Universidade Estadual de Londrina

JULIO CESAR LIVIERO DELLA FLORA

MOTOR DE BUSCA ABERTO COMO ESTRATÉGIA DE INDEXAÇÃO E MINERAÇÃO DE DADOS

Londrina
2015

JULIO CESAR LIVIERO DELLA FLORA

MOTOR DE BUSCA ABERTO COMO ESTRATÉGIA DE INDEXAÇÃO E MINERAÇÃO DE DADOS

Dissertação apresentada ao Programa de Pós graduação em Ciência da Informação da Universidade Estadual de Londrina (PPGCI-UEL), como requisito parcial para obtenção do título de Mestre.

Orientador: Prof. Dr. Benjamin Luiz Franklin

Londrina
2015

Dados Internacionais de Catalogação-na-Publicação (CIP)

D357m Della Flora, Julio Cesar Liviero
Moto de busca aberto como estratégia de indexação e mineração de dados / Julio Cesar Liviero Della Flora. – Londrina, 2015.
80 f.

Orientador: Benjamin Luiz Franklin.
Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Centro de Educação Comunicação e Artes, Programa de Pós-Graduação em Ciência da Informação, 2015.
Inclui bibliografia.

1. Mineração de dados (Computação) – Teses. 2. Indexação – Teses. 3. Internet – Teses. 4. Ciência da Informação – Teses. I. Franklin, Benjamin Luiz. II. Universidade Estadual de Londrina. Centro de Educação Comunicação e Artes. Programa de Pós-Graduação em Ciência da Informação. III. Título.

CDU 025.4

JULIO CESAR LIVIERO DELLA FLORA

MOTOR DE BUSCA ABERTO COMO ESTRATÉGIA DE INDEXAÇÃO E MINERAÇÃO DE DADOS

Dissertação apresentada ao Programa de Pós graduação em Ciência da Informação da Universidade Estadual de Londrina (PPGCI-UEL), como requisito parcial para obtenção do título de Mestre.

BANCA EXAMINADORA

Orientador: Prof. Dr. Benjamin Luiz Franklin
Universidade Estadual de Londrina - UEL

Prof. Dr. Silvana Drumond Monteiro
Universidade Estadual de Londrina - UEL

Prof. Dr. Rodrigo Duarte Seabra
Universidade Federal de Itajubá - UNIFEI

Londrina, 3 de outubro de 2015.

Dedico este trabalho à minha família, pelo apoio e amor incondicional.

AGRADECIMENTOS

Agradeço a Deus, razão das minhas conquistas!

Ao meu orientador, Prof. Dr. Benjamin Luiz Franklin, pelo constante e valioso acompanhamento acadêmico neste trabalho.

Ao Prof. Dr. Rodrigo Duarte Seabra, pela contribuição acadêmica e profissional dispensada durante todo o período acadêmico.

À Prof. Dr. Silvana Drumond Monteiro, pelo profissionalismo e incentivo constantes.

Aos colegas, pelas inestimáveis trocas de experiências.

Gostaria de agradecer, também, às pessoas que contribuíram, direta e indiretamente, para que este trabalho se concretizasse.

Muito obrigado!

A ciência, como um todo, não é nada mais do que um refinamento do pensar diário.

Albert Einsten

DELLA FLORA, Julio Cesar Liviero. **Motor de busca aberto como estratégia de indexação e mineração de dados**. 2015. 80 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Londrina, 2015.

RESUMO

A importância da informação para o desenvolvimento da Ciência é inquestionável. Por isso, a Ciência da Informação é relevante para a comunidade científica ao desenvolver formas de organização e recuperação da informação possibilitando que o conhecimento produzido possa ser divulgado, acessado e utilizado por outros pesquisadores. Nesse sentido, ocorre a necessidade de que a organização da informação possa ser realizada livremente e busque a melhor resposta independente de patrocínio e censura. O objetivo desse trabalho é analisar o motor de busca, baseado em *software* livre *YaCy/Solr* como estratégia de indexação e mineração de dados. Foi utilizada pesquisa qualitativa, exploratória, bibliográfica e aplicada com a técnica de análise temática. A consequência desse estudo mostra que o *software* proporcionou ampla gama de resultados no âmbito da mineração de dados, evidenciado pelo experimento apresentado. Porém, é importante salientar que os mesmos não configuram a real contribuição do *YaCy*/*Solr*, mas sim, a capacidade que o *software* dispõe para realizar tais experimentos, sejam eles concentrados em qualquer gama de documentos ou metadados. Conclui-se que a partir do *YaCy/Solr* obteve-se acesso à estratégia de indexação como um diferencial para produzir informações não fornecidas pelos motores de busca proprietários. Futuras investigações poderão incluir versões mais precisas do experimento realizado, assim como a utilização de novos metadados e *filter queries.* Apesar das limitações apresentadas na exposição do *software*, considera-se que o trabalho proporcionou uma visão diferenciada e pouco ortodoxa acerca dos mecanismos de busca e da mineração de dados, visão esta que poderá constituir um ponto de partida para futuras pesquisas na área.

Palavras-chave: *Internet.* Motores de Busca. Mineração de Dados. Estratégias de Indexação.

DELLA FLORA, Júlio Cesar Liviero. **Open search engine like indexing strategy and data mining**. 2015. 80 p. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Londrina, 2015.

ABSTRACT

The importance of information for the development of science is unquestionable. Therefore, the Information Science is relevant to the scientific community to develop forms of organization and information retrieval enabling the knowledge produced can be disseminated, accessed and used by other researchers. In this sense, there is the need for the organization of information can take place freely and seek the best response independent of patronage and censorship. The aim of this study is to analyze the search engine, based on free software YaCy/Solr as indexing strategy and data mining. It used qualitative, exploratory, literature and applied to thematic analysis. The result of this study shows that the software provided wide range of results within the data mining, evidenced by the presented experiment. However, it is important to note that they do not constitute the actual contribution of YaCy/Solr, but rather the ability of the software has to carry out such experiments, whether concentrated in any range of documents, or metadata. We conclude that from the YaCy/Solr was obtained access indexing strategy as a differential to produce information not provided by the owners search engines. Future research may include more accurate versions of the experiment conducted, and the use of new metadata and filter queries. Despite the limitations presented in the software exhibition, it is considered that the work provided a different and unorthodox view about the search engines and data mining, a view that could be a starting point for future research in the area.

Keywords: Internet. Search Engines. Data Mining. Indexing Strategies.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ARPA	*Advanced Research Projects Agency*
CI	Ciência da Informação
DM	*Data Mining*
DOAJ	*Directory of Open Acess Journals*
GPL	*General Public Licence*
HTML	*Hyper Text Markup Language*
INFOBILA	*Información Y Bibliotecología Latino Americana*
IP	Protocolo de Intra-Rede
IPTO	*Information Processing Techniques*
ISO	*International Association for Standardization*
LISA	*Library an Information Science Abstracts*
P2P	*Peer To Peer*
PKP	*Public Knowledge Projects Metadata Archive*
RI	Recuperação de Informação
SEER	Sistema Eletrônico de Editoração de Revistas
SRI	Sistema de Recuperação de Informação
TCP	Protocolo de Controle de Transmissão
UEL	Universidade Estadual de Londrina
URI	*Uniform Resource Identifer*
URL	*Uniform Resource Locator*
WWW	*World Wide Web*

SUMÁRIO

1 INTRODUÇÃO

A importância da informação para o desenvolvimento da ciência é inquestionável, pois, conforme afirma Le Coadic (1996, p. 27), “[...] a informação é o sangue da ciência. Sem informação, a ciência não pode se desenvolver nem viver” visto que o conhecimento resultante das atividades científicas e técnicas deve ser registrado e organizado para impulsionar novas pesquisas e conhecimentos. “Ela só existe [...] no contato efetivo entre uma mensagem e o usuário.” (ARAÚJO, 1995).

Segundo Baptista *et al.* (2007, p. 2) “[...] a comunicação do conhecimento científico tem sido uma das questões mais estudadas pela Ciência da Informação (CI), o que demonstra a importância desse tema para a área.” Isso porque “a busca do conhecimento científico se constitui como um grande e permanente empreendimento, um fator determinante para a ampliação da capacidade de assimilação e desenvolvimento de novas tecnologias.” (AMBINDER; MARCONDES, 2011, p. 1). Faz parte do escopo da CI o estudo de processos de indexação e Recuperação da Informação (RI). (SOUZA; ALVARENGA, 2004, p. 139).

Pode-se entender a importância que a CI tem para a comunidade científica ao desenvolver formas de organização e RI científica e possibilitar que o conhecimento produzido possa ser divulgado, acessado e utilizado por outros pesquisadores.

Em complemento, Kuramoto (2006, p. 91) evidencia de forma clara que “[...] a informação científica é o insumo básico para o desenvolvimento científico e tecnológico de um país. Esse tipo de informação, resultado das pesquisas científicas, é divulgado à comunidade por meio de revistas.”

Diante deste cenário, é possível afirmar que os motores de buscas são tidos, atualmente, como ferramentas necessárias para o acesso à informação, uma vez que é por meio deles que o conhecimento se efetiva, no sentido de se atualizarem novos conhecimentos.

A grande maioria das instituições de ensino no país possui um portal *Web* cujo propósito é facilitar a troca de informações entre docentes, discentes, colaboradores e comunidade externa.

Em face da natureza dinâmica da Universidade, periódicos, artigos e documentos em geral são acrescidos, diariamente, à base de dados do *Website* os

quais são inseridos e categorizados pelos diversos departamentos existentes. Este fluxo de conteúdo considerado por demasiado período de tempo tornará a base de dados extensa, a ponto de impossibilitar a busca manual por informações contidas no site.

Colimando à facilitação ao acesso à informação é imprescindível a adoção de um mecanismo de busca integrado ao portal, pois, o mais comum é a utilização de uma barra de pesquisa de uma versão personalizada do buscador oferecido pelo *Google*[1].

Desse modo, o portal acadêmico tem suas pesquisas condicionadas a restrições como o devido acesso à massa de dados, a forma como é ranqueado e, consequentemente, impossibilitando que os usuários dessa massa de dados possam se beneficiar de um leque mais amplo de informações, o que lhes renderia economia de tempo e esforço.

É importante argumentar que as ferramentas de busca, em sua maioria constituídas por serviços oferecidos por grandes companhias como o *Google* e *Microsoft*[2], cujos sistemas são essencialmente de código-fonte[3] sigiloso, resultam em uma tecnologia de indexação e classificação desconhecidas pelos usuários deste serviço.

Ocorre que ao utilizar buscadores comerciais, não é possível mensurar quanta informação está sendo censurada, bloqueada ou removida do resultado, pois apenas a entidade detentora do *Software* possui esse controle. Com isso, ao optar pela indexação de uma página *Web*, o desenvolvedor torna-se complacente às regras e termos de uso impostos pela empresa subsidiária do buscador. O descumprimento dessas regras resulta em punições severas, podendo ocorrer, inclusive, a sua não indexação em casos extremos. Assim, entende-se que buscas empreendidas mediante mecanismos comerciais são, primordialmente, facciosas, seja por políticas organizacionais, privilégios a patrocinadores ou determinações judiciais.

Fragoso (2007, p. 15) menciona que “ano após ano, *Google*, Yahoo! e MSN figuram entre os dez sites mais visitados em todas as nações [...] mais de 80% das buscas se concentram sobre essas mesmas empresas.”

---

1 <https://www.google.com.br/>
2 <https://www.microsoft.com/pt-br/>
3 Instruções de programação implícitas ao sistema.

A empresa *Google*, sozinha, “[...] detém uma grande parcela do mercado de buscas (cerca de 50% nos Estados Unidos e 90% no Brasil), o que a torna uma empresa-paradigma que dita normas de fato para o setor.” (TIGRE; NORONHA, 2013, p. 123).

Diante deste contexto, é possível argumentar, então, que o usuário do serviço pode ser induzido de forma a consumir informações que não sejam, necessariamente, as mais adequadas em sua busca, pois a indexação e a organização de conteúdo atende a benefícios comerciais, categorizando os usuários de seu serviço não como clientes, mas como produto.

A problemática, ora vislumbrada, está diretamente relacionada à necessidade de que a organização da informação, de modo geral, possa ser realizada livremente e busque a melhor resposta independente de patrocínio, censura, interesses políticos e/ou comerciais.

É relevante, ainda, ressaltar que ao utilizar os mecanismos de buscas comerciais, o usuário se abstém da estratégia de indexação em *corpus*, porém, com um mecanismo de código aberto, ele tem o poder de minerar os dados, pois estão abertos para o usuário, que pode, inclusive, mudar a estratégia de indexação.

Desta forma, optar por um *Software* de código aberto mostra-se uma excelente alternativa na obtenção de resultados livres, por exemplo, por meio do mecanismo *YaCy*/*Solr*, tendo em vista que sua execução é facilitada por ser um sistema aberto, gratuito e não oferecer censura ao índice compartilhado.

Esse *Software* consiste em um motor de busca no qual qualquer indivíduo dotado de capacidades básicas em operação de microcomputadores está apto a construir e utilizar sua própria página de busca, podendo indexar conteúdo em sua intranet, ou mesmo, colaborar para a busca pública na rede mundial de computadores, de forma que a busca pela informação possa ser realizada local e globalmente.

Vale destacar que a discussão acerca da mudança entre sociedades nas máquinas contemporâneas afeta os motores de busca. O *Software* em questão é construído para operar em redes descentralizadas, um reflexo do câmbio de paradigmas na contemporaneidade.

Assim, esta pesquisa se justifica ao expor a facciosidade dos resultados nos motores de busca proprietários em contraste com os motores de

busca abertos. Justifica-se também ao empoderar o usuário, introduzindo uma solução que objetiva a maior abrangência de conteúdo pesquisado a partir do próprio acervo de documentos digitalizados.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Analisar o motor de busca baseado em *software* livre *YaCy*/*Solr* como estratégia de indexação e mineração de dados.

1.1.2 Objetivos Específicos

a) apontar as características de um motor de busca de código aberto em contraste com os mecanismos de busca comerciais;
b) descrever a arquitetura de um motor de busca de código aberto, tendo o *YaCy*/*Solr* como sistema referência;
c) indicar as propriedades de mineração de dados do motor de busca aberto *YaCy*/*Solr*;
d) produzir uma aplicação que evidencie uma estratégia de indexação contida no motor de busca *YaCy*/*Solr*.

1.2 ORGANIZAÇÃO DO TRABALHO

O texto foi estruturado em cinco capítulos, da forma a seguir:

Capítulo 1 – apresenta a introdução do trabalho abrangendo o tema de modo geral, sua justificativa e os respectivos objetivos.

Capítulo 2 – destina-se à apresentação da revisão bibliográfica no que se refere à teoria científica publicada sobre a Internet, informação, Big Data e mineração de dados, indexação e buscadores, para contribuir no alcance dos objetivos propostos.

Capítulo 3 – busca caracterizar a pesquisa, conforme a metodologia utilizada para a abordagem do problema, quanto à natureza dos objetivos, ao método de investigação e aos procedimentos de coleta e análise de dados e ao contexto da pesquisa.

Capítulo 4 – visa analisar os dados levantados a partir da prática aplicada por meio de experimento, como resposta aos objetivos propostos.

Capítulo 5 – engloba as considerações finais desta pesquisa e as recomendações para trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

As seções a seguir levaram em consideração os objetivos propostos, em consonância com os descritores utilizados no levantamento teórico para embasar a fundamentação apresentada.

2.1 HISTÓRICO DA *INTERNET*

É oportuno iniciar a revisão bibliográfica a partir da abordagem da *Internet* que consiste no "[...] conjunto de inúmeras redes de computadores conectadas entre si que permite a comunicação, partilha de informações, programas e equipamento entre seus usuários." (BRANSKI, 2004, p. 71). Ela está inserida em um contexto, segundo Ferneda (2003, p. 91), que "[...] pode ser contado a partir da Guerra Fria, período histórico que teve seu início no pós-guerra."

Em 1969, uma rede de computadores conhecida como *Arpanet* foi projetada e implementada pela *Advanced Research Projects Agency* (ARPA), agência instituída em 1958 pelo Departamento de Defesa Americano cujo principal objetivo era desenvolver superioridade militar e tecnológica em relação à União Soviética (CASTELLS, 2003).

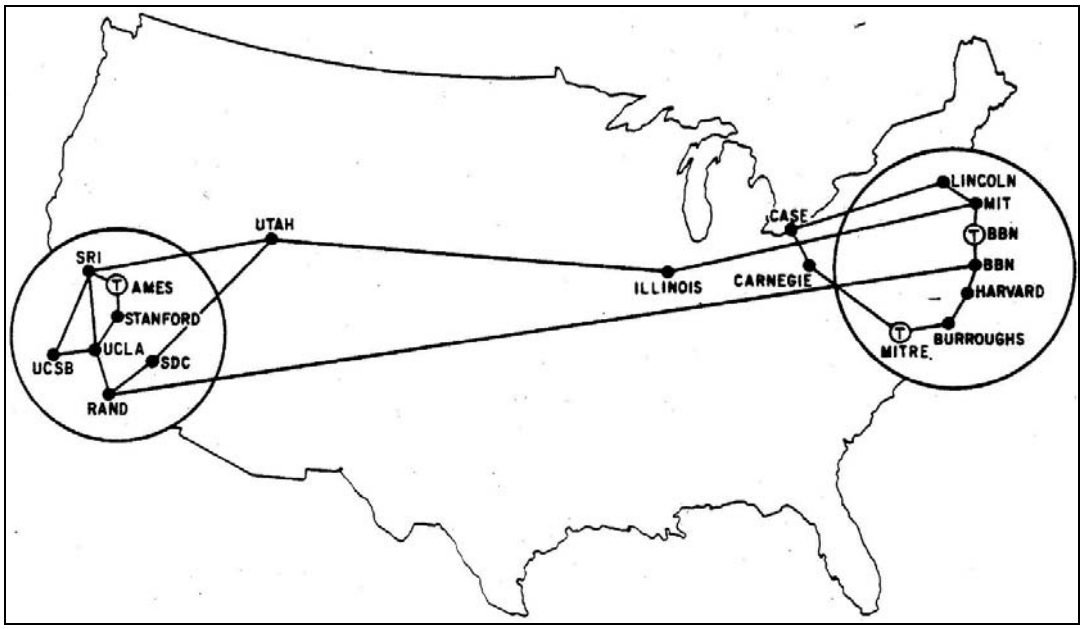
Conforme Castells (2003), a *Arpanet* surgiu a partir de um pequeno departamento da ARPA, o *Information Processing Techniques Office* (IPTO), visando estimular pesquisas em computação interativa. A iniciativa de montagem dessa rede era justificada como meio para assegurar a intercomunicação entre os vários centros de computadores e grupos de pesquisas que trabalhavam para a agência.

Ainda segundo o autor, o IPTO, ao implantar a rede interativa de computadores, utilizou a tecnologia de comutação por pacotes desenvolvida por Paul Baran (*Rand Corporation*) e Donald Davies. Essa tecnologia revolucionária de comunicação descentralizada foi proposta ao Departamento de Defesa dos Estados Unidos como um sistema militar capaz de sobreviver a ataques nucleares, apesar desse nunca ter sido o objetivo para o desenvolvimento da *Arpanet*.

2.1.1 Estratégia de Descentralização das Comunicações

Os primeiros nós da rede estavam localizados em Los Angeles, na Universidade da Califórnia, em Santa Bárbara, no *Stanford Research Institute* (SRI) e na Universidade de Utah. Em 1971, em sua primeira apresentação bem sucedida, a Arpanet possuía 15 nós (Figura 1), sendo grande parte em centros universitários de pesquisa.

Figura 1 – Arpanet em 1971



Fonte: http://som.csudh.edu/fac/lpress/history/arpamaps/f8sep1971.jpg

Como próxima etapa, a necessidade de conexão entre a *Arpanet* e outras redes de computadores foi alcançada por meio de protocolos de comunicação padronizados em 1973, com o projeto do Protocolo de Controle de Transmissão (TCP)[4] (CASTELLS, 2003).

Segundo Pinho (2003, *apud* VERGILI, 2012, p. 42), o TCP/IP é um "[...] conjunto de protocolos da *Internet*, que define como se processam as comunicações entre os vários computadores. É a linguagem universal da *Internet* e pode ser implementada virtualmente em qualquer tipo de computador, pois é independente do *hardware*." Em síntese, pode-se dizer que o protocolo TCP usa o IP para enviar pacotes por meio da *Internet* (VERGÍLIO, 2012).

[4] Em 1978, o TCP foi dividido em duas partes acrescentando um Protocolo de Intra-Rede (IP) que culminou no protocolo TCP/IP, padronização em que a *Internet* opera atualmente.

Voltando ao foco deste tópico, resta acrescentar que, preocupado com possíveis incidentes de segurança, o Departamento de Defesa Norte Americano optou pela criação de uma rede independente para usos militares em 1983, batizada como *MILNET*. A *Arpanet,* então, exclusivamente dedicada às pesquisas tornou-se *ARPA-INTERNET* (VERGÍLIO, 2012).

Tecnologicamente obsoleta em 1990, a *Arpanet* foi retirada de operação e libertada de seu ambiente militar. O Departamento de Defesa havia decidido comercializar a tecnologia da *Internet* financiando fabricantes de computadores para que o protocolo TCP/IP fosse incluído em seus projetos. Por meio dessa iniciativa, nessa mesma década, grande parte dos microcomputadores de uso pessoal nos EUA possuía a capacidade de se comunicar via rede, firmando os alicerces para a operação comercial da *Internet* (CASTELLS, 2003).

Após a retirada dos militares, a *Internet,* que foi concebida como máquina de combate, teve seu uso redirecionado, atuando como máquina de cooperação social por intermédio dos grupos de discussão na *Usenet.* Isso culminou em um mecanismo (meio) de produção de afetos, relações e cooperação mútua, distanciando seu propósito original em que a importância se pautava apenas no transporte de informações científicas, financeiras e militares (MALINI; ANTOUN, 2013).

2.2 *SOFTWARE* LIVRE

É importante esclarecer em que contexto se deu o surgimento dos *softwares* livres. De acordo com Evangelista (2009, p. 80) “[...] a origem do movimento *software* livre é atribuída aos Estados Unidos da América (EUA).”

Quadro 1 – Linha do tempo do *Software* Livre

ANO	EVENTO
1950s e 1960s	Códigos fontes são distribuídos sem restrição entre empresas (como IBM), centro de pesquisas (como os laboratórios Bells e MIT) e universidades.
1969	Ken Thompson desenvolve a primeira versão do UNIX. O código-fonte desse sistema é distribuído livremente.
1978	Donald Knuth (Standford) publicou o TEX como software livre.
1979	Após a AT&T's anunciar a comercialização do UNIX, a Universidade de Berkeley dá início ao desenvolvimento da sua própria versão do UNIX: o BSD (*Berkeley Software Distribution*). Eric Allmann, um estudante da mesma Universidade de Berkeley, desenvolveu um programa que transfere mensagens entre computadores por meio da ARPANET, que posteriormente evolui para o *Sendmail*.
1983	Richard Stallman publica o Manifesto GNU buscando a difusão do software livre e cria a *Free Software Foundation*.
1987	O desenvolvedor Andrew Tanenbaum lança o Minix - a versão do UNIX para PCs, Mac, Amiga e Atarai ST, disponibilizando completamente o código-fonte.
1991	Linus Torvalds publica a versão 0.2 de uma variação do kernel do Minix para o projeto GNU, que ele chamou de "Linux".
1993	É lançado o FreeBSD 1.0, baseado no BSD Unix. Ian Murdock cria uma nova distribuição do GNU-Linux chamada de "Debian".
1994	Marc Ewing forma a empresa *Red Hat* Linux e cria uma distribuição própria para prestar serviços com este software livre.
1995	O Grupo de hackers denominado de "Apache" constrói um novo software (livre) para servidores Web que, atualmente, é o mais usado em todo o mundo.
1996	O *desktop* KDE é lançado para usuários do GNU-Linux por Matthias Ettrich, porém com alguns aplicativos proprietários.
1997	O Projeto GNOME é iniciado por Federico Mena e Miguel de Icaza como *desktop* livre oficial do Projeto GNU.
1999	O número de usuários GNU-Linux é estimado em 7.5 milhões de usuários.
2000	Novas empresas multinacionais de TI (como a Novel e Real) lançam versões de seus produtos que rodam no GNU-Linux.
2001	O número de usuários GNU-Linux é estimado em, pelo menos, 30 milhões em todo o mundo.
2007	Mais de 140.000 projetos de softwares livres estão registrados em *apenas um* dos maiores repositórios de código aberto do mundo - o site *SourceForge.net*.

Fonte: *Open Source Time Line* In: Hars e Ou (2002 *apud* AGUIAR, 2009, p. 13)

Até a década de 80 (Quadro 1), os tecnólogos computacionais almejavam, apenas, o desenvolvimento colaborativo de um sistema operacional que executasse em qualquer tipo de computador e pudesse ser conectado à *Internet* (AGUIAR, 2009, p. 10).

Neste contexto,

> O *UNIX* tornou-se um ambiente de *Software* para todo tipo de sistema, libertando assim, os programadores da necessidade de inventar linguagens específicas para cada máquina: o *Software* tornou-se portátil, o que permitiu a comunicação entre computadores cumulativa. (CASTELLS, 2003, p. 39).

O movimento do *software* livre surgiu em 1984, quando Richard Stallman[5] decide, em conjunto com um grupo de programadores, desenvolver o sistema operacional batizado como *GNU*, após a empresa *AT&T* reivindicar direitos de propriedade sobre o *UNIX* (AGUIAR, 2009, p. 10).

Stallman escolheu o nome *GNU* por ser um acrônimo recursivo de: *GNU is Not Unix* (em português: *GNU* Não é *Unix*).

GNU é a base do sistema em que está quase a totalidade dos programas necessários para o funcionamento. O *LINUX* designa uma pequena parte desse sistema operacional, o chamado *Kernel* (cerne), mas de vital importância para o funcionamento do sistema, uma vez que realiza o gerenciamento dos dispositivos (mouse, teclado, monitor e outros) (EVANGELISTA, 2009, p. 84).

2.2.1 O Projeto *GNU*

O Projeto *GNU* teve como foco o desenvolvimento de um sistema operacional completo, compatível com o *Unix*, que fosse *software* livre: o sistema *GNU*. Sua grande arma é a *General Public Licence* (*GPL*).

A *GPL* foi criada para preservar as quatro liberdades básicas, de acordo com Campos*a* (2006):

1) a liberdade de executar o programa, para qualquer propósito (liberdade nº 0);
2) a liberdade de estudar como o programa funciona e adaptá-lo para as suas necessidades (liberdade nº 1). Acesso ao código-fonte é um pré-requisito para esta liberdade;

[5] Programador do Laboratório de Inteligência Artificial do MT nos EUA.

3) a liberdade de redistribuir cópias de modo que você possa ajudar ao seu próximo (liberdade nº 2);
4) a liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie (liberdade nº 3). Acesso ao código-fonte é um pré-requisito para esta liberdade.

Como se percebe, "Stallman protagonizou a abertura do código-fonte, plena liberdade de uso, aperfeiçoamento e distribuição dessa tecnologia [...]." (AGUIAR, 2009, p. 11).

A mobilização pelo *software* livre descrita no tópico a seguir pratica a ideia de que "[...] códigos bons não se desperdiçam, devem ser compartilhados, apoiados e melhorados por uma comunidade." (SILVEIRA, 2009, p. 235).

2.2.2 *Software* Livre na *Web*

O modelo de *software* livre é chamado de *open source software* ou *software* de código aberto (EVANGELISTA, 2009, p. 88). Esse mesmo autor explica que,

> O que define um *software* como livre ou proprietário não está dado em sua arquitetura, mas em sua forma de licenciamento[6], isto é, no modo como é regulamentado juridicamente, regulamentação que configura/autoriza determinadas relações na sociedade, e não outras (EVANGELISTA, 2009, p. 81).

Assim, *software* livre é o *software* que pode ser usado, copiado, estudado, modificado e redistribuído sem restrição e com a disponibilização do seu código-fonte (CAMPOS*a*, 2006). Entretanto, é importante ressaltar que,

> Ao contrário do que muitos pensam, Código Aberto não quer dizer simplesmente ter acesso ao código-fonte dos *softwares* (e não necessariamente acompanhado das (4 liberdades) do *software* livre). Para uma licença ou *software* ser considerado como Código Aberto pela *Open Source Initiative*, eles devem atender aos 10 critérios da Definição de Código Aberto [http://www.opensource.org/docs/definition.php], que

[6] Licença é o documento de valor jurídico no qual estão descritos os direitos e deveres dos usuários.

> incluem itens como Livre Redistribuição, Permissão de Trabalhos Derivados, Não Discriminação, Distribuição da Licença e outros. (CAMPOS*a*, 2006, p. 4).

O *software* livre não necessariamente precisa ser gratuito (CAMPOS*a*, 2006, p. 3). Stallman deixa claro que o *free* de *free software* (do termo original em inglês) não significa “grátis”, mas “livre” (EVANGELISTA, 2009, p. 107).

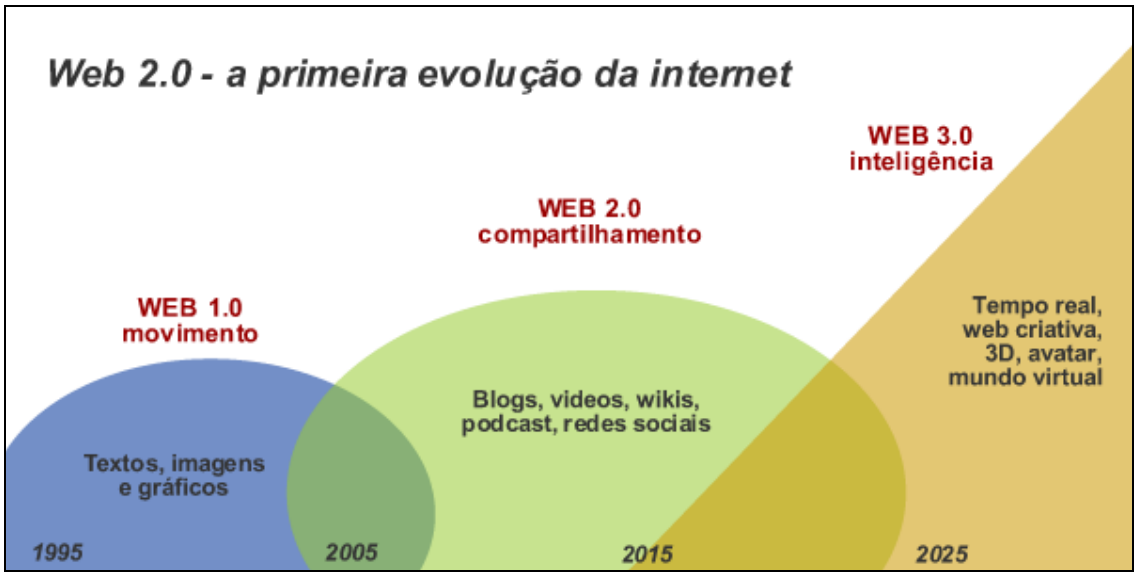
Essa forma de produção de *software* – cooperativa, descentralizada e “anárquica” – foi chamada, por Eric S. Raymond, de “método bazar”, como contraponto ao “método catedral”, “forma centralizada e controlada de se desenvolver *software*” e que “necessita de um arquiteto central” (GONÇALVES JR.; SILVA, 1999, p. 2-3 *apud* APGAUA, 2004, p. 223).

Alguns *softwares* livres notáveis são o *Linux*, o ambiente gráfico KDE, o compilador GCC, o servidor *Web Apache*, o *OpenOffice.org*, o navegador *Web Firefox*, o *YaCy/Solr*, entre muitos outros (CAMPOS*a*, 2006).

2.2.3 As três Fases da *Internet*

Visando uma melhor compreensão do contexto que envolve a *Internet*, principalmente quando da abordagem dos sub-tópicos que versam sobre as *Web’s*, é apresentada uma breve síntese das suas três fases (Figura 2).

Figura 2 – As três fases da *Internet*



Fonte: Carvalho (2012)

De acordo com Vergili (2012, p. 49), “[...] a primeira geração da *Internet* ficou conhecida como *Web* 1.0, cujo formato teve como característica a comunicação de apenas uma via, de um para muitos.”

Na *Web* 2.0 ou *Web* Social, criada em 2004, estão inseridas as redes sociais conectadas, sendo consideradas também, como produtoras de conteúdo, as pessoas com acesso à rede. Momento em que a comunicação passou a ser de muitos para muitos (VERGILI, 2012, p. 50).

Em 2001, surge a *Web* 3.0, também conhecida como *Web* Semântica, criada para fazer com que as informações pudessem ser lidas por pessoas e por máquinas (VERGILI, 2012, p. 52).

Segundo Bax (2012, p. 6), “essa tecnologia *Web* mais recente representa um grande avanço no tratamento e na busca por informações sobre entes do mundo real, pessoas, eventos, coisas etc. e organização de suas propriedades”. Tem o escopo de transformar a *Web* de documentos em uma *Web* de dados.

2.3 *WEB´S*

O advento da *Web* mudou o mundo de um modo que poucas pessoas conseguiriam prever (BAEZA-YATES; RIBEIRO NETO, 2011, p. 11).

Proposta por Tim Berners-Lee no início da década de 90, a *World Wide Web* (rede mundial de computadores) entremeada ao hipertexto é um espaço de deslocamento e comunicação[7].

Atualmente, a *Web* é o ambiente para o qual estão voltados os maiores esforços de desenvolvimento na área de RI devido à grande quantidade de informações nela disponível (AMBINDER; MARCONDES, 2011, p. 2). Ela é uma rede universal que nomeia recursos informacionais usando *Uniform Resource Identifer* (URI)[8].

[7] A comunicação entre cliente e servidor na *Web* se dá pela utilização do protocolo HTTP.
[8] Cadeia de caracteres compacta usada para identificar ou denominar um recurso na *Internet*.

Desse modo,

> A *Web* é um espaço virtual construído sobre a infra-estrutura física da *Internet*. São os protocolos que realmente enviam os *bits* que compõem os *hipertextos* que fluem da aplicação servidora para a aplicação cliente, o *browser* ou o navegador. (BAX, 2012, p. 8).

Assim, um documento da *Web* é composto por uma mistura de dados e metadados (SOUZA; ALVARENGA, 2004, p. 134).

Segundo Ferneda (2003, p. 92) “[...] a *Web* é a face hipertextual da *Internet*, considerada como a maior fonte de informação nas principais áreas do conhecimento.”

Seu surgimento contribuiu para o “crescente valor da informação e das técnicas relacionadas à pesquisa, coleta, armazenamento e difusão da informação.” (AMBINDER; MARCONDES, 2011, p. 1).

Dessa forma, a *Web* constitui o principal “lugar” da *Internet*, o centro para todas as possibilidades de interface (MONTEIRO; PICKLER, 2007).

2.3.1 *Web´s Visíveis*

Ao acessar a *Internet*, grande parte dos usuários de computadores opta por consultar uma unidade centralizadora de informação devido à massiva quantidade de páginas disponíveis na *Web*.

Os mecanismos centralizadores são denominados “buscadores” e sua função consiste em indexar o maior número possível de endereços *Uniform Resource Locator* (*URL*), que é o “[...] endereço de um arquivo acessível através da *Internet* [...] uma cadeia de caracteres formada por componentes padronizados, em uma ordem específica.” (FERNEDA, 2003, p. 93), como pode ser observado na Figura 3.

Figura 3 – Partes de uma *URL*

<http://www.gnu.org/software/wget/manual/wget.html>

Onde:

<http://	= Protocolo
/www.gnu.org/	= Computador (servidor, *host*)
/software/wget/manual/	= Caminho (diretório, pasta)
/wegt.html>	= arquivo

Fonte: Adaptado de Ferneda (2003, p. 93)

Em síntese, a *URL* descrita na Figura 3 identifica um arquivo, o qual, para ser acessado, é necessário utilizar o protocolo apontado e que está armazenado no computador “gnu”, cujo domínio “org” indica que é uma organização de caráter não comercial. Neste computador está o arquivo “wget”, cujo formato é “html” (FERNEDA, 2003). Quando indexado por um mecanismo de busca, o *Website* em questão pertence à porção visível ou indexável da *Web*.

Em seu trabalho, Vignoli (2014) discorre a respeito da *Web* Visível, no qual outros desdobramentos são possíveis tais como a *Web* Social, *Web* Semântica, *Web* Pragmática, entre outras. A autora ressalta a evolução para *Web* Social a partir da *Web* 1.0 ou estática, em que os sujeitos controlam sua própria inserção de dados. Em suma, uma *Web* centrada no usuário onde a qualidade do serviço é incrementada paralelamente à quantidade de utilizadores.

Já a *Web* Semântica, para Beernes-Lee (2001), tem como proposta “[...] permitir que aplicações combinem e processem dados e informações disponíveis na rede.”

As buscas realizadas pelos usuários seriam semelhantes, entretanto, a *Web* com características semânticas centralizaria dados de diversas fontes tornando os resultados mais precisos e inteligentes. Nesse novo modelo, a *Web* Semântica traria refinamento às buscas, discernindo palavras a partir do contexto. “Seu maior desafio é permitir que a *Web* passe de uma plataforma de divulgação para uma plataforma de informação.” (AMBINDER; MARCONDES, 2011, p. 7).

Nesse sentido, as *URIs* são usadas também para identificar coisas no mundo, e não apenas recursos de informações (BAX, 2012, p. 2-3).

Dessa forma, um processo novo e diferenciado da atual indexação realizada pelos robôs[9] de busca seria necessário, bem como o uso de metadados e ontologias[10] formando um complexo sistema para a organização do conhecimento. Destarte, a *Web* Semântica pertence também à porção indexável, e, por consequência, visível da *Web* (VIGNOLI, 2014).

2.3.2 *Web´s* Invisíveis

As porções da *Internet* nas quais a indexação é dificultada ou impossibilitada recebem o nome de *Web* Invisível e/ou *Web* Profunda.

Vignoli (2014) aponta que o termo *Web* Invisível foi cunhado por Dr. Jill Ellsworth, em 1994, para designar os conteúdos de custosa recuperação por parte dos indexadores.

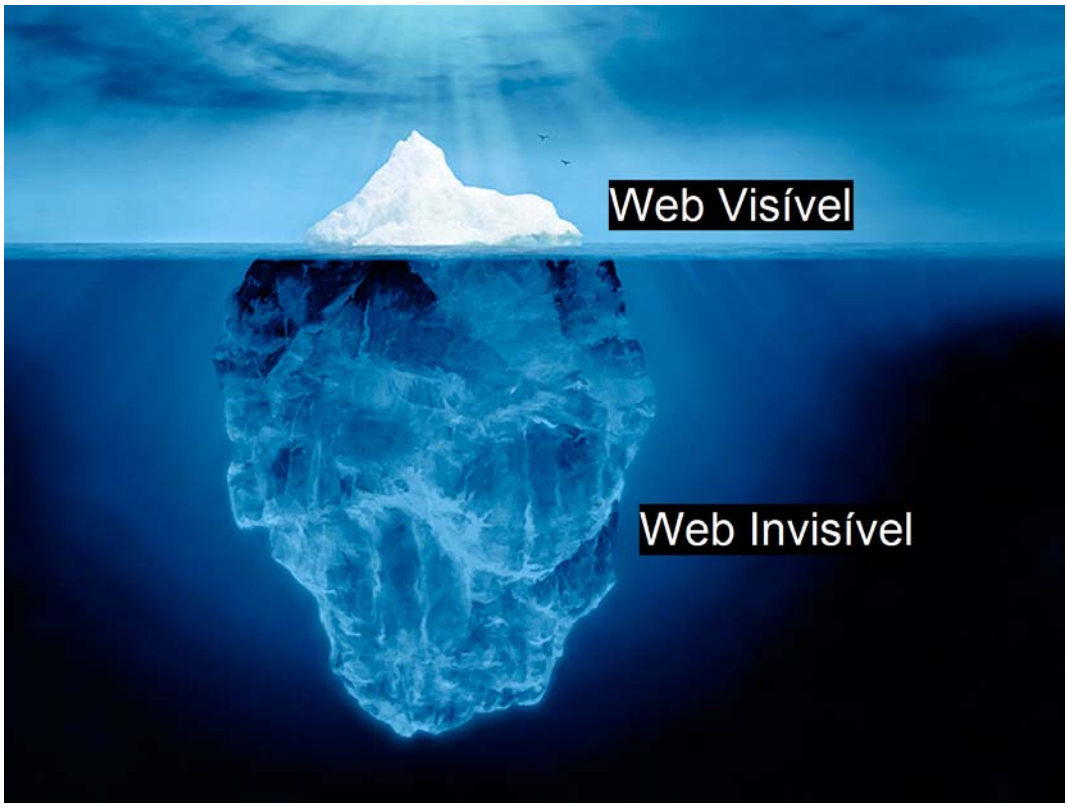
O conteúdo invisível deve ser analisado em duas frentes: inicialmente, pela incapacidade dos buscadores em indexar informação; e, consecutivamente, pela supressão intencional dos algoritmos de busca por meio de técnicas computacionais empregadas pelo desenvolvedor do *Website*.

Em Franco (2013), a *Web* Profunda é comumente associada a um *iceberg*, analogia cujo objetivo é demonstrar a dissonância entre a *Web* Visível e Invisível no que tange sua dimensão. A Figura 4 busca demonstrar uma proporção entre as *Web's*.

[9] Os robôs não estão programados para entender a estrutura de um banco de dados ou as linguagens utilizadas para recuperar a informação.

[10] Modelo de relacionamento de entidades e suas interações, em algum domínio particular do conhecimento ou específico a alguma atividade.

Figura 4 – Visibilidade nas *Web*'s



Fonte: Bergman (2001)

O conceito de visibilidade na *Web* se apresenta de maneira turva em grande parte das situações. Para Bergman (2001), o conteúdo da *Web* Profunda é de 1000 a 2000 vezes maior que o da *Web* Visível. Araújo (2012) aponta que a diversidade nos formatos de hipermídia[11] constitui uma grande dificuldade na varredura e recuperação da informação por parte dos mecanismos de busca. Em suma, caso uma página não possua *hiperlink*s[12] de referência, um buscador não seria capaz de identificá-la em sua varredura (BRANSKI, 2004).

Vignoli (2014), baseado em Sherman e Price (2001), traz em sua bibliografia algumas subclassificações para a *Web* Invisível. Nesse contexto, cinco subdivisões são apresentadas, de modo rudimentar, cujo objetivo prima por não estender em demasia o tema, sendo elas: (I) *Web* Opaca ou *Web* Oculta; (II) *Web* Privada; (III) *Web* Proprietária; (IV)*Web* Verdadeiramente Invisível e (V) *Dark Web*.

A *Web* Opaca flutua entre Visível e Invisível devido ao fato do conteúdo, por vezes, não ser incluído em índices de motores de busca. Os diversos

[11] A *Web* utiliza a hipermídia na sua formação básica, pois ela une os conceitos de não-linearidade, interface e multimídia em uma só linguagem.
[12] Conexões entre *sites*.

tipos de conteúdo estão literalmente opacos ou ocultos, entretanto, não necessariamente invisíveis. Monteiro e Fidêncio (2013) explicam que na *Web* Opaca os *sites*, diversas vezes, associam arquivos e mídias ao seu conteúdo, alguns incompreensíveis aos rastreadores do motor de busca.

Segundo Yamaoka (2002, p. 52), fazem parte da *Web* oculta:

a) conteúdo de banco de dados que formam páginas dinâmicas montadas pelos usuários, como o *Orkut*, por exemplo;
b) conteúdos protegidos por *firewall* em redes privadas;
c) conteúdos protegidos por *sites* protegidos por senhas de acesso;
d) documentos isolados da *Web* (que não recebem hiperligações de outros documentos); e
e) páginas com *frames* e *image-maps* também não são indexados por alguns mecanismos de busca.

Para Branski (2004, p. 82), há duas razões para estes *sites* estarem fora dos bancos de dados de grande parte dos buscadores:

a) questões técnicas que impedem o acesso dos *spiders*[13] a alguns tipos de *sites*; e
b) por decisão dos administradores dos motores de busca.

Na *Web* Privada, páginas da *Web* são tecnicamente indexáveis, entretanto, deliberadamente mantidas fora do índice por parte do mecanismo de busca. Dois motivos para esse feito podem ser elencados: (i) a ocorrência de credenciais para o acesso (usuário e senha) ou (ii) convenções computacionais entre *Website* e buscador.

Vignoli (2014) menciona a pertinência da nomenclatura “privada” na qual Redes Sociais e Fóruns de Discussão se enquadram nesse quesito por resguardar a privacidade de certas páginas.

Quanto à *Web* Proprietária, são páginas que necessitam de cadastro ou acesso aprovado pelo administrador, permanecendo, também, alheias aos buscadores (SHERMAN; PRICE, 2001). Embora indexável, o conteúdo pertence

[13] Sistema responsável pela recolha automática de conteúdos da *Web*.

aos seus mantenedores, que decidem as regras para o acesso à informação conforme exposto por Monteiro e Fidêncio (2013).

Para Sherman e Price (2001), páginas com pouco conteúdo textual ou conteúdo irrelevante aos motores de busca, páginas dinâmicas, bases de dados restritas, arquivos comprimidos ou executáveis caracterizam a *Web* Verdadeiramente Invisível. Monteiro e Fidêncio (2013) reconhecem ainda os buscadores híbridos como o *Google*, que recuperam arquivos em múltiplos formatos e apresentam constante aprimoramento de seus algoritmos de busca.

Monteiro e Fidêncio (2013) consideram, ainda, a *Dark Web* como uma nova ramificação da *Web* Invisível que dispõe de conteúdo ilegal e se beneficia da não indexação por parte dos mecanismos de busca. Everett (2009) argumenta que a *Dark Web* se faz presente por meio de qualquer servidor *Web* que não pode ser encontrado por buscadores como o *Google*. Ainda conforme os autores, a *Dark Web* proporciona a liberdade de publicar e comercializar conteúdo sem nenhum tipo de censura, visto que o ambiente é criptografado, concedendo alto nível de anonimato aos utilizadores.

Pelo exposto, é possível afirmar que, “[...] apesar da sua difusão, a lógica, a linguagem, os limites da *Web* não são bem compreendidos além da esfera de disciplinas estritamente tecnológicas.” (CASTELLS, 2003, p. 8).

2.4 Sistemas de Recuperação de Informação

Na década de 50 iniciou-se o desenvolvimento da área de Recuperação de Informação (RI) por meio das pesquisas elaboradas por Hans Peter Luhn, Eugene Garfield, Philip Bagley e Calvin Moores, este último, considerado o criador do termo “recuperação da informação” (BAEZA-YATES; RIBEIRO NETO, 2011, p. 2). Atualmente, “ [...] os sistemas de informação viabilizam, virtualmente, todas as atividades humanas.” (SOUZA, 2006, p. 162).

Sistemas humanos de processamento da informação, sistemas eletrônicos de processamento de dados e sistemas de recuperação da informação constituem exemplos de mecanismos especificamente planejados para possibilitar a RI (ARAÚJO, 1999, p. 15).

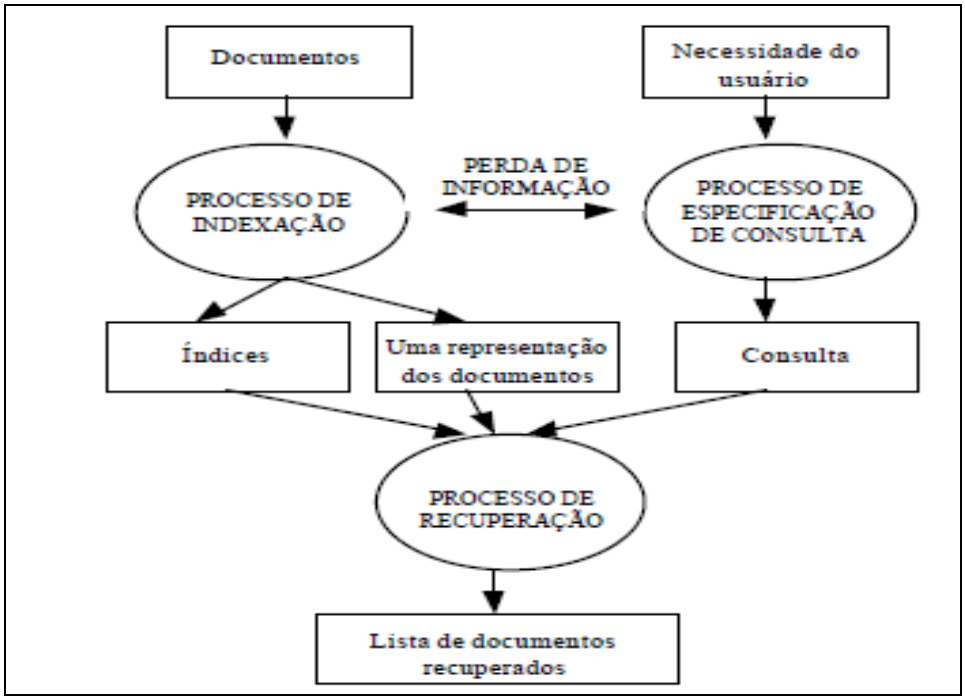
Igualmente, este trabalho optou por discorrer sobre os processos e recursos voltados para a recuperação e disseminação de informações no que diz respeito aos sistemas de busca da *Web*.

A obra de Holanda e Braz (2012, p. 44) traz a expressão "recuperação da informação", definida por Calvin Mooers, como sendo o "englobamento dos aspectos intelectuais da descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação", citada anteriormente na obra de Saracevic (1996).

Para Araújo (1995, p. 15) "SRI são tipos de sistemas de comunicação que, entre outras funções, visam dar acesso às informações neles registradas."

Um SRI pode ser estruturado conforme a Figura 5 a seguir:

Figura 5 – Estrutura de um sistema de informação



Fonte: Cardoso (2003, p.1)

Os componentes do sistema descrito na Figura anterior incluem documentos, necessidades do usuário, geração da consulta formulada e, finalmente, o processo de recuperação que, a partir das estruturas de dados e da consulta formulada, recupera uma lista de documentos considerados relevantes (CARDOSO, 2003). Esses sistemas possibilitam,

> O planejamento de estratégias de busca com maior nível de complexidade envolvendo vários conceitos na mesma estratégia; permitem a utilização de busca de palavras apenas dos títulos e resumos dos documentos, isto é, termos da linguagem natural; buscam os termos específicos de linguagens controladas, nos campos de descritor; buscam por autores; por ano de publicação; por títulos de periódicos; por classificação; permitem, também, a busca de conceitos compostos ou simples e a possibilidade de truncagem de raízes de palavras e de substituição de caracteres no meio dos termos, dentre outros recursos de recuperação. (LOPES, 2002, p. 60).

Conforme Souza (2006, p. 162), "[...] desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários." Nesse sentido, é oportuno comentar que

> A recuperação de dados, no contexto de um sistema de recuperação de informação, consiste na identificação de quais documentos da coleção contêm as palavras-chave da consulta do usuário, o que, com frequência, não é suficiente para satisfazer a necessidade de informação do usuário. (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 6).

A eficiência de um SRI está diretamente ligada ao modelo que o mesmo utiliza (FERNEDA, 2003, p. 18).

Cardoso (2003) explica que os modelos clássicos utilizados no processo de RI (booleano, vetorial e probabilístico) apresentam estratégias de busca de documentos relevantes para uma consulta (*query*).

a) Modelo booleano: o índice atribuído aos documentos deve indicar qual documento é mais relevante que outro, estabelecendo uma ordem de relevância. Os documentos recuperados são aqueles que contêm os termos que satisfazem a expressão lógica da consulta;
b) Modelo vetorial: o vetor resultado para uma consulta é montado por meio de um cálculo de similaridade;
c) Modelo probabilístico: descreve documentos considerando pesos binários que representam a presença ou ausência de termos. O vetor resultado gerado pelo modelo tem como base

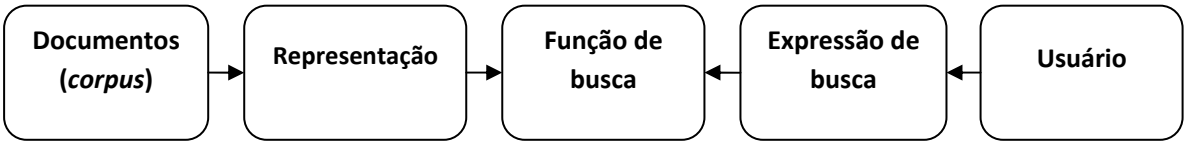
o cálculo da probabilidade de que um documento seja relevante para uma consulta.

De acordo com Ferneda (2003), tais modelos são quantitativos, de caráter empírico, presentes, ainda, na maioria dos SRIs, principalmente, pelo fato de "os motores de busca da *Web* terem introduzido características específicas para tratar a quantidade de informação disponível na *Internet*".

Segundo Teixeira e Schiel (1997) "[...] compreende basicamente três etapas: indexar, armazenar e recuperar."

Diante desse contexto, "[...] o processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário." (CARDOSO, 2003, p. 2). Ou seja, "[...] consiste em identificar, no conjunto de documentos (*corpus*) de um sistema, quais atendem a necessidade de informação do usuário." (FERNEDA, 2003, p. 14). A Figura 6 mostra um esquema de como acontece o processo de RI.

Figura 6 – Processo de RI



Fonte: Ferneda (2003, p. 15)

a) documentos: envolve quaisquer signo que represente textos, imagens, sons etc.;
b) representação: busca descrever ou identificar cada documento do *corpus* por meio de seu conteúdo e do processo de indexação;
c) função de busca: compara as representações dos documentos com a expressão de busca dos usuários e recupera os itens que supostamente fornecem a informação que o usuário procura;
d) expressão de busca: decorrente da necessidade de informação do usuário, resulta na recuperação de um número

> de documentos que possibilite a verificação de cada um deles a fim de selecionar os que são mais úteis.

No que diz respeito à indexação em SI “[...] ela é reconhecida como a parte mais importante do sistema porque condiciona os resultados de uma estratégia de busca.” Além disso, “[...] a indexação pode ser observada em dois momentos distintos dentro do sistema: na entrada (no tratamento temático da informação); e na saída (na busca e RI).” (RUBI; FUJITA, 2003, p. 69).

A eficiência de um processo de RI está diretamente ligada à estratégia de busca[14] elaborada pelo usuário. Nesse sentido, Lopes (2002, p. 60) relata que “[...] o alcance da qualidade na informação recuperada requer o planejamento de estratégias de busca específicas para cada base de dados.”

A busca na *Web* corresponde à aplicação mais proeminente de RI e suas técnicas (BAEZA-YATES; RIBEIRO NETO, 2011, p. 12).

Na próxima subseção, o foco do estudo se delimitou a abordagem dos motores de busca.

2.5 MOTORES DE BUSCA

A *Web* surgiu como uma ferramenta de comunicação e evoluiu, rapidamente, para o compartilhamento de informações. Se, a princípio, era necessário digitar corretamente o endereço dos *sites* para acessá-los, a possibilidade de busca aumentou, exponencialmente, a gama de usuários capazes de localizar informações.

Levando-se em consideração que “[...] as primeiras máquinas de busca na *Web* eram, fundamentalmente, sistemas de RI [...].” (BAEZA-YATES; RIBEIRO NETO, 2011, p. 100), a que se argumentar, segundo Battelle (2006, p. 7) que “[...] a busca foi um dos primeiros serviços úteis a habitar a *Internet* [...]”, o que a tornou atraente, usável e acessível.

Buscadores, ferramentas de busca, mecanismos de buscas ou motores de busca (*search engines*) são sistemas especializados na RI na *Internet* (BRANSKI, 2004, p. 72).

---

[14] Visa uma recuperação de informação de acordo com as necessidades dos usuários.

Neste trabalho, entretanto, optou-se por utilizar o termo “motores de busca”. Em adiantada explicação, é possível afirmar que “[...] os motores de busca [...] permitem ao usuário submeter sua expressão de busca e recuperar uma lista de *URLs* que, presumivelmente, são relevantes para a sua necessidade de informação.” (FERNEDA, 2003, p. 96). Segundo Holanda e Braz (2012), as características destes buscadores “[...] influenciam no modelo de indexação utilizado e na forma que os assuntos serão recuperados.”

2.5.1 Contexto Histórico

É importante salientar que os primeiros buscadores baseados em motores de busca começaram a ser usados em 1993 com o objetivo de medir o crescimento da *Web* (FRAGOSO, 2007; FUENTES, ORDUÑA, 2010), porém, não houve continuidade do propósito inicial devido ao aumento exponencial de *sites* (BATTELLE, 2006).

A inquietação em localizar informação na *Internet* foi apresentada por Cendón (2001, p. 39) ao enfatizar que:

> Desde os primórdios da *Internet*, houve a preocupação de se criarem ferramentas para localização de seus recursos informacionais. [...]. Com o advento da *Web* e a consequente explosão das publicações disponibilizadas por meio dela, começaram a surgir as ferramentas específicas para pesquisa de suas páginas. Existem hoje centenas destes instrumentos que fornecem meios para localizar o que se busca entre as cerca de um bilhão de páginas *HTML*[15], que se estimam.

Durante vários anos, os sistemas de banco de dados existiram para buscar dados estruturados, e a RI se restringia à busca de documentos simples e previamente ordenados (MELO, 2009, p. 5).

Para Broder (2002, p. 18), a busca na *Web* foi impulsionada pela necessidade de informação para cumprir uma tarefa, e não somente pelo acesso à informação por ela mesma. Na visão do autor, as buscas na *Web* podem ser classificadas em três categorias:

[15] A utilização do HTML faz com que os motores de busca entendam melhor as páginas dos *sites*, sendo capazes de indexar o conteúdo de forma mais eficaz.

1) Navegativa: quando a intenção de busca é pontual para localizar e acessar determinado *site*;
2) Informativa: quando a intenção é adquirir informações que podem estar presentes em uma ou mais páginas da *Web*;
3) Transacional: a intenção é identificar locais em que se possam executar determinadas transações, como fazer compras ou *download*s.

Para Fuentes e Orduña (2010), os buscadores atuam no sentido de facilitar e incentivar a organização da extensa quantidade e variedade de informação disponível na *Internet*, convertendo-a em universalmente acessível.

Dessa forma, a preocupação em desenvolver ferramentas que facilitem a busca por informação cresce à medida que o volume de informação aumenta.

Segundo Cendón (2001), os diretórios foram ofertados como primeira alternativa para organizar e localizar conteúdo na *Web,* precedendo o modelo atual baseado em motores de busca. É válido salientar que aquele modelo foi introduzido quando a quantidade de informação disponível na *Internet* ainda era modesta, se comparada ao conteúdo que se pode ter acesso atualmente.

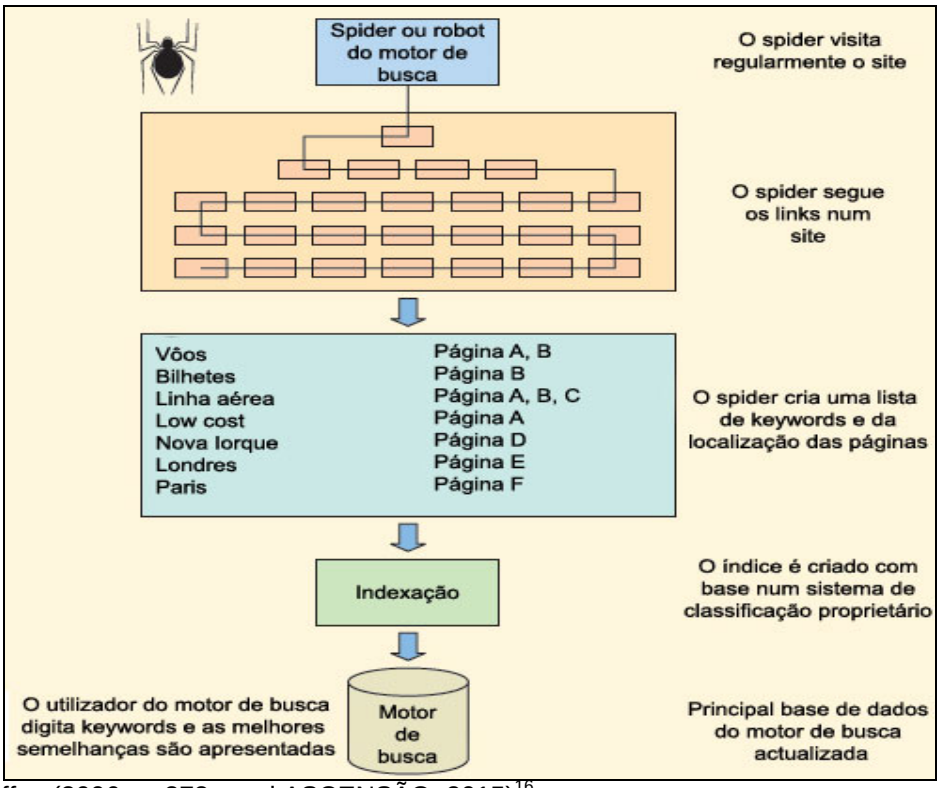
O autor complementa explicando que os diretórios possuíam como método a divisão do conteúdo eletrônico por categorias, que, por sua vez, poderiam se desdobrar em subcategorias, isto é, sua organização era de forma hierárquica. Portanto, o usuário navegaria entre as páginas transitando entre categorias e subcategorias até alcançar o conteúdo desejado. Em contrapartida,

A busca na *Web* pode ser motivada não somente pela carência de informação, mas também pela necessidade de localizar bens, produtos e serviços. Battelle (2006) aborda a personalização da busca, ou melhor, a individualização dos resultados de acordo com as necessidades de quem pesquisa, à resposta perfeita. Ainda não é possível alcançar resultado tão pontual, porém, muitos são os esforços para atingir esse objetivo.

2.5.2 Funcionamento de um Motor de Busca

Na sequência, sintetizado na Figura 7, para melhor compreensão, é apresentado o funcionamento de um motor de busca.

Figura 7 – Funcionamento de um motor de busca



Fonte: Chaffey (2006, p. 378 apud ASCENSÃO, 2015)[16]

Dentre alguns exemplos apresentados por Monteiro *et al.* (2009, p. 8) para a tipologia de organização e busca do conhecimento na *Internet*, pode-se destacar a categoria anatomia, a qual é classificada em: (I) *Crawling* (varrer); (II) *Indexing* (indexar ou gerar o índice a partir da base de dados); e (III) *Searching* (buscar através da interface de busca).

Primordialmente, o programa denominado *Crawler* navega de forma autônoma na *Internet*, reunindo o maior número possível de páginas Web, gerando, desta forma, uma base de dados e, por conseguinte, um índice[17], que será apresentado ao usuário por meio de uma interface amigável.

[16] <http://www.portalwebmarketing.com/MotoresdeBusca/Comofuncionaummotordebusca/Default.>
[17] Se um termo não estiver incluído no índice ele não será encontrado, portanto, os critérios utilizados para indexação influenciarão nos resultados.

Segundo Battelle (2006), nas primeiras versões de *Crawlers,* apenas o título das páginas da *Web* era indexado, porém, atualmente, pode-se indexar todo o conteúdo, inclusive diversos tipos de documentos.

Holanda e Braz (2012, p. 47) explicam que “[...] os motores de busca não indexam os *sites* em si, mas sim os conteúdos das páginas que os compõem.” E ainda, “[...] isso significa que uma página de um determinado site pode estar entre os primeiros resultados de uma busca, como pode estar entre os últimos em outra consulta.”

Subsequentemente, a geração do índice (*Indexing*) associa as palavras presentes na página *Web à* URL[18], gerando metadados que serão tratados de acordo com o algoritmo implementado no motor de busca. Conforme Battelle (2006), o índice representa uma enorme base de dados onde se encontram informações importantes a respeito de diversos sites na Web.

Por último, no *Searching* é apresentado o “[...] motor de busca propriamente dito” citado por Cendón (2001), ao qual a interface propicia ao usuário consultar de maneira intuitiva a base de dados indexada pelo *Software*. Nota-se que o processo executado pela máquina de busca está intimamente atrelado à maneira como o *Software* foi arquitetado por seu desenvolvedor.

Como se percebe, numa visão mais sintetizada, os motores de busca apresentam três funções principais: um robô, que localiza os documentos; um indexador, que extrai as informações dos documentos, e uma interface com o usuário (YAMAOKA, 2002).

Em tempo, é importante mencionar que,

> Os motores de busca diferem entre si em relação ao tamanho de suas bases de dados, critérios para indexação e inclusão de páginas, a ordenação dos resultados, além de suas interfaces, recursos de buscas, frequência de atualização de suas bases de dados e como apresentam o resultado. (CENDÓN, 2001, p. 42).

Monteiro *et al.* (2011, p. 2546) explicam que “[...] o fato é que, quanto maior a base do buscador, maior a probabilidade de os resultados apresentarem-se relevantes. Não se deve esquecer, também, que atrás de cada grande interface de busca há um motor que produz resultados.”

[18] Cada página de um *site* tem um endereço único denominado URL.

Assim, Branski (2004, p. 72) ressalta que “[...] quando se realiza uma busca não se está pesquisando diretamente na *Internet*, mas no banco de dados do buscador escolhido.”

2.5.3 Estratégias de Indexação

Fatores como o fácil acesso às Tecnologias de Informação e Comunicação (TICs), a disponibilização maciça de informação na *Web* e o imensurável crescimento da frequência de busca na *Internet* podem ter estimulado mudanças significativas na forma de indexação.

Conforme Monteiro (2006, p. 34), “[...] a organização do conhecimento na *Internet*, hoje é possível a partir da indexação realizada pelas máquinas de busca, ou seja, motores de busca e metabusca[19] na *Internet.*” Esses mecanismos permitem que a informação na *Internet* seja passível de indexação e recuperação. Nesse sentido,

> A indexação em linguagem controlada é relevante em unidades de informação, bibliotecas virtuais e outras fontes de informação científica, pois possibilita uma padronização nos termos indexados e favorece a recuperação. Porém, [...] o astronômico crescimento da *Internet*, seu imenso número de documentos e relações entre esses documentos nos obrigam a encontrar novas formas de orientação e busca. (SANTAELLA, 2007, p. 183).

A indexação é um procedimento que tem como objetivo selecionar, em documentos, conceitos que o representem da melhor forma, para facilitar seu armazenamento e recuperação (GIL-LEIVA; ALONSO-ARROYO, 2007, p. 1175). Deve produzir uma correspondência precisa com o assunto pesquisado em índices (RUBI; FUJITA, 2009, p. 100).

Para Holanda e Braz (2012, p. 46), seu objetivo principal é “[...] assegurar a recuperação de qualquer documento ou informação no momento em que houver solicitação em um sistema de informações.”

Seu processo envolve a criação de estruturas de dados associados à parte textual dos documentos, as quais podem conter dados sobre características

[19] Dispõem de mecanismos que acessam a vários índices simultaneamente, economizando tempo e aumentando as chances de encontrar o que se está procurando.

dos termos na coleção de documentos, tais como a frequência de cada termo em um documento (CARDOSO, 2003).

No momento da indexação são extraídos conceitos do documento por meio da análise de seu conteúdo e traduzidos em termos de uma linguagem de indexação, tais como tesauros, cabeçalhos de assunto etc. (FERNEDA, 2003, p. 16).

De acordo com a Norma 5963/1985 da *International Association for Standardization* (ISO), o processo de indexação deve se atentar, principalmente, ao título, resumo, tabelas de conteúdo, apresentações, início de parágrafos e capítulos, conclusões, ilustrações, diagramas, tabelas e legendas, sublinhados e negritos de palavras ou frases (GIL-LEIVA; ALONSO-ARROYO, 2007).

A indexação permite a busca por palavras localizadas em qualquer parte do documento. Segundo Ferneda (2003, p. 16), também pode ser efetuada tendo em vista a sua recuperação. Neste caso, a análise do documento é realizada com a preocupação de tornar o seu conteúdo visível para os usuários de um sistema de informação. Quanto aos métodos automáticos de indexação,

> Geralmente utilizam “filtros” para eliminar palavras de pouca significação (*stop words*), além de normalizar os termos reduzindo-os a seus radicais [...]. essa forma de indexação seleciona formas significantes (termos ou frases) dos documentos desconsiderando os significados que os mesmos podem possuir de acordo com os textos [...]. (FERNEDA, 2003, p. 17).

Quando a indexação é realizada manualmente, por seres humanos, cabe a estes descobrir conceitos que sirvam de termos-índices para serem vasculhados durante as consultas de usuários. Na indexação automática existem dezenas de estratégias para a correta ponderação do valor do documento de acordo com uma explicitação de necessidade de informação (SOUZA, 2006, p. 165).

Desse modo, para alcançar a resposta pretendida pelo usuário de informação, faz-se necessária a execução de movimentos e operações, ora restringindo os resultados alcançados, ora ampliando-os para a obtenção de informações mais relevantes, conforme o pedido de busca demandado (LOPES, 2002, p. 61).

Souza e Alvarenga (2004, p. 133) afirmam que “[...] não há, na *Web*, nenhuma estratégia abrangente e satisfatória dos documentos nela contidos [...].” Entretanto, de acordo com Monteiro (2009, p. 69), “pensar estrategicamente na

organização do conhecimento na atualidade significa compreender como as máquinas estão operando a indexação na *Internet.*"

2.6 *BIG DATA* E MINERAÇÃO DE DADOS

A obra de Taurion (2003) traz uma definição cunhada pelo Instituto McKinsey Global, na qual *Big Data* refere-se ao conjunto de dados com tamanhos que excedem a capacidade dos atuais *software*s de banco de dados nos quesitos captura, armazenamento, gerenciamento e análise.

É importante salientar a particularidade nesta definição, pois, onde não são citados parâmetros numéricos como escala visto que, conforme a tecnologia avança, o montante de dados necessários para caracterizar o *Big Data* também aumenta.

A massa de dados (consequência da arquitetura já descrita) produzida por meio dos dispositivos eletrônicos na contemporaneidade é indiscutível, todavia, apesar de volumes colossais, outras variáveis igualmente importantes compõem o *Big Data*.

A variedade de dados coletados em fontes diferenciadas como mídias sociais, sensores e até mesmo padrões de navegação na *Internet* altera de maneira radical a análise dos dados e a consequente extração de informações[20] úteis.

Atualmente, a tecnologia permite que, ao contrário de abordagens mais conservadoras em que a análise era constituída por meio de amostras, seja viável analisar toda uma massa de dados proporcionando maior acuidade nas decisões.

A possibilidade de analisar volumes inéditos de dados digitais terá um impacto tão grande em processos de negócio quanto a popularização da *Internet* segundo Taurion (2003). O autor ressalta que dados obtidos por meio de sistemas transacionais (visíveis comumente) representam uma diminuta parcela dos dados que circulam em uma organização. Segundo o autor, o *Big Data* pode ser visto de

[20] Extração de informação é a tarefa de identificar fragmentos específicos que constituem o núcleo semântico de um documento em particular e construir modelos de representação da informação (conhecimento) a partir dele.

maneira análoga à invenção do microscópio, tornando visível um enorme montante de informações que passavam desapercebidas pela sociedade.

O *blog* TechCrunch, através de uma entrevista com o Facebook em 2012, divulgou que esta rede social processa, em média, 500 terabytes de dados por dia. Taurion (2013) estima, tendo em vista os 1,8 zetabytes de dados (10^{21} bytes) gerados em 2012, que haverá um salto para 7,9 zetabytes em 2015. O autor afirma ainda que 90% dos dados existentes hoje foram criados nos últimos dois anos.

Apenas 25% dos dados produzidos mundialmente eram armazenados em formato digital no ano 2000. Esse percentual aumentou drasticamente em 2007 totalizando 94%, extrapolando os 99% em 2013 (TAURION, 2013). O volume de dados cresce de forma espantosa na Sociedade da Informação e boa parte dessa contribuição se deve à miniaturização das tecnologias, bem como o aumento da capacidade de armazenar e processar dados.

Ainda nesse panorama, a alta capacidade de processamento e miniaturização dos componentes eletrônicos tornam possíveis a criação da "*Internet* das Coisas"[21], que contribuirá exponencialmente para o crescimento da massa de dados. Um novo conceito em objetos identificáveis e interconectáveis vai aglutinar o mundo físico com o digital, trazendo consigo um montante torrencial de dados e expandindo a possibilidade na análise do *Big Data*.

Taurion (2013) afirma que há uma relação simbiótica entre o mundo físico e o mundo digital, relação trazida à tona neste trabalho, na qual as entidades físicas possuem identidade digital única, interagindo com outras entidades do mundo virtual, sejam objetos ou pessoas.

O *Big Data* constitui uma nova forma de investigar o imenso volume de dados que circula nas empresas e demais instituições, logo, não se constitui apenas de *Software*. Ao empreender iniciativas nesse campo, estratégias[22] bem definidas de atuação devem reger a instituição.

A fase inicial do processo é composta pela coleta dos dados. O volume e variedade são características importantes nesse quesito, pois cada empreendimento busca dados diferentes para obter as informações desejadas.

[21] Revolução tecnológica com o objetivo de conectar os itens usados do dia a dia (eletrodomésticos, meios de transporte, tênis, roupas, maçanetas, etc.) à rede mundial de computadores.

[22] Técnica ou conjunto de regras para tornar possível o encontro entre uma pergunta formulada e a informação armazenada em uma base de dados.

Posteriormente, há um trabalho de limpeza e adequação dos dados, eliminando entradas incompletas ou inconsistentes (TAURION, 2013).

A etapa seguinte traz a integração e agregação dos dados obtidos anteriormente, diferentes tipos e formatos demandam tratamento específico, categorizar os dados e definir critérios para sua validação são aspectos importantes dessa fase. Por conseguinte, a etapa analítica pode ser considerada a mais visível, na qual os resultados serão interpretados. Destaca-se pelo oneroso trabalho tendo em vista o montante de dados muitas vezes superior a *terabytes*.

2.6.1 O Impacto do *Big Data*

Ao considerar o impacto causado pelo *Big Data* nas instituições de ensino, empresas privadas e órgãos governamentais, constatam-se padrões e relacionamentos anteriormente opacos, cujos dados situados, não somente na infraestrutura de tecnologia das organizações, mas também na rede mundial de computadores, constituem um fator diferencial nas mais variadas aplicações, entre elas: transparência; segmentação acurada; análise preditiva; algoritmos automatizados; modelos de negócio (TAURION, 2013).

A disponibilização de um número cada vez mais elevado de dados, até então inacessíveis, propicia a diversos setores o cruzamento de informações que antes restavam isoladas. Uma oportunidade de integração e melhoria de gestão nos mais diversos segmentos seria fruto dessa transparência na interoperabilidade dos dados. Fontes de informação tendem a ser consideravelmente ampliadas com a instauração do *Big Data,* em especial no âmbito da segmentação precisa. Entretanto, dados capturados na *Web* propiciam maior amplitude de associação entre vida real e hábitos virtuais, culminando na habilidade de prever padrões de consumo, comportamento que vêm desmantelando as fronteiras entre público e privado.

Analisando padrões em grandes volumes de dados é admissível a previsão de fenômenos fisiológicos ou comportamentais, um exemplo é a capacidade de prever epidemias por meio de palavras-chave no campo de pesquisa dos buscadores (Figura 8).

Figura 8 – *Google t*endências da gripe

Fonte: *Google* tendências da gripe

Ao procurar por sintomas ou medicamentos o usuário dos motores de busca também informa que, possivelmente, sofre de alguma enfermidade. Essa informação é coletada e, posteriormente, utilizada como iniciativa na prevenção de doenças.

O *Google Flu Trends* é um serviço que apresenta, em tempo real, tendências de gripe (Influenza) ao redor do mundo, estendendo em caráter experimental para casos de dengue (*Aedes aegypti*). Essa técnica expande o potencial das políticas de controle sanitário introduzindo análises preditivas mais apuradas.

Outro diferencial apresentado pelo *Big Data* é a capacidade de subsidiar algoritmos para completar ou substituir decisões humanas. Nessa abordagem, volumes elevados de dados ao serem submetidos a algoritmos sofisticados permitem a automatização de funções em variadas áreas, como gerenciamento de processos ou controle de tráfego (TAURION, 2013).

Ainda segundo o autor, novos modelos de negócio surgem através do *Big Data.* Ao possuírem condições para predizer eventos cotidianos empresas de diversos setores podem utilizar técnicas de análise preditiva com o intuito de evitar desperdícios em manutenções preventivas por exemplo.

Fundamentalmente, grandes massas de dados fornecem matéria prima para o reconhecimento de padrões comportamentais. A identificação e coleta desses padrões constitui um novo segmento de mercado que vem crescendo de maneira exponencial, atrelando-se a todos os setores da sociedade humana com o objetivo de entender demandas ou mesmo se posicionar de maneira mais competitiva no mercado.

O *Big Data* transpõe as barreiras da economia inteligente produzindo um fluxo continuo de informações derivadas de diversas fontes como *desktop's*, *tablet's*, *smartphones*, mídias sociais e milhões de sensores espalhados pela infraestrutura das cidades que podem ser monitorados e analisados. Esse novo cenário de oportunidades permite que empreendedores criativos analisem dados de forma inimaginável até o momento, criando vantagens competitivas de grande expressão. O tratamento de dados será, segundo Taurion (2003), tão importante quanto qualquer outro recurso organizacional, não sendo possível proceder sem a sua análise contínua.

2.6.2 Mineração de Dados

Nas últimas décadas, em que a maioria das operações e atividades das instituições privadas e públicas é registrada computacionalmente e se acumula em grandes bases de dados, "[...] a técnica da mineração de dados – *Data Mining* (DM) – é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados [...]. (GALVÃO; MARIN, 2009, p. 687).

A mineração de dados "[...] não é um processo trivial; consiste na habilidade de identificar, nos dados, os padrões válidos, novos, potencialmente úteis e compreensíveis [...]. (GALVÃO; MARIN, 2009, p. 688).

A DM possui várias etapas: a definição clara do problema; a seleção de todas as fontes internas e externas de dados e a preparação dos dados, que inclui o pré-processamento, reformatação dos dados e análise dos resultados [...]. (GALVÃO; MARIN, 2009, p. 688). "[...] é uma das tecnologias mais promissoras da atualidade. (CAMILO; SILVA, 2009, p. 2).

Desse modo, é possível afirmar que, "[...] a mineração contribui de forma significativa no processo de descoberta de conhecimento, permitindo aos

especialistas concentrarem esforços apenas em partes mais significativas dos dados.” (CAMILO; SILVA, 2009, p. 8).

Com isso, é possível afirmar que “[...] os estudos sobre a DM comprovam o pressuposto da transformação de dados em informação, e posteriormente em conhecimento.” (GALVÃO; MARIN, 2009, p. 688).

2.7 MOTORES DE BUSCAS COMO SISTEMAS PROPRIETÁRIOS

Atualmente, os *software*s de busca são detentores e difusores de todo o conhecimento veiculado na *Internet*, em sua maioria, oferecidos por grandes companhias como o *Google* e *Microsoft*, cujos sistemas são essencialmente fechados, resultando em uma tecnologia de indexação e classificação enevoada aos utilizadores da aplicação.

Ao se utilizar buscadores comerciais, não se pode arguir quanta informação será censurada, bloqueada ou removida do resultado, ficando este a critério apenas da entidade detentora do *Software*.

Neste tipo de *Software* é impossível visualizar, estudar e modificar o código-fonte (EVANGELISTA, 2009). Assim, “[...] a essência do modelo proprietário de licenciamento de *Software* está no controle do conhecimento/estratégia de indexação que o modelo proprietário bloqueia. Não o uso.” (SILVEIRA, 2009, p. 206).

Caso o detentor da página *Web* queira indexá-la por meio de um mecanismo de busca, deverá aceitar suas regras e termos de uso, assim como a sua insubordinação acarretará em punições severas ao *Website* em questão, ou seja, sua não indexação.

Em uma empresa responsável por administrar um mecanismo de busca, supõe-se que em um possível conflito de interesses ou exposição de informações indesejadas, a balança penderá em seu próprio benefício, de forma que impactos nocivos à sociedade empresarial sejam minimizados. Esta questão norteia um argumento interessante que fundamenta a base desta problematização.

Buscas efetuadas mediante mecanismos comerciais são, fundamentalmente, tendenciosas, seja por políticas organizacionais, privilégios a patrocinadores do serviço ou determinações judiciais (FRAGOSO, 2007). Nesse contexto, a indexação e a organização de conteúdo atende a benefício próprio,

categorizando os usuários de seu serviço não como clientes, mas como produtos. Os motores de busca comerciais contradizem o espírito descentralizador da *Internet*, impondo hierarquias convenientes a um determinado grupo corporativo.

Na subseção a seguir são apresentados exemplos de motores de busca padrão (*Google*), semântico (*Wolfram Alpha*) e misto, que indexa porções de *Web* invisível (*Duck Duck Go*).

O motor de busca *Google* é um projeto que foi iniciado em 1996 com base nos estudos de Larry Page e Sergey Brin, enquanto desenvolviam sua tese de doutorado na Universidade de Stanford. O objetivo principal dos seus criadores, [...] era "[...] organizar as informações do mundo todo e torná-las acessíveis e úteis em caráter universal." (HOLANDA; BRAZ, 2012, p. 43).

Como mencionado no início desse trabalho, o *Google* apresenta a maior base de dados existente na *Web* e, ainda, "[...] após a apresentação dos itens resultantes de uma busca, permite especificar uma nova expressão a busca apenas nesses itens recuperados." (FERNEDA, 2003, p. 103).

Os motores de busca presentes em meados dos anos 90 classificavam seus resultados baseando-se no número de ocorrências da palavra buscada em uma página da *Web*. Essa lógica de classificação indica que quanto maior a incidência da palavra buscada em uma página, melhor seria seu posicionamento no motor de busca. Entretanto, os fundadores do *Google* teorizavam uma maneira mais eficiente para o retorno da informação, na qual as relações de *hiperlink*s entre os *sites* determinavam suas posições na busca, tecnologia que ficou conhecida como *PageRank*[23]. "Essa forma de estruturar a informação tem como premissa a ideia de que os *sites* mais populares oferecem informações de melhor qualidade." (BRANSKI, 2004, p. 73). A performance do *PageRank* funciona da seguinte maneira:

> O usuário realiza uma pesquisa na página principal do *Google* e as máquinas fazem uma busca no índice das páginas que correspondem e retornam os resultados que "parecem" ser os mais relevantes. Essa relevância é julgada por mais de duzentos fatores, mas é o *PageRank* que ministra toda a protuberância dos resultados. Ou seja, para cada link gerado de uma página, em outro *site* é adicionado um *PageRank* ao *site* "linkado", mas nem todos os *links* são iguais, o sistema identifica *spams* e outras ameaças ao resultado da pesquisa. (HOLANDA; BRAZ, 2012, p. 49).

[23] Fórmula matemática que rastreia os *site*s em busca dos *links* gerados e ordena a importância que cada página tem na *Internet*.

Em detrimento desta nova ótica, a relevância dos *sites*, bem como seu posicionamento, era determinada pela importância[24] dos *hiperlinks* e sua incidência em outros *sites* (BRANSKI, 2004).

Baeza-Yates e Ribeiro Neto (2011, p. 7) explicam que “[...] o propósito do ranqueamento é identificar os documentos que têm maior probabilidade de serem considerados relevantes pelo usuário.”

O *Wolfram Alpha,* desenvolvido pela empresa *Wolfram Research,* se diferencia dos demais buscadores por possuir um mecanismo de conhecimento computacional que efetivamente responde as consultas mediante o processamento de uma extensa base de dados estruturados.

Este curioso mecanismo foi apresentado em 2009 pelo físico Stephen Wolfram, sendo capaz de acumular informações em diversas áreas do conhecimento. Os dados utilizados são previamente processados e limitados pela empresa desenvolvedora, levantando questões quanto à confiabilidade das respostas apresentadas. Apesar da questão levantada, é possível requerer a fonte utilizada pelo *Wolfram Alpha* nas buscas (*WOLFRAM ALPHA, 2015*).

Desenvolvido por cientistas especializados em diversas áreas do conhecimento, o buscador mostra seus resultados de modo diferenciado em forma de dados, tabelas e informações úteis, e apresenta uma distinção evidente com relação aos demais buscadores. “Estabelece sofisticados diálogos com o usuário.” (BAX, 2012, p. 7).

Nota-se um aprimorado tratamento semântico por trás da interface, que é capaz de codificar a pergunta feita originalmente em linhagem natural e interpretá-la (BAX, 2012, p. 7).

O buscador *Duck Duck Go* tem como principal particularidade a utilização de fontes baseadas em conteúdo colaborativo (como a *Wikipédia*), enfatizando também a privacidade de seus clientes, logo, nenhuma informação de navegação é armazenada pela plataforma.

Fundado por Gabriel Weinberg, em 2008, é caracterizado como um motor de busca híbrido por ser construído sobre uma Interface de Programação de

[24] O mecanismo exato para determinação da importância das páginas varia de motor para motor e, geralmente, não é revelado, porque os algoritmos de ordenação por relevância são um dos maiores fatores de competição entre os motores.

Aplicações (API) de grandes fornecedores, entre eles o *Yahoo!*. O núcleo do *software* é codificado na linguagem de programação Perl (*DUCK DUCK GO, 2015*).

O *Duck Duck Go* exibe os resultados pesquisados de forma mista, evidenciando informações de maior relevância, todavia, apresentando *hiperlinks* úteis para futuras referências. Aponta um índice de busca com retornos satisfatórios para um buscador com menor subsídio de capital em relação aos seus concorrentes.

Diante do exposto, cabe apontar algumas considerações preliminares, pois, analisando em espectro amplo, o buscador comercial da *Google* cumpre sua proposta de maneira eficaz, no entanto não exibe informações pesquisadas de maneira direta, mas sim, por meio de *links* que devem ser acessados a fim de obter alguma informação útil.

Os destaques vão para os buscadores *Wolfram Alpha* e *Duck Duck Go*. O primeiro consegue sintetizar um emaranhado de dados em inúmeras fontes, agregando-os com grande relevância. O *Duck Duck Go* faz uma junção salutar entre informação de relevância semântica e resultados de busca orgânica (modo tradicional), mas falha ao buscar termos genéricos exibindo resultados de forma clássica quando não verifica sentido na sentença buscada.

Percebe-se que algumas ferramentas merecem ser evidenciadas por apresentar informações de forma diferenciada e se destacando ao recuperar informações verdadeiramente proveitosas.

2.7.1 Agenciamento da Escassez

Os mecanismos de busca proprietários produzem, a partir do estabelecimento de um critério proprietário de hierarquização, um mecanismo de escassez.

Para Franklin *et al.* (2013), o ponto de destaque é o agenciamento da escassez e o conceito de repetição. A escassez é condicionada pelo entorno político, raciocínio que acompanha o pensamento marxiniano. Ao transformar objetos materiais em informação, ocorre também uma mudança no estatuto da escassez que o regulava.

Uma cópia não significa mais a degeneração ou corrupção do original, visto que nesse novo modelo os mecanismos de fluxo econômico possuem

atributos de não rivalidade, diferentemente do antigo modelo de produção, distribuição e consumo de mídia (ANDERSON, 2009).

Não somente o conceito de escassez, mas também a propriedade intelectual exerce severas mudanças na relação entre a máquina e a lei. É oportuno esclarecer que a propriedade intelectual nasceu como direito de reprodução ofertado pela realeza aos donos de papelarias, cujo intuito era o controle do material impresso, um método sintético de produção de escassez (FRANKLIN *et al.*, 2013).

Não obstante, a lei produz escassez de modo artificial (assevera o autor supramencionado), assim como as tecnologias e técnicas que sistematizam o mundo sensível por intermédio de máquinas que cooperam na produção e distribuição das cópias perfeitas.

Negroponte (1995) aponta uma incoerência no modelo de negócio produzido pela indústria de entretenimento e comunicação. Franklin *et al.* (2013) ressalta que para Foucault não existe tal incoerência, mas uma forma substanciada das relações de poder que compõem a produção do par escassez/abundância. A crise em tela seria um sintoma atribuído aos empreendimentos que perderam seu controle biopolítico, restando apenas artifícios coercitivos oferecidos pelo Estado. Em suma, a mudança no estatuto da escassez danifica o aparelho penal em seu entorno.

Franklin *et al.* (2006), em sua abordagem sobre a bibliografia de Foucault, discute as sociedades disciplinares e de controle e aponta o trânsito de uma sociedade disciplinar para uma sociedade de controle, em que a rigidez hierárquica cede lugar a um modelo acêntrico na gestão do poder.

A fabricação de escassez gerada pela máquina operava, até então, em conformidade com o aparelho estatal, dando lugar a uma posterior dissonância compreendida como mudança da forma discursiva. A perspectiva disruptiva da revolução digital é analisada consoante Franklin *et al.* (2013) como um conflito entre enunciados que, até certa época, sustentaram a indústria da mídia, porém não mais.

A informação se sujeita às instituições que lhe facultam sustento, adjacente ao conceito de dispositivo em que a soberania oferece um obstáculo. Para Deleuze (2006), a soberania se apóia em uma estrutura arraigada ao modelo platônico de simulacro. Logo, capaz de distinguir entre a informação apta a circular pelo território, daquela ilegal onde pelas ordens do soberano deve ser aprisionada e abandonada (FRANKLIN *et al.*, 2013).

2.7.2 Trânsito entre Sociedades

Nota-se um trânsito entre sociedade disciplinar para sociedade de controle ao diferenciar as máquinas produzidas na modernidade e na contemporaneidade. Em estado inicial, a modernidade pode ser associada à sociedade disciplinar, pois, para Foucault (2008), o panóptico corresponde a um modelo arquitetônico hábil em refluir os enunciados para um controle centralizado, no qual os subordinados são observados a partir de um ponto invisível.

Deleuze aborda em sua bibliografia a sociedade de controle, sucessora da disciplinar, na qual o poder é exercido de modo co-extensivo, não havendo necessidade para fórmulas centralizadoras de produção de sentido. O panóptico cede seu lugar ao rizoma de Deleuze, uma rede fluida que promove a libertação dos corpos (FRANKLIN *et al.*, 2013).

A herança institucional para Franklin *et al.* (2013), alcançada na modernidade, evidencia o divórcio entre a máquina contemporânea e a lei, um atrito entre diferentes fantasias ideológicas. A instauração de um novo tipo universal com alta capacidade de estruturação é a grande mudança na estrutura simbólica contemporânea (FRANKLIN; MONTEIRO, 2012).

Quando compreendida como uma circularidade entre os dispositivos compatíveis a informação encontra, na sociedade de controle, consistência fora dos dispositivos que lhe eram familiares na sociedade disciplinar. É importante salientar, ainda congruente à visão de Franklin *et al.* (2013), que a informação possui adesão dos sujeitos contemporâneos por compatibilidade de fantasias ideológicas, entretanto, incoadunável à circulação social autorizada pelas antigas instituições modernas. Logo, as atuais instituições hierarquizadas e disciplinadoras não possuem compatibilidade com os sujeitos contemporâneos. Indícios dessa incompatibilidade podem ser encontrados no ciberespaço e na forma descentralizada com que as redes de computadores se organizam.

A *Internet* é a representação do modelo contemporâneo, que desafia a máquina moderna. Nesse contexto, os motores de busca funcionam como uma retomada do modelo moderno num modelo contemporâneo. Há, assim, uma tensão revelada pelos motores de busca.

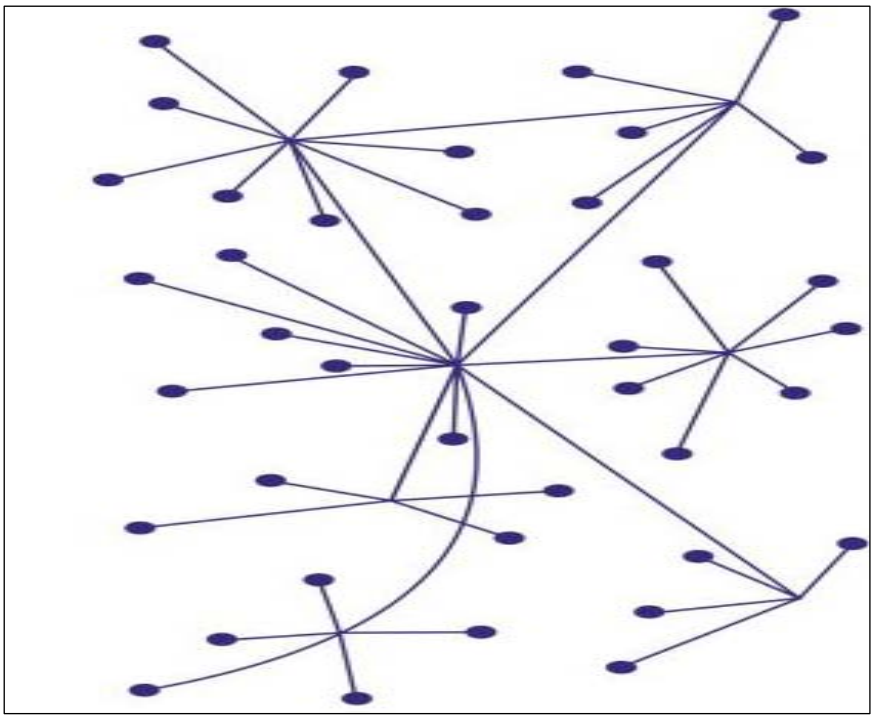
2.8 MOTORES DE BUSCAS COMO *SOFTWARES* LIVRES

No Brasil, o termo “código aberto” é pouco utilizado, assim, neste trabalho, optou-se pelo termo “*software* livre”.

2.8.1 Redes Descentralizadas

Em resposta ao câmbio de sociedades que ocorre com grande veemência na *Internet*, surgem também redes de computadores subservientes ao modelo da sociedade de controle, redes descentralizadas que não obedecem a um modelo de poder hierárquico. A Figura 9 demonstra como é uma rede descentralizada.

Figura 9 – Rede descentralizada



Fonte: Adaptado de Baran (1964, p. *apud* VERGILI, 2012, p. 40)

A Figura 9 mostra que,

> Uma rede descentralizada possui múltiplos *hosts* centrais, cada um com seu próprio conjunto de nós do satélite. Um nó do satélite pode ter conectividade com um ou mais *hosts*, mas não com os outros nós. A comunicação, geralmente, viaja unidirecionalmente dentro de ambas as redes, centralizadas e descentralizadas: a partir do tronco central até as folhas radiais. (VERGILI, 2012, p. 40).

A tecnologia P2P (*Peer-to-Peer*) pode ser conceituada como uma arquitetura de rede de computadores na qual cada ponto (microcomputador) possui características híbridas, isto é, agindo simultaneamente como cliente e servidor. Esta função assegura o compartilhamento de informações dispensando servidores de processamento centralizados.

Gonçalves (2012) explica que em um modelo tradicional de transmissão existem centrais computacionais denominadas “servidores”, cuja função é prover conteúdo aos clientes que delas solicitam dados ou informações.

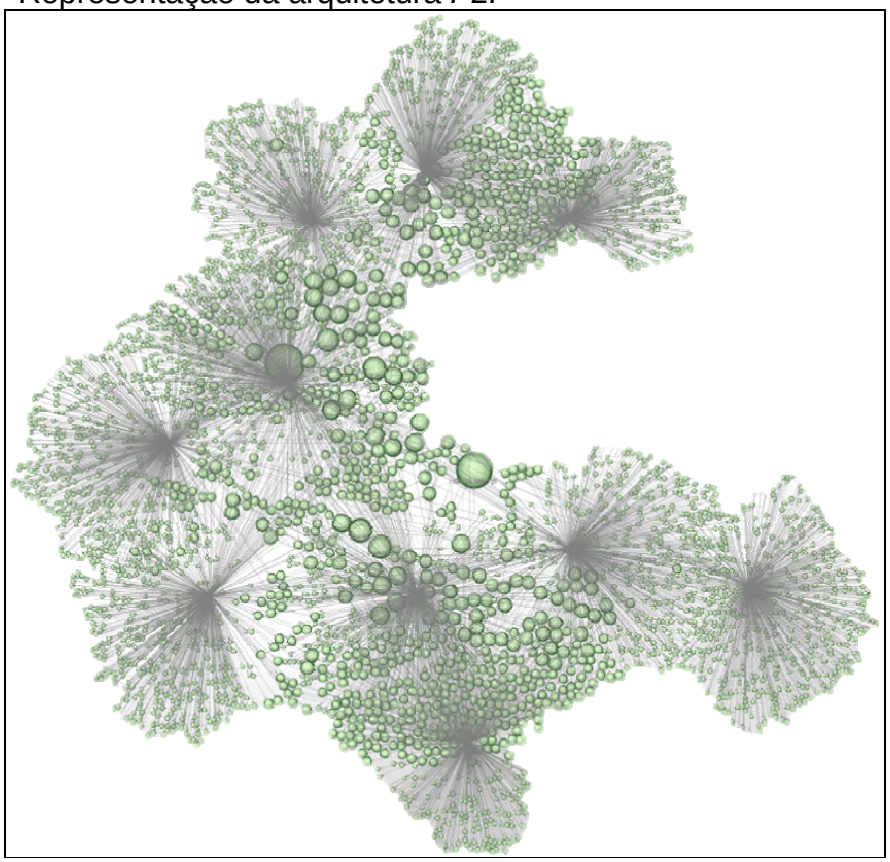
É importante salientar que o modelo retromencionado requer investimentos contínuos em sua infraestrutura, representando, portanto, um grande desafio à medida que o número de clientes aumenta.

Destarte, os integrantes de redes consubstanciadas em tecnologia P2P possuem participação igualitária dividindo tarefas entre si, diferentemente do modelo tradicional ou centralizado cujos servidores são responsáveis por fornecer toda informação requisitada pelos clientes.

De acordo com Pourebrahimi *et al.* (2005), o modelo P2P pode ser classificado em três grupos principais, a saber: (i) puramente descentralizado; (ii) parcialmente centralizado; (iii) híbrido descentralizado.

As arquiteturas puramente descentralizadas possuem a capacidade de auto-organização, permitindo grande escalabilidade e tolerância a falhas. Esses benefícios são alcançados justamente pela ausência de unidades fiscalizadoras, conforme a Figura 10.

Figura 10 – Representação da arquitetura *P2P*



Fonte: o próprio autor

Ao observar a representação gráfica desta arquitetura é possível identificar significativa semelhança com o Rizoma de Gilles Deleuze e Félix Guattari em sua obra Mil Platôs, sistema filosófico que se corporifica por meio do rizoma botânico, cujos brotos se ramificam em qualquer ponto, tornando-se multifuncionais, conforme asseveram, afirmando que,

> Um rizoma não começa nem conclui, ele se encontra sempre no meio, entre as coisas, inter-ser, intermezzo. A árvore é filiação, mas o rizoma é aliança, unicamente aliança. A árvore impõe o verbo "ser", mas o rizoma tem como tecido a conjunção "e... e... e..." Há nesta conjunção força suficiente para sacudir e desenraizar o verbo ser. (DELEUZE; GUATTARI, 1995, p, 63).

Devido, portanto, seu caráter descentralizador, a arquitetura P2P é encontrada em uma ampla gama de *software*s comumente associados ao compartilhamento de informações e dados.

2.8.2 *Software YaCy/Solr*

O *YaCy*, em sua definição principal, pode ser caracterizado como um “motor de busca *Web* descentralizado” que se baseia na tecnologia P2P para disseminar informações entre seus pares. O mecanismo opera de maneira similar aos aplicativos de compartilhamento de arquivos, enviando índices *Web* a todos os usuários do serviço.

Outra função exercida pelo *software YaCy* é constituída pela capacidade de obter informações valiosas por meio da mineração de dados não estruturados. Esta funcionalidade expande, potencialmente, seus benefícios de uso proporcionando ao administrador da ferramenta a possibilidade de analisar e reconhecer padrões de dados de maneira ímpar.

Esse *software* corresponde a uma solução de busca completa que utiliza o Apache *Solr* para armazenar seu índice de busca local. O índice do *Solr* apresenta-se profundamente incorporado ao *YaCy* que herda, por sua vez, diversas funcionalidades derivadas dessa associação.

Concebido em 2007 como projeto subsidiado pela fundação de *software* Apache, o *Solr* constitui um servidor de busca baseado na linguagem de programação Java, idealizado, a princípio, como uma aplicação *Web* que forneceria amplas capacidades de busca textual. Ademais, o *Solr* é uma tecnologia madura e flexível que não oferece apenas um poderoso buscador de texto, mas também diversas funcionalidades como busca por localização geográfica, auto-sugestões, filtros avançados, entre outras.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa “[...] é um procedimento formal, com método de pensamento reflexivo, que requer tratamento científico e se constitui no caminho para se conhecer a realidade ou para descobrir verdades parciais.” (LAKATOS; MARCONI, 2003, p. 155).

Gil (2009, p. 17) também entende a pesquisa como sendo “[...] um processo que envolve inúmeras fases, desde a adequada formulação do problema até a satisfatória apresentação dos resultados.”

Nos sub-tópicos a seguir estão definidos os tipos de pesquisa adotados para o alcance dos objetivos propostos neste estudo.

3.1 TIPO DE PESQUISA

Como abordado por Silva e Menezes (2005), existem diversas formas quanto à classificação de pesquisas científicas, a partir de pontos de vista variados:

a) Ponto de vista da sua natureza;
b) Ponto de vista da forma de abordagem do problema;
c) Ponto de vista de seus objetivos;
d) Ponto de vista dos procedimentos técnicos.

Quanto à natureza, se trata de pesquisa aplicada, ou seja, “[...] aquela que busca gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos.” (SILVEIRA; CÓRDOVA, 2009, p. 35).

Quanto à abordagem, essa pesquisa é qualitativa, pois, segundo Silveira e Córdova (2009, p. 31), “[...] trabalha com fenômenos que não podem ser reduzidos à operacionalização de variáveis.”

De acordo com tais autoras, esse tipo de abordagem apresenta características relacionadas à “[...] hierarquização das ações de descrever, compreender, explicar, precisão das relações entre o global e o local em determinado fenômeno [...].”

Quanto aos objetivos a pesquisa se configura como de caráter exploratório por ter como finalidade o desenvolvimento, esclarecimento e

modificação de conceitos e ideias, no intuito de formular problemas específicos ou hipóteses pesquisáveis para estudos posteriores (GIL, 2009). Também, optou-se por esta pesquisa por oferecer maior familiaridade com o problema, tornando-o explícito.

Por conseguinte, pode ser classificada como pesquisa bibliográfica, a qual, quanto aos procedimentos, corresponde à "pesquisa feita a partir do levantamento de referências teóricas já analisadas, e publicadas por meios escritos e eletrônicos, como livros, artigos científicos, páginas de *Websites* [...]." (SILVEIRA; CÓRDOVA, 2009, p. 37).

3.2 TÉCNICA DE ANÁLISE DOS DADOS

Os dados qualitativos foram analisados a partir da técnica de análise temática por ser considerada a mais simples e apropriada para investigações qualitativas. Tal análise ocorreu em três fases, corroborando com o entendimento de Minayo (2007): (I) pré-análise a partir da organização e exploração do material analisado; (II) codificação, classificação e organização em categorias teóricas e empíricas dos dados; e, (III) tratamento dos resultados por meio da interpretação à luz do quadro.

3.3 TRATAMENTO DOS RESULTADOS

Para que fosse alcançado o objetivo de evidenciar as capacidades de mineração de dados nos motores abertos de busca esse estudo fez uso, como base de análise, do banco de dados da Revista Institucional Informação & Informação que faz parte do *Website*[25] da Universidade Estadual de Londrina – UEL, a qual coleta material publicado por membros da Universidade em bases de dados de indexação.

A revista Informação & Informação é um periódico científico com publicação desde 1996, e tem como objetivo "[...] disseminar a informação científica na área da Ciência da Informação e difundir o diálogo intelectual entre pesquisadores, profissionais e estudantes que atuam em diferentes regiões do país e no exterior." Está vinculada ao Programa de Pós-Graduação em Ciência da

[25] <http://www.uel.br>

Informação da UEL e disponível em *Open Access* no Sistema Eletrônico de Editoração de Revistas (SEER) (REVISTA INFORMAÇÃO & INFORMAÇÃO, 2013).

Atualmente, sua periodicidade é quadrimestral, e apresentou *Qualis* CAPES B1 em 2015. Está indexado nas bases *Library and Information Science Abstracts* (LISA), *Directory of Open Acess Journals* (DOAJ), *Ulrich's Periodicals Directory*, Sistema regional de informaciónen línea para revistas científicas de América Latina, el Caribe, España y Portugal (LATINDEX), Información y Bibliotecología Latino Americana (INFOBILA), e *Public Knowledge Project's Metadata Archive* (PKP).

A indexação em importantes bases de informação científica proporciona visibilidade, mas a indexação local em *software* livre tem conotação particular ao periódico. Além da interface amigável ao usuário, também possibilita a interatividade e identificação das preferências de busca, visando melhorias no sistema.

Le Coadic (1996, p. 89) exalta que “[...] a revista possui qualidades: valida as prioridades, serve de repositório dos trabalhos científicos e os torna públicos.” Nesse sentido também Meadows (1999) destaca a importância do periódico e seu impacto no fluxo da comunicação científica, além de ressaltar que sua organização estrutural traz ao leitor facilidades de acesso e compreensão.

Desse modo, entende-se que os periódicos científicos são fundamentais para o desenvolvimento da ciência. Portanto, uma das formas de favorecer esse desenvolvimento é organizar melhor a informação produzida para que o usuário possa localizá-la com mais eficiência e rapidez.

Assim, optou-se pelo *YaCy/Solr*, *software* escolhido para a implantação do servidor de busca na revista retromencionada, tendo em vista a viabilidade técnica de sua execução.

4 RESULTADOS EXPERIMENTAIS

Os sub-tópicos a seguir evidenciam a descentralização da *Internet* e seus movimentos homônimos analisando tendências rizomáticas no mecanismo de busca rumo à despolarização de suas bases de dados.

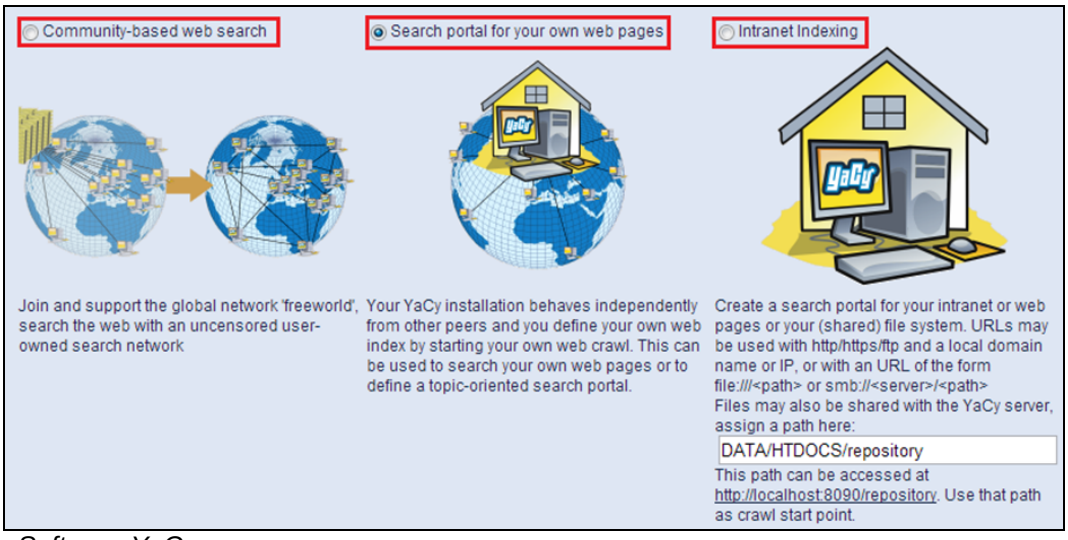
A análise experimental empreendida se deu por meio da simulação de uma estratégia de mineração de dados utilizando o *software YaCy/Solr*.

4.1 INDEXAÇÃO COM O *SOFTWARE YACY/SOLR*

Neste estudo, o *YaCy*/*Solr* foi o *software* escolhido para a implantação do servidor de busca por se tratar de um mecanismo de busca com a funcionalidade de *crawler*, o que proporciona a capacidade de obter todas as páginas *Web* apontadas pelo utilizador. Entretanto, para que essa função trabalhe de maneira apropriada são necessárias configurações específicas no *software*.

Inicialmente, é necessário escolher o perfil de operação do *software* (Figura 11).

Figura 11 – Escolha do perfil de operação do *software*



Fonte: *Software YaCy*

O propósito de uso do *software* é fator determinante nessa escolha, ao apresentar três opções principais de funcionamento:

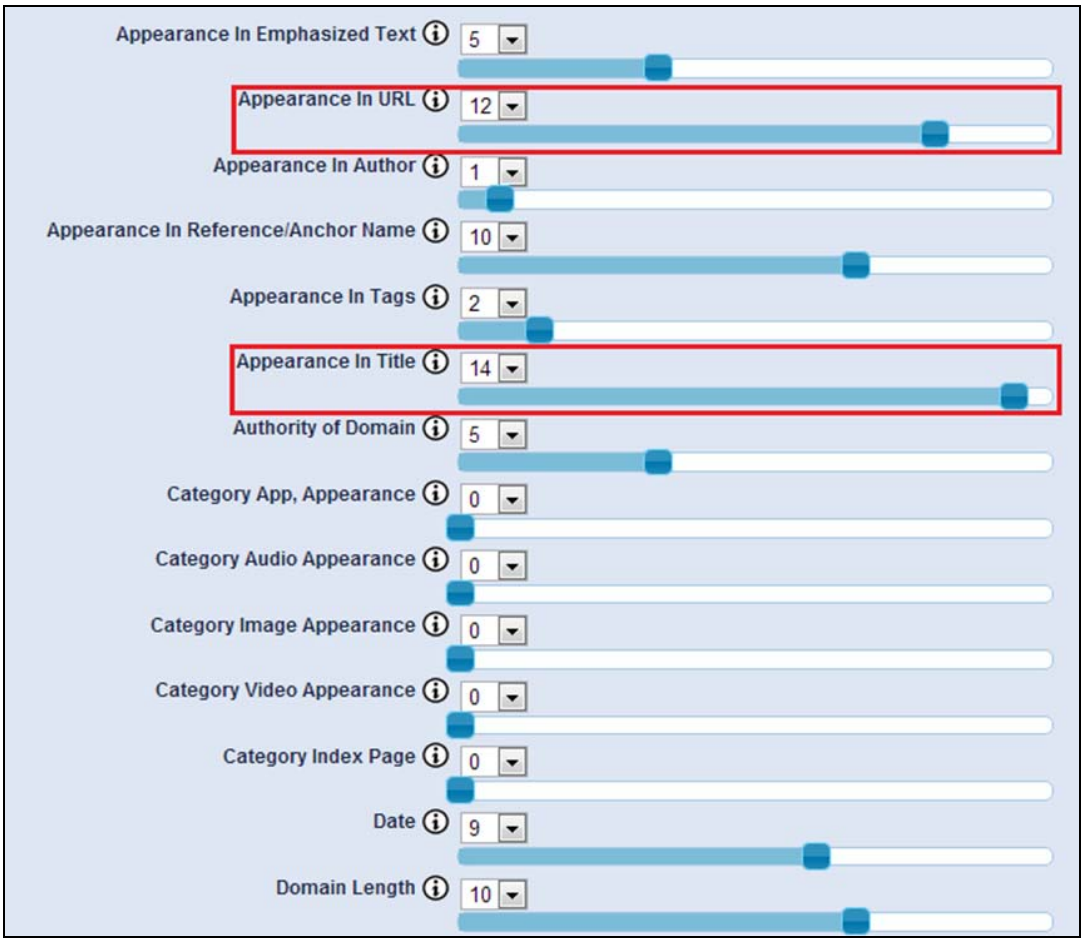
1) *Community-based Web search*: insere o servidor *YaCy/Solr* recém-criado em uma rede global livre de censura denominada *freeworld*, replicando seu índice e disponibilizando-o para consultas futuras;
2) *Search portal for your own Web pages*: apresenta em sua base de dados apenas conteúdo indexado pelo usuário, cujo funcionamento ocorre de maneira independente à rede de busca global (*freeworld*). É comumente utilizado na criação de portais de busca orientados por assunto;
3) *Intranet Indexing*: empregado na concepção de buscadores para *intranet* pode trabalhar de maneira integrada com servidores de troca de arquivo, o que beneficia organizações que possuem grande quantidade de documentos dispostos de maneira não estruturada.

Este aplicativo proporciona o rastreamento e indexação do conteúdo de forma descomplicada ao usuário, possibilitando conhecer e modificar todo o processo de pesquisa.

Admite-se, também, a possibilidade de modificar a ordem dos resultados, conferindo ênfase diferenciada às palavras com ocorrência em trechos distintos de um documento.

A Figura 12 apresenta o sistema de *ranking*, no qual é possível observar uma atribuição numérica em seus campos. Cada opção possui influência direta na ordenação do resultado pesquisado.

Figura 12 – Sistema de *ranking* do buscador *YaCy*/*Solr*



Fonte: *Software YaCy*

Para melhor compreensão do mecanismo de *ranking* é notado que os campos *Appearence In URL* (responsável pela incidência de palavra buscada no endereço Web) e *Appearance In Title* (responsável pela incidência de palavra buscada no título do documento) possuem os numerais inteiros 12 e 14, respectivamente. Logo, o arranjo de classificação priorizará resultados nos quais a palavra buscada ocorra no título do documento e, posteriormente, no endereço URL.

Desta forma, alguns dos principais problemas na utilização de soluções corporativas como a censura do conteúdo e a classificação obscura dos resultados podem ser evitados.

Definido como “buscador distribuído de código aberto”, o projeto *YaCy*/*Solr* foi desenvolvido utilizando como base para sua implementação a arquitetura de redes P2P, e tem seu núcleo codificado por meio da linguagem de programação Java, o que confere maior escalabilidade entre sistemas operacionais distintos. Encontra-se disponível em distribuições para *Windows*, *Linux* e *OSX*. Para

os desenvolvedores do projeto (*YACY*, 2011), este *software* possui como grande diferencial a utilização de um modelo baseado na tecnologia P2P para transferência de arquivos.

Conforme Gonçalves (2012), em um modelo de transmissão tradicional o servidor fica responsável por prover todo conteúdo aos clientes solicitantes. Todavia, esta topologia pode representar grandes desafios à medida que o número de clientes aumenta. Modelos de transmissão fundamentados na tecnologia de redes P2P proporcionam uma capacidade híbrida, em que cada nó (usuário) poderá atuar tanto como cliente, quanto como servidor.

As arquiteturas puramente descentralizadas são capazes de se auto-organizar e possuem a mesma importância na rede. Essa topologia permite grande tolerância a falhas e escalabilidade do sistema, pois, devido à inexistência de uma unidade centralizadora, qualquer nó pode ser compensado por outro nó na rede.

A rede *YaCy*/*Solr* detém uma arquitetura descentralizada, o que denota a ausência de um servidor de processamento central e confere direitos iguais a todos os pontos de acesso (*YaCy*-peer's). Concebido para ser executado em computadores pessoais, assim que a instalação do aplicativo é concluída, este é considerado um *YaCy*-peer ou ponto de acesso *YaCy*, tornando-se, então, viável utilizar os recursos disponíveis na ferramenta.

Cada *YaCy*-peer possui a capacidade de varrer (*crawl*) a *Internet*, analisar e indexar páginas *Web* encontradas, bem como armazenar o resultado desta tarefa em bancos de dados comuns (chamados de índices). Os índices são compartilhados com outros *YaCy*-peer's utilizando os princípios das redes P2P.

Para que seja concebível acessar as funcionalidades do programa, existe um servidor Web, que é executado junto ao *software*, cuja função é prover uma interface de busca. Termos de pesquisa podem ser inseridos, tendo os resultados exibidos em formato similar a outros buscadores.

Em sua arquitetura, o *YaCy*/*Solr* se baseia nos quatro elementos principais apresentados (*crawler, indexer*, *web interface* e *data storage*).

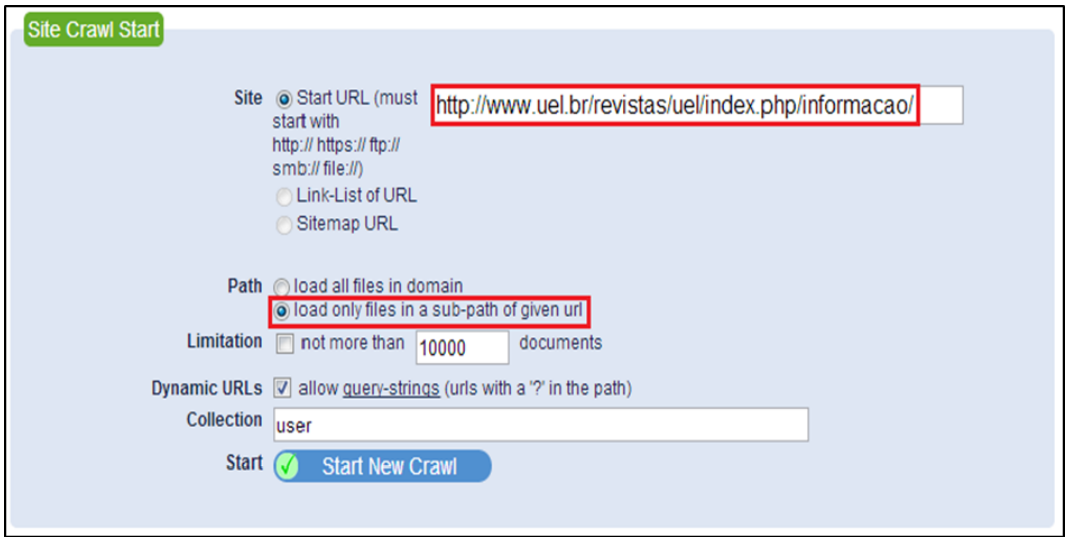
O uso de modelos de busca descentralizados, além dos benefícios já citados, proporciona uma drástica redução no consumo de energia destinada a abastecer servidores de busca e indexação, visto que todo o processamento passa a ocorrer no computador do usuário.

Conforme os desenvolvedores do projeto, se apenas um entre mil usuários da *Web* operasse como indexador de conteúdo, este montante já seria capaz de substituir completamente os portais de pesquisa centralizados (YACY, 2012).

Para as atividades de indexação são necessários alguns passos de configuração do *software*.

Primeiramente, é essencial apontar o endereço *Web* correto por meio da opção "*Site*" (Figura 13), selecionando a opção de carregar apenas os subdiretórios do endereço informado. Desta forma, somente as páginas *Web*, hierarquicamente a seguir, serão indexadas.

Figura 13 – Configurações iniciais para a indexação do *Website*



Fonte: *Software YaCy*

Ao final da indexação, realizada de maneira autônoma, o *software* identificou um total de 1982 *URL's* disponíveis para busca. Esta soma reflete o montante total de páginas *Web* publicadas pelo periódico em seu *Website*.

A partir da navegação nos diretórios indexados foi possível observar que a revista (alvo da indexação) conta com 412 artigos publicados em seu subdiretório de *downloads* até o momento, todos disponíveis para serem copiados de maneira individual.

O *software* disponibiliza uma interface semelhante aos buscadores comerciais como o *Google*, *Yahoo!* e *Bing*, o que possibilita fazer testes após os procedimentos para indexação.

A Figura 14 apresenta um exemplo de pesquisa que conta com uma ampla gama de customizações, busca por imagens, nuvem de *tags*, entre outros benefícios para o usuário.

Segundo Fragoso (2007, p. 17) os mecanismos de busca comerciais não disponibilizam qualquer tipo de customização em sua interface gráfica ou mesmo em seu algoritmo de classificação, visto que operam com base em critérios estritamente comerciais que são mantidos em total sigilo.

Figura 14 – Exemplo de tela de pesquisa no *YaCy/Solr*

Fonte: *Software YaCy*

Além de possuir uma interface *Web* que proporciona experiência de uso amigável ao utilizador, é possível, mediante código *HTML* fornecido pelos próprios desenvolvedores do projeto, efetuar a inclusão de uma caixa de busca em qualquer *Website* (Figura 15).

Figura 15 – Código *HTML* para a inclusão de caixa de busca

```
<form method="get" accept-charset="UTF-8" action="http://127.0.0.1:8090/yacysearch.html">
  <div style="text-align:center; padding:5px; backgrounnd-color:#eeeeee; border:1px solid
  #cccccc; -webkit-border-radius:5px; -moz-border-radius:5px; border-radius:5px;
  display:block; float:left; margin-right:5px;">
    <div style="font-family:Arial,Helvetica,sans-serif; font-size:16px; display:block;
    float:left; padding-top:3px; padding-right:5px;">
      MySearch
    </div>
    <input type="text" name="query" value="" maxlength="80"
           style="width:300px; font-size:16px; float:left;" />
    <input type="hidden" name="verify" value="cacheonly" />
    <input type="hidden" name="maximumRecords" value="10" />
    <input type="hidden" name="meanCount" value="5" />
    <input type="hidden" name="resource" value="local" />
    <input type="hidden" name="urlmaskfilter" value=".*" />
    <input type="hidden" name="prefermaskfilter" value="" />
    <input type="hidden" name="display" value="2" />
    <input type="hidden" name="nav" value="all" />
    <div style="font-size:16px; display:block; float:right; padding-top:1px;">
      <input type="submit" name="Enter" value="Search" />
    </div>
  </div>
  <p style="clear:both;"></p>
</form>
```

This would look like:

MySearch [ ] Search

Fonte: *Software YaCy*

Este recurso possibilita adicionar a opção de pesquisa do *Website* indexado em outros *sites* relacionados ou de interesse, por exemplo, na instituição em que está vinculado, grupo de pesquisa, e outro que se fizer adequado.

4.2 MINERAÇÃO DE DADOS COM O *SOFTWARE YACY/SOLR*

Focando, especificamente, nos filtros avançados disponibilizados pelo projeto *Solr*, uma nova gama de capacidades no contexto da DM é oferecida ao *YaCy*/*Solr*, sendo que, com eles, o *software* pode retornar dados conclusivos sobre a massa de dados indexada.

Consultas podem ser feitas ao *Solr* utilizando o protocolo HTTP mediante qualquer *browser* através do parâmetro “q=” (*query*), dispondo o resultado dos dez primeiros documentos na tela (Figura 16).

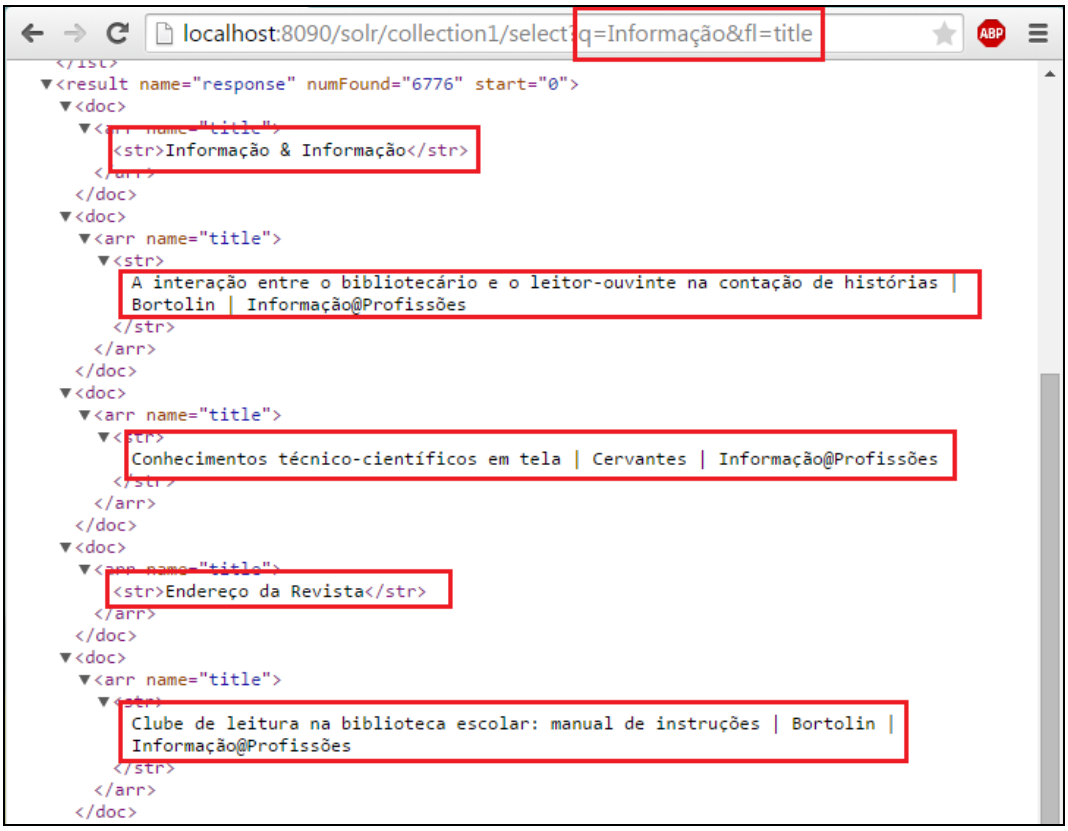
Figura 16 – *Query* no *YaCy/Solr*

Fonte: *Software YaCy*

No exemplo, quando a palavra “Informação” é adicionada através do parâmetro de busca “q=” observa-se um retorno de 6776 documentos contendo essa palavra, os quais podem ser arquivos de texto, nomes ou metadados de figuras, páginas *HTML* entre outros.

Outro importante parâmetro ao filtrar dados no *Solr* é o comando “fl=” (*fieldlist*) usado, em geral, para restringir os campos exibidos na tela de resultados, que podem ser futuramente exportados para arquivos em ampla gama de extensões. Como exemplo, a Figura 17 lista apenas o título dos documentos encontrados pela *query* “Informação”.

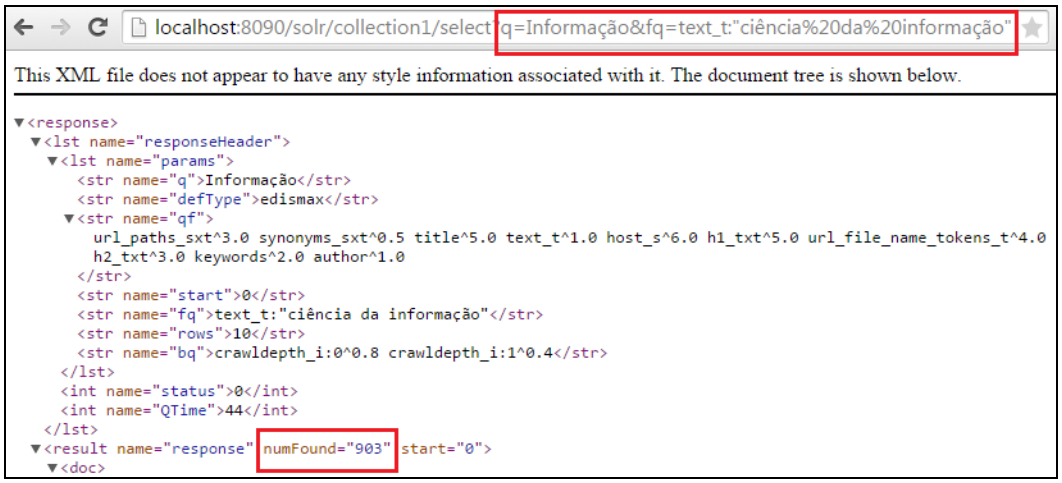
Figura 17 – Parâmetro fl inserido na *query* “Informação”



Fonte: *Software YaCy*

Qualquer informação pode ser listada através desse parâmetro estendendo suas funcionalidades a conjuntos de campos separados por vírgula ou espaço. O parâmetro “fl=” planifica o retorno da busca, entretanto não altera a quantidade de documentos encontrados. Com o intuito de afunilar o contexto de busca o parâmetro “fq=” (*filter query*) pode ser introduzido. Esse parâmetro define restrições no retorno da busca quando associado a conjuntos de metadados. A Figura 18 exemplifica a função do comando *filter query*.

Figura 18 – Parâmetro *filter query* afunilando a busca



Fonte: *Software YaCy*

Disposto na figura anterior, o parâmetro “fq=” em conjunto com o metadado “text_t” restringe a busca da *query* para 903 resultados. Nessa consulta, apenas documentos detentores das palavras “ciência da informação” no corpo do texto principal serão contabilizados, tendo em vista que os documentos retornados precisam conter também a palavra “Informação”, especificada pelo parâmetro “q=”. Em condições adequadas os filtros e listas do projeto *Solr* constituem uma ferramenta valiosa que, atrelada às funcionalidades do *YaCy,* fundamentam um meio singular de aquisição no âmbito da mineração de dados.

4.3 Experimento: Explorando a Mineração de Dados

Como ponto principal no experimento de mineração, atentou-se em verificar a incidência de um conceito no decorrer das edições publicadas pela revista alvo da indexação, desenvolvendo o processo metodológico adjacente.

Ao iniciar o processo, e visando um índice reduzido de falsos positivos na pesquisa, optou-se pelo *download* de todos os artigos publicados pela revista desde a sua concepção em 1996, entretanto a obtenção automatizada dos 412 documentos presentes em seu diretório de acesso público apresentou peculiaridades que serão descritas no decorrer do experimento.

Com o intuito de agilizar a obtenção dos arquivos de texto hospedados pela revista, fez-se necessária a utilização do *software GNU Wget,*

cujas funcionalidades incluem a aquisição de arquivos em massa por meio do protocolo de transferência de hipertexto[26] (HTTP).

O *GNU Wget* é um programa gratuito comumente utilizado em sistemas operacionais Linux, com a capacidade de realizar *download* recursivo em *Website*s. O *software* também pode ser utilizando em sistemas operacionais Windows (GNU, 2014). Para obter de maneira automatizada os documentos almejados pelo experimento optou-se pelo seguinte comando:

Quadro 3 - Linha de comando

wget –r –P C:/ -np http://www.uel.br/revistas/uel/index.php/informacao/

Fonte: o próprio autor

As opções dispostas na linha de comando supracitada dizem respeito aos parâmetros informados ao programa que habilitam a capacidade de recuperação recursiva e impedem que se estenda aquém do diretório corrente. Essa configuração permite buscar arquivos a partir do diretório "../informação/", que se qualifica como diretório raiz da revista Informação & Informação.

O número de documentos extraídos totaliza quatrocentos e doze compondo a massa de dados alvo do experimento, que deve ser indexada posteriormente utilizando o *software YaCy*/*Solr* através da opção *Intranet Indexing*.

Após executar a indexação automatizada é possível verificar através do parâmetro "*numFound*", na aba *Solr Default Core,* que 412 documentos foram inseridos à base de dados.

Com a massa de dados devidamente inserida no sistema dá-se início ao processo de mineração dos textos contidos nos arquivos, que é provido pelo software Apache *Solr*, considerada uma popular plataforma de busca responsável pela indexação e recuperação do texto alvo.

O objetivo do experimento é verificar a incidência do termo "Gestão da Informação" no decorrer dos anos nas edições da revista indexada. Para tal, é necessária a confecção de *queries* em linguagem de marcação XML (*Extensible*

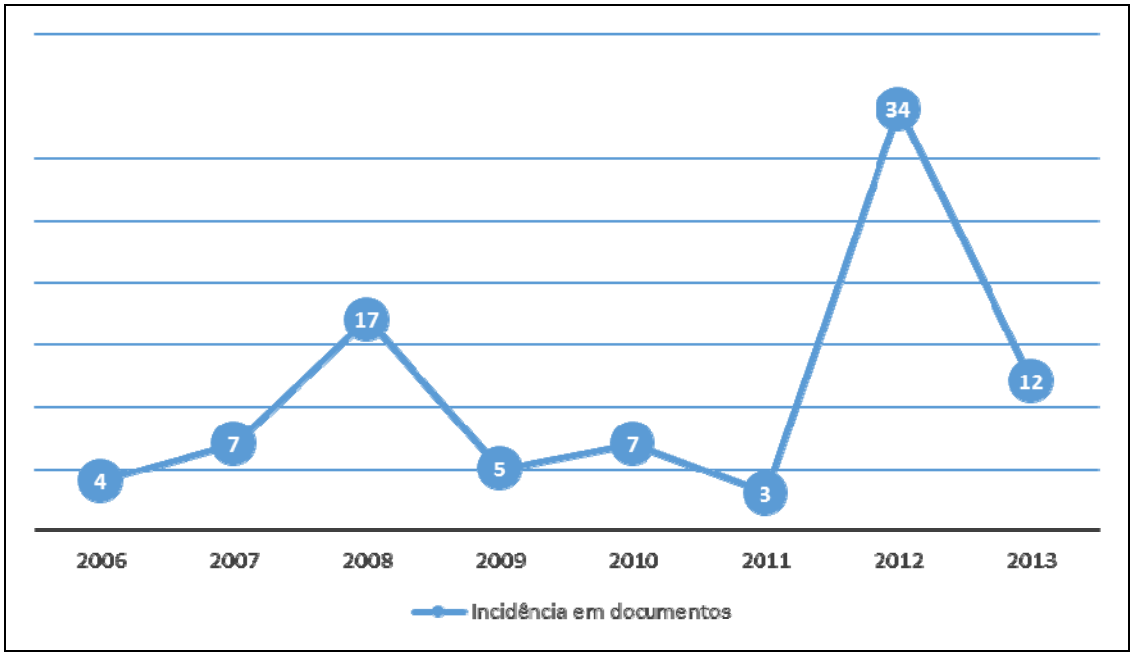
[26] Os vínculos de hipertexto são uma rica fonte de informações a ser explorada, pois dentre outras coisas, ajudam no processo de ranqueamento de páginas pelos motores de busca e na identificação de micro-comunidades na *Web*.

Markup Language) que devem apresentar em intervalos anuais o número de documentos cujo termo buscado se faz presente.

Através da *query* abaixo foram obtidas as informações necessárias para construir o Gráfico 1 que apresenta os resultados obtidos após a aplicação das *queries* em *XML*. Insta salientar que a questão em foco diz respeito ao número de documentos cujo termo se faz presente, e não o número de ocorrências do termo em cada documento.

"http://localhost:8090/solr/collection1/select?q=*:*&defType=edismax&start=0&rows=9999&fq=text_t:"gestão da informação"&fq=last_modified:[2006-01-01T00:00:00Z%20TO%202007-01-01T00:00:00Z]".

Gráfico 1 – Termo "Gestão da Informação" nos Documentos Pesquisados



Fonte: o próprio autor

Verifica-se, então, a ocorrência do termo buscado a partir do ano de 2006 com 4 documentos listados, evoluindo para 17 incidências em 2008 e atingindo um pico de 34 resultados positivos em 2012[27].

O experimento concluso permite verificar o comportamento de um conceito em relação ao tempo, seu crescimento e como se correlaciona com demais

[27] No momento em que essa pesquisa foi encerrada (set./out. 2014) a revista não havia publicado o semestral de 2014, motivo pelo qual os dados utilizados limitaram-se ao ano de 2013.

eventos, mostrando-se uma formidável técnica para coleta e devido ordenamento de dados, concedendo-os relevância e propósito.

No entanto, o experimento em si não reflete o universo de possibilidades onde encontra-se a técnica, visto que a maneira de se conduzir o experimento oferece maior produção de conhecimento científico que o próprio resultado.

Evidenciou-se que técnicas de mineração de dados baseadas em *queries* podem ser utilizadas em uma grande variedade de pesquisas destinadas a inúmeros fins.

É importante ressaltar, nesse contexto, o empoderamento do usuário que pode gerar informações estatísticas a partir do *corpus* indexado pelo mecanismo de busca aberto, algo impensado nos mecanismos de busca convencionais e, também, a capacidade de alterar a estratégia de indexação por meio do algoritmo de "*ranking*" retornando buscas que, verdadeiramente, atendem aos interesses do utilizador, e não visam obter lucro para uma empresa.

Por fim, técnicas de DM, baseadas em buscadores, constituem uma interessante forma de manipular conteúdo semiestruturado e não estruturado (porção majoritária da Web), sendo que este experimento apresenta uma maneira para a extração de informação em meio a extensas massas de dados.

5 CONSIDERAÇÕES FINAIS

A dissertação apresentada avocou em seu objetivo primário explorar o motor de busca baseado em *software* livre *YaCy*/*Solr* como estratégia de indexação e mineração de dados, contrastando-o com os principais buscadores comerciais na *Web*.

Com intuito de compreender o *software* conduziu-se uma exploração abordando suas principais funcionalidades, tendo em vista a carência de documentações e referências.

Realizou-se, em primeiro lugar, uma revisão da literatura iniciando a partir do histórico da *Internet*. A Arpanet justificava-se como meio para assegurar a comunicação entre vários centros de pesquisa norte-americanos onde as informações em trânsito na rede poderiam ser descentradas evitando perda de dados.

Por conseguinte, as subseções do referencial teórico abordam o surgimento das Web's, bem como do movimento de *software* livre que reforçam o cerne descentralizador e colaborativo da rede mundial de computadores.

A subseção 2.4 trata acerca dos sistemas de recuperação de informação oferecendo aporte posterior subseção sobre motores de busca onde a dissertação apesenta maior ênfase.

Os motores de busca comerciais facilitam o acesso ao grande conteúdo disponível na Web, trazendo (supostamente) resultados que satisfaçam as necessidades do utilizador.

É notado que o ímpeto descentralizador nasceu com o projeto Arpanet, trazendo posteriormente as Web's como reflexo evolutivo dessa qualidade. De maneira oposta, os motores de busca comerciais apresentam tendências de aglomerar sua massa de dados indexada, decidindo sua estratégia de indexação sem interferência externa.

Observa-se uma tensão entre os motores de busca centralizadores (comerciais) e os descentralizadores (*YaCy*/*Solr*), partindo do pressuposto que a *Internet* foi originalmente concebida de modo a descentralizar a informação.

Os buscadores centralizadores vão contra o ímpeto distribuído da *Internet* impondo uma sociedade disciplinar onde a narrativa aponta para uma

sociedade de controle. Como alternativas aos modelos de indexação centralizados o trabalho apresenta o *YaCy*/*Solr*.

O *software YaCy*/*Solr* se mostrou produtivo por possibilitar reunir em uma página local todo o conteúdo indexado. O administrador desse *software* tem autonomia para aperfeiçoar nuances de pesquisa, selecionando elementos que evidenciem a atuação de seus usuários, podendo, inclusive, contribuir para o gerenciamento e melhorias no serviço de disponibilização da informação.

A indexação distribuída não possui a mesma visibilidade das bases de dados internacionais, entretanto coaduna com o modelo rizomático da *Internet*, proporcionando aos buscadores libertação do modelo disciplinar.

O *software YaCy*/*Solr* contempla a discussão teórica supramencionada uma vez que contrasta com os modelos tradicionais de buscadores e apresenta uma alternativa viável para indexação compartilhada adequando-se ao modelo fluido da *Internet*. Contudo, a exploração dos mecanismos de busca, em especial daqueles com propósito de busca distribuída, não se finda.

Ademais, o *software* proporcionou ampla gama de resultados no âmbito da mineração de dados, evidenciado pelo experimento apresentado. Porém, é importante salientar que os resultados do experimento não configuram a real contribuição do *YaCy*/*Solr*, mas, sim, a capacidade que o *software* dispõe para realizar tais experimentos, sejam eles concentrados em qualquer gama de documentos ou metadados.

É fundamental que a partir do *YaCy*/*Solr* obteve-se acesso a estratégia de indexação como um diferencial para produzir informações que os motores de busca proprietários não oferecem.

As estratégias que podem ser utilizadas na formulação de buscas devem ser tema de forma mais recorrente na Ciência da Informação, dada sua importância acadêmica e social, principalmente pelo aumento da demanda de informações disponibilizadas na *Internet*, evitando, com isso, a recuperação de documentos não relevantes e o desperdício de tempo.

Por fim, o trabalho constituiu um contributo para o conhecimento acerca dos mecanismos de busca abertos. Conteúdo didático em forma de vídeo-

aula foi disponibilizado[28] no decorrer dos estudos para a confecção dessa dissertação.

Futuras investigações poderão incluir versões mais precisas do experimento realizado, assim como a utilização de novos metadados e *filter queries*. Apesar das limitações apresentadas na exposição do *software*, considera-se que o trabalho proporcionou uma visão diferenciada e pouco ortodoxa acerca dos mecanismos de busca e da mineração de dados, visão esta que poderá constituir um ponto de partida para futuras pesquisas da área.

[28] <https://www.youtube.com/watch?v=EkQ9AEh1h2I>; <https://www.youtube.com/watch?v=yBeqwUYEWec>; <https://www.youtube.com/watch?v=9-ikmL6YwgI>.

REFERÊNCIAS

AGUIAR, V. M. (org.). ***Software livre*, cultura *hacker* e o ecossistema da colaboração**. São Paulo: Momento Editorial, 2009.

AMBINDER, D. M.; MARCONDES, C. H. As potencialidades da *Web* semântica e *Web* 2.0 para a ciência da informação e os novos formatos de publicações eletrônicas para a pesquisa acadêmico-científica. **Revista EDICIC**, v. 1, n. 4, 2011.

ANDERSON, C. **Free**: the future of a radical price. New York: Hyperion 2009.

ANDRÉ, F. **scientific research output**: from gutenberg to the *web*. In: _____. Apostila. 2005.

APGAUA, R. O Linux e a perspectiva da dádiva. **Horizontes antropológicos**, Porto Alegre, ano 10, n. 21, p. 221-240, jan./jun. 2004.

ARAÚJO, J.P. **Invisível, oculta ou profunda?:** a *Web* que poucas ferramentas enxergam. 2012. Disponível em: <www.comunicar.pro.br/artigos/*Web*oculta.htm>. Acesso em: 22 fev. 2015.

ARAÚJO, V. M. R. H. Sistemas de informação: nova abordagem teórico-conceitual. **Ciência da Informação**. v. 24, n. 1, 1995.

ASCENSÃO, C. P. **Como funciona um motor de busca?** 2015. Disponível em: <http://www.portalwebmarketing.com/MotoresdeBusca/Comofuncionaummotordebusca/tabid/435/Default.aspx>. Acesso em: 28 jul. 2015.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de informação:** conceitos e tecnologia das máquinas de busca. São Paulo: Bookman, 2011.

BAPTISTA, A. *et al.* Comunicação científica: o papel da *open archives initiative* no contexto do Acesso Livre.**EncontrosBibli:**Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, n. esp., jan./jul. 2007. Disponível em: <http://journal.ufsc.br/index.php/eb/article/view/1518-2924.2007v12nesp1p1/435>. Acesso em: 10 fev. 2013.

BATTELLE, J. **A busca**. Campinas: Campus; Rio de Janeiro: Elsevier, 2006.

BAX, M. P. A evolução da *Web* rumo à *Web* Semântica. **Revista Prisma.com**, n. 19, 2012.

BERGMAM, M K. White paper: the deep *Web* surfacing hidden value. **Journal of Eletronic Publishing**, v. 7, n. 1, 2001.Disponível em:<http://quod.lib.umich.edu/cgi/t/text/textidx?c=jep;view=text;rgn=main;idno=3336451.0007.104>. Acesso em: 23 fev. 2015.

BERNERS-LEE, T. **O físico Tim Berners-Lee, o "pai da www", prevê novo salto da rede com intercâmbio maior de dados.** 2006. Disponível em:<http://veja.abril.com.br/especiais/tecnologia_2006/p_040.html>. Acesso em: 25 fev. 2015.

BRANSKI, R. M. Recuperação de informações na *Web*. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 9, n. 1, p. 70-87, jan./jun. 2004. Disponível em:<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/351>. Acesso em: 23 fev. 2015.

BRODER, A. **A taxonomy of *Web* search**. Special Interest Group on Information Retrieval, 2002. Disponível em: <http://www.sigir.org/forum/F2002/broder.pdf>. Acesso em: 02 jun. 2014.

CAMILO, C. O.; SILVA, J. C. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. **Relatório Técnico**, Instituto de Informática. Universidade Federal de Goiás, ago. 2009.

CAMPOS*a*, A. **O que é software livre**. BR-Linux. Florianópolis, março de 2006. Disponível em <http://br-linux.org/linux/faq-softwarelivre>. Acesso em: 15 jul. 2015.

CAMPOS*b*, A. **O que é Linux**. BR-Linux. Florianópolis, março de 2006. Disponível em: <http://br-linux.org/linux/faq-linux>. Acesso em: 14 jul. 2015.

CARDOSO, O. N. P. **Recuperação de informação**. Departamento de Ciência da Computação - Universidade Federal de Lavras, 2003. Disponível em: < http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>. Acesso em: 14 jul. 2015.

CASTELLS, M. **A Galáxia *Internet*:** reflexões sobre a *Internet*, negócios e a sociedade. Zahar, 2003.

CENDÓN, B. Ferramentas de busca na *Web*. **Ci. Inf**., Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001. Disponível em: <http://www.scielo.br/pdf/%0D/ci/v30n1/a06v30n1.pdf>. Acesso em: 03 jul. 2014.

DELEUZE, G. ; GUATTARI, F. **Mil Platôs. Capitalismo e esquizofrenia**. v.1. Trad. Aurélio Guerra neto e Celia Pinto Costa. Rio de Janeiro: Ed. 34, 1995.

DELEUZE, G. **Lógica do sentido**. São Paulo: Perspectiva, 2006.

DUCK DUCK GO. **About Duck Duck Go**. 2015. Disponível em: <https://duckduckgo.com/about>. Acesso em: 20 jul. 2014.

EVANGELISTA, R. Política e linguagem nos debates sobre software livre. In: AGUIAR, V. M. (org.). ***Software livre*, cultura *hacker* e o ecossistema da colaboração**. São Paulo: Momento Editorial, 2009.

EVERETT, C. **Moving across to the *Dark* side**. Network Security, set. 2009.

FERNEDA, E. **Recuperação de informação**: análise sobre a contribuição da ciência da computação para a ciência da informação. 147p. Tese (Doutorado em Ciências da Computação). Universidade de São Paulo, São Paulo, 2003.

FOUCAULT, M. **Vigiar e punir nascimento da prisão**. Petrópolis: Vozes, 2008.

FRAGOSO, S. Quem procura acha? O impacto dos buscadores sobre o modelo distributivo da World Wide *Web*. R**evista de Economía Política de lasTecnologías de laInformación y Comunicación**, v. 9, n. 3, set./dez. 2007. Disponível em: <http://seer.ufs.br/index.php/eptic/article/view/255/245>. Acesso em: 14 jul. 2013.

FRANCO, D. P. Deep *Web*: mergulhando no sub-mundo da *Internet*. **Revista Segurança Digital**, n. 10, abr. 2013.

FRANKLIN, B. *et al.* **Informação Ilegal**: o divórcio entre a máquina e a lei. Encontro Nacional de pesquisa em Ciência da Informação – ENANCIB, 14, 2013. Disponível em:http://enancib2013.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/126/174 Acesso em: 20 jan. 2015.

_____. Neutralidade da Rede: crise da mídia e sociedade de controle*.* In: **Terceira edição do Congresso *ONLINE* do Observatório para a Cibersociedade**, 2006. Conhecimento aberto, sociedade livre, 2006. Disponível em: <https://docs.*Google*.com/file/d/0BxBTIRSbeQn_VzRGZkd4a2xmZm8/edit>. Acesso em: 22 jan. 2015.

FRANKLIN, B.; MONTEIRO, S. **Por uma economia do sentido**. Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB, 13, 2012.

FUENTES, M.; ORDUÑA, O. **SEO**: cómo triunfar en buscadores. Madrid: ESIC, 2010. Disponível em: <http://books.*Google*.com.br/books?hl=pt-BR&lr=&id=rJ0bIFsEcjcC&oi=fnd&pg=PA15&dq=buscadores&ots=b2ZNv1b73d&sig=9W119OSAh0j3qF81-4XFVxwjwkk#v=onepage&q=buscadores&f=false>. Acesso em: 25 jul. 2013.

GALVÃO, N. D.; MARIN, H. de F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paul Enferm**., v. 22, n. 5, p. 686-90, 2009.

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2008.

_____. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2009.

GIL-LEIVA, I.; ALONSO-ARROYO, A. **Keywords given by authors of scientific articles in database descriptors**. J. Am. Soc. Inf. Sci**.**, v. 58, p. 1175–1187. 2007. Disponível em: <http://*online*library-wiley-com.ez78.periodicos.capes.gov.br/doi/10.1002/asi.20595/abstract>. Acesso em: 22 jul. 2013.

GNU. **Manual.** Disponível em: <http://www.gnu.org/*Software*/wget/manual/wget.html>. Acesso em: 07 de nov. de 2014.

GONÇALVES, K. C. **Caracterização das propriedades dinâmicas da rede sobreposta em uma aplicação de transmissão par-a-par de vídeo ao vivo**, 2012. 66 f. Dissertação (Mestrado em Ciência da Computação). Universidade Federal de Minas Gerais, Belo Horizonte, 2012. Disponível em: <http://www.dcc.ufmg.br/pos/cursos/defesas/1474M.PDF>. Acesso em: 26 mai. 2013.

GOOGLE. **CSE**. Disponível em: < http://www.*Google*.com/cse/>. Acesso em: 26 mai. 2013.

_____. ***Google* History**. *Google*. Disponível em: <https://www.*Google*.com.br/about/company/history/>. Acesso em: 20 jul. 2014.

HOLANDA, C.; BRAZ, M. I. Indexação automática de conteúdos na web: análise de *sites* de museus. **Biblionline**, João Pessoa, v. 8, n. 1, p. 42-59, 2012.

KURAMOTO, Hélio. Informação científica: proposta de um novo modelo para o Brasil. **Ci. Inf.**, Brasília, v. 35, n. 2, p. 91-102, maio/ago. 2006.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas 2003.

LE COADIC, Yves-François. **A ciência da informação**. Brasília, DF: Briquet de Lemos, 1996.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002.

MALINI, F.; ANTOUN, H. **A *Internet* e a rua**: ciberativismo e mobilização nas redes sociais. Porto Alegre: Sulina, 2013.

MCKINSEY. ***Big Data***: the next frontier for innovation, competition and productivity. Disponível em: <http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx>. Acesso em: 05 mar. 2015.

MEADOWS, A. J. **A comunicação científica**. Brasília: Briquet de Lemos, 1999.

MELO, V. *Bing* – **O novo motor de pesquisas da *Microsoft***. pplware, 2009. Disponível em: <http://pplware.sapo.pt/informacao/*Bing*-o-novo-motor-de-pesquisas-da-*Microsoft*/>. Acesso em: 20 jul. 2014.

MONTEIRO, S. D.; FIDÊNCIO, M. V. As dobras semióticas do *ciberespaço*: da *Web* visível à invisível. **TransInformação**, Campinas-SP, v. 1, n. 25, p. 35-46, jan./abr. 2013. Disponível em: <https://www.puc-campinas.edu.br/periodicocientifico>. Acesso em: 24 fev. 2015.

MONTEIRO, S. D. O *ciberespaço* e os mecanismos de busca: novas máquinas semióticas. **Ci. Inf**., v. 35, n. 1, Brasília, jan./abr. 2006. p.31-38. Disponível em: <http://www.scielo.br/pdf/ci/v35n1/v35n1a04.pdf>. Acesso em: 12 jul. 2015.

MONTEIRO, S. D. *et al.* Em busca da compreensão da “busca” no *ciberespaço*. In: Encontro Nacional de Pesquisa e Ciência da Informação, 12, 2011. **Anais...** Brasília, DF, 2011. p. 2536-2551. Disponível em: <http://enancib.ibict.br/index.php/xii/enancibXII/paper/view/823>. Acesso em: 29 maio 2013.

_____. As múltiplas sintaxes dos mecanismos de busca no *ciberespaço*. **Ci. Inf.**, Londrina, v. 14, n. esp, p. 68-102. 2009. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/viewArticle/2027>. Acesso em: 18 jun. 2013.

MONTEIRO, S. D.; PICKLER, M. E. V. O *ciberespaço*: o termo, a definição e o conceito. **Datagramazero,** Rio de Janeiro, v. 8, p. 1-18, 2007.

MINAYO, M. C. de S. **O desafio do conhecimento**. 10. ed. São Paulo: HUCITEC, 2007.

NEGROPONTE, N. **A vida digital**. Trad. Sérgio Tellaroli. São Paulo: Companhia das Letras, 1995.

POUREBRAHIMI, B.; BERTELS, K.; VASSILIADIS, S. **A survey of peer-to-peer networks**, 2005. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/*download*?doi=10.1.1.64.1218&rep=rep1&type=pdf>. Acessoem: 28 mai. 2013.

REVISTA INFORMAÇÃO & INFORMAÇÃO. **Histórico**. Londrina: UEL, 1996. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/index>. Acesso em: 12 jul. 2013.

RIBEIRO, V. G. *et al.* O emprego de técnicas de mineração de dados para definição de estratégias em processos de divulgação científica em periódicos de design. **Strategic Design Research Journal**, v. 6, n. 2, maio/ago. 2013.

RUBI, M. P.; FUJITA, M. S. L. Elementos de política de indexação em manuais de indexação de sistemas de informação especializados. **Perspect. cienc. inf.,** Belo Horizonte, v. 8, n. 1, p.66-77, jan./jun. 2003./2009

SANTAELLA, L. **Linguagens líquidas na era da mobilidade**. São Paulo: Paulus, 2007.

SENSEBOT. **The Search Engine that finds sense in a heap of *Web* pages**. Sense Bot. Disponível em: <http://www.sensebot.net/about.htm>. Acesso em: 20 jul. 2014.

SHERMAN, C.; PRICE, G. **The invisible *Web*:** uncovering information sources: searches engines cant´s see. Medford: Cyberage Books, 2001.

SILVA, E. L.; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**, 2005. Disponível em: <http://www.tecnologiadeprojetos.com.br/banco_objetos/%7B7AF9C03E-C286-470C-9C07-EA067CECB16D%7D_Metodologia%20da%20Pesquisa%20e%20da%20Disserta%C3%A7%C3%A3o%20%20UFSC%202005.pdf>. Acesso em: 27 mai. 2013.

SILVEIRA, S. A. Mobilização colaborativa, cultura *hacker* e a teoria da propriedade imaterial. In: AGUIAR, V. M. (org.). ***Software livre*, cultura *hacker* e o ecossistema da colaboração**. São Paulo: Momento Editorial, 2009.

SILVEIRA, D. T.; CÓRDOVA, F. P. A pesquisa científica. In: GERHARDT, T. E.; SILVEIRA, D. T. (orgs.). **Métodos de pesquisa**. Porto Alegre: UFRGS, 2009.

SOUZA, R. R; ALVARENGA, L. A *Web* semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 1-16, jan./abr. 2004.

SOUZA, R. R. Sistemas de Recuperação de Informações e Mecanismos de Busca na web: panorama atual e Tendências. **Perspect. ciênc. inf**., Belo Horizonte, v.11 n.2, p. 161 -173, mai./ago. 2006.

TAURION, C. **Big data**. Rio de Janeiro: Brasport, 2003.

TEIXEIRA, C. M. S.; SCHIEL, U. A internet e seu impacto nos processos de recuperação da informação. **Ciência da Informação**, v. 26, n. 1, 1997. Disponível em:< http://revista.ibict.br/ciinf/index.php/ciinf/article/view/421/380>. Acesso em: 5 jun. 2015.

TIGRE, P. B.; NORONHA, V. B. Do mainframe à nuvem: inovações, estrutura industrial e modelos de negócios nas tecnologias da informação e da comunicação. **R. Adm**., São Paulo, v. 48, n. 1, p. 114-127, jan./fev./mar. 2013.

VERGILI, R. **Premissas deontológicas de Relações Públicas e exigências do mercado:** relacionamento entre grandes empresas e *stakeholders* por meio de redes sociais conectadas. 293 f. Dissertação (Mestrado em Comunicação). Faculdade Cásper Líbero. São Paulo, 2012.

VIGNOLI, R. **A topografia da *Dark Web* e seus não lugares**: por um estudo das dobras invisíveis do *ciberespaço*. 2014. Dissertação (Mestrado em Ciência da Informação) – Universidade Estadual de Londrina, Londrina. 2014. Disponível em: <http://www.bibliotecadigital.uel.br/document/?code=vtls000191992>. Acesso em: 23 fev. 2015.

WOLFRAM ALPHA. **FAQ**. 2014. Disponível em: <http://www.wolframalpha.com/faqs.html>. Acesso em: 20 jul. 2014.

YACY. ***Philosophy***. 2007. Disponível em: <http://YaCy/Solr.net/en/Philosophy.html>. Acesso em: 28 mai. 2013.

_____. ***IntroSearch.*** 2011. Disponível em: <http://www.YaCy/Solr-*Webs*uche.de/wiki/index.php/En:IntroSearch >. Acesso em: 28 mai. 2013.

_____. **FAQ**. 2012. Disponível em: <http://www.YaCy/Solr-*Webs*uche.de/wiki/index.php/En:FAQ >. Acesso em: 28 mai. 2013.

YAHOO!. ***Yahoo!*! Media Relation**. *Yahoo!*. 2014. Disponível em: <http://archive.today/puqz#selection-67.360-73.54>. Acesso em: 20 jun. 2014.

YAMAOKA, E. J. **Recuperação da informação na *web*:** cenário atual e perspectivas para o futuro. Brasília, [S.n], 2002. 19 p.