não existem transmissões unilaterais de um ator para outro, mesmo que existam mensagens que circulam entre eles. De salientar que existem redes orientadas cujas relações entre os atores se consideram não orientadas. É o caso de algumas redes acadêmicas, em que maior parte das mensagens emitidas pelos atores que ocupam as posições estratégicas acaba exercendo algum tipo de influência ou controle sobre os demais, mesmo que a base do relacionamento seja o simples compartilhamento de informações ou pesquisas.

***Figura 9**: Relações assimétricas entre os principais atores da política de ensino superior no Brasil*



***Fonte:** Criada pelo autor com base em Lemieux e Ouimet (2004, p.19)*

A figura 9 ilustra uma relação orientada ou assimétrica, em função de determinados traços de troca de informação. Por exemplo, o governo retém mais informação na relação com a CAPES, por conseguinte, exerce algum tipo de influência ou detém certa vantagem sobre ela. O mesmo acontece com a CAPES, na relação com as universidades ou instituições de ensino superior. Já as universidades ou instituições de ensino mantêm entre si relações não orientadas, baseadas na troca de informações ou cooperação.

Algumas redes sociais na Web (Twitter, Facebook e Linkedin) apresentam-se como relações não orientadas ou simétricas, em função da aparente igualdade e liberdade entre os diversos atores. Todavia, na prática podem desenvolver relações orientadas, ou seja, com base nas características de determinado ator, pode-se obter vantagem que se configure no poder de influência sobre os demais. Um exemplo elucidativo é de estudantes que seguem um renomado pesquisador ou docente no Twitter, ou ainda de funcionários que mantêm o relacionamento com o presidente do conselho de administração da empresa. Neste tipo de relações sempre ocorre certo tipo de influências. No caso do Facebook, algumas influências são demonstradas pelo compartilhamento ou comentários favoráveis das publicações.

Quanto ao tipo de conexidade, Lemieux e Ouimet (2004) classificam as redes sociais em: não conexas, de conexidade quase forte, de conexidade semiforte e de conexidade forte. Uma rede é não conexa quando um ou mais atores estão isolados, representando uma estruturação desintegrada das relações.

A conexidade quase forte acontece quando existe no mínimo um ator dominante, de um conjunto de atores que não tenham nenhuma conexão entre si. Este tipo de conexidade representa uma estruturação hierárquica (total ou parcial), pelo fato de pelo menos dois atores não terem qualquer conexão entre si.

A conexidade semiforte acontece quando existe no mínimo um ator dominante, de um conjunto de atores que tenham pelo menos uma conexão entre si. Por isso esta conexidade representa uma estruturação estratificada, pois é possível distinguir no mínimo dois estratos de atores.

A conexidade forte ocorre quando todos os atores são dominantes. Este tipo de conexidade é extremamente raro nas redes sociais.

Lemieux e Ouimet (2004, p.24) acrescentam que nas relações orientadas, um ator pode ser de posição:

- “Dominante” quando é o emissor de uma conexão com cada um dos outros atores;
- “Dominada” quando num conjunto de atores em que existe pelo menos um ator dominante, não é o emissor de nenhuma uniconexão;
- “Semidominante” quando, apesar de não ocupar uma posição dominante, é o emissor de uma uniconexão com um ou mais atores e

igualmente o destinatário de uma uniconexão proveniente de um ou mais atores;

- "Subdominante" quando, na ausência de um ator dominante e apesar de não se encontrar numa posição semidominada, é o emissor de uma uniconexão com um ou vários atores;
- "Subdominada" quando, num conjunto em que não haja qualquer ator em posição dominante, não é o emissor de nenhuma conexão e é o destinatário de uma uniconexão proveniente de pelo menos outro ator;
- "Isolada" quando, num conjunto de atores, não é nem o emissor nem o destinatário de qualquer uniconexão com outro ator.

Para relações não orientadas, Lemieux e Ouimet (2004, p.26) consideram as seguintes medidas de centralidade: grau, proximidade e intermediariedade. A centralidade do grau "mede o número de conexões diretas de cada ator num grafo". Assim, o ator que ocupa a posição mais central é aquele que possui o maior número de conexões diretas com outros atores. Já a centralidade de proximidade considera a capacidade de autonomia ou de independência dos atores, isto é, "quanto mais um ator se encontra afastado dos outros atores, mais autônomo será no que se refere às suas ecolhas de ações".

Por último, a centralidade de intermediariedade tem o objetivo de avaliar a capacidade que os atores têm na coordenação e controlo. Assim, de acordo com Lemieux e Ouimet (2004, p.28) "quanto mais se encontrar numa posição intermediária, ou seja, quanto mais se encontrar numa situação em que os atores têm de passar por ele para chegar aos outros atores, mais capacidade de controlo terá sobre a circulação da informação entre esses atores".

A análise estrutural, conforme se referiu anteriormente, é fundamental para a compreensão do modelo de organização e funcionamento de uma rede social, bem como para a definição de estratégias para determinadas áreas como (política, educação, comércio, etc.). Por exemplo, a posição ocupada por cada ator ou o grau de domínio sobre os demais pode ser crucial para a implementação da publicidade de produtos na área comercial, ou mesmo para uma campanha de angariação de novos membros de um partido político ou clube esportivo.

Na tabela 7, Lemieux e Ouimet (2004) apresentam algumas teorias adotadas e técnicas utilizadas para certos domínios.

Tabela 7: *Teorias adotadas e técnicas utilizadas em sete análises estruturais*

Domínio da análise estrutural	Teorias adotadas	Técnicas utilizadas
1. As relações de parentesco	Teoria da grupabilidade Teoria dos laços	Observação direta
2. As redes sociométricas	Teoria da grupabilidade (Teoria da coordenação) (Teoria dos buracos estruturais)	Questionário
3. O capital social	Teoria dos laços (Teoria dos buracos estruturais)	Análise documental Entrevista
4. As redes de apoio descritivo	Processo essencialmente Teoria da troca Teoria dos laços	Entrevista
5. As redes de mobilização	Processo essencialmente descritivo (Teoria dos laços)	Observação direta
6. As redes de empresas	Processo essencialmente descritivo (Teoria dos laços)	Análise documental
7. As redes de política pública	(Teoria da coordenação)	Análise documental Entrevista

Fonte: *Lemieux e Ouimet (2004, p.66)*

No caso específico da Web, algumas teorias e técnicas sugeridas por Lemieux e Ouimet (2004) podem ser de difícil execução, devido à complexidade das redes sociais e à sofisticação tecnológica. Daí a necessidade da mineração de dados e de recursos de visualização, como o *RGraph, Graphviz e WP-Cumuls*[59], para integrá-los às teorias e técnicas sugeridas pelos autores.

6.5 Mineração de dados da Web social (Facebook)

A Web social ou as redes sociais, conforme anteriormente se referiu, são ambientes de compartilhamento de informações entre usuários na Web. Com o crescente uso destes espaços, muitas potencialidades dos dados produzidos não são exploradas pelos atuais métodos de processamento de informação na CI. No caso do Facebook, a proposta inicial que era de atender aos estudantes da Universidade de Harvard foi expandida para outras universidades e para o público em geral. Com isso, esta rede social teve um crescimento explosivo e em 2012 tinha mais de 1 bilhão de usuários ativos que podiam adicionar outros usuários como amigos, trocar mensagens, postar e visualizar informações no mural (status, fotos, notificações, etc.) e participar de grupos de interesse comum. Atualmente a rede Facebook também é usada para a difusão e compartilhamento de notícias, por parte de grandes canais de mídia e propaganda de produtos e serviços.

A contribuição do Facebook é notável em muitas áreas de interesse como: educação, comércio, política, notícias, relacionamentos, saúde, entre outras.

[59] Alguns destes recursos de visualização são descritos no ponto 6.5.

Contudo, o fluxo de informações de caráter público neste ambiente ocorre de forma contínua, através do mural e o seu acompanhamento requer que o usuário esteja permanentemente conectado na rede. Assim, essas informações são de difícil recuperação, na medida em que estão preservadas em suportes intangíveis tanto para o usuário comum, como para os sistemas atualmente em uso no campo da Ciência da Informação. Do ponto de vista científico, por exemplo, este problema afeta grande parte das pesquisas com foco neste tipo de dados.

O outro aspecto que se prende à difícil recuperação das informações do Facebook consiste na especificidade da rede social em si. Ora, o princípio do Facebook é de relacionamento entre os usuários e o acesso à informação é permeabilizado nesse contexto. Por outras palavras, o usuário só pode acompanhar informações sobre outros usuários se tiver algum vínculo com os mesmos, adicionando-os como amigos ou seguindo-os através do recurso "seguir" ou *"follow"*. Este é mais um entrave na construção do conhecimento, na medida em que nem sempre as necessidades informacionais dos usuários são explícitas. Por exemplo, usuários que não conhecem as suas necessidades sobre notícias e curiosidades, dificilmente criam vínculos com os órgãos de informação, como Folha de São Paulo, Globo News, G1, The New York Times, CNN, para acompanhar as respectivas publicações.

O Facebook não tem o mecanismo de busca de informações, o que supõe que os usuários conhecem as informações que precisam para criarem vínculos de acesso. Uma das limitações da MD do Facebook é a condição imposta pela privacidade, pois nem todas as informações estão disponíveis para os aplicativos.

As restrições no âmbito da política de privacidade do Facebook vêm sendo reformuladas em função das críticas sociais, algumas das quais estão relacionadas à violação da privacidade e manipulação dos usuários. Uma matéria publicada na revista norte-americana The Atlantic (2014) indica um estudo efetuado pelo Facebook com o intuito de alterar deliberadamente o estado emocional dos usuários, através de notícias com palavras positivas, para aumentar a popularidade da rede social. O problema desta pesquisa é que além de interesses comerciais e de manipulação, violou os princípios éticos na medida em que não observou o consentimento das pessoas que participaram da pesquisa e viciou as notícias propagadas com termos que fossem preferidos pelos usuários. Ademais, a própria política é controversa porque permite que o Facebook utilize os dados dos usuários

na forma que entender. Estas controvérsias em relação à rede social não eliminam o potencial de dados produzidos diariamente, cujos detalhes são imprescindíveis para o conhecimento. Neste sentido, a mineração de dados do Facebook é fundamental para a extração desses detalhes e fornecimento de opções de visualização.

Segundo Russel (2011, p.300), para a mineração de dados do Facebook é preciso criar aplicativos que utilizem OAuth, observando os princípios e políticas de uso. Estes aplicativos são hospedados em próprios ambientes de servidor.

Com base na própria documentação do Facebook, OAuth é um padrão do protocolo aberto que permite a autorização segura e simples de aplicações Web, móveis e *desktop*. A estrutura de autorização OAuth 2.0 permite que aplicativos de terceiros tenham acesso limitado a um serviço HTTP, tanto usando o nome do proprietário do recurso, como através do nome desses terceiros.

OAuth 2.0 substituiu OAuth 1.0, no qual a autenticação era cliente-servidor, ou seja, o cliente solicitava um recurso protegido no servidor usando as credenciais do proprietário desse recurso. Para tal, o proprietário do recurso era obrigado a compartilhar as suas credenciais com terceiros, o que criava vários problemas e limitações. Por exemplo, os aplicativos de terceiros guardavam as credenciais do proprietário (senhas) para uso futuro, os servidores eram imprescindíveis para suportar a autenticação de senhas, os proprietários não conseguiam restringir o acesso de terceiros aos seus recursos e o comprometimento do aplicativo de terceiros resultava no comprometimento da senha do usuário final e dos dados protegidos.

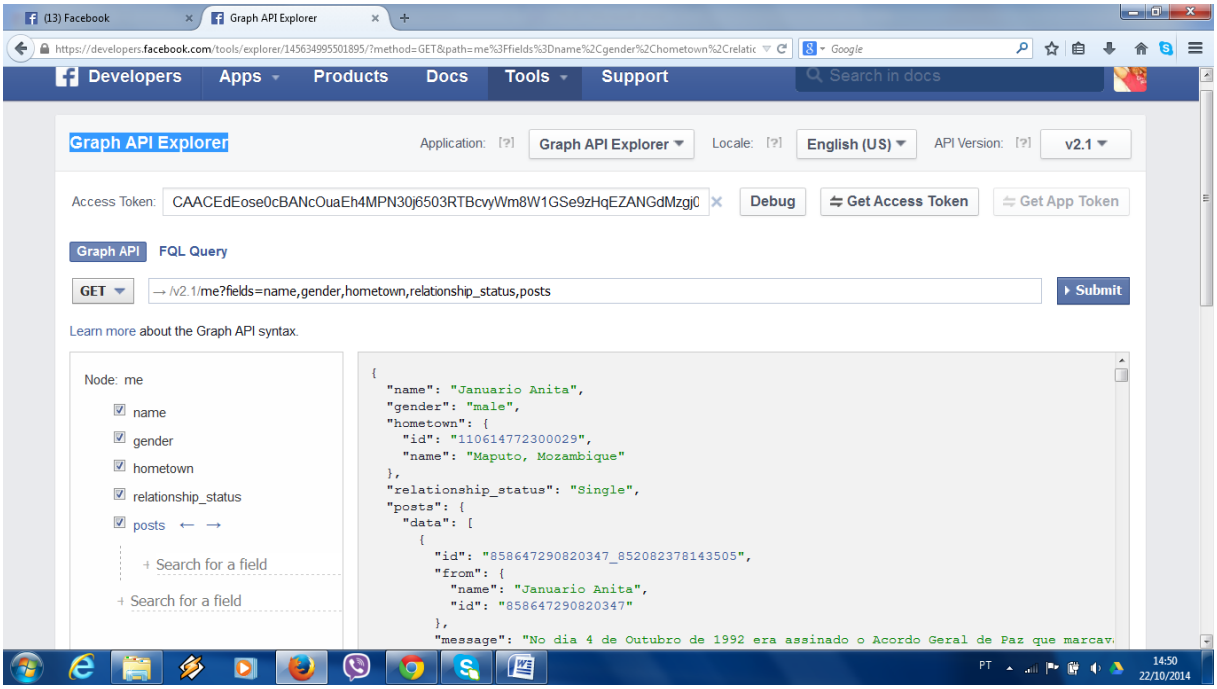
O Auth 2.0 permite que o usuário tenha acesso aos recursos protegidos por diferentes proprietários e hospedados em servidores, através de um *token* de acesso ou string indicando o âmbito específico, o tempo de vida e outros atributos de acesso. Esses *tokens* de acesso são emitidos por um servidor de autorização e dependem da aprovação do proprietário do recurso.

A MD do Facebook pode ser feita através da Interface de Programação de Aplicativos – *Application Programming Interface (API)*, disponível para desenvolvedores da Web. Antigamente, a aplicação Graph API do Facebook permitia o acesso a todas as informações sobre o usuário e as pessoas a ele relacionadas.

A Graph API é um aplicativo que permite obter dados dentro e fora do gráfico social do Facebook. É uma API baseada em HTTP para a consulta de dados, envio de mensagens, upload de fotos, entre outras tarefas.

Com a mudança da política de privacidade e, consequentemente, da documentação sobre os aplicativos do Facebook, apenas é possível trabalhar no ambiente do próprio usuário, usando a interface Explorer para o aplicativo Graph API conforme ilustra a figura abaixo. Assim, o primeiro passo é a criação de uma conta ou registro no Facebook. Em seguida, a instalação da aplicação *"Developer"* em http://www.facebook.com/developers, autenticando a senha e aceitando as políticas de privacidade. Neste caso, é necessário verificar a conta e confirmar o código enviado pelo telefone.

Figura 10: *Mineração de dados com Graph API Explorer*



Fonte: *Produzido pelo autor com base na Graph API Explorer do Facebook*

A Graph API Explorer permite testar o aplicativo Graph API e outros aplicativos criados pelo próprio usuário. A interface Explorer permite:

- Fazer diferentes pedidos para a API Graph através da sintaxe *"GET"* e visualizar os resultados formatados em linha;
- Postar ou apagar informações por meio das sintaxes *"POST"* e *"DELETE"*, respectivamente;

- Explorar as conexões para cada objeto e visualizar as descrições de campos para entender o significado de cada resposta;
- Obter um *access_token* com as permissões específicas e necessárias para acessar os dados;
- O *access_token* segundo o OAuth 2.0 fornece o âmbito específico, o tempo de vida e outros atributos de acesso. Neste caso, é preciso especificar as “permissões” básicas e “entendidas” que o aplicativo terá acesso.

Conforme a figura 10, depois de especificar o aplicativo Graph API e de obter o *access_token* com as devidas permissões especificadas, é possível recuperar a informação sobre vários nós do usuário, como: nome, gênero, cidade de proveniência, estado civil, informações postadas, entre outros. De realçar que mesmo que permita a recuperação de informações do usuário, o Graph API Explorer permite minerar e agregar maior valor às informações do Facebook, pois contém uma riqueza de detalhes ou granularidade que não aparecem na interface do usuário comum. Por exemplo, numa publicação é possível visualizar o seguinte: a própria publicação; as ações feitas em relação à publicação (comentários, “likes” e compartilhamentos); a privacidade da publicação (descrição como pública ou privada, o valor da disponibilidade, os amigos que podem visualizar, os relacionamentos, as permissões e restrições); o tipo de publicação; o nome e a identificação da aplicação usada para a publicação; a data e hora em que a publicação foi feita; a data e hora em que a publicação foi atualizada; o nome e identificação das pessoas que comentaram, curtiram ou compartilharam, etc. Se for uma informação compartilhada, também é possível visualizar os dados do respectivo autor.

A visualização e o processamento dos dados na Graph API Explorer podem ser feitos rapidamente, integrando diferentes informações na mesma planilha.

Com a mudança da política do Facebook, atualmente com a Graph API Explorer é impossível minerar os dados de outros usuários, além do próprio proprietário ou criador da API. Deste modo, depois dos mesmos passos necessários para a Graph API, é preciso criar o próprio aplicativo do usuário no *Software Development Kit (SDK)*, dando o nome e a respectiva categoria. Esse aplicativo, conforme Russell (2011, p.300), pode ser criado no *“Developer”* do Facebook ou no

Google App Engine (GAE), pois os aplicativos do Facebook devem ser hospedados externamente em seu próprio ambiente de servidor.

Após a verificação da segurança, o aplicativo terá um identificador - ID e um token de acesso que permitem a implementação OAuth 2.0. Depois é preciso configurar o Web Site do aplicativo, digitando o URL nos referidos campos.

Para que o aplicativo criado pelo usuário acesse os dados do Facebook é necessário escrever um script para a autenticação e obtenção do *token* de acesso na linguagem Python. Segundo Russell (2011, p. 302), a obtenção de um *token* de acesso OAuth 2.0 para uma aplicação desktop *(facebook_login.py)* pode ser feita através da seguinte linguagem:

```
# _*_ coding: utf-8 _*_
import os
import sys
import webbrowser
import urllib
def login():
        CLIENT_ID = 'valor da API criada no Facebook'
        REDIRECT_URI = \
                'URL na qual se encontra hospedada a API do Facebook'
EXTENDED_PERMS = [
        'user_about_me',
        'friends_about_me',
        ]
args = dict(client_id=CLIENT_ID, redirect_uri=REDIRECT_URI,
        scope=','.join(EXTENDED_PERMS), type='user_agent', display='popup')
webbrowser.open('https://graph.facebook.com/oauth/authorize?'
        + urllib.urlencode(args))
access_token = raw_input('Enter your access_token:')
if not os.path.isdir('out'):
        os.mkdir('out')
filename = os.path.join('out', 'facebook.access_token')
f = open(filename, 'w')
f.write(access_token)
f.close()
print >> sys.stderr, \
        "Access token stored to local file: 'out/facebook.access_token'"
return access_token
if _name_ == '_main_':
        login()
```

O termo *"EXTENDED_PERMS"* se refere a permissões estendidas que podem ser atribuídas ao aplicativo, através do respectivo *token* de acesso, de acordo com as necessidades da mineração. Algumas dessas permissões encontram-se sintetizadas na tabela 8, na medida em que se tratam das mesmas que podem ser atribuídas ao token de acesso da API Graph. Os detalhes sobre o token de acesso e as permissões se encontram na própria documentação de autenticação do Facebook para desenvolvedores.

Tabela 8: *Permissões de token de acesso e os dados a serem minerados*

Cate-goria	Permissões do Token de Acesso (usuário e amigos)	Alguns Nós	Dados colhidos
Sobre	User/Friend_about_me	Id, name	ID e nome (usuário/amigos)
		First/Middle/Last_name	Primeiro nome/Nome do meio/Último nome (usuário/amigos)
		Name_format	Formato do nome (nome/nome do pai/sobrenome ou outros)
		Gender	Gênero (usuário/amigos)
		Adress	Endereço (usuário/amigos)
		Bio/Age_range	Biotipo (detalhes sobre a entidade)/Faixa etária
		Languages	Idiomas (ID e nome)
		Cover	Foto de capa (deslocamento, ID, fonte, data e hora da criação ou postagem; metadados: coordenadas geográficas ou URL, altura e largura; ID e nome de quem comentou ou curtiu a foto, data e hora do comentário ou curtida)
		Family	Família (ID e nome de cada membro e tipo de relação familiar)
		Devices	Dispositivos em uso para a navegação
		Installed	Situação do usuário ou amigos (ativo ou inativo)
		Link/Website	Endereço do perfil/Website do usuário ou amigos
	User/Friend_birthday	Birthday	Data, mês e ano de nascimento
	User/Friend_relationship_details	Status/Context	Estado civil/Amigos, amigos em comum (ID e nome), curtidas mútuas (categoria, nome e ID)
	User/Friend_location	Location/Timezone	Localização (ID e nome da cidade da cidade e do país)/Fuso horário
	User/Friend_hometown	Hometown	Cidade natal (ID, nome da cidade e do país de naturalidade)
	User/Friend_education_history	Education	Educação (ID, nome e tipo de cada escola; ID e nome de amigos vinculados)
	Email	Email	Endereço eletrônico do usuário ou amigos
	User/Friend_work_history	Work	Trabalho (ID, nome de cada local; ID e nome de amigos vinculados)
	User/Friend_interests	Interested_in/sports	Preferências (mulheres/homens)/esportes preferidos (ID e nome)
		Favorite_athletes	Atletas favoritos do usuário ou amigos (ID e nome)
		Favorite_team	Equipe favorita do usuário ou amigos (ID e nome)
		Inspirational_people	Pessoas que servem de inspiração (ID e nome)
	User/Friend_religion_politics	Religion/Political	Religião e interesses políticos do usuário ou amigos
Relações	User/Friend_friends	Friendlists/friends	Lista de amigos (ID, nome e tipo de lista ou relação)/Amigos (ID e nome)
	User/Friend_groups	Suggested/Tested	Grupos sugeridos/Grupos de participação
		Groups	Grupos (ID e nome, publicações lidas e não lidas pelo usuário no grupo)
	User/Friend_likes	Likes	Curtidas (ID, nome, categoria e data de criação da entidade ou do evento curtido)
	User/Friend_events	Events	Eventos (ID e nome, data e hora de criação e modificação)
Ações na rede	User/Friend_status	Feeds ou statuses	Publicações: ID da publicação, ID e nome do proprietário; mensagem publicada e ações feitas (nome e link de comentários, curtidas e compartilhamentos); configurações de publicidade; tipo de mensagem e dispositivo na qual foi produzida; datas de criação e modificação
		Updated_time	Hora, dia, mês e ano da última atualização de perfil
	User/Friend_activities	Home	Ações feitas na rede (publicações, comentários, curtidas – detalhes de cada ação)
		Inbox	Mensagens (ID e nome do emissor e receptor; data e hora do envio e da leitura, conteúdo da mensagem, data e hora da criação; detalhes (visualizada, lida, comentada, etc)
		Notifications	Notificações (ID da notificação, ID e nome do emissor e do receptor,data e hora da criação e modificação, título, link e detalhes)
	User/Friend_photos	Albums	Álbuns (ID do álbum; código e nome do proprietário; nome, link, tipo, privacidade, datas de criação e modificação do álbum; ID e nome dos comentários e curtidas)
		Photos	Fotos (ID e data de criação, ID e nome do criador, detalhes: tamanho, extensão, link, fonte, coordenadas geográficas do local da foto)
	User/Friend_website	Website	Website (ID, nome e link)
	User/Friend_actions.books	Books	Livros (ID e nome)
	User/Friend_actions.music	Music	Músicas (categoria, ID, nome, data e hora de criação; link)
	User/Friend_tagged_places	Places	Lugares (ID, nome e data da visita)
	User/Friend_videos	Videos	Vídeos (categoria, ID, nome, data e hora de criação; link)
	User/Friend_actions.video	Video_upload_limits	Tamanho e duração dos vídeos adicionados
	User/Friend_games_activity	Games	Jogos (ID e nome)

Fonte: *Criada pelo autor com base na Graph API Explorer do Facebook*

Com o *token* de acesso, Russell (2011, p. 310) recomenda a execução do *"easy_intall facebook-python-sdk"*, que é um SDK Python oficial para a API Graph que permite a consulta de informações, por meio da sintaxe *"Get"*. Tal como na figura 9, os resultados das consultas são apresentados em *links* que podem ser rastreados, associando diferentes níveis de grafos de cada objeto. A API usa o *Facebook Query Language (FQL)*, que relaciona o usuário aos dados da conexão.

A linguagem FQL é semelhante à Structured Query Language - linguagem SQL. Por exemplo, uma consulta FQL para o nome, sexo e estado civil do usuário do Facebook é baseada na seguinte sintaxe, segundo Russell (2011, p.315):

> Select name, sex, relationship_status from user where uid in
> (select target_id from connection where source_id = me() and target_type = 'user')

Enquanto a maioria das conexões pode existir entre usuários, outras conexões envolvem usuários e outros objetos de dados, como páginas, e isso se consegue através do filtro *"target_type"*.

Além da recuperação, segundo Russell (2011, p.319), a MD também permite a visualização de dados. Algumas das visualizações podem ser feitas com o *RGraph*[60] da biblioteca do *JavaScript Info Vis Toolkit (JIT)*, através de uma estrutura com listas de objetos que representam os diferentes nós e suas respectivas adjacências. Através de uma consulta FQL, é possível calcular os IDs dos amigos numa rede de relacionamento, conectar esses amigos no gráfico radial e capturar detalhes como nomes e datas de nascimento, para visualizar as pessoas mais populares, melhor conectadas ou agrupar os amigos de acordo com o que gostam. Com base na análise estrutural das redes sociais, também é possível analisar a conexidade da rede e a posição ocupada por cada ator dentro da rede (dominante, dominada, semidominante, subdominante, subdominada, ou isolada). Este tipo de informações é essencial em estratégias de marketing ou da política, pois os usuários que ocupam posições de destaque numa rede exercem algum tipo de influência sobre os demais que compartilham, curtem ou comentam as suas publicações.

Além do *RGraph*, o *WP-Cumulus* é outro recurso de código aberto que permite a visualização de dados do mural do Facebook como uma nuvem giratória de tags. Este recurso pode ser usado para capturar diversas páginas de notícias e calcular uma estrutura JSON de listas de tuplas *(term, URL, frequency)*, que pode

[60] RGraph é uma visualização de rede que organiza a exibição, dispondo os nós em círculos concêntricos (RUSSELL, 2011, p.319).

ser alimentada em um template HTML. Deste modo, através da nuvem de tags que representa a frequência dos termos mais usados, é possível visualizar os assuntos das conversas no mural ou as *trends* (RUSSELL, 2011).

A mineração de dados na CI traduz a essência da complexidade do campo, herdada da tecnologia e da Revolução Científica, enquanto movimento caracterizado pela dissociação da ciência no geral, da filosofia atrelada à busca pelo entendimento ou essência do universo. Assim, o surgimento de novos campos como a CI, justificava-se pela necessidade de um viés científico voltado para os problemas da humanidade. Neste contexto e, conforme anteriormente se referiu, a CI é um campo científico pós-moderno, com foco em todos os processos informacionais, visando a acessibilidade e usabilidade.

A MD amplia o universo da CI, na medida em que através do uso de aplicativos, permite a recuperação de informações ricas em detalhes de metadados que podem ser preservadas e reutilizadas em diferentes unidades ou contextos informacionais. De igual modo, amplia o objeto da CI, permitindo o acesso e representação de informações, a priori inacessíveis, em suportes tangíveis para os atuais métodos e processos em uso na área.

A MD do Facebook, por exemplo, possibilita a recuperação de informações publicadas por qualquer integrante da rede, associado à API. Deste modo, é possível obter informações com granularidade maior ou ricas em detalhes que propiciem novas inferências no conhecimento. Além da simples recuperação da foto no Facebook, pode se saber quem postou, quem compartilhou, comentou ou curtiu, qual foi o dispositivo usado para a sua criação e publicação, qual foi a data e hora da publicação, em que lugar foi tirada (latitude e longitude), entre outros detalhes. Estes detalhes podem ser agregados pelo usuário, tanto para aperfeiçoar a recuperação, como para produzir associações sobre dados na Web ou na situação que norteia a busca.

A MD é um recurso bastante fundamental na análise estrutural das redes sociais. Muitos detalhes que a priori escapam da percepção pela limitação das capacidades humanas, podem ser explorados por diferentes áreas, de modo a subsidiar a produção do conhecimento.

7

Considerações finais

7.1 Visão sobre o Capítulo

Nesta seção apresenta-se a síntese do trabalho, elucidando-se os principais resultados alcançados pela pesquisa em relação aos antecedentes históricos que nortearam a origem da CI, as características do campo e suas abordagens e as perspectivas futuras sobre o conhecimento. Igualmente, se apresentam as limitações da pesquisa, e se faz um delineamento para pesquisas afins que possam enriquecer a temática da abordagem.

7.2 Considerações

A pesquisa partiu do objetivo sobre a proposta da mineração de dados na CI como solução para a recuperação da informação intangível em ambientes da Web Social. Assim, ao longo do trabalho adotou-se o conceito da mineração de dados voltado para a perspectiva da CI sobre a recuperação da informação, e não da Ciência da Computação, voltado para a descoberta do conhecimento.

Em correlato, a abordagem visou outros objetivos a seguir: contextualizar a gênese e os problemas que a CI se predispôs a resolver através do objeto "informação" e delinear a fragmentação da informação nos ambientes informacionais digitais colaborativos. Estes objetivos foram alcançados pelos Capítulos 2 e 3, respectivamente. Os resultados atingidos por estes objetivos foram esmiuçados a partir do contexto da revolução das ciências.

Os primeiros conceitos sobre aquilo que atualmente se conhece como ciência estavam somente atrelados à Filosofia, com enfoque na busca do entendimento sobre a essência do universo. A Revolução Científica que ocorreu entre os séculos XVI e XVIII, foi caracterizada pelo desmembramento de outros campos científicos, da Filosofia, e pelo abandono da busca incessante pelo entendimento para a incidência aos problemas humanos. Neste contexto, o pensamento iluminista que na modernidade desempenhou um papel importante no conhecimento, através do culto da razão para a liberdade, emancipação e autonomia, contrapondo-se à religião e ao poder monárquico, teve a queda decretada pelas grandes narrativas totalitaristas e universais, às quais o pós-modernismo se contrapôs.

O advento da pós-modernidade, fundamentalmente entrelaçado com a tecnologia, provocou a queda das grandes narrativas e da ideologia prevalecente na cultura, nas artes, na moda, na religião e, principalmente, na ciência. Com os

constantes questionamentos sobre os critérios de razão e verdade, o conhecimento perdeu a validade universal e ganhou a circunstancialidade e o subjetivismo. Neste cenário, a atividade de produção do conhecimento que, no passado era reservado aos eruditos filosóficos, passou a ser de todos os sujeitos, enaltecendo a explosão informacional, igualmente fortalecida pela tecnologia.

A relação entre a tecnologia e o subjetivismo na construção do conhecimento propicia a fragmentação, o fugidio, e a efemeridade das informações. Estas características, aliadas às descontinuidades que refutam quaisquer precedentes históricos modernistas, perfazem o conhecimento pós-moderno. Um conhecimento mediado pela tecnologia e fragmentado pela complexidade tecnológica de produção, observação, apresentação e diversificação dos modelos de representação, comprometendo, deste modo, a racionalidade. Mas, se por um lado a tecnologia provoca a fragmentação, por outro, permite a despersonalização do conhecimento e a criação de ambientes para o compartilhamento de informações.

A explosão informacional e o surgimento de novas linguagens além do textual desafiavam tanto os métodos de processamento de informação, como os tradicionais campos científicos. Assim, os tradicionais sistemas de representação e recuperação da informação que vinham sendo desencadeados por Dewey e Otlet foram consolidados e institucionalizados numa nova área de conhecimento, a CI. Por isso a CI nasceu no paradigma da complexidade dos problemas humanos e na vertente tecnicista e interdisciplinar da Biblioteconomia, Ciência da Computação, Documentação, Arquivologia e Museulogia. Do mesmo modo, as ideias sobre a extensão dos computadores de simples cálculos para a interação e integração aos problemas humanos foram sintetizadas através da Internet e da Web.

Com base na noção de campo proposta por Pierre Bordieu, concluiu se que a CI é um campo científico. Mas, ao lado das relações de disputa entre o capital temporal, institucional e institucionalizado e o poder específico ou de prestígio pessoal, também existem relações de harmonia.

Ainda existem divergências sobre a origem da CI. Enquanto alguns autores situam a CI no período pós-guerra, outros consideram que os olhares e fazeres da CI são anteriores a esse período. O trabalho concluiu que a CI institucionalizou-se como campo científico no período pós-guerra, principalmente com a Conferência da Royal Society de 1948, que debateu as questões propostas por Vannevar Bush sobre o *memex*. Mas, a maioria das suas abordagens é anterior ao período da

explosão informacional pós-guerra e, conforme se referiu, foi incorporada a partir dos campos tradicionais da Biblioteconomia, Ciência da Computação, Documentação, Arquivologia e Museulogia.

O objetivo sobre a perspectiva histórica da Web e os desdobramentos da sua complexidade foi alcançado pelo Capítulo 2. Sobre este objetivo, concluiu-se que a maior parte das ações que nortearam a criação da Web foi desencadeada antes do período pós-guerra ou da explosão informacional. Na área da Documentação, Paul Otlet criou os sistemas complexos de organização para integrar dados bibliográficos, imagens e textos, numa hibridização de linguagens que, atualmente, se usa na Web. Para a recuperação desses recursos, Otlet criou o modelo de nós associativos das facetas do documento que, mais tarde, subsidiou o hipertexto. Além disso, há muito tempo Otlet mostrou a necessidade de um sistema de tratamento de informações para reunir no mesmo espaço as novas publicações, a gestão de bibliotecas, de arquivos e dos museus, bem como da enciclopédia universal e da rede universal, para condensar todo o conhecimento humano.

Outras ações que nortearam a criação da Web foram desenvolvidas por Douglas Engelbart na área da computação. Engelbart revolucionou a área da computação através do sistema online, que passou a integrar os recursos de hardware e software, permitindo o desenvolvimento de interfaces gráficas, a criação de bibliotecas digitais, o armazenamento e a recuperação de documentos, e o processamento do texto. Através desta computação interativa, as ideias de Engelbart sobre o uso de computadores para aumentar o intelecto humano e permitir a colaboração na produção do conhecimento, subsidiaram a criação da Web.

O pós-modernismo também esteve na origem da Web, com a ideologia revolucionária da liberdade, subjetividade, e contextualidade, tanto na construção do conhecimento, como no papel desse conhecimento para indivíduos, grupos e sociedade. Por isso, à semelhança da CI, a Web nasceu no paradigma da complexidade e herdou muitas manifestações pós-modernistas.

A Web permitiu a hibridização das linguagens ou convergências de mídias numa única estrutura e cultura, a cultura digital. Assim, a produção e o compartilhamento de informações ganharam novos contornos que exigem novas abordagens para o objeto da Ciência da Informação. No campo científico, a CI notabiliza-se através de contribuições teóricas, conceituais e metodológicas sobre a

natureza, manifestações e efeitos da informação e do respectivo processo de comunicação, isto é, sobre o ciclo informacional (construção, uso e comunicação).

Os Capítulos 4 e 5 foram desenvolvidos em resposta ao objetivo sobre a análise dos sistemas de representação e recuperação da informação, dos pontos de vista histórico, conceitual e funcional. Através deste objetivo concluiu-se que, no âmbito das contribuições, a CI destaca-se com uma das suas linhas de atuação – a representação e a recuperação da informação com enfoque social, ou seja, pelo conjunto de abordagens que visam organizar e tornar a informação acessível e útil, em observância às especificidades contextuais e situacionais de cada usuário final. Atualmente, esta linha de atuação do campo incide sobre as estruturas híbridas de linguagens que impõem novos desafios enleados à complexidade tecnológica, fruto da pós-modernidade e da dimensão humana.

Um dos desafios e limitações do campo se prende à dificuldade da representação e recuperação de informações disponíveis nessa nova estrutura híbrida, com destaque para a Web Social. Tal dificuldade é condicionada pelas manifestações pós-modernistas de fragmentação e efemeridade que atuam sobre a informação, dando a sensibilidade de imaterialidade, e comprometendo o acesso ou a recuperação. Estas manifestações também incidem sobre os ambientes de produção, comunicação e uso da informação, tornando-a intangível para os processos e métodos atuais em uso na CI. Assim, esta problemática suscita o recurso à interdisciplinaridade que esteve na gênese e desdobramento da área da CI.

O trabalho defende a tese sobre o uso da mineração de dados como processo complementar aos sistemas de representação e recuperação, de modo a dar condições à tangibilidade de informações em ambientes da Web Social. A tese foi validada pelo Capítulo 6, que demonstra o processo da mineração de dados do Facebook, com base na API e na linguagem de programação *Python*. O mesmo processo pode ser usado no Twitter e no Linkedin.

A mineração de dados é uma técnica da Ciência da Computação que permite o processamento, a descoberta de informações e a visualização de detalhes de informações em grandes volumes de dados. Por isso, concluiu-se que face ao problema da intangibilidade ou inacessibilidade de algumas informações na Web, a sua adoção na CI vai complementar o processo de representação e recuperação atualmente empregue na área. Com isso, será possível reforçar o uso de dados com

granularidade maior, agregando-lhes valores que contribuam para o crescimento do conhecimento humano no todo.

Além dos problemas referenciados, também se concluiu que o crescente volume de informações e a constante “desprofissionalização” de alguns métodos no ciclo informacional constituem outros desafios da CI. Estes problemas, em parte, são influenciados pela liberdade e pelo subjetivismo que caracterizam a Web. Por exemplo, em alguns ambientes da Web não existem padrões para a estrutura do conteúdo e a indexação das páginas é feita pelos próprios usuários, sem a observância estrita dos requisitos da representação, como o uso de metadados. Assim, muita informação se perde ou se torna incoerente, inconsistente ou mesmo inconfiável.

Os problemas em relação aos quais a CI tem de lidar não param por ai. A área da representação e recuperação da informação é complexa e exige um elevado grau de conhecimento, pois envolve variações linguísticas que impõem indeterminados jogos de linguagem, nos quais ocorrem as intersecções e os modelos de representação de cada sujeito.

7.3 Limitações e recomendações

O trabalho teve limitações, maior parte das quais, sobre a impossibilidade de testar o código que permite a mineração de dados na linguagem Python. Conforme se referiu anteriormente, a MD é oriundo da Ciência da Computação e, por isso, o seu uso na Ciência da Informação exige certo grau de conhecimentos sobre a programação e a estatística. Além disso, impõe outras terminologias que podem resultar em conflitos; por exemplo, em vez da descoberta do conhecimento, o trabalho adotou o conceito da recuperação da informação, de modo a evitar controvérsias no campo da CI. Isso pode suscitar críticas de caráter interpretativo para o significado e alcance do termo “mineração de dados”.

A CI é complexa porque está enleada à tecnologia e aos problemas informacionais humanos. A tecnologia e a sociedade são dinâmicas e cada estágio do seu desenvolvimento reflete-se nos aspectos de cada indivíduo ou grupo, por isso, o tema sobre o campo da CI é infinito. Pesquisas afins, com vieses interdisciplinares, se mostram cada vez necessárias para este contexto. Ademais, mesmo com muitos anos de pesquisa, os conceitos de informação, conhecimento e comunicação ainda revestem o caráter de subjetividade, fruto das clivagens da

dimensão humana. Tratando-se de problemas de ordem epistemológico, essas vicissitudes estendem-se aos sistemas de representação e recuperação da informação, pela ambiguidade e imprecisão.

A representação e a recuperação da informação são processos eminentimente individuais, cognitivos e culturais. Por isso, nenhum sistema baseado no fragmento do conhecimento por meio de relacionamentos entre nós informacionais e tipificações de relevância será capaz de refletir com exatidão as variações sintáticas e semânticas de cada indivíduo. Mesmo com a mineração de dados, algumas informações na Web permanecerão inacessíveis devido às barreiras sociais, limitações humanas, interesses corporativistas, políticas institucionais, entre outras. Daí, recomendam-se mais pesquisas para aprofundar o tema.

Algumas sugestões para investigações futuras recaem sobre o diálogo entre a Mineração de Dados e a Ciência da Informação, de modo a enquadrar os respectivos conteúdos temáticos e as respectivas terminologias. De igual modo, recaem sobre outras opções de tratamento, representação, acesso e recuperação de informações da Web Social, que estejam dissociados da estrutura complexa da mineração de dados.

Referências

ANGLO-AMERICAN CATALOGUING RULES (AACR). Disponível em: <http://www.aacr2.org/>. Acesso em: 26 fev. 2011.

BAPTISTA, S. G.; CUNHA, M. B. da. Estudo de usuários: visão global dos métodos de coleta de dados. **Perspectivas em Ciência da informação**, vol.12, n.2, 2007, p. 168-184. Disponível em: <http://www.scielo.br/pdf/pci/v12n2/v12n2a11.pdf>. Acesso em: 28 mar. 2011.

BAEZA-YATES, R.; RIBEIRO NETO, B. **Modern information retrieval**. Nova Iorque: The ACM Press, 1999.

BARRETO, A. A. Uma quase história da ciência da informação. **Datagramazero** - Revista de Ciência da Informação, Rio de Janeiro, v. 9, n. 2, Abr. 2008. Disponível em: <http://www.dgz.org.br/abr08/Art_01.htm>. Acesso em: 17 abr. 2013

BERNERS-LEE, T. **The World Wide Web: past, present and future**. Disponível em: <http://www.w3.org/People/Berners-Lee/1996/ppf.html>. Acesso em: 2 set. 2014.

BOURDIEU, P. **Os usos sociais da ciência: por uma sociologia clínica do campo científico**. São Paulo: Editora UNESP, 2004.

BORKO, H. Information science: what is it? **American Documentation**, v.19, n.1, p.3-5, jan. 1968.

BROADBAND COMMISSION FOR DIGITAL DEVELOPMENT – The state of broadband 2014: Broadband for all – a report by the Broadband commission, 2014. Disponível em: <http://www.broadbandcommission.org/Documents/reports/bb-annualreport2014.pdf>. Acesso em: 20 set. 2014.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science** (JASIS), v.45, n.5, p.351-360, 1991.

BUSH, V. As we may think. **Atlantic Monthly**, v.176, 1, 1945. Disponível em: <http://www.theatlantic.com/magazine/archive/1969/12/as-we-may-think/3881/1/> Acesso em: 03 mar. 2010.

BUSINESS DICTIONARY. Disponível em: <http://www.businessdictionary.com/definition/infoglut.html>. Acesso em: 12 set. 2009.

CAPURRO, R.; HJORLAND, B. O conceito de informação. Tradução de CARDOSO, A. M. P.; FERREIRA, M. G. A.; AZEVEDO, M. A. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 12, n. 1, Jan./Abr. 2007. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/54/47> Acesso em: 13 jun. 2013

CHU, H. **Information representation and retrieval in the digital age**. 3ª Tiragem, New Jersey: Asist&T, 2007.

COSTA, L. F. da; SILVA, A. C. P. da; RAMALHO, F. A. (Re)visitando os estudos de usuário: entre a tradição e o alternativo. **Datagramazero – Revista de Ciência da Informação**, v. 10, n. 4, p.1-12, 2009. Disponível em: <http://www.dgz.org.br/ago09/Art_03.htm> Acesso em: 27 mar. 2011.

DEBORD, G. **A sociedade do espetáculo**. Rio de Janeiro: Contraponto, 1997.

DYSON, G. M. A new notation and enumeration system for organic compounds. **Journal of chemical education**, Out. 1950. Disponível em: <http://pubs.acs.org/doi/abs/10.1021/ed027p581.2>. Acesso em: 16 nov. 2010.

DUBLIN CORE METADATA INITIATIVE. Disponível em: < http://dublincore.org/>. Acesso em: 26 fev. 2011.

EAGLETON, T. **As Ilusões do pós-modernismo**. Tradução de BARBOSA, E: Oxford: Blackwell Publishers, 1996.

ELMASRI, R.; NAVATHE, S. B. **Sistema de banco de dados**, 4. ed., São Paulo: Pearson Addison Wesley, 2005.

EMIRBAYER, M.; GOODWIN, J. Network analysis, culture, and the problem of Agency. **The American Journal of Sociology**, Chicago, v.99, n.6, Mai. 1994. Disponível em: <http://depts.washington.edu/methods/readings/emirbayer.pdf >. Acesso em: 03 out. 2014.

ENGELBART, D. C. Improving our ability to improve: a call for investment in a new future. **Simpósio Co-evolution da IBM**, Set. 2003. Disponível em: <http://www.almaden.ibm.com/coevolution/pdf/engelbart_paper.pdf>. Acesso em: 04 mai. 2013.

FADEL, B. et al. Gestão, mediação e uso da informação. In: VALENTIM, M. (Org.) **Gestão, mediação e uso da informação**. São Paulo: Cultura Acadêmica, 2010. p.13-31.

FEYNMAN, R. P. **O significado de tudo**. Lisboa: Gradiva, 2001, p. 11-37.

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003. 147f. Tese (Doutorado em Ciências da Comunicação) – Universidade de São Paulo, São Paulo, 2003.

FIGUEIREDO, N. M. de. **Estudos de uso e usuários da informação**, Brasília: IBICIT, 1994.

FOULONNEAU, M.; RILEY, J. **Metadata for Digital Resources.** Implementation, Systems Design and Interoperability, Oxford: Chandos Publishing, 2008.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 4 ed. São Paulo: Atlas, 1994.

GILL, T. Metadata and the Web. In: **Introduction to Metadata** 2. ed., Los Angeles: Getty Research Institute, 2008. p. 20-37. Disponível em:

<http://www.getty.edu/research/publications/electronic_publications/intrometadata/pdf.html>. Acesso em: 25 fev. 2011.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GROMOV, G. **History of Internet and World Wide Web - Roads and Crossroads of the Internet History**. Disponível em: <http://history-of-internet.com/>. Acesso em: 25 fev. 2014.

HARPRING, P. **Introduction to controlled vocabularies:** terminology for art, architecture, and other cultural works, 1.ed., Los Angeles: Getty Research Institute, 2010. Disponível em: <http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/pdf.html>. Acesso em: 25 fev. 2011.

HARVEY, D. **Condição pós-moderna: uma pesquisa sobre as origens da mudança cultural**. Tradução de SOBRAL, A. U; GONÇALVES, M. S.: São Paulo: Edições Loyola, 2004.

JAMESON, F. **Pós-modernismo: a lógica cultural do capitalismo tardio**. Tradução de CEVASCO, M. E. São Paulo: Editora Ática, 2004.

JUSTIN'S LINKS. Disponível em: <http://www.links.net/>. Acesso em: 25 ago. 2014.

LAKATOS, E. M.; MARCONI, A. **Metodologia do trabalho científico**: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos. 6. ed. São Paulo: Atlas, 2001.

LE COADIC, Y. F. **A Ciência da Informaçã**o. Brasília, DF: Briquet de Lemos/Livros, 2004.

LEMIEUX, V.; OUIMET, M. **Análise estrutural das redes sociais**. Tradução de PEREIRA, S. Lisboa: Instituto Piaget, 2004.

LÉVY, P. **Cibercultura**. 1.ed., São Paulo: Ed. 34, 1999.

LIBRARY OF CONGRESS. **What is a MARC record, and why is it important?**, 2009. Disponível em: <http://www.loc.gov/marc/umb/um01to06.html>. Acesso em: 10 mai. 2015.

LYOTARD, J. F. **O pós-moderno**. 3.ed. Tradução de BARBOSA, R. C. Rio de Janeiro: José Olympio J.O. Editora, 1988.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 2, Ago. 2002 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200007&lng=en&nrm=iso>. Acesso em: 12 nov. 2010.

MEDIABISTRO. Disponível em: <http://www.mediabistro.com/Social-Media-profile.html>. Acesso em: 14 nov. 2014.

MELO JÚNIOR, C. S. de. **Web 2.0 e Mashups: reinventando a internet**. Rio de Janeiro: Brasport, 2007.

MORIN, E. A comunicação pelo meio (teoria complexa da comunicação). In: MARTINS, F. M.; SILVA, J. M. da (Orgs.). **A genealogia do virtual**: comunicação, cultura e tecnologias do imaginário. Porto Alegre: Sulina, 2004, p. 11-19.

MORIN, E. Da necessidade de um pensamento complexo. In: MARTINS, F. M.; SILVA, J. M. da (Orgs.). **Para navegar no século XXI**: Tecnologias do imaginário e cibercultura. Porto Alegre: Sulina, 2003.

MORVILLE, P. **Ambient findability**. Sebastopol: O'Really, 2005.

NATIONAL INFORMATION STANDARDS ORGANIZATION – NISO: **understanding metadata**, 2004. Disponível em: <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>. Acesso em 10 Abr. 2014.

NELSON, T. Ted Nelson's Computer Paradigm, expressed as One-Liners. Disponível em: <http://xanadu.com.au/ted/TN/WRITINGS/TCOMPARADIGM/tedCompOneLiners.html>. Acesso em: 23 ago. 2014.

PAGE, L. et al. The PageRank citation ranking: bringing order to the Web, 1998. Disponível em: <http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf>. Acesso em 18 abr. 2015.

PYTHON. Disponível em: <http://www.python.org/>. Acesso em: 12 ago. 2013.

POPOVIČ, M.; WILLETT, P. The effectiveness of stemming for natural-language access to Slovene textual data. **Journal of the American Society for Information Science**, vol. 43, n.5, 1992, p.384-390. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-4571%28199206%2943:5%3C384::AID-ASI6%3E3.0.CO;2-L/abstract>. Acesso em: 20 nov. 2014.

QUÉAU, P. O tempo do virtual. In: PARENTE, A. (org.) **Imagem-máquina:** a era das tecnologias do virtual. Rio de Janeiro: 34 Letras, 1999.
RAYWARD, W. B. Visions of Xanadu: Paul Otlet (1868-1944) and Hypertext. **Journal of the American Society for Information Science**, vol. 45, n.4, 1994, p.235-250. Disponível em: <http://people.lis.illinois.edu/~wrayward/Visions%20of%20Xanadu_JASIS.pdf>. Acesso em: 15 set. 2014.

ROCHA, R. P. Metadados, Web Semântica, Categorização Automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web. **Em Questão**, v. 10, n. 1, Porto Alegre, 2004 (p. 109-121). Disponível em:

<http://www6.ufrgs.br/emquestao/pdf_2004_v10_n1/EmQuestaoV10_N1_2004_art07.pdf>. Acesso em: 21 Ago. 2009.

RUSSEL, M. A. **Mineração de dados da web social**. Tradução de ZANOLLI, R. São Paulo: Novatec Editora, 2011.

SANTOS, J. F. dos. **O que é pós-moderno**. São Paulo: Brasiliense, 2004.

SARACEVIC, T. Information science: origin, evolution and relations. **In**: VAKKARI, P., CRONIN, B. (ed.). Conceptions of library and information science: historical, empirical and theoretical perspectives. London: Taylor Graham, 1992, p. 5 – 27.

SARACEVIC, T. Ciência da Informação: origem, evolução e relações.Tradução de CARDOSO, A. M. P. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 1, n. 1, p. 41-62, Jan./Jun. 1996.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados**, 5.ed., Rio de Janeiro: Elsevier, 2006.

SIQUEIRA, M. A. **XML na Ciência da Informação:** uma análise do MARC 21. 2003. 133f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2003.

SMIRAGLIA, R. P. et al. **Metada: A cataloger´s primer**, v.40, Binghamton: The Haworth Information Press, 2005a.

SMIRAGLIA, R. P. **Instantiation: Toward a Theory**, 2005b. Disponível em: <http://www.cais-acsi.ca/proceedings/2005/smiraglia_2005.pdf>. Acesso em: 22 Ago. 2009.

STATISTA INC. Disponível em <http://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/>. Acesso em: 1 mai. 2015.

STENBERG, D. **History of IRC (Internet Relay Chat)**. Disponível em: <http://daniel.haxx.se/irchistory.html>. Acesso em 9 set. 2014.

STEWART, D.L. **Building Enterprise Taxonomies**. Mokita Press, 2008.

STOKES, D. E. **O quadrante de Pasteur: a ciência básica e a inovação tecnológica**. Tradução de MAIORINO, J. E. Campinas, SP: Editora da UNICAMP, 2005.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao DATAMINING Mineração de dados**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2009.

TEXT RETRIEVAL CONFERENCE (TREC). Disponível em: <http://trec.nist.gov/>. Acesso em: 28 out. 2010.

TERUEL, A. G. **Los estúdios de necessidades y usos de la información:** fundamentos y perspectivas actuales, Valência: Ediciones Trea, S. L., 2005.

THE OAUTH 2.0 AUTHORIZATION FRAMEWORK. Disponível em: <http://tools.ietf.org/html/rfc6749>. Acesso em: 22 out. 2014

WERSIG, G. Information Science: the study of postmodern knowledge usage. **Information Processing and Management**: an International Journal, Tarrytown-Nova Iorque, v.29, n.2, p.229-239, Mar./Abr.1993.

WILSON, T. D. On user studies and information needs. **Journal of Librarianship**, 37(1), p.3-15, 1981. Disponível em: <http://informationr.net/tdw/publ/papers/1981infoneeds.html>. Acesso em: 2 mar. 2011.

WITTEN, I., H.; FRANK, E.; HALL, M., A. **Data mining: practical machine learning tools and techniques**. Burlington: Elsevier, 2011.

8 PRINCIPLES of Open Government Data. Sebastopol: Open Government Data, 2007. Disponível em: <http://www.opengovdata.org/home/8principles>. Acesso em 11 jul. 2011.