UNIVERSIDADE FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO

FRANK COELHO DE ALCANTARA

RECUPERAÇÃO E CLASSIFICAÇÃO DE INFORMAÇÕES PROVENIENTES DA WEB E DE REDES SOCIAIS

CURITIBA
2013

FRANK COELHO DE ALCANTARA

RECUPERAÇÃO E CLASSIFICAÇÃO DE INFORMAÇÕES PROVENIENTES DA *WEB* E DE REDES SOCIAIS

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência, Gestão e Tecnologia da Informação do Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação, Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Profª. Drª. Denise Fukumi Tsunoda.

CURITIBA
2013

Alcantara, Frank Coelho de

Recuperação e classificação de informações provenientes da Web e de redes sociais virtuais / Frank Coelho de Alcantara. – Curitiba, 2013.

118 p. : il. ; 30 cm.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação, Universidade Federal do Paraná, 2013.

1. Inteligência artificial – Redes sociais. 2. Recuperação da informação. 3. Redes sociais virtuais. Título.

CDU 004.8:301

CDU

TERMO DE APROVAÇÃO

Frank Coelho de Alcantara

"RECUPERAÇÃO E CLASSIFICAÇÃO DE INFORMAÇÕES PROVENIENTES DA WEB E DE REDES SOCIAIS VIRTUAIS"

DISSERTAÇÃO APROVADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE MESTRE NO PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO DA UNIVERSIDADE FEDERAL DO PARANÁ, PELA SEGUINTE BANCA EXAMINADORA:

Prof.ª Dr.ª Denise Fukumi Tsunoda
(Orientadora/UFPR)

Prof.ª Dr.ª Patrícia Zeni Marchiori
(Examinadora/UFPR)

Prof. Dr. Orlando Alcantara Soares
(Examinador/PUCPR)

28 de fevereiro de 2013

Dedico

Aos meus saudosos pais, Wilson Menezes de Alcantara e Dalmira Virgínia Coelho de Melo por todo amor, cuidado, incentivos, exemplos, valores e oportunidades que me proporcionaram, sem os quais não teria chegado até aqui.

A minha adorada esposa, Francisca Mary Magalhães de Alcantara pelas horas e mais horas de amor, carinho, incentivo e compreensão que norteiam minha vida.

Ao meu amado filho Matheus Magalhães de Alcantara pelas tardes em que não brincamos, os dias em que não passeamos, os jogos que não jogamos, as aulas que assistimos juntos e a compreensão, muito além da sua idade, que me incentiva e estimula.

AGRADECIMENTOS

A minha querida orientadora Profa. Dra. Denise Fukumi Tsunoda pela compreensão, cumplicidade e principalmente por ser este porto seguro de conhecimento e inspiração.

Aos professores doutores Deborah Ribeiro Carvalho, Denise Maria Woranóvicz Carvalho, Helena de Fátima Nunes Silva, José Simão de Paula Pinto e Sônia Maria Breda pela atenção e carinho além do dever com que sempre me trataram e, principalmente pela paciência quase infinita com este perguntador inveterado.

Aos professores doutores Celso Yoshikazu Ishida e Cícero Aparecido Bezerra e demais professores do programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação da Universidade Federal do Paraná pela atenção dispensada e pelos conhecimentos passados.

Aos meus colegas e amigos do programa de Pós-Graduação em Ciência, Gestão e Tecnologia da Informação da Universidade Federal do Paraná pelos incentivos, apoio e paciência.

A todos os meus alunos pelas brincadeiras, perguntas e incentivo e, principalmente por entenderem a pressa e o mal humor. Principalmente ao meu amigo e ex-aluno Laércio Santo de Araújo pela ajuda inestimável durante o curso.

A minha esposa e filho que a todos os momentos estiveram sorrindo, estimulando e apoiando o desenvolvimento deste estudo.

A verdadeira genialidade é a capacidade de avaliar informações imprecisas, perigosas e conflitantes.
Winston Churchill

RESUMO

Esta dissertação apresenta um estudo, exploratório e metodológico, sobre o problema de recuperação e classificação de informações na *Internet*: (*i*) avalia a possibilidade de usar a infraestrutura de conexão entre os documentos na *web* acrescida do processo de interação social online, para a criação de gráficos sociais virtuais e para a análise da opinião expressa; (*ii*) utiliza os gráficos sociais virtuais e a análise de opinião como fatores de classificação da informação recolhida; (*iii*) desenvolve uma pesquisa bibliográfica na ciência da análise de redes sociais, destacando a história desta ciência e suas métricas, expande este estudo para as redes sociais virtuais observando seu funcionamento e sua história; (*iv*) estuda as características estruturais da *web*, seus documentos e os protocolos nela utilizados. (*v*) avalia métodos utilizados na mineração de opinião com um levantamento dos algoritmos utilizados para avaliação de opinião em fragmentos de texto; (*vi*) apresenta um sistema de recuperação e classificação de informação da *Internet* considerando domínio específico que utiliza algoritmos e técnicas recentes para definir uma arquitetura de busca, recuperação e classificação de informação baseado em um léxico característico fornecido pelo usuário, na criação e avaliação de gráficos sociais virtuais e na análise de opinião. Os resultados encontrados indicam que o uso de gráficos sociais virtuais e mineração de opinião como fatores de classificação podem ser utilizados como complementos ao Pagerank minimizando o ruído nos resultados de buscas.

Palavras chaves: gráficos sociais virtuais; mineração de opinião, redes sociais.

ABSTRACT

This paper presents a exploratory and methodological study about the recovery and classification of information's problem on the Internet: (i) evaluates the possibility of using the infrastructure connection between web documents plus the process of online social interaction to create a virtual social graphs and to opinion analysis; (ii) uses the virtual social graphs and opinion analysis as collected information's classification; (iii) develops a literature review on the science of social network analysis, highlighting the history, its metrics, and expand the concepts to virtual social networks observing its operation and its history; (iv) study the structural features of world wide web it's documents and protocols; (v) evaluates methods used in mining opinion with a survey of the algorithms used to assess opinion on text fragments; (vi) presents a system of classification and retrieval of information from the internet considering specific a domain that uses recent algorithms and techniques to define an architecture for search, retrieval and classification of information based on a lexicon characteristic supplied by the user in the creation of virtual social graph and opinion analysis. The results indicate that the use of virtual social graphs and opinion mining as ranking factors can be utilized as complements to Pagerank minimizing noise in search results.

Keywords: virtual social graphs; opinion mining; social networks.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE SIGLAS

AJAX - *Asynchronous Javascript and XML*

API - *Application Programming Interface*

CERN - *Conseil Européen pour la Recherche Nucléaire*

CSS - *Cascading Style Sheets*

DHTML - *Dynamic Hypertext Markup Language*

DNS - *Domain Name Service*

DOM - *Document Object Model*

ECMA - *European Computer Manufacturers Association*

FTP - *File Transfer Protocol*

GSV - *Gráficos Sociais Virtuais*

HITS - *Hyperlink-Induced Topic Search*

HTML - *Hypertext Markup Language*

HTTP - *Hipertext Transfer Protocol*

IP - *Internet Protocol*

JSON - *Javascript Object Notation*

MIME - *Multipurpose Internet Mail Extensions*

MMORPG - *Massively-Multiplayer Online Role-Playing Games*

MVC - *Model View Controller*

OSI - *Open Systems Interconnection*

PHP - *Hypertext Preprocessor*

REST - *REpresentational State Transfer*

RFC - *Request For Comments*

RPC - *Remote Procedure Call*

SGML - *Standard Generalized Markup Language*

SOAP - *Simple Object Access Protocol*

SVM - *Support Vector Machines*

TCP - *Transfer Control Protocol*

URI - *Uniform Resource Identifier*

URL - *Universal Resource Locator*

WSDL - *Web Service Definition Language*

XML - *eXtended Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Talvez a mais antiga atividade humana seja o convívio social. Pesquisadores encontraram registros arqueológicos de grupos humanos interagindo em sociedade, desenvolvendo e utilizando tecnologias para esculpir e pintar rochas, no que parece ser um esforço para a construção de templos, datados entre oito e dez mil anos antes de Cristo (SCHMIDT, 2000). O que parece indicar a existência de capacidade de organização criativa e construtiva três mil anos antes do que se acreditava possível. Outros registros, com vários milhões de anos, de grupos de hominídeos, parecem corroborar o conceito de vida em sociedade (MASSEY, 2002).

Do ponto de vista neurológico, estudos recentes indicam o convívio em sociedade como um fator determinante tanto para a estruturação quanto para a funcionalidade de alguns circuitos cerebrais em primatas. Acredita-se que o convívio social tenha sido um elemento decisivo no desenvolvimento da linguagem e que esta, por sua vez, tenha sido fundamental para o surgimento do conhecimento (SALLET; MARS *et al.*, 2011).

Nesta pesquisa o termo informação é utilizado em uma forma ampla significando qualquer forma de dado contextualizado e que possa ser transformado em conhecimento, incluindo texto, bancos de dados, vídeo, áudio, gráficos, revistas, filmes e qualquer outro artefato informacional que possa ser armazenado, ou registrado, e que sobre a análise humana tenha a possibilidade de ser transformada em conhecimento. Este por sua vez, baseado no estudo de Vail (1999), é aqui entendido como a forma tácita da informação que tem propósito ou utilidade.

Dada a sua abrangência e importância, o estudo das interações entre indivíduos e entre grupos de indivíduos, ocupa lugar de destaque em vários campos da ciência, desde a compreensão dos eventos históricos e urbanísticos até a previsão de preços de ativos em bolsas de valores, passando pela economia e medicina. Ao longo do tempo a humanidade desenvolveu formas e ferramentas para facilitar este convívio. Uma das mais recentes, a *Internet*, é objeto dos estudos apresentados realizados durante a pesquisa e apresentados neste estudo.

Em 1990 Tim Berners Lee, e a equipe do *Conseil Européen pour la Recherche Nucléaire*[1] (CERN), disponibilizaram na *Internet*, a primeira página *web,* criando uma estrutura para a interligação de informação de forma simples e direta. Desde então, utilizando as ferramentas fornecidas pela *Internet*, a humanidade está criando formas novas para o relacionamento social e distribuição de informações. Entre as novas ferramentas, e ambientes de interação, estão *sites* como o Facebook®, Twitter® e o Google+®, além de milhões de fóruns, *sites* de revistas e jornais, *sites* de compra e venda e páginas estáticas.

Com um tamanho estimado de 50 bilhões de páginas (KUNDER, 2012), é possível considerar a *web* como sendo o maior repositório de informações já criado pelo ser humano. Em sua maioria, disponível a qualquer pessoa que tenha interesse ou necessidade. Todavia, o Google®, a mais popular ferramenta de busca indexa pouco menos de sete bilhões de páginas (FURCHE; GOTTLOB *et al.*, 2012) indicando que, aproximadamente, 86% de toda informação produzida na *web* não pode ser localizado e, consequentemente, tem sua utilidade reduzida aos proprietários deste conhecimento explicitado e a uns poucos eleitos que conhecem sua localização. A parte não indexada, que não pode ser localizada facilmente, esta camada obscura da *web*, denominada de *deep web*[2] (ou *hidden web*[3]), pode conter informações de valor inestimável e merece atenção.

O termo *deep web* representa o conjunto formado pelas páginas que os programas dos *sites* de busca não conseguem acessar. Torna-se necessário o distanciamento entre o termo *deep web*, com intenção científica, que se refere a páginas que são dinamicamente geradas e não estão disponíveis por meio de um *link*[4], do termo *deep web*, divulgado em filmes e séries de televisão, que se refere a um universo oculto de criminalidade. Este estudo concentra-se na *web* profunda do ponto de vista meramente científico.

O uso desta informação oculta requer que ela seja localizada. Mais que isso, é necessário que ela seja indexada. Permitindo a criação de rotinas de recuperação que, por sua vez, facilitem a utilização da informação pelo usuário final, ou por algoritmos especiais de classificação. As empresas de busca têm sugerido soluções alternativas para tentar expandir o alcance dos seus sistemas de recolhimento e

[1] Do francês: Centro Europeu de Pesquisa Nuclear.
[2] *Deep web* pode ser traduzido como teia profunda em português.
[3] *Hidden web* pode ser traduzido como *web* oculta.
[4] Do inglês: ligação.

classificação, como o *sitemaps*[5] proposto pelo Google® (GOOGLE, 2011), que permite a localização de todas as páginas geradas de forma dinâmica, mas previsível e estruturada, em um *site*.

Dado ao caráter de interligação dos documentos na *web*, e a ineficiência dos métodos de recolhimento de informação baseados nessa interligação estas soluções não parecem ser suficientes. Pesquisas sugerem que as técnicas utilizadas para a análise de redes sociais possuem potencial para ser uma alternativa eficiente à localização de informações na *web*. A forma como os usuários de redes sociais virtuais explicitam o conhecimento indica a possibilidade do uso destas interações sociais virtuais como ferramenta de localização e classificação da informação. (JACKSON, 1997)

Neste contexto, o presente estudo apresenta os resultados de uma pesquisa bibliográfica e metodológica com a pretensão de explorar a *internet*, as redes sociais virtuais e propor um modelo de classificação de informações tendo, como suporte, em última instância, o convívio social *online*. Para tal, a pesquisa explora a *web* por meio da criação de um sistema específico capaz de recuperar páginas *web* e interações sociais, utilizando-se de um domínio específico caracterizado por um léxico próprio e de algoritmos específicos de recuperação e classificação. Propõe um sistema de classificação utilizando as métricas de análise de redes em gráficos socias virtuais e a mineração de opinião para criar um conjunto de informações classificadas em uma forma que reflete as interações sociais entre os autores da informação.

1.1 PREMISSA FUNDAMENTAL E PROBLEMAS DE PESQUISA

A quantidade de informação na *web* é provavelmente finita. Contudo considerando-se que páginas dinamicamente geradas podem levar a outras páginas dinamicamente geradas, esta recorrência tem o potencial de gerar um volume infinito de informação (BAEZA-YATES; CASTILLO, 2004). Um exemplo, fornecido por Baeza-yates e Castillo (2004) de páginas dinamicamente geradas é um aplicativo de agenda típico. Um ser humano percebe as limitações do aplicativo e sua relação com a realidade humana e, provavelmente não procurará compromissos em um prazo maior que 50 anos. Os aplicativos que percorrem a *web* recolhendo informações não estão

[5] Uma estrutura em XML que contém todas as páginas de um site, em uma ordem hierárquica.

adequados para atender este tipo de sutileza (BAEZA-YATES; CASTILLO, 2004). Contudo, do ponto de vista dos algoritmos de busca baseados em hipertexto, a descoberta de um conteúdo que é gerado de forma dinâmica, de acordo com as interações e vontades dos usuários é uma tarefa monumental e praticamente impossível (BAEZA-YATES; CASTILLO, 2004). Para tal finalidade seria necessário criar um algoritmo capaz de gerar todas as interações possíveis e recuperar e armazenar cada resultado obtido (BAEZA-YATES; CASTILLO, 2004).

Mesmo usando apenas a *web* indexada pelos mecanismos de busca, para a recuperação de informações, resta o problema de classificar as informações recolhidas ou de recuperar a informação com maior grau de utilidade. Algoritmos como o Pagerank (PAGE, BRIN, *et al.*, 1999), o *Hyperlink-Induced Topic Search*[6] (HITS) (KLEINBERG, 1999) ou o Cheirank (SHEPELYANSKY, 2011) fazem uso da infraestrutura de conexão entre documentos na *web* como fator de classificação da informação, usando para isso as características da própria estrutura de *hyperlinks* e atribuindo valores diferentes a cada página devido a estas características. Indubitavelmente esta aproximação obteve resultados significativos na recuperação de informações de forma genérica (BAEZA-YATES; CASTILLO, 2004), destaca-se como exemplo o *site* de buscas do Google®. Pecam, no entanto, na recuperação de informações referente a um domínio específico e nas informações necessárias ao suporte à tomada de decisão (BAEZA-YATES; CASTILLO, 2004).

A maior parte dos *web sites* que oferecem, por exemplo, livros ou passagens aéreas, são baseados no uso de banco de dados acessados por interfaces complexas de busca e navegação definidas para uso humano (FURCHE, GOTTLOB; SCHALLHART, 2012). Os sistemas de recuperação de informação atual não conseguem acessar estes dados já que não existem *hyperlinks* para todas as informações disponíveis. Quando conseguem algum acesso, não conseguem responder perguntas específicas tais como: Seria esta empresa confiável? Este livro é bom? Qual o menor preço para este produto? Perguntas que permitiriam entender um determinado domínio, tempo ou interesse (CHANDRAMOULI; GAUCH; ENO, 2010).

Os métodos completamente automatizados parecem não ser suficiente para o suporte a tomada de decisão em situações específicas como, por exemplo, as

[6] Do inglês: busca de tópicos induzida por hyperlinks.

decisões de compra, ou contratação (FENSEL; LAUSEN *et al.*, 2007). Em contrapartida, ferramentas sociais já foram utilizadas como suporte à tomada de decisão em situações tão críticas como a compra e venda de ações, por meio da medição das opiniões externadas pelos usuários de redes sociais virtuais e mineradas por algoritmos específicos (BOLLEN; MAO; ZENG, 2011).

A percepção desta deficiência provocou a criação de *sites* de busca com arquiteturas voltadas a tomada de decisão. A versão original do Bing®[7] da Microsoft® talvez seja o mais conhecido deles. Trata-se de uma classe de *sites* de busca, conhecidos como *Decision Engines*[8], que procuram aumentar a precisão do resultado oferecendo um conjunto de opções para refinar o resultado da busca. Caso o usuário pesquise por televisão, por exemplo, na primeira página de resultados, o serviço de busca irá ofertar um conjunto de opções complementar tais como: marca tamanho, voltagem, sistema, região (WAI; LAU, 2002). Mediante a seleção destas opções o resultado da busca é atualizado na tentativa de interpretar a vontade do usuário. Este conjunto extra de informações aumenta a probabilidade da localização da informação desejada. Porém, não permite a localização de informações complementares, ou conflitantes, indispensáveis à tomada de decisão (BLACK, 2010).

Já foi provado que existem informações de valor nas redes sociais (FURCHE; GOTTLOB *et al.*, 2012), e que os sistemas de recuperação de informação são ineficientes para a recuperação de informações voltadas a classificação de um domínio específico (BAEZA-YATES; CASTILLO, 2004) (FURCHE; GOTTLOB; SCHALLHART, 2012). Parece não existirem dúvidas que a informação tem valor estratégico para as empresas e para os indivíduos (DRUCKER, 1992).

Considerando os problemas encontrados na busca e classificação de informações na *web* esta pesquisa parte da premissa de que é possível gerar estruturas de classificação adequadas a cada domínio de informação e procura comprovar a validade desta premissa respondendo a pergunta: **De que forma é possível utilizar a estrutura da *internet* como suporte para recuperação e classificação da informação necessária a um usuário específico?**

1.2 OBJETIVOS DA PESQUISA

[7] MICROSOFT BING. Disponível em: http://www.bing.com, acessado em 12 fev. 2012.
[8] Do inglês: motores de decisão.

Recuperar e classificar informações na *web* parece ser uma necessidade pungente na sociedade contemporânea. No entanto, este campo é extenso, complexo e novo. Esta pesquisa limita seu escopo a pretensão de atender os objetivos explicitados nas seções 1.2.1 e 1.2.2, respectivamente objetivo geral e objetivos específicos.

1.2.1 Objetivo geral

A *web* e as redes sociais nela contidas e a informação espalhada nestas redes sociais e na *web* em geral, constituem o objeto desta pesquisa. Desta forma o objetivo geral pode ser expresso como sendo explorar um conjunto de técnicas de recuperação e classificação de informações, na *web* e em redes sociais virtuais, relativas a um determinado domínio, caracterizado por um léxico pessoal de acordo com as necessidades e especificações de um usuário final, em um ambiente de relacionamento social virtual, disponível na *web*.

1.2.2 Objetivos específicos

Os objetivos específicos constituem os passos necessários à pesquisa para formação do conhecimento necessário para a criação do sistema de busca, recuperação e classificação da informação que é utilizado para validar a questão de pesquisa e podem ser sumarizados nos seguintes pontos:

a) caracterizar a análise de redes sociais, sob uma perspectiva histórica e metodológica, destacando as métricas de qualificação de nós e ligações;

b) estudar a arquitetura típica de *web sites, web services* e da interação social virtual aplicando os algoritmos de busca de informação genéricos atualmente em uso;

c) identificar os sistemas de mineração de opinião adequados à classificação das informações disponíveis na estrutura social da *web*;

d) utilizar as interações sociais disponíveis na *web* para criar um gráfico social virtual que permita a classificação da informação recolhida de acordo com as características topológicas deste gráfico.

1.3 JUSTIFICATIVA DA PESQUISA

.Sabe-se que a informação tem valor. Contudo este valor é diferente para pessoas diferentes, situações diferentes e tempos diferentes. Algumas informações tem valor para o entretenimento, outras para negócios, de uma forma ou de outra e, independente da sua origem ou uso, as pessoas estão dispostas a pagar pela informação (SHAPIRO; VARIAN, 1999). Este valor da informação parece justificar o esforço para o desenvolvimento de ferramentas mais eficientes para a localização, indexação e classificação de informações na *web*.

Kotler (2006), um pesquisador da ciência do marketing, ressalta que a informação adequada confere vantagens competitivas uma vez que as empresas ficam habilitadas a escolher seus mercados, dimensionar suas ofertas e entender a competição. Drucker (1998), por sua vez, afirma que a vantagem competitiva é criada em organizações baseadas em informação. A vantagem competitiva, seja em diferenciação ou custo, é função da cadeia de valor da empresa (PORTER; MILLAR, 1985). A informação transforma esta cadeia de valor, permitindo o cumprimento de uma função primordial para sua sobrevivência, modificando a forma como as atividades de valor são interligadas, alterando produtos ou criando oportunidades de inovação. Estes efeitos, básicos e intrínsecos, parecem explicar por que a informação é diferente do outros ativos usados na empresa (PORTER; MILLAR, 1985).

A tecnologia da informação e a disponibilidade de novas informações estão constantemente alterando as regras da competição de três formas: primeiro, a velocidade com que a tecnologia avança está alterando a estrutura básica dos negócios; segundo, a tecnologia da informação é o expediente que as empresas podem usar para criar vantagens competitivas, baixando custos e aumentando a diferenciação; finalmente, estas novas tecnologias estão espalhando negócios completamente novos por todas as áreas do mercado (PORTER; MILLAR, 1985).

Hardware e software são meros mecanismos usados para criar e gerenciar a informação. Esta última é o verdadeiro ativo oculto que permeia a infraestrutura de produção moderna. A informação provê a habilidade de melhorar resultados, ganhar vantagens competitivas e, além disso, pode ser vendida separadamente na forma de produtos independentes (MOODY; WALSH, 1999).

Parece estar claro que a informação é fundamental para o sucesso dos negócios. Restam questões relacionadas à qualidade da informação, disponibilidade e atualidade. O estudo de Shapiro e Varian (1999) parece ajudar a responder algumas destas questões quando afirmam que o valor da informação é diferente para pessoas diferentes, situações diferentes e tempos diferentes. A informação que é útil para o ator A pode não o ser para o ator B. Concordando com Shapiro e Varian (1999), Moody e Walsh (1999), Portter e Millar (1985), Kotler (2006) e Drucker (1998) não é difícil inferir que a informação tem valor, é fundamental para o sucesso dos negócios e deve ser individualizada para garantir vantagem competitiva, seja ela individual ou empresarial (MOODY; WALSH, 1999).

Independentemente da quantidade de informação disponível na *web*, algumas pesquisas indicam que os sistemas de busca de informação, de característica genérica, disponíveis atualmente se mostram ineficientes quando se trata exatamente da informação individualizada e específica (BAEZA-YATES; CASTILLO, 2004) (FURCHE; GOTTLOB; SCHALLHART, 2012). O valor da informação e a falta de sistemas eficientes para a recuperação e classificação de informação para domínios específicos sustentam e justificam a necessidade de estudos nesta área.

1.4 LIMITAÇÃO DA PESQUISA

O sistema para a pesquisa foi desenvolvido de forma limitada, considerando a necessidade de funcionar adequadamente em *hardware* caracteristicamente encontrado em computadores pessoais. Optou-se por ignorar a avaliação em tempo real das informações na *web* já que a quantidade de memória e banda necessárias está além dos limites possíveis em computadores pessoais.

Como o objetivo é explorar as técnicas de recuperação de informações relativas a um domínio específico, o sistema foi configurado para buscar informações relativas

a este domínio, referenciadas pela palavra *notebook* e um léxico limitado as palavras utilizadas pelos respondentes de um questionário especialmente criado para este fim.

Mesmo com o risco de perdas de velocidade na recuperação de informação da *web* optou-se pelo uso de *script*s, programas e bibliotecas desenvolvidas em *software* livre e código aberto, que estejam disponíveis gratuitamente na *Internet*. Reduzindo, desta forma, o custo de desenvolvimento da pesquisa aos *script*s necessários a integração das diversas funcionalidades dos artefatos de *software* disponíveis. O uso de diversos artefatos de *software*, desenvolvidos em linguagens de programação e arquiteturas independentes, impediu a criação de uma arquitetura de *software* uniforme. Assim sendo, o sistema desenvolvido para a prova de conceito baseia-se nos preceitos da programação orientada a objetos em que um modelo arquitetônico é obtido de modo derivativo, com uma formalização mais acurada deixada para trabalhos posteriores.

1.5 ESTRUTURA DO DOCUMENTO

O presente estudo divide-se em cinco capítulos. A introdução aborda o contexto do estudo, suas limitações, seus objetivos, o problema, as premissas e a pergunta que nortearam a pesquisa.

O capítulo dois descreve o referencial teórico composto da análise de redes sociais, virtuais ou não, da análise da estrutura da *web*, da análise de algoritmos de mineração de opinião e das técnicas envolvidas no processo de mineração da *web*.

No capítulo três estão descritos os procedimentos metodológicos que levaram a criação do sistema de recuperação e classificação criado na intenção de avaliar as dificuldades envolvidas no processo de recuperação e classificação de informações originadas da *web*, e na criação dos gráficos sociais virtuais. No capítulo quatro está a análise dos resultados obtidos destacando-se a classificação final dos usuários, a dimensão do conjunto de informações recuperadas, os resultados dos processos de validação e classificação das informações e os problemas encontrados com os formatos utilizados para a externalização da opinião em redes sociais virtuais e na *web* em geral.

2 REFERENCIAL TEÓRICO

O estudo das interações sociais *online*, e sua influência na recuperação de informações, voltadas para facilitar a tomada de decisões ou não, implica no entendimento da estrutura das redes sociais, suas características intrínsecas; o entendimento do meio, suas características especiais e operacionalidade; e no entendimento das ferramentas necessárias a recuperação de informações, de característica social, neste meio.

2.1 ORIGEM DA ANÁLISE DE REDES SOCIAIS

Em 1954 J. A. Barnes publicou um artigo, "*Class and Committeess in a Norwegian Island Parish*"[9], onde utilizou o termo rede, do inglês: *network*, para definir a estrutura do relacionamento social entre membros de um ou mais grupos. Em uma nota de rodapé, aqui em tradução livre, explicou:

> Anteriormente, usara o termo teia (*web*) retirado do título do livro *The web of Kinship* de M. Fortes. Contudo, aparentemente, a maior parte das pessoas associa a palavra teia a teia de aranha, com duas dimensões. Enquanto eu estava tentando criar uma imagem para um conceito multidimensional. Trata-se apenas de uma generalização da convenção pictográfica que os genealogistas têm usado por séculos em seus gráficos de *pedigree*. (BARNES, 1954, p. 43)[10]

Para representar esta rede social, Barnes (1954) utilizou-se de uma estrutura de pontos, chamados nós ou, na análise de redes sociais, de atores. Neste mesmo modelo, estes atores foram interligados por linhas representando o relacionamento interpessoal, um conjunto de atores e ligações define a estrutura de uma rede social.

[9] Do inglês: Classes e comitês em uma paróquia de uma ilha norueguesa.
[10] *Earlier I used the term web, taken from the title of M. Fortes' book, The Web of Kinship. However, it seems that many people think of a web as something like a spider's web, in two dimensions, whereas I am trying to form an image for a multi-dimensional concept. It is merely a generalization of a pictographic convention which genealogists have used for centuries on their pedigree charts*. (BARNES, 1954, p. 43)

O estudo da estrutura desta rede é o estudo das interações sociais e pode definir a importância dos atores e dos seus relacionamentos.

O conceito de redes sociais de Barnes pode ser facilmente estendido para incluir toda a humanidade em uma mesma rede ou intensificado a ponto de separar apenas um pequeno grupo de indivíduos e interações particulares para estudo (NEWMAN; BARABASI; WATTS, 2006). Mesmo antes da criação do termo rede social, a abrangência, poder e oportunidades criadas por estas ligações interpessoais já eram percebidos e discutidos (NEWMAN; BARABASI; WATTS, 2006).

A possibilidade do uso prático das redes sociais aparece de forma definida pela primeira vez na peça teatral Laços (em húngaro: *Láncszemek*), do dramaturgo húngaro Karinthy Frigyes, em 1929 (NEWMAN; BARABASI; WATTS, 2006). Usando a voz de um dos seus personagens Frigyes propõe um jogo para determinar a possibilidade de contatar qualquer pessoa da Terra, usando apenas cinco conhecidos de conhecidos, partindo dos seus próprios conhecidos.

A hipótese de Frigyes, ainda que no campo da dramaturgia e sem formalismo científico, se verdadeira, permitiria inferir que a Terra é na verdade um mundo pequeno (*small world)* onde seria possível, com pouco ou nenhum esforço contatar qualquer pessoa do planeta. Este conceito é tão popular e intuitivo que praticamente todos os idiomas do planeta possuem expressões para descrevê-lo (BARABÁSI, 2002). Se verdadeiro, restaria uma questão a ser elucidada: quão pequeno é este mundo?

Em 1958, Sola Pool e Kochen (1978) escreveram um artigo sobre a relação entre os contatos pessoais de um indivíduo e a qualidade da capacidade de influência deste indivíduo devida a esta rede de contatos. Este artigo circulou durante 20 anos sendo publicado apenas em 1978, no primeiro volume da revista Social Networks. Talvez o mérito do estudo de Sola Pool e Kochen esteja no uso da expressão *small world*, e na criação de um modelo que relaciona a probabilidade de dois indivíduos quaisquer se conhecerem dado às relações que estes mantêm com seus próprios conhecidos.

Modelos semelhantes ao de Sola Pool e Kochen (1978), utilizando cadeias de Markov, teoria das filas ou caminhadas aleatórias, abordam o problema da transferência de informação em redes sociais utilizando a probabilidade de conhecimento entre dois indivíduos quaisquer, seus relacionamentos e a topologia da rede (RAPOPORT; HORRATH, 2007).

Em 1967, Stanley Milgram publicou o artigo *The small-world problem* (MILGRAM, 1967) estudando, de forma científica, o problema central da análise de redes sociais: tomando-se duas pessoas quaisquer, de forma aleatória, na população mundial, denominados de indivíduos originais, qual a probabilidade de que eles se conheçam? Ou, em termos mais modernos: existem indivíduos que eventualmente façam parte das relações pessoais dos indivíduos originais e, possam, entre seus próprios conhecidos, traçar uma rede de conhecimento tal que permita a criação de um elo de relacionamento entre os indivíduos originais?

Milgram (1967) explorou esta hipótese criando uma experiência com a distribuição de mensagens e confirmou a existência de uma relação entre o número de ligações entre os indivíduos em uma determinada população e o tamanho desta população, além de características específicas relativas à distribuição de informação, específicas de cada indivíduo. Encontrando um valor médio de cinco graus de separação entre os indivíduos estudados.

O resultado de outra experiência feita com os residentes dos EEUU que encontrou uma média de 5,2 graus de separação (mínimo de 4,6 e máximo de 6,1) em um universo de 296 indivíduos espalhados fisicamente entre as duas costas dos EEUU (TRAVERS; MILGRAM, 1968). É possível supor que sejam estes 5,2 graus que tenham dado origem aos famosos seis graus de separação, que ocupam a mídia e o imaginário popular (BARABÁSI, 2002).

A peça teatral de John Guare[11], *Six Degrees of Separation*[12], de 1990 é, provavelmente, o principal fator para a popularização da expressão seis graus de separação. Segundo a peça, os indivíduos da espécie humana estão separados por no máximo seis graus. Para entender a abrangência desta afirmação, é necessário considerar cada ligação entre atores na rede de Barnes (1954) como um grau de separação. Se todos os habitantes do planeta estiverem separados por seis graus de separação, os indivíduos originais estariam separados por não mais que seis ligações.

Uma simples equação exponencial parece explicar o uso do número seis. Considere um indivíduo A que tenha relacionamento com outros 100 indivíduos e estes, por sua vez também tenham relacionamento com outros 100 indivíduos. Com estas condições pode-se afirmar que o indivíduo A está a dois graus de separação de

[11] JOHN GUARE (1938) Dramaturgo estadunidense, indicado para o prêmio Pulitzer em 1991 na categoria Drama pela peça *Six Degrees of Separation* (ANDERSON, 2002).

[12] Do inglês: Seis Graus de Separação

10.000 indivíduos. Estenda esta equação em cinco graus, mantendo o valor de 100 relacionamentos e o número de conhecidos atinge os 10 bilhões. Número mais que suficiente para englobar a população mundial no começo do século XXI.

Infelizmente a distribuição de relacionamentos não é constante nem uniforme (RAPOPORT; HORRATH, 2007; MILGRAM, 1967; TRAVERS; MILGRAM, 1968). Como as experiências de Milgram (1967; 1968) mostraram, a quantidade de relacionamentos depende de características únicas de cada indivíduo, da distância física entre eles e, principalmente, do interesse despertado a cada interação social.

Desde que Tim Berners Lee colocou a primeira página *web* no ar em 1990 (BERNERS-LEE; FISCHETTI, 2002), e mesmo antes disso, a humanidade está trabalhando de forma caótica na criação de uma rede universal de distribuição de informação e estabelecimento de relações pessoais. Com centenas de milhões de pessoas diariamente explicitando conhecimento em todos os idiomas e formas existentes e imagináveis. Despretensiosamente, é possível declarar que a *Internet* é por si só, uma rede social que engloba todo o planeta. Ainda assim, seria possível intensificar os conceitos de Barnes (1954) e reduzir a *Internet* a fragmentos menores, pequenas redes sociais. Cada uma com características únicas de inter-relacionamento e funcionamento.

Em busca de novas interações sociais e da determinação da distância média entre dois indivíduos quaisquer no mundo, Jure Leskovec e Eric Horvitz (2008) publicaram um estudo onde analisaram, de forma anônima, as possíveis ligações entre 240 milhões de pessoas, usuários do sistema de mensagens instantâneas da Microsoft®. A análise desta população, diversificada geograficamente, determinou que estes indivíduos estivessem separados, em média, por 6,6 graus (LESKOVEC; HORVITZ, 2008).

Pode-se observar que o número de graus de separação, é suscetível ao crescimento da população mundial, a tecnologia disponível e ao interrelacionamento humano:

a) interrelacionamento: a simples existência de tribos indígenas totalmente isoladas invalida, por si só, a possibilidade de que toda humanidade esteja separada por seis graus. Considere também a possibilidade de existirem grupos de pessoas isoladas por catástrofes naturais, desastres ou quarentena médica. Nenhuma destas pessoas teria qualquer tipo de ligação com a rede

social mundial. Milgram (1967) observou o efeito deste isolamento na criação de redes;

b) tecnologia: o número de pessoas, que integram o circulo de relacionamentos de qualquer outro indivíduo, cresceu exponencialmente em consequência de tecnologias de relacionamento em ambiente virtuais (MCCORMICK; SALGANIK; ZHENG, 2010). Existe a hipótese que este crescimento no número de relações sociais de cada indivíduo original provoque uma redução no número de graus de separação. Em busca da comprovação desta hipótese, em 2011 foi publicado um estudo sobre as ligações entre os usuários do Twitter®[13]. Neste estudo foi possível perceber que quaisquer dois indivíduos originais, usuários dos serviços providos pelo Twitter estão ligados por apenas 3,88 graus de separação (BAKHSHANDEH; SAMADI et al., 2011). Estudo semelhante, realizado usando-se a população de usuários do site Facebook®[14], quando este possuía apenas 721 milhões de usuários e aproximadamente 69 bilhões de relacionamentos, encontrou um grau médio de separação de 4,74 (BACKSTROM; BOLDIY et al., 2012);

c) crescimento populacional: se por um lado à tecnologia opera facilitando os relacionamentos, o crescimento da população mundial torna a rede mais complexa incluindo milhares de novos pontos a cada dia. Milgram, em seu estudo seminal (1967), encontrou indivíduos sem disposição para interação social, com poucos conhecidos, incapazes de dar continuidade a um processo de interação. Quanto maior o número destes indivíduos maior é a dificuldade de uniformização da rede social.

A análise estrutural de redes sociais tem seu mérito para as ciências sociais, urbanísticas e comportamentais. Entretanto, esta análise não esgota o assunto, Granovetter (1973) extrapola a análise da estrutura da rede social e se dedica a análise do efeito da intensidade das ligações entre indivíduos. Seu objetivo é entender como relações pessoais em nível micro podem afetar características sociais como organização política, mobilidade social e revoluções em nível macro.

[13] TWITTER. Disponível em: http://www.twitter.com, acessado em 04 jan. 2012.
[14] FACEBOOK. Disponível em: http://www.facebook.com, acessado em 04 jan. 2012.

Granovetter (1973) classifica as ligações pessoais, segundo o relacionamento, em: ausentes, fracas e fortes. E demonstra que, em algumas situações específicas, ligações interpessoais fracas são mais eficientes que ligações fortes. Notadamente para a propagação de informação. A informação parece propagar-se mais eficientemente nas redes sociais fazendo uso das ligações fracas que das ligações fortes (CSERMELY, 2006).

Estas ligações fracas, no inglês *weak links* são definidas por Granovetter de forma qualitativa considerando os sentimentos envolvidos na relação, o tempo dedicado à relação e o valor pessoal, intrínseco e subjetivo, que cada indivíduo atribui ao relacionamento ou interação social (GRANOVETTER, 1973). Os *weak links* foram estudados de forma quantitativa por Berlow (1999) considerando o efeito que a eliminação de uma determinada ligação, do inglês *link*, causa no valor médio objetivado em uma rede determinada. Este valor objetivado pode ser: uma propriedade do relacionamento ou uma resposta que a rede apresente a um dado estímulo (BERLOW, 1999). Chegando a conclusões que demonstram que tanto os laços fracos quanto os fortes são fundamentais para que a rede se mantenha e atinja seus objetivos, Berlow (1999) suporta as conclusões de Granovetter (1973) e elucida o funcionamento da sociedade explicitando as consequências das relações pessoais para a continuidade, ou sucesso, da rede sejam estas relações fortes ou fracas.

A popularização da *Internet*, no começo dos anos noventa do século XX colocou a disposição da humanidade um conjunto completamente novo de ferramentas para a interação social. Aplicativos, protocolos, mensagem instantânea e *sites web* tornaram possível que a interação humana seja realizada à distância, de forma instantânea, eventualmente anônima, sem sujeição aos limites impostos pelas fronteiras políticas e econômicas, custos, distância e tempo. A facilidade promovida pela *Internet*, invariavelmente, modificou as formas com que são realizadas as interações sociais (BERNERS-LEE; FISCHETTI, 2002).

Em 1997 surge o *site* Six Degrees[15] considerado o primeiro *site* total e unicamente voltado para a criação e uso de redes sociais (BOYD; ELLISON, 2007). Ainda que muitos acreditem que, antes da formalização de um serviço apenas para redes sociais, as pessoas já o faziam de forma espontânea utilizando suas listas de

[15] SIX DEGREES. Disponível em: http://www.sixdegrees.com, acessado em 04 mar. 2012.

e-mail, listas de mensagens instantâneas e *sites* pessoais ou de afiliação (BERNERS-LEE; FISCHETTI, 2002).

Apesar das facilidades e abrangência das ferramentas *online*, as mesmas divisões sociais observadas por Milgram (1967), Granovetter (1973) e Pool e Kochen (1978) foram identificadas e parecem explicar as divisões, ou grupos de afinidade, que existem em todas as ferramentas *online*. Indivíduos procuram indivíduos parecidos consigo mesmos (MCPHERSON; SMITH-LOVIN; COOK, 2001). Indicando que esta característica hemofílica tem potencial de utilização como uma forma de classificar grupos de indivíduos de acordo com suas similaridades e comportamentos mesmo em redes sociais virtuais (WENG; LIM *et al.*, 2010). Esta homofilia, segundo Weng, Lim *et al.* (2010) indica interesses em comum e pode ser percebida nas relações de indivíduos em redes sociais sejam elas virtuais ou não.

2.2 ANÁLISE DE REDES SOCIAIS

Define-se rede social como um grupo de pessoas conectadas por um conjunto de relações sociais, tais como: amizade, parentesco, coleguismo ou simples troca de informação expandindo os conceitos de Barnes (1954) e Milgram (1967) do ponto de vista sociológico.

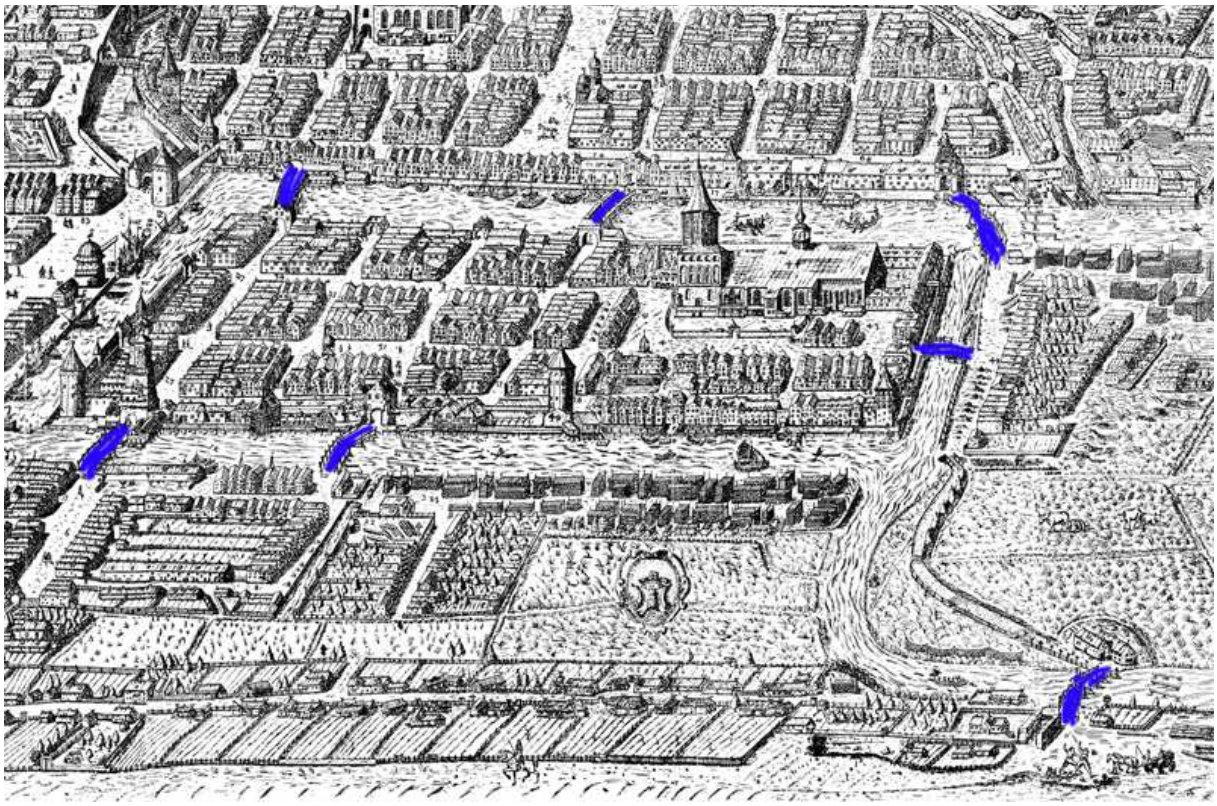
Para esta pesquisa redes sociais virtuais são entendidas como aquelas redes sociais nas quais tanto os relacionamentos quanto a troca de informações é realizada por meio de programas de computadores interligados a outros programas de computadores por meio da *Internet*.

O objetivo da análise de redes sociais, sejam elas reais ou virtuais, é entender o conjunto destes relacionamentos, a eficiência, duração, intensidade dos mesmos, seus pontos de estabilidade e instabilidade (WASSERMAN; FAUST, 1994).

A matemática que dá suporte a análise de redes sociais iniciou com os estudos do matemático Leonhard Eüler em 1753 através da análise de um problema popular na sua época: a cidade de *Königsberg* que em 1753 era a capital da Prússia e hoje se chama Kaliningrad e fica na Rússia, é cortada pelo Rio Preguel. No centro do rio, a certa altura, no meio da cidade, existiam duas ilhas.

As ilhas de *Königsberg* eram interligadas à cidade por sete pontes e havia a dúvida se seria possível criar um roteiro de passeio pela cidade atravessando as sete pontes, passando em cada ponte apenas uma vez (BARABÁSI, 2002). A Figura 1 mostra a distribuição das pontes, pintadas em azul, sobre um mapa datado de 1613 da cidade *Königsberg*.

FIGURA 1 – PONTES DE *KÖNIGSBERG*



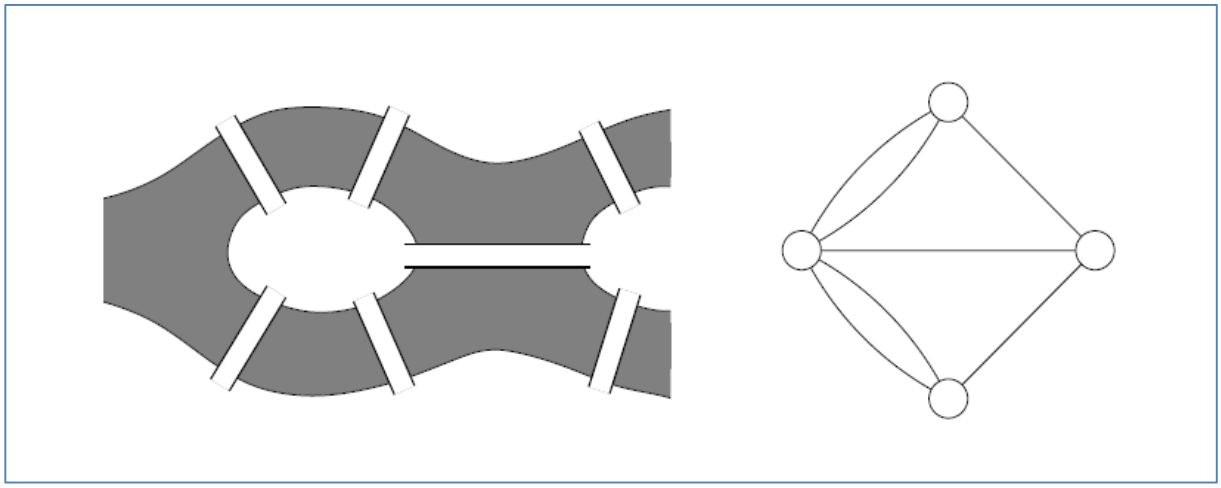
FONTE: adaptado do Mapa de Königsberg de Bering (MERIAN-ERBEN, 1652).

Para solucionar o problema Eüler fez uso de grafos. Uma representação matemática composta de pontos, também chamados de nó, vértice ou ator e linhas de conexão, também chamados de *links* ou bordas. No problema das pontes de *Königsberg* o grafo é composto de nós interligados por sete linhas, ou *links*. O problema consiste em descobrir se existe um caminho *Eüleriano*[16] possível. *Eüler* provou que não (NEWMAN; BARABASI; WATTS, 2006) este caminho não é possível. A ciência que *Eüler* iniciou, estudando o problema das pontes de *Königsberg* é a principal ferramenta matemática para o estudo de redes interligadas sejam elas sociais ou não e iniciou todo um ramo novo da matemática.

[16] Um caminho eureliano é um caminho que passa por todos os arrestas de um grafo sem repetir arestas.

Na Figura 2, é possível ver o diagrama simplificado da distribuição de pontes na cidade do *Königsberg* e a estrutura de nós e linhas, criada por *Eüler* para estudar o problema. *Eüler* resolveu o problema, demonstrando que só seria possível atravessar todas as pontes uma única vez, se e somente se, não existisse um nó com dois caminhos ou se existisse um número impar de nós de onde saíssem dois caminhos (BARABÁSI, 2002), ou seja, mantendo as pontes atuais seria necessária a construção de pelo menos outra ponte para permitir que o caminho fosse percorrido sem a repetição.

FIGURA 2 - DIAGRAMAS DAS PONTES DE *KÖNIGSBERG*



FONTE: (NEWMAN, BARABÁSI e WATTS, 2006, p. 3)

No processo de solução, Eüler criou os arcabouços do que viria a ser a teoria dos grafos, fundamental para o estudo das redes sociais. A análise de redes sociais compreende o entendimento da topologia, dos nós, das ligações entre eles, da representação em dados, do levantamento destas informações para que seja possível construir um modelo representativo e, usando este modelo, gerar conhecimento a partir da rede (NEWMAN; BARABÁSI; WATTS, 2006).

2.2.1 Conceitos matemáticos da análise de redes sociais

A análise de redes sociais requer o entendimento de uma taxonomia própria, adequada às necessidades únicas do problema. O estudo de redes sociais se ocupa da análise das entidades engajadas no relacionamento social (WASSERMAN;

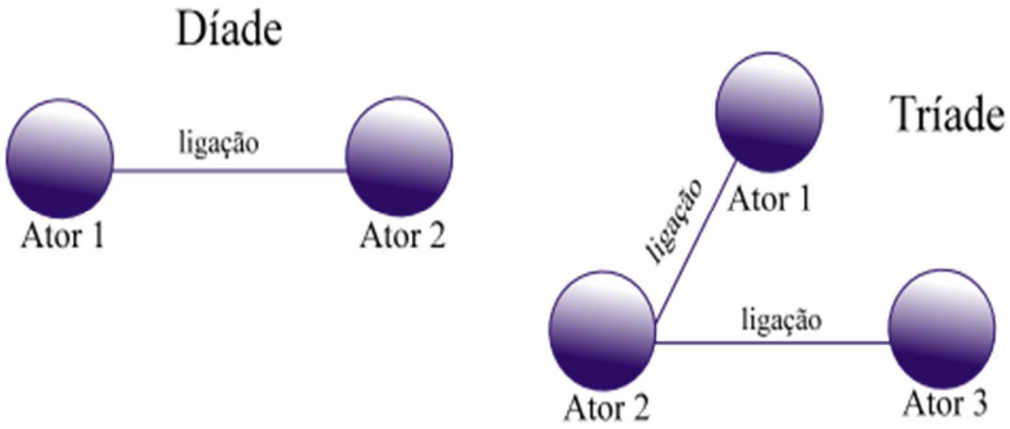
FAUST, 1994). Denominam-se de atores as entidades engajadas em relacionamentos pessoais.

Um ator pode representar um indivíduo independente, um grupo, uma empresa ou qualquer conjunto de entidades. Os atores são conectados entre si por meio dos relacionamentos sociais, ou laços. Um laço é um canal para a troca de informação entre os atores (WASSERMAN; FAUST, 1994).

Estes laços caracterizam relações diferentes. Podendo variar desde um contrato social, no caso de relação entre sócios a uma declaração de amizade em redes sociais virtuais, de uma relação de parentesco à afiliação em clubes.

A rede social mais simples (Figura 3) é chamada de díade e é constituída por apenas dois atores que podem estar ligados ou não. A existência da ligação constitui uma propriedade específica do par e, consequentemente não pertence a nenhum dos atores (WASSERMAN; FAUST, 1994).

FIGURA 3 - DÍADE E TRÍADE



FONTE: o autor (2013)

A ligação, mesmo não pertencendo a nenhum dos atores só existe entre um par de atores. Assim não há díade sem ligação ou ligação sem díade (FRISBY, 2002).

O estudo da topologia da ligação entre três ou mais atores é possível e atende necessidades específicas de cada caso. Destaca-se neste estudo, a tríade (Figura 3), esquema de ligações entre três atores, consistindo na existência de três pares potenciais, de grande valia para a análise sociológica por esquematizar de forma

qualitativa relacionamentos que não podem ser reduzidos à díade ou a um único ator (FRISBY, 2002).

Dois outros conceitos merecem destaque: sistema e grupo. Um sistema consiste nas ligações que existem entre atores de um mesmo grupo (WASSERMAN; FAUST, 1994). A existência de díades e tríades permite inferir a definição de um grupo com qualquer número de atores e sua respectiva ligação. O levantamento das características específicas de tais grupos, sua localização e estudo compõem um dos principais objetivos da análise de redes sociais (FRISBY, 2002).

O conjunto de ligações de um tipo específico entre os membros de um mesmo grupo é chamado de relação. Assim, por exemplo: chama-se relação ao conjunto de diferentes níveis de amizade entre crianças no jardim de infância ou ao conjunto de ligações diplomáticas entre nações (WASSERMAN; FAUST, 1994).

Uma vez definido grupo, sistema e relação, é possível definir redes sociais de uma forma mais precisa como sendo um conjunto finito de atores e suas relações (WASSERMAN; FAUST, 1994). Ampliando este conceito é possível definir redes sociais virtuais como sendo um conjunto finito de atores utilizando um serviço *web* que permite a criação de grupos, privativos ou públicos, permitindo a criação de relacionamentos entre eles. Parecendo indicar que a diferença entre redes sociais reais e virtuais consiste apenas nos mecanismos utilizados para permitir a ligação entre os atores.

2.2.2 Representação matemática de redes sociais

A teoria dos grafos consiste em um conjunto de ferramentas matemáticas utilizadas para análise de redes sociais desde os primeiros gráficos de Barnes (1954) e Milgram (1967). O uso desta matemática requer a adaptação da taxonomia originada das ciências sociais à linguagem matemática. Atores doravante são referidos pelo termo nó, os gráficos ou diagramas sociais são referidos como sociogramas e as linhas de relação são referidas como linhas de ligação, elos ou *links* - em inglês.

Segundo a nomenclatura mais utilizada (WASSERMAN; FAUST, 1994), a letra n será utilizada para representar um nó (ator) enquanto a letra *N* representará um conjunto de nós. Uma ligação será representada pela letra l enquanto o conjunto de ligações será representado pela letra *L*. Um sociograma será definido por G de tal

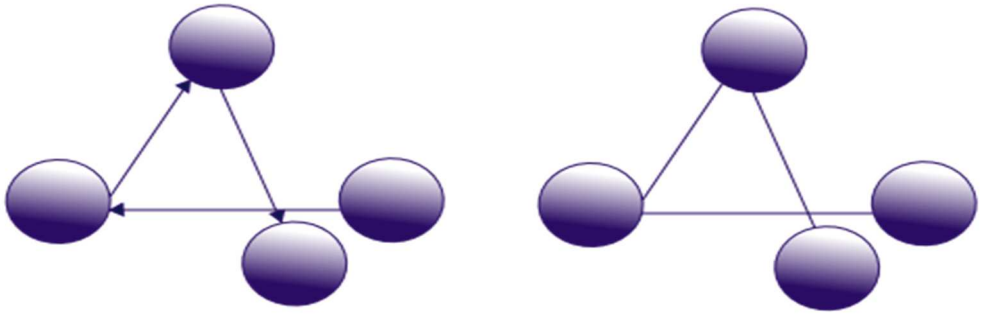
forma que G= (*N*, *L*) desde que $L \neq \emptyset$. Se o número de elementos em $\mathfrak{L}$ e $\mathfrak{N}$ forem respectivamente determinados por N e K os elementos destes conjuntos podem ser expressos por:

$$N = \{\mathbf{n_1}, \mathbf{n_2}, \mathbf{n_3} \ldots \mathbf{n_N}\} \quad (1)$$

$$L = \{\mathbf{l_1}, \mathbf{l_2}, \mathbf{l_3} \ldots \mathbf{l_K}\} \quad (2)$$

Normalmente um nó é referido como sendo o elemento de ordem *i* do conjunto *N*. Sendo assim uma ligação deve ser definida por um par de nós (*i*, *j*) ou $\mathrm{l_{ij}}$. Esta ligação é dita incidente entre os nós *i* e *j*; os dois nós são referidos como nós terminais da ligação $\mathrm{l_{ij}}$ e os nós que são ligados pela ligação $\mathrm{l_{ij}}$ são referidos como nós adjacentes ou vizinhos (BOCCALETTI; LATORA *et al.*, 2006).Em sociogramas não direcionais a ordem de *i* e de *j* é irrelevante. Em gráficos direcionais $\mathrm{l_{ij}} \neq \mathrm{l_{ji}}$. Na Figura 4 podem ser vistas as duas formas possíveis: direcional e não direcional, de representar um sociograma com quatro nós e três ligações.

FIGURA 4 - DIRECIONAIS E NÃO DIRECIONAIS



FONTE: o autor (2013)

O posicionamento dos nós no sociograma é totalmente irrelevante. Por outro lado é indispensável observar quais pares formam ligações e quais não formam.

Para um dado sociograma G de tamanho N, o número de ligações K será no mínimo zero e no máximo igual a $\mathrm{N}(\mathrm{N}-1)/2$. G será dito esparso se $\mathrm{K} \ll \mathrm{N}$ e denso se $\mathrm{K} = \mathrm{O}(\mathrm{N}^2)$.

Diz-se que um sociograma é vazio quando $\mathrm{N} = \mathrm{K} = 0$.

Um sociograma simples é aquele onde não existem laços (ligações simultâneas entre dois nós em direções opostas). Um sociograma é dito trivial quando existe

apenas um nó e dito completo quando, em um sociograma simples. Observa-se que todos os nós distintos são também adjacentes.

Para cada sociograma existe um único grafo completo denotado por K_n (WASSERMAN; FAUST, 1994).

Wasserman e Faust (1994) definem uma série de conceitos da matemática dos grafos indispensáveis ao entendimento de redes sociais:

a) dois nós, ainda que não adjacentes, podem ser alcançados seguindo-se as ligações de outros nós;

b) um caminho é uma sequência de nós alternados que começa em i e termina em j sem que haja repetição de nós;

c) o comprimento do caminho é definido como o número de ligações na sequência;

d) uma trilha é um caminho no qual nenhum ligação é repetida;

e) o caminho mais curto é a sequência mínima entre dois nós distintos;

f) um ciclo é um caminho fechado, com no mínimo três nós, no qual nenhuma ligação é repetida;

g) um sociograma é dito conectado se para cada par (i, j) do seu grafo existir um caminho entre i e j, caso contrário este sociograma é dito desconectado.

2.2.3 Graus de incidência e separação

Diz-se que um grau *ki* de um nó é o número de ligações incidentes neste nó. Define-se o grau de um nó utilizando a matriz adjacente dada por:

$$k_i = \sum_{j \in N} a_j \qquad (3)$$

Se o grafo representativo é direto, diz-se que existem dois componentes no grau: o grau de entrada (ligações que terminam no nó - *in* em inglês) e o grau de saída (ligações que se originam no nó - *out* em inglês). O grau *ki* é dado por:

$$\mathrm{k_i} = \mathrm{k_i^{out}} + \mathrm{k_i^{in}} \tag{4}$$

A caracterização topológica mais elementar de um grafo G é dada na forma da distribuição de graus *P(k)* definida como sendo a probabilidade de certo nó, escolhido de forma randômica e uniforme, tenha o grau *k* (BOCCALETTI, LATORA, *et al.*, 2006). A Chamada de relação entre os nós com *links in* e *out*, em redes não direcionais, como é o caso da *Internet*, o grau de distribuição pode ser calculado pelo momento da distribuição, determinado por:

$$(\mathrm{k^n}) = \sum_{\mathrm{k}} \mathrm{k^n P(k)} \tag{5}$$

O primeiro momento k determina o grau médio. O segundo mede as flutuações na conectividade da rede. Este segundo momento k^2 é fundamental para o cálculo do grau de distribuição em redes sociais virtuais.

2.2.4 O caminho mais curto ou caminho mínimo

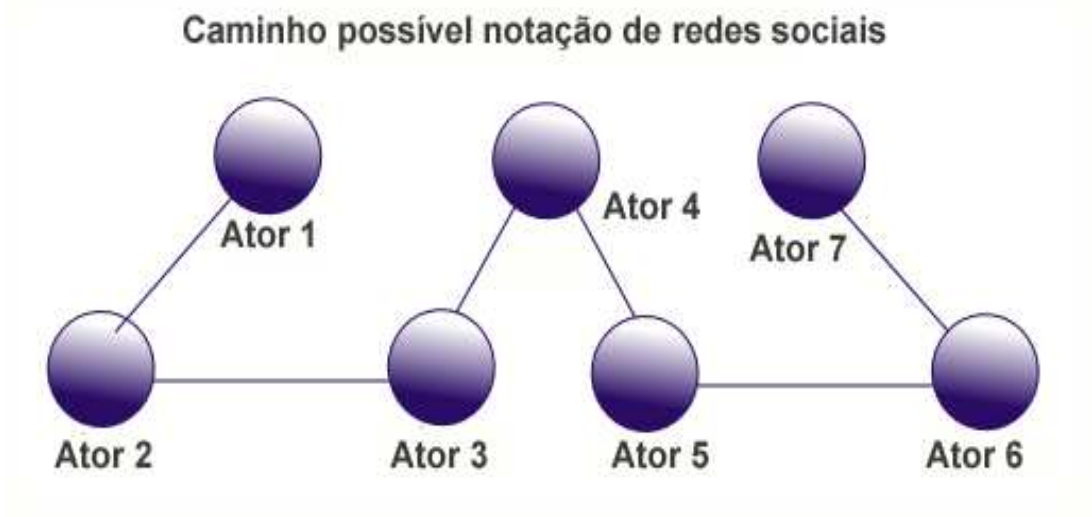
O caminho é um grafo da forma:

$$\{\mathbf{n_1}, \mathbf{n_2}, \mathbf{n_3} \ldots \mathbf{n_{iN}}\}, \{\mathbf{n_i}, \mathbf{n_{i+1}} : \mathbf{1} \leq \mathbf{i} < \mathrm{n}, \mathbf{n_i} \in \mathcal{N}\} \tag{6}$$

Entende-se que o caminho $\mathcal{C}$ representa um grafo que admite a permutação de nós, de tal forma que o $\mathbf{n_1}$ e o $\mathbf{n_i}$ são os extremos do caminho.

A Figura 5 mostra um caminho possível, com a notação adequada à análise de redes sociais, substituindo a nomenclatura nós por atores.

FIGURA 5 - CAMINHO POSSÍVEL NOTAÇÃO DE REDES SOCIAIS



FONTE: o autor (2013)

O número de arestas em um caminho determina o comprimento deste caminho (BOCCALETTI, LATORA, *et al.*, 2006).

O caminho mais curto é de suma importância na análise de redes sociais. Determina, por exemplo, o caminho que a informação deverá percorrer entre o nó *i* e *j,* permitindo a avaliação dos custos de tempo, processamento e vulnerabilidade. Muitas vezes, o caminho desejado é o caminho mais curto, não só dadas as limitações de custo, como principalmente, devido à redução do tempo gasto. Esta prioridade dada ao caminho mais curto não é exclusividade da tecnologia da informação nem do estudo das redes sociais virtuais e tem importância em ciências tão dispares quanto a engenharia eletrônica e a logística (WASSERMAN; FAUST, 1994).

Representam-se todos os caminhos mínimos de um sociograma G na matriz $\mathcal{D}$, na qual uma entrada *dij* representa a distância entre os nós *i* e *j*. A maior distância *dij* de um dado sociograma é chamada de diâmetro (BOCCALETTI; LATORA *et al.*, 2006).

Para sociogramas representados por grafos conectados, onde não existam elementos desconectados, a distância típica entre dois nós é determinada pelo comprimento do caminho mínimo médio dado por:

$$\mathrm{L} = \frac{1}{\mathrm{N}(\mathrm{N}-1)} \sum_{\mathrm{i,j}\in\mathrm{N},\mathrm{i}\neq\mathrm{j}} \frac{1}{\mathrm{d_{ij}}} \qquad (7)$$

Observe que a existência de um único nó não conectado implicará na divergência de L. Esta situação pode ser evitada, para redes não conectadas, calculando o comprimento do caminho mínimo médio, apenas para nós conectados. (BOCCALETTI; LATORA *et al.*, 2006).

2.2.5 *Clusters*, *motifs* e comunidades

O estudo dos *clusters* ou agrupamentos, *motifs* e comunidades são de suma importância para o entendimento da estrutura social da rede. Cada um apresenta características específicas e determinam um comportamento distinto para a rede (BOCCALETTI; LATORA *et al.*, 2006).

2.2.5.1 Clusters

Clusters, ou agrupamentos, constituem uma propriedade de redes sociais de conhecimento onde dois indivíduos originais, mutuamente desconhecidos, tem maior probabilidade de vir a se conhecerem se tiverem um amigo comum.

De forma matemática, os agrupamentos constituem uma propriedade que indica a probabilidade de um ator ter uma conexão com outro ator se ambos estiverem conectados a um terceiro ator em comum, mas não entre si.

Define-se o coeficiente de agrupamento de um dado grafo C como sendo a média do coeficiente de agrupamento individual de todos os nós de um determinado grupo G indicado por $\mathrm{c_i}$ (BOCCALETTI; LATORA *et al.*, 2006):

$$\mathrm{C} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i} \in \mathrm{N}} \mathrm{c_i} \qquad (8)$$

Por sua vez *ci*, o coeficiente de agrupamento local de um dado nó i, que expressa qual a probabilidade de *ajm* = 1 para duas vizinhanças, *j* e *m* do no *i*, pode ser calculado pela contagem das ligações (representadas por *ei*) do subgrupo local (G*i*) do nó *i*. Desta forma *ci* é definido como sendo a razão entre *ei* e o número máximo possível de ligações em um determinado grupo G*i*:

$$c_i = \frac{2e_i}{k_i(k_i - 1)} \quad (9)$$

2.2.5.2 Motifs

O *motif* representa uma pequena parte do todo que contém informações relevantes sobre este todo (HORNBY, 2010). Na teoria musical a palavra motif representa um pequeno conjunto de notas que, quando executado, permite identificar a música.

O conceito de motifs também foi introduzido em outras ciências como a biologia (KASHTAN; ITZKOVITZ *et al.*, 2004) para a identificação de redes em proteínas. Na análise de redes sociais um motif M é um subgrupo do grafo G (BOCCALETTI; LATORA *et al.*, 2006).

A pesquisa e localização de motifs em um grafo G são baseadas em um algoritmo que conta o número total de ocorrências de cada subgrupo M composto de *n* nós no gráfico original e em gráficos randômicos. A significância estatística de M é descrita por Z_M pode ser vista na Equação 10:

$$Z_M = \frac{n_M - (n_M^{rand})}{\sigma_{n_M}^{rand}} \quad (10)$$

Onde n_M é o número de vezes que o subgrupo M aparece em G e (n_M^{rand}) e $\sigma_{n_M}^{rand}$ são respectivamente a média e o desvio padrão do número de aparências em uma rede randômica equivalente (BOCCALETTI; LATORA *et al.*, 2006).

2.2.5.3 Comunidade

Uma comunidade, ou subgrupo coeso, é um subgrafo G'(*N*', *L*') de um grafo G(*N*,*L*), no qual os nós estão fortemente conectados (WASSERMAN; FAUST, 1994).

As comunidades podem ser definidas como subgrupos que não possuem nós, ou regiões, de interseção. Contudo, a definição mais aceita define uma comunidade como sendo um subconjunto de nós, interligados, de forma que todos os nós são adjacentes entre si (WASSERMAN; FAUST, 1994).

A teoria dos grafos extrapola o conteúdo deste estudo. Os conceitos aqui apresentados servem como base para o entendimento dos problemas e das soluções

encontradas, notadamente para indicar o funcionamento dos algoritmos utilizados para o levantamento do gráfico social usado na classificação das informações recuperadas. Um estudo mais profundo desta teoria pode ser encontrado em Balakrishnan (1997), Diestel (2000).

2.2.6 Propriedades dos atores

Algumas propriedades (ou métricas) características dos atores são fundamentais para a classificação da informação recuperada:

a) centralidade: uma métrica que indica a distância que um ator está do centro do grafo, considerando a distância máxima entre este ator e todos os outros atores e o diâmetro total do grafo. Em geral, atores localizados na periferia do grafo são considerados menos influentes;

b) adjacência: esta métrica considera a centralidade como uma medida da distância entre cada um dos atores do grafo. Quanto maior o valor, mas próximo um ator estará do outro;

c) centralidade de grau: percentual de atores que estão ligados ao ator estudado. Se o grafo é direcional é possível definir duas centralidades de grau, de entrada e saída;

d) centralidade de *betweenness*: se um ator está em muitos dos caminhos mais curtos dos entre vários pares de atores, este ator terá um alto grau de *betweenness*.

2.3 REDES COMPLEXAS, PROPRIEDADES E MODELOS

Algumas pesquisas indicam ser possível construir modelos para a representação de comportamentos sociais, e de comportamentos naturais, usando-se redes regulares ou aleatórias, notadamente com o uso da teoria dos grafos (NEWMAN; BARABÁSI; WATTS, 2006) (BARABÁSI; DEZS *et al.*, 2003). Todavia, observou-se que algumas redes apresentam comportamentos e propriedades

organizacionais distintos dos comportamentos apresentados pelas redes puramente aleatórias ou regulares.

Os pesquisadores Albert, Jeong e Barabási (1999), enquanto tentavam aplicar o estudo de Milgram (1967) à *Internet* na busca do diâmetro total desta rede composta de documentos *online*, perceberam que os nós desta rede estavam separados por menos de 20 cliques, confirmando o estudo de Milgram (1967). Perceberam também, e talvez aqui esteja o maior mérito do estudo, que a criação de *links* entre os nós não se dava de forma completamente aleatória, parecendo indicar a existência de uma regra de formação (BARABÁSI, 2002). A existência de uma regra de formação, por si só, ainda que desconhecida, invalida a hipótese da rede ser aleatória (ALBERT; JEONG; BARABÁSI, 1999).

Para distinguir estas redes, cuja formação, comportamento e organização não obedecem as regras da aleatoriedade, ou da regularidade, é necessário criar uma nova denominação: redes complexas.

Como exemplos de redes complexas é possível citar as redes de mundo pequeno, ou *small world* (WATTS; STROGATZ, 1998), as redes de transmissão de energia e as redes invariantes por escala, também denominadas de Redes Barabási-Albert (METZ; CALVO *et al.*, 2007). O modelo de redes complexas pode ser utilizado para o estudo do comportamento de diversas redes encontradas naturalmente entre as quais as redes sociais virtuais (BARABÁSI, 2002).

O que primeiro chamou atenção para as redes complexas foram as observações feitas por Albert, Jeong e Barabási (1999) quanto a não aleatoriedade durante a criação de conexões entre os atores, ou nós, da *Internet*. Esta propriedade, recém-descoberta, suscitou a necessidade de maiores estudos para este tipo de rede (ALBERT; JEONG; BARABÁSI, 1999).

Em outros estudos foram observadas e classificadas propriedades importantes para a caracterização das redes complexas (BARABÁSI; DEZS *et al.*, 2003; CSERMELY, 2006; METZ; CALVO *et al.*, 2007), que são discutidas nos subtópicos seguintes:

a) coeficiente de aglomeração: os agrupamentos, ou subgrafos, de uma rede complexa são caracterizados pelo seu coeficiente de aglomeração, ou fenômeno de transitividade. Em uma rede onde o ator A está conectado ao ator B que, por sua vez está conectado ao ator C. Existe uma probabilidade de

que o ator A também esteja conectado diretamente ao Ator C. O coeficiente de aglomeração indica a presença de um número elevado de tríades em uma determinada rede. É possível calcular o coeficiente de aglomeração pode ser calculado por:

$$\mathrm{CA} = \frac{3\mathrm{n}\Delta}{\mathrm{nV}} \tag{11}$$

Onde $\mathbf{n\Delta}$ representa o número de tríades e nV representa o número de vértices. O fator de multiplicação, três, garante a obediência às regras da probabilidade, garantido que CA ficará entre zero e um (METZ; CALVO *et al.*, 2007);

b) distribuição de graus: o grau de um nó, ou autor, qualquer, define o número de conexões, ou links, que incidem sobre este nó. A distribuição de graus é uma função da distribuição probabilística que indica a probabilidade de um determinado nó ter grau fixo. A quantificação desta distribuição pode ser feita por meio da Equação 12:

$$\mathrm{P_k} = \sum_{\mathrm{k'=k}}^{\infty} \mathrm{pk'} \tag{12}$$

Onde $\mathrm{pk'}$ é o subconjunto dos atores com grau k e $\mathrm{P_k}$ é a função cumulativa de distribuição de propriedades. A distribuição de graus em uma rede aleatória segue a distribuição de Poison (METZ; CALVO *et al.*, 2007). Mostrou-se, no entanto, que as redes complexas, notadamente as redes do tipo Albert-Barabási, seguem a lei de potencia em que $\mathrm{p_{k'}} = \mathrm{k}^{-\alpha}$ sendo α uma constante característica (BARABÁSI; DEZS *et al.*, 2003);

c) robustez: indica a capacidade de manter a estabilidade, e a própria existência contra a remoção de atores e conexões. A robustez esta diretamente relacionada à distribuição de graus característica (METZ; CALVO et al., 2007). Um alto grau de robustez indica a que a capacidade de comunicação entre os atores permanece mesmo em face de um grau elevado de perda de conectividade (ALBERT; JEONG; BARABÁSI, 1999). A robustez das redes complexas está baseada no alto grau de heterogeneidade da distribuição de conexões. Já que uma distribuição

baseada em uma lei de potência implica que a grande maioria dos atores terá poucas conexões. A remoção de um destes atores, ou de suas conexões não afeta a estabilidade da rede (ALBERT; JEONG; BARABÁSI, 1999);

d) padrões: existe uma relação de probabilidade de conexão que indica que atores do mesmo padrão, ou tipo, têm maiores probabilidades de estarem interligados (MCPHERSON; SMITH-LOVIN; COOK, 2001). Este aumento na probabilidade de conexão, devido à homofilia, parece ser uma característica das redes sociais (NEWMAN; BARABÁSI; WATTS, 2006). Ainda assim, é possível encontrar uma mistura de padrões, característica de redes complexas, quando são considerados fatores de distinção (METZ; CALVO et al., 2007) (NEWMAN, 2003);

e) correlação de graus: a correlação de graus indica que atores se associam a atores com graus parecidos. Esta propriedade é utilizada para investigar a probabilidade de conexão entre atores de tipos diferentes (METZ; CALVO et al., 2007). Parecendo indicar a existência do mesmo fenômeno de homofilia observado por McPherson, Smith-Lovin e Cook (2001) no estudo de redes sociais virtuais. Aparentemente as redes sociais virtuais apresentam um alto grau de associação por correlação de grau, diferindo das outras redes complexas, em um comportamento ainda não totalmente compreendido (NEWMAN, 2003).

2.3.1 Modelos de redes complexas

O estudo de redes complexas é recente, dinâmico e encontra respaldo nos estudos de Barabási, Dezs *et al.* (2003), Barabási e Reka (1999) e Boccaletti e Moreno (2006). Para esta pesquisa a relevância se encontra na identificação da *web* como rede complexa (BARABÁSI; DEZS *et al.*, 2003). As redes complexas podem ser classificadas em:

a) redes aleatórias: o modelo mais simples que pode ser aplicado às redes complexas foi proposto por Erdös e Rény, conhecido como modelo de redes

aleatórias (BARABÁSI, 2002). Neste modelo as conexões são acrescentadas de forma aleatória a um número fixo de atores. No modelo de redes aleatórias cada conexão é representada de forma independente com base em uma equação de probabilidade p que segue a distribuição de Poison com limite máximo N. O grau esperado de um vértice qualquer pode ser determinado pela Equação 13:

$$k = p(N - 1) \quad (13)$$

Onde N representa o número de atores da rede, e k é o número de conexões de um determinado ator (METZ; CALVO *et al.*, 2007). Erdös e Rény concluíram que todos os atores de uma determinada rede possuem aproximadamente o mesmo número de conexões e possuem a mesma probabilidade de receber novas conexões (BARABÁSI; DEZS *et al.*, 2003);

b) redes mundo pequeno: Watts e Stogratz (1998) observaram que muitas redes tendem a formar padrões com alto grau de interconexão ainda que mantendo um número pequeno de conexões em cada nó. Baseados nessa observação foi proposto um modelo de rede aleatória onde a maior parte das conexões é estabelecida entre nós adjacentes. O efeito “mundo pequeno” ocorre em redes onde a maioria dos atores se conecta a outros atores usando o caminho mínimo da rede (METZ; CALVO et al., 2007), o que corrobora o experimento de Milgram (1967). Este efeito “mundo pequeno” tem implicações claras nas redes sociais virtuais exacerbando a propagação de informações em toda a rede o que pode ser visto nos estudos relacionados a propagação de infeções (CHRISTLEY; PINCHBECK et al., 2001) ou boatos (MORENO; NEKOVEE; PACHECO, 2006) ou ainda a informação de um modo geral (PALLA; DERÉNYI et al., 2005);

a) redes Barabási-Albert: Barabási e Albert (1999) demonstram que em algumas redes existe uma tendência de conexão, denominada de conexão preferencial. Neste caso, um novo ator apresenta um grau maior de probabilidade de se conectar com atores que já possuam muitas conexões. Atores estes denominados hub (BARABÁSI; ALBERT, 1999) s. As redes que apresentam

esta característica de conexão preferencial são denominadas de redes de escala livre, ou Barabási-Albert. Esta característica de conexão cria uma topologia em que poucos atores estão altamente conectados e muitos atores possuem um número ínfimo de conexões (METZ; CALVO et al., 2007). A distribuição de conexões neste tipo de redes segue uma lei de potência, implicando em robustez (BARABÁSI; ALBERT, 1999). Esta robustez, por sua vez, determina um alto grau de imunidade a falhas aleatórias, paradoxalmente estas redes são suscetíveis a ataques coordenados (BARABÁSI; ALBERT, 1999). Neuman (2003) destacou que redes deste tipo foram observadas em áreas tão dispares quanto redes de citação científica, estruturas metabólicas celulares, a própria *Internet* e as redes sociais virtuais (GARTON; HAYTHORNTHWAITE; WELLMAN, 1997).

2.4 REDES SOCIAIS VIRTUAIS

Quando um computador conecta pessoas ou organizações entre si, cria uma rede social virtual (GARTON; HAYTHORNTHWAITE; WELLMAN, 1997). A inexorável popularização da *Internet* provocou a criação, e uso, de diversas ferramentas específicas para o contato entre seres humanos, de forma completamente virtual (BERNERS-LEE; FISCHETTI, 2002). Estas novas ferramentas, e novos protocolos, trouxeram novos hábitos e provocaram a criação de novas tecnologias que, por sua vez, em um ciclo virtuoso criaram novas ferramentas e novos protocolos (GARTON; HAYTHORNTHWAITE; WELLMAN, 1997).

A onipresença da tecnologia na vida cotidiana está mudando a sociedade por meio da criação de novos hábitos sociais, ou de novas estruturas operacionais, para suprir a necessidade de socialização e as formas tradicionais de estar em contato com outros seres humanos (GARTON; HAYTHORNTHWAITE; WELLMAN, 1997). A influência da tecnologia nos hábitos sociais é tão grande que uma pesquisa de 2011, entre jovens de 16 e 22 anos nos EEUU, indicou que 53% deles abririam mão do sentido do olfato para não perder o acesso à tecnologia (NEWSWIRE, 2011). Esta influência afeta o cotidiano as pessoas em tal intensidade que está provocando a mudança de regimes políticos e governos com revoluções orquestradas *online*

(LUGANO, 2011), ajudando a salvar vítimas de catástrofes naturais (BLACKBURN, 2011) ou permitindo a organização de protestos contra a descriminação sexual (BERGAMIM JR., 2011).

2.4.1 História das redes sociais virtuais

No começo dos anos noventa do século XX a equipe do CERN, liderada por Tim Berners-Lee (BERNERS-LEE; FISCHETTI, 2002) criou uma forma de interligar arquivos de texto. Com a intenção original de resolver o problema das citações em artigos científicos, utilizou uma estrutura de etiquetas, chamada de *hypertext*[17] que permitiu atribuir características adicionais a fragmentos de texto parcialmente baseados no projeto Xanadu de 1969 (NELSON, 1999). Berners-Lee criou também os protocolos necessários à edição, armazenamento, localização e uso destes documentos especiais (BERNERS-LEE; FISCHETTI, 2002).

As tecnologias desenvolvidas pelo CERN permitiram a criação de uma rede de informação interligada por *links* (ligações), virtuais - dentro de uma rede de computadores - que deu origem a uma nova dimensão de relacionamentos. Conhecida como *World Wide Web*, teia de alcance mundial, em inglês, ou simplesmente *web* (BERNERS-LEE; FISCHETTI, 2002).

A *web* (abreviatura de *world wide web*, teia de alcance mundial) inaugurada por Tim Berners-Lee em 1990 (BERNERS-LEE; FISCHETTI, 2002) é composta de bilhões de documentos interligados por laços formados com hipertexto que obedecem as regras do protocolo *Hipertext Transfer Protocol*[18] (HTTP), também definido por Berners-Lee (BERNERS-LEE; FIELDING *et al.*, 1994; FIELDING; GETTYS *et al.*, 1999) que permite que um documento possua uma ligação direta, e unidirecional, com qualquer outro documento disponível (CHAKRABARTI, 2003).

Antes da *web*, a humanidade já utilizava a *Internet* como canal para a troca de informações sociais e mesmo para a criação de pequenas redes (BERNERS-LEE; FISCHETTI, 2002). A diferença entre a *web* original e esta nova, que surge baseada em interações sociais está na organização. A *web* foi concebida com o objetivo de tornar a informação organizada e fácil de encontrar, foi concebida em torno da

[1717] Do inglês: hipertexto
[18] Do inglês: protocolo de transferência de hipertextos.

informação (BERNERS-LEE; FISCHETTI, 2002). As redes sociais virtuais foram concebidas em torno dos usuários (MISLOVE; MARCON, *et al.*, 2007).

O *e-mail*, um dos mais antigos serviços criados para o relacionamento entre atores *online*, criado em março de 1972, indicando que pares aplicativo/protocolo para a troca de informações *online* já eram utilizados quando a *web* foi criada (LEINER; CERF *et al.*, 1999). As listas de *e-mail*s pessoais, e as mensagens enviadas para vários usuários ao mesmo tempo constituíam os pilares de uma rede social: grupo, contato e troca de informações. Ainda que fosse possível a interação social individual, o protocolo *File Transfer Protocol*[19] (FTP), constituiu, durante vários anos, a única alternativa viável para a troca e compartilhamento de documentos (LEINER; CERF *et al.*, 1999).

Em 1991 uma equipe de pesquisadores da Universidade de *Minnesota* apresentou ao mundo o *Gopher* (FENSEL; LAUSEN *et al.*, 2007). Outro par aplicativo/protocolo que unia algumas vantagens do *e-mail* e do FTP com uma interface gráfica que permita navegar em árvores de arquivos, públicas e privadas, em um ambiente com interface semelhante ao Windows 3.11. As universidades adotaram o protocolo *gopher* de forma quase instantânea. Apesar disso, este protocolo caiu em desuso graças a maior facilidade de navegação e uso da estrutura de hipertextos proposta por Berners-Lee (2002) e dos aplicativos de navegação que foram criados para o uso destes hipertextos.

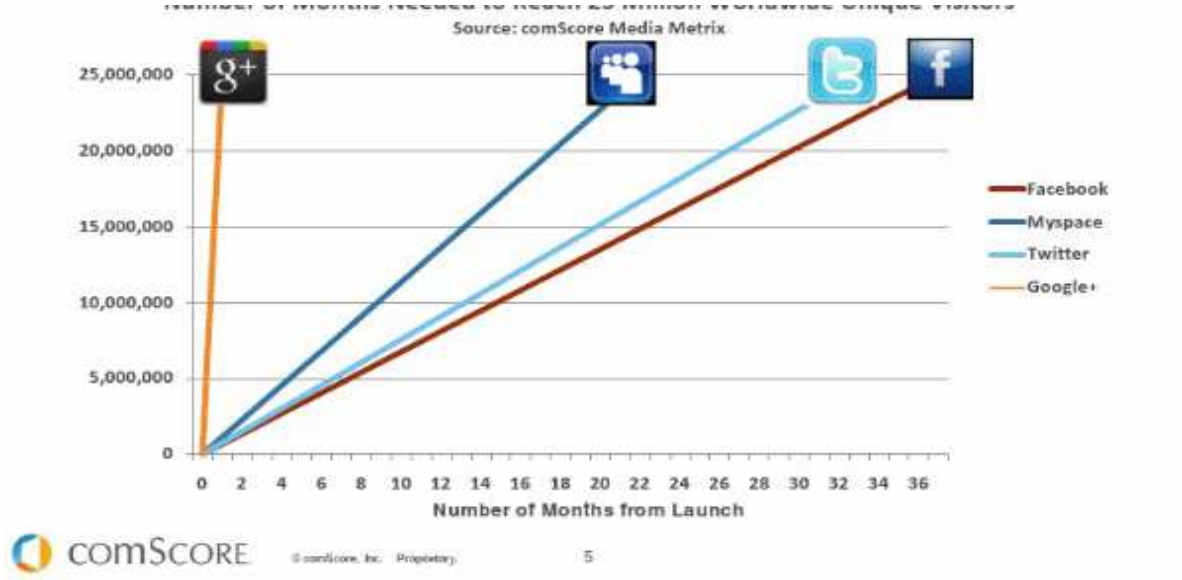
A popularidade da *web* e as novas tecnologias desenvolvidas permitiram que em 1997 surgisse o *site* Six Degrees considerado como o primeiro *site* total e unicamente voltado para a criação e uso de redes sociais (BOYD; ELLISON, 2007).

Em maio de 2003 o Linkedin começou suas operações com a intenção de ser uma rede social para negócios (WILSON, 2010); em 2004 surgem o Orkut e o Facebook (FRAGOSO, 2006). Nesta sequência, a primeira mensagem do Twitter® foi enviada por seu fundador, Jack Dorsey, em 2006 (PICARD, 2011). Hoje existem milhares de *sites* dedicados exclusivamente às redes sociais virtuais atendendo comunidades com necessidades específicas (KALLAS, 2011). No ano de 2011 o Google® lançou o Google+® sua rede social (GOOGLE, 2011) sem tirar do ar o Orkut®, que permanece, até o momento, um serviço com características de rede social independente.

[19] Do inglês: protocolo para a troca de arquivos.

Poucas semanas depois do seu lançamento o Google+® já era considerado o *site* com maior taxa de crescimento da história da *Internet* atingindo a marca de 25 milhões de usuários no seu primeiro mês de operação. Dados da Comscore, uma empresa especializada em estatísticas de tráfego *web* indicam permitem observar a taxa de crescimento do Google+® em relação as redes podem ser vistos na Figura 4.

FIGURA 6 - DADOS DO CRESCIMENTO DO GOOGLE+®



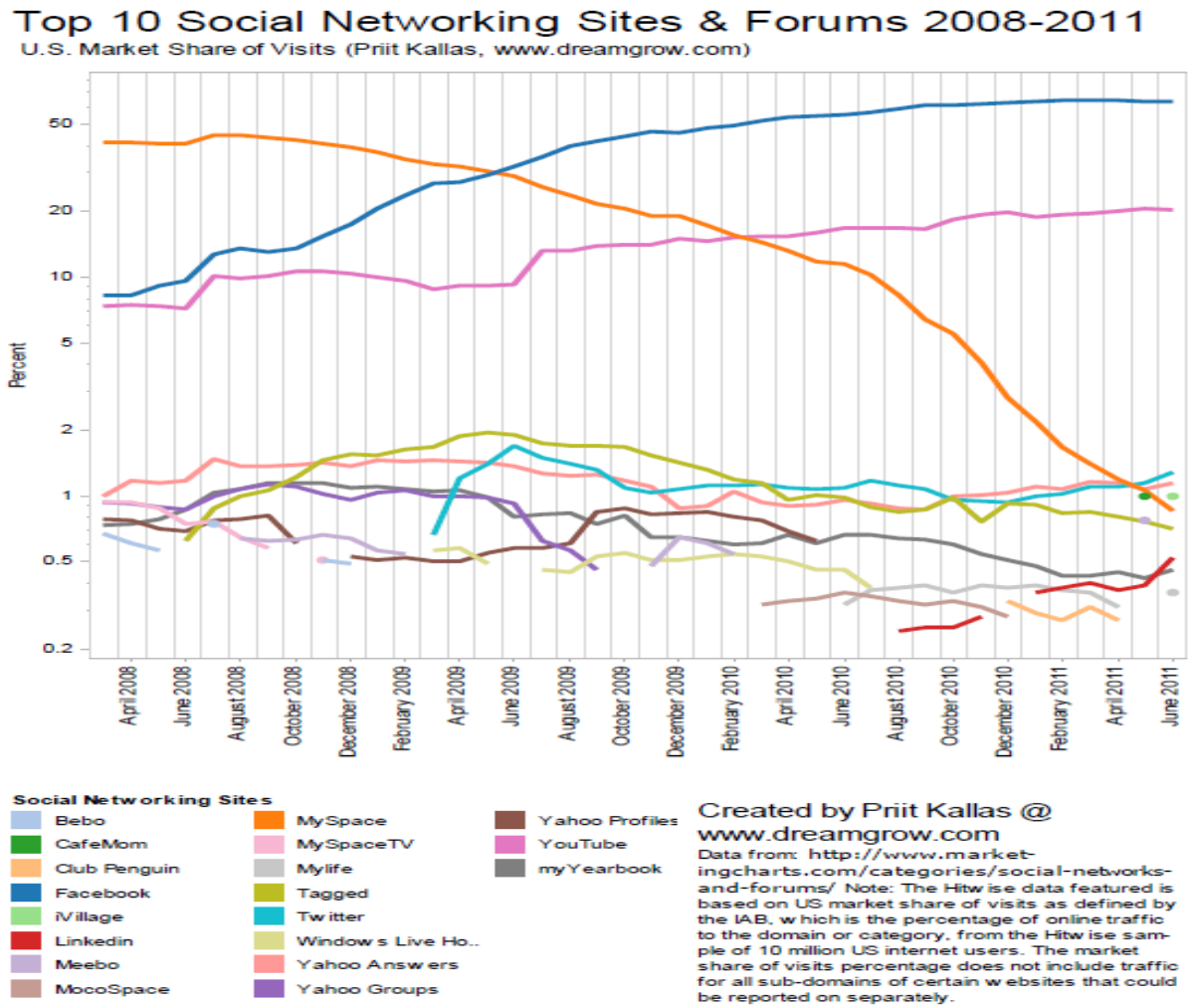
FONTE: (BARR, 2011)

A medida da popularidade de uma rede social virtual requer especialização e monitoramento constante. O número de *page views*[20] e o número de usuários únicos[21] frequentes determinam o faturamento destas empresas, sendo informações de grande importância para empresas de mídia social, especializadas na realização de campanhas de marketing nestes *sites* e como tal, mantidos em sigilo (MISLOVE; MARCON *et al.*, 2007). A *Dreamgrow Social Média* divulgou um gráfico mostrando a participação de mercado das dez maiores redes sociais dos EEUU.

[20]Page views, do inglês: visualização de páginas. Trata-se do número de vezes que uma determinada página é visitada em uma dada unidade de tempo.

[21] Usuários únicos número de usuários singulares, identificados por meio eletrônico automatizado que visitam uma determinada página, ou site, em uma dada unidade de tempo.

FIGURA 7 - PRINCIPAIS *SITES* DE REDES SOCIAIS



FONTE: (KALLAS, 2011)

Durante os primeiros anos dos *sites* de redes sociais no Brasil, o Orkut® reinou absoluto, mais como fruto de um choque cultural entre brasileiros e americanos que devido à qualidade do serviço prestado (FRAGOSO, 2006).

Em destaque, imprensa especializada e os *sites* de estatísticas de acesso já indicam o Facebook® como sendo o *site* de rede social mais visitado do Brasil (AGUIARI, 2011; ALEXA, 2011).

Durante o período de redação deste estudo, o mercado mundial de redes sociais virtuais esteve dominado pelo Facebook® com 750 milhões de usuários únicos (FACEBOOK, 2011), seguido pelo Twitter® com 200 milhões de usuários únicos, do Linkedin® com 100 milhões de usuários únicos. Acrescente-se a estes o Google+® que já chegou entre os dez maiores (EBIZMBA, 2011).

2.4.2 Serviços disponíveis em redes sociais virtuais

A principal força motora do sucesso das redes sociais virtuais está na estrutura desenvolvida em torno do usuário (PALLIS; ZEINALIPOUR-YAZTI; DIKAIAKOS, 2011). Para estas empresas, os interesses dos usuários são relevantes e, em sua maioria, a principal FONTE de informação. Consequentemente, o principal objetivo de um serviço de rede social virtual é fornecer facilidades de relacionamento pessoal para os seus usuários (PALLIS; ZEINALIPOUR-YAZTI; DIKAIAKOS, 2011).

Segundo Pallis, Zeinalipou-Yazti e Dikaikos (2011), classificam-se, como rede social virtual, os serviços *online* capazes de:

a) atuar como concentrador de informações para estabelecer relacionamento entre usuários (amigos, conhecidos, colegas, etc.). Cada usuário é capaz de, utilizando os recursos e politicas disponíveis, criar uma lista de usuários com os quais possui algum tipo de ligação;

b) fornecer ferramentas que permitam a criação de um senso de comunidade entre os usuários de forma informa e voluntária. Os usuários devem ser capazes de interagir entre si, contribuir com informações a um espaço comum, e participar em atividades interativas tais como jogos, divulgação de fotos, divulgação de opinião ou voto;

c) permitir que o usuário crie um perfil *online* contendo, dados pessoais, dados de localização, fotos, lista de ligações pessoais e afiliações. Mediante este perfil *online* os usuários são conhecidos, identificados e localizados. Receberão notificações e requisições de conexão, para participação de atividades em grupo ou para a troca de informações privadas;

Estes serviços de socialização incluem desde ferramentas de *e-mail* e mensagem instantânea até ferramentas para edição de filmes e fotos, passando por ferramentas específicas de agendamento (aniversários, promoções, etc.) e divulgação de status social e pessoal. Com o intuito de trazer para o mundo virtual todas as formas de relação pessoal existente sobre uma estrutura tecnológica que facilite este contato (PALLIS; ZEINALIPOUR-YAZTI; DIKAIAKOS, 2011).

Além dos serviços de relacionamento, os serviços de redes sociais virtuais possuem ferramentas especificamente desenvolvidas para exacerbar comportamentos sociais naturais de forma a aumentar o número de pessoas *online*. Um dos exemplos mais interessantes nesta categoria de ferramenta são os jogos *online*, também conhecidos como jogos sociais. Exceção deve ser feita para o Linkedin® e Twitter® que, apesar de permitir a criação de ferramentas específicas para jogos, não os possuem integrados no próprio serviço.

Os jogos do tipo *Massively-Multiplayer Online Role-Playing Games*[22] *(*MMORPG*)* podem ser definidos como jogos onde existe interação entre os atores de uma rede social, na forma de contribuição indireta ou participação direta. Este tipo de interação social torna o jogo mais atrativo. O apelo lúdico, a diversão pura e simples, parece irresistível (YEE, 2006). Os jogos fazem parte do processo de socialização humana desde os primórdios da humanidade (DELL'AMORE, 2010) as redes sociais virtuais apenas tornaram mais simples, e barato, encontrar parceiros para jogar.

Os dez jogos de maior sucesso, oito são jogos onde um jogador precisa construir alguma coisa (por exemplo: *Farmville* e *Cityville*) e esta construção depende, em maior ou menor grau da interação com outros usuários (GAMASUTRA, 2012).

Na popularidade dos jogos que permitem construir algo se percebe um fenômeno conhecido pelos profissionais de marketing: aparentemente o ser humano desenvolve um apreço irracional por aquilo que constrói (SHAPIRO, 2004). As redes sociais parecem utilizar este apreço pelo que construímos para crescer e prosperar.

2.5 ESTRUTURA DA *WEB*

Os documentos distribuídos na *web* são escritos em uma linguagem de marcação, especialmente desenvolvida para este fim, chamada *Hypertext Markup Language*[23] (HTML). Que consiste em uma série de etiquetas de texto, ou comandos, que são inseridas em um documento redigido em texto claro, que permitem a montagem do documento em um programa específico no formato desejado pelo autor (BERNERS-LEE; FISCHETTI, 2002).

[22] Do inglês: jogos de regras *online* massivamente multijogadores
[23] Do inglês: linguagem de marcação de hipertexto.

Esta linguagem também permite a inserção de ligações, os *links*, com outros documentos e a inserção de conteúdo rico, como áudio, vídeo e animações. A estrutura do HTML, apesar de rigidamente definida e especificada, permite que cada documento seja escrito e estruturado da forma que o autor desejar. Esta flexibilidade torna a *web* um meio caótico e desestruturado (CHAKRABARTI, 2003).

Uma das características mais importantes do formato HTTP é permitir a inclusão de *links* para outros documentos. Esta inclusão é feita por meio da especificação de um endereço único relativo ao documento que se deseja incluir. Este endereço, a *Universal Resource Locator*[24] (URL), é o indicador que os aplicativos de navegação utilizarão para localizar um documento (BERNERS-LEE; MASINTER; MCCAHILL, 1994). A URL é um caso especial de *Uniform Resource Identifier*[25] (URI), especificado na *Request For Comments*[26] (RFC) 3986, cuja função é atender os objetivos de identificar, endereçar e nomear recursos computacionais disponíveis *online* mediante o uso do protocolo HTTP (BERNERS-LEE; FIELDING; MASINTER, 2005).

2.5.1 A identificação de artefatos de informação na *web*

Cada informação na *web,* não importando seu formato ou conteúdo, pode ser localizado segundo um endereço específico: a URI (BERNERS-LEE; FIELDING; MASINTER, 2005) o qual consiste de uma sequência de caracteres dividida em cinco partes características: *scheme, authority, path, query* e *fragment*[27]*.* Destaca-se, no entanto, que apenas as partes *scheme* e *path* são indispensáveis. O *scheme* declara o tipo de URI e, consequentemente determina o significado das outras partes da URI. A *authority* indica o responsável pelo recurso apontado. No caso da *web* e do protocolo HTTP, a *authority* é indispensável e obrigatória, contendo o nome do servidor que mantém o recurso apontado. Neste caso a *authority* pode conter também o endereço de uma porta *TCP/IP*. Ainda que, na maior parte das vezes, o endereço desta porta seja desnecessário e automaticamente atribuído de acordo com as convenções de uso do protocolo *TCP/IP*.

---

[24] Do inglês: localizado universal de recursos.
[25] Do inglês: identificador uniforme de recursos.
[26] Do inglês: requisição de comentário. Trata-se do sistema usado pelo W3C para o processo de registro de protocolos.
[27] Do inglês, na ordem: esquema, autoridade, caminho, seleção e fragmento.

O *path*[28], obrigatório, é utilizado para referenciar em relação ao escopo indicado pelo *scheme*, no caso do protocolo HTTP o *path* é utilizado para permitir a localização do recurso dentro do servidor onde este está hospedado.

O *query* permite a recuperação de dados não hierárquicos dentro do recurso apontado enquanto o *fragments* permite a localização de partes específicas do recurso. A Figura 8 mostra os constituintes da URI usando um *site* aleatoriamente escolhido.

FIGURA 8 - ELEMENTOS DE UMA URI

FONTE: o autor (2013)

2.5.2 O Protocolo HTTP

O HTTP é um protocolo executado na camada de aplicação conforme esta está definida no modelo estrutural *Open Systems Interconnection*[29] (OSI) (INTERNATIONAL TELECOMMUNICATION UNION, 1994) e, como tal, os aplicativos responsáveis por este protocolo são também responsáveis por algum tipo de interação com os usuários, como os aplicativos de navegação na *web* como o Mozilla Firefox e o Microsoft *Internet* Explorer.

O HTTP é um protocolo adequado a arquitetura cliente-servidor que atende as especificações de um modelo de comunicação baseado no par requisição/resposta (FIELDING; GETTYS *et al.*, 1999) e, para permitir a comunicação precisa de um aplicativo cliente (um navegador *web*) e um aplicativo servidor (o HTTPD da Fundação Apache® ou o *Internet* Information Server da Microsoft®).

O modelo de comunicação baseado em requisição/resposta implica que os padrões de troca de mensagem sejam compostos apenas de ciclos de requisição e resposta que devem ser, obrigatoriamente, iniciados pelo cliente.

[28] Do inglês: caminho
[29] Do inglês: sistemas de interconexão aberta.

O protocolo HTTP inclui um componente opcional entre o cliente e o servidor denominado em inglês de *proxy*[30] que combina as funções de cliente e servidor executando funções de filtragem e armazenamento de dados.

O *proxy* ideal deve ser transparente tanto para o cliente quanto para o servidor (FIELDING; GETTYS *et al.*, 1999). Tanto as mensagens de requisição quanto as mensagens de resposta possuem a mesma estrutura. Devem iniciar com uma linha de inicialização seguida de linhas de cabeçalho opcionais, uma linha em branco e outra parte opcional contendo informações sobre a requisição ou sobre a resposta desejada.

Existem verbos específicos para semânticas específicas. O mesmo recurso pode ser tratado de forma diferente de acordo com o verbo HTTP utilizado na requisição.

O Quadro 1 apresenta os principais métodos do protocolo HTTP e uma breve descrição de suas características. A coluna Cache indica se os resultados da requisição podem ser armazenados em um *proxy* ou precisam ser atualizadas a cada requisição realizada.

Quando a emissão de requisições múltiplas de um mesmo método para um mesmo recurso resulta em uma única resposta o método é dito idempotente (FIELDING; GETTYS *et al.*, 1999).

QUADRO 1- VERBOS DO PROTOCOLO HTTP

Método	Uso	Seguro	Idempotente	Cache
GET	Recolhe informações sobre o recurso e o recurso em si.	Sim	Sim	Sim
HEAD	Essencialmente idêntico ao GET, exceto que o recurso não é devolvido.	Sim	Sim	Sim
PUT	Cria ou atualiza recursos.	Não	Sim	Não
DELETE	Remove um recurso do servidor.	Não	Sim	Não
POST	Usado para criar um novo recurso ou, em alguns casos, para disparar ações remotas.	Não	Não	Não
OPTIONS	Retorna informações sobre o recurso e suas representações disponíveis	Sim	Sim	Não

FONTE: o autor (2013)

[30] Do inglês: pProcurador

As primeiras especificações do HTTP exigiam a abertura de uma conexão TCP/IP para cada para requisição/resposta. Esta exigência provocou uma sobrecarga de tráfego desnecessária. A eliminação deste desperdício foi possível com a padronização de uma estrutura de conexão persistente. Uma conexão existente entre cliente e servidor, na camada TCP/IP é reutilizada para todas as requisições e respostas subsequentes. Posteriormente este tipo de conexão foi aprimorado permitindo a emissão de requisições subsequentes antes que a requisição que abriu a conexão seja respondida e o ciclo seja encerrado.

2.5.3 O Padrão HTML

O protocolo HTTP não restringe nem especifica o formato da entidade que é devolvida, nem dos dados que possam ser enviados. Desde sua origem a *web* foi voltada para o uso de mídias diversas. Sendo assim, o HTTP permite o uso dos mais diversos recursos, do texto simples a recursos ricos compostos de áudio, vídeo e dados de geoposicionamento. Entretanto, a maior parte destes recursos está contida, ou pode ser acessado, por meio de um arquivo HTML.

O HTML é uma linguagem de marcação de texto (BERNERS-LEE, 1992), cujo objetivo é permitir a formatação deste texto no dispositivo de leitura, baseada no - *Standard Generalized Markup Language*[31] (SGML) (BERNERS-LEE, 1992). O HTML fornece um conjunto de etiquetas, ou *tags* em inglês, que permitem que o aplicativo responsável pela leitura do arquivo HTML, chamado de navegador no caso da *web*, possa formatar o texto de acordo com as decisões de formatação tomadas pelo autor durante a redação. Garantindo que um texto específico seja visto sempre da mesma forma por diferentes leitores, independente do aplicativo utilizado para a sua leitura. Estes documentos de texto marcado suportam hipermídia, interação com o usuário e, graças a uma camada de visualização baseada em folhas de estilo, podem ter sua apresentação determinada e personalizada para uso em ambientes diversos.

As folhas de estilo, ou *Cascading Style Sheets*[32] (CSS) permitem que cada elemento HTML possa ter suas propriedades gráficas personalizadas de acordo com a vontade do autor de forma que seja possível obter um documento final graficamente

[31] Do inglês: padrão de linguagem de marcação genérica.
[32] Do inglês: folhas de estilo em cascata.

perfeito independente das características da mídia onde será exibido (WORLD WIDE WEB CONSORTIUM, 2012).

De forma a melhorar a interação com o usuário, documentos HTML permitem o uso de linguagem de *script*, definida pela *European Computer Manufacturers Association*[33] (ECMA), denominada de *javascript* (ECMA, 2011). Os *scripts* definidos com esta linguagem não só permitem uma melhor interação com os usuários, como também permitem que partes do documento HTML possam ser requisitados, sob demanda, em um processo conhecido como *Asynchronous javascript and XML*[34] (AJAX).

A quinta revisão do HTML, o HTML5, inclui no padrão um conjunto de novas *tags*, melhorou o uso de conteúdos multimídia e definiu um conjunto de camadas de interação chamadas de *Application Programming Interface*[35] (API) com o objetivo de aperfeiçoar o uso de dados, sensores, leitura fora de linha, velocidade de carregamento e estabilidade. Uma API denominada *Web* Sockets complementa o sistema de mensagens utilizado pelo protocolo HTTP, mediante o uso de uma camada de conexão bidirecional de baixa latência, f*ull-duplex*[36] baseada no protocolo *WEB*SOCKET (FETTE; MELNIKOV, 2011) o que aumenta a velocidade de carregamento das páginas e diminui a quantidade de dados trocados em cada interação provocando uma sensação de fluidez nas interações entre os clientes e os servidores, principalmente em aplicativos *web* de tempo real (FETTE; MELNIKOV, 2011).

2.5.3.1 Formatos genéricos

Além dos formatos especificados pelos órgãos padronizadores, não é incomum encontrar formatos de codificação, comunicação ou exibição, definidos por usuários, grupos de usuários ou empresas que, devido a sua eficiência e praticidade se tornam extensivamente usados e, com o passar do tempo, acabam sendo padronizados.

[33] Do inglês: associação dos fabricantes de computadores Europeus.
[34] Do inglês: javascript e XML assíncronos.
[35] Do inglês: interface de programação de aplicativo. Conjuntos de funções, ou métodos, que permitem acrescentar funcionalidades, ou utilizar funcionalidades disponíveis, em um programa por outro.
[36] Full-duplex é um termo da área de telecomunicações que indica que a comunicação pode ser feita em dois sentidos ao mesmo tempo e que não possui tradução em português.

Exemplos de protocolos em uso e de formato genérico são o *eXtended Markup Language*[37] (XML), e o *Javascript Object Notation*[38] (JSON). Utilizados para a troca de informações entre os servidores e os aplicativos clientes estes dois formatos competem em igualdade de condições.

O XML, já padronizado (WORLD WIDE WEB CONSORTIUM, 2008), apesar de fornecer um conjunto de ferramentas e opções mais extenso, requer uma quantidade extra de caracteres para transmitir as mesmas mensagens. O JSON, apesar de mais limitado, permite uma significativa economia de banda no canal de transmissão. O XML é mais versátil, enquanto o JSON é mais rápido e econômico (NURSEITOV; PAULSON *et al.*, 2009). A escolha depende das características que o desenvolvedor do serviço deseja ou precisa

2.5.4 *Web sites*

Cada endereço *web* aponta para um local, ou *site*, onde pode existir um simples texto, um conjunto de textos e imagens, apenas imagens, apenas textos, ou aplicativos completos e complexos.

A *web*, originalmente concebida para a troca de arquivos de textos, simples e estáticos, evoluiu de forma a permitir a criação de verdadeiras soluções corporativas e pessoais, socialmente integradas, dinâmicas e ricas capazes de executar arquivos de mídias diversas (BERNERS-LEE; FISCHETTI, 2002).

Web sites são aplicativos baseados na utilização da *web* como meio de comunicação e interação projetados para uso humano com o auxílio de aplicativos específicos de leitura conhecidos como navegadores. O Google Chrome©, o Mozilla Firefox© e o Microsoft *Internet* Explorer© são exemplos de navegadores utilizados para acessar, ler e interagir com *web sites* (BERNERS-LEE; FISCHETTI, 2002). Este estudo usará o termo *web site* de forma genérica para representar um conjunto finito de páginas *web* que pode ser acessado a partir de um mesmo endereço de domínio.

Os *web sites*, diversos em complexidade e funcionalidade, evoluíram de uma estrutura baseada na busca e distribuição de informação mediante o uso de hipertexto para aplicativos interativos e colaborativos com as mais diversas aplicações (BERNERS-LEE; FISCHETTI, 2002).

---

[37] Do inglês: linguagem de marcação estendida.
[38] Do inglês: notação de objetos em javascript.

Inicialmente a expressão *web site*, em inglês, representava apenas um lugar virtual, na *web* que poderia ser acessado por meio de um endereço específico que continha um conjunto de hipertextos conectados (HORNBY, 2010). Alguns dos tipos mais comuns de *web sites* incluem:

a) redes sociais: também chamadas de redes sociais virtuais, ou redes sociais *online*, são sites especificamente desenhados para permitir a interação social entre os seus usuários. Estão entre os sites mais visitados na *web* segundo a Nielsen *Online* (NIELSEN ONLINE, 2011);

b) blogs: originalmente criados como uma espécie de diário eletrônico, do inglês *web*log. Os blogs evoluíram como plataforma de difusão de informação e atendem desde a necessidade pessoal do autor até campanhas de marketing, relatórios de pesquisa e eventos institucionais (WIJNIA, 2004). Estão, também segundo a Nielsen *Online* (2011), rivalizando com as redes sociais virtuais em utilidade e quantidade de visitas;

c) aplicativos colaborativos: são sites exclusivamente desenhados para permitir algum tipo de colaboração entre seus usuários (NEUMANN; PRUSAK, 2007). O exemplo mais clássico são os WIKIS39 e seu expoente mais representativo a Wikipedia[40];

d) sites de comércio eletrônico: sites especialmente desenhados para permitir a venda de produtos usando a *web* como infraestrutura de marketing, atendimento e vendas. Segundo a (NIELSEN ONLINE, 2011), a Amazon destaca-se como exemplo deste tipo de site.

2.5.5 Serviços *web*

Serviços *web*, ou no inglês *web services*, fornecem acesso a aplicativos funcionais utilizando o HTTP, e a infraestrutura da *web* para a troca de informações

[39] O termo wiki deriva de uma expressão, wiki wiki, tipicamente havaiana que significa rápido (CUNNINGHAM, 2004).
[40] Disponível em: http://www.wikipedia.org acesso em 09 de outubro de 2011

no formato de mensagens ou chamadas remotas de funções, com o objetivo promover a interação entre aplicativos diferentes (CURBERA; DUFTLER *et al.*, 2002). Diferentemente dos *web sites*, os *web services* são voltados, única e exclusivamente, à interação entre aplicativos. Para permitir tal interação os *web services* são divididos em três áreas distintas: protocolos de comunicação, descrição de serviços e descoberta de serviços (CURBERA; DUFTLER *et al.*, 2002). Cada *web service* deve adotar um dos protocolos ou desenvolver um protocolo próprio.

A adoção dos protocolos existentes apresenta a vantagem da interoperabilidade (CURBERA; DUFTLER *et al.*, 2002). Os protocolos mais utilizados são:

a) XML-RPC: trata-se de um protocolo de chamadas de procedimentos remotos (*Remote Procedure Call*) que opera sobre a infraestrutura existente na *web* usando o protocolo HTTP. Uma mensagem XML-RPC é uma requisição HTML onde o corpo da mensagem é uma estrutura XML (WINER, 1999);

b) SOAP: o Simple Object Access Protocol41 (SOAP), (WORLD WIDE WEB CONSORTIUM, 2007) foi inicialmente desenvolvido pela Microsoft® e depois expandido com o auxílio da IBM©. Assim como o XML-RPC o SOAP utiliza os protocolos existentes na *web* para transferir mensagens em uma estrutura XML entre dois aplicativos distintos. Além das mensagens em si, o protocolo SOAP define um ator, responsável pelo processamento da mensagem e determina de que forma esta mensagem deve ser processada (CURBERA; DUFTLER et al., 2002);

c) WSDL: uma linguagem de definição de serviços *web*, o *Web Service Definition Language*42 (WSDL) foi inicialmente desenvolvida pela IBM para complementar algumas deficiências encontradas no SOAP e para aprimorar a troca de informações entre aplicativos distintos utilizando a infraestrutura da *web* (CURBERA; DUFTLER et al., 2002). Proposto pelo World Wide *Web* Consortium (2001) inclui a definição de quais mensagens devem ser trocadas

[41] Do inglês: protocolo simples de acesso a objetos.
[42] Do inglês: linguagem de definição de serviços *web*.

entre os aplicativos para a que a comunicação possa ser considerada como bem sucedida (CURBERA; DUFTLER et al., 2002);

d) RESTful: a transferência de estado representacional ou protocolo REpresentational State Transfer43 (REST) representou uma completa troca de paradigmas. Enquanto os protocolos anteriores se basearam em uma arquitetura baseada em envio e recebimento de mensagens, o REST em uma arquitetura totalmente diferente. O termo foi criado por Roy Thomas Fielding (2000). Os sistemas que adotam a arquitetura REST são frequentemente chamados de RESTful. Nesta arquitetura os *web* services são vistos como um recurso disponível na *web* e como tal, podem ser acessados por meio de uma URL. Assim sendo, os verbos HTTP GET e POST[44] podem ser utilizados para realizar as ações de criar, apagar, ler e atualizar um recurso. Esta coleção de quatro ações simplifica e, ao mesmo tempo, padroniza a arquitetura. O protocolo HTTP 1.1 está intimamente relacionado a arquitetura REST sem que exista interdependência entre este protocolo e a arquitetura (FIELDING, 2000).

Como estes protocolos são flexíveis e abrangentes, cada *site*, ou aplicativo *web*, determina um conjunto de regras de uso para o protocolo para atender uma arquitetura específica adotada, ou sua própria API[45]. O *site Programmable Web*[46] contém um diretório com milhares de APIs que permitem a interação entre dezenas de milhares de aplicativos *web* diferentes.

2.6 *WEB* MINING

Web mining, ou mineração *web*, é a ciência de recuperar informação da *web* com o uso da estrutura de hipertextos e baseada em técnicas de mineração de dados e textos (CHAKRABARTI, 2003). A *web* é, sem dúvida, o maior repositório de textos conectados por hipertexto disponível no planeta. A crescente complexidade da *web* e

[43] Do inglês: transferência do estado representacional.
[44] Do inglês: pegar e enviar.
[45] Do Inglês: camada de interação entre programas.
[46] *Programmable WEB*. Disponível em: http://www.programmable*web*.com, acessado em 04 de ago. 2012.

a riqueza do seu conteúdo tornam cada vez mais complexas as tarefas de recuperar informações deste ambiente (CHAKRABARTI, 2003).

Mesmo contendo uma grande quantidade de informação na forma de artigos postados em *sites* e *blogs*, comentários, fóruns e redes sociais virtuais, a extração automatizada de informações de páginas *web* não é uma tarefa trivial (PASTERNACK; ROTH, 2009). As páginas *web* são conhecidas por sua estrutura complexa, até mesmo caótica, e pelo uso indiscriminado de componentes de texto, tais como menus, logotipos, anúncios, cabeçalhos e rodapés que dificultam a recuperação de informações (THOMSEN; BRABRAND, 2012).

Uma das características que pode ser utilizada para separar a informação desejada é a própria estrutura sintática da página *web*. Mesmo que a informação em si não seja estruturada, em uma página *web,* esta informação está contida em uma estrutura fixa, determinada pela sintaxe do HTML e pela árvore *Document Object Model*[47] (DOM) (NICOL; WOOD *et al.*, 2001) que a espelha (PASTERNACK; ROTH, 2009).

As técnicas de extração de conteúdo de páginas *web*, em inglês *scrapping*, são, em grande parte, baseadas na varredura da árvore DOM. Estas técnicas adotam uma abordagem supervisionada e se baseiam em regras manualmente especificadas que indicam qual parte da página deve, ou não, ser considerada, utilizando-se a frequência ou o uso de determinadas *tags*, ou padrões de *tags* (BAR-YOSSEF; RAJAGOPALAN, 2002). Estudos anteriores mostraram que esta técnica é eficiente em um volume que varia entre 40% e 50% das páginas *web* (CHEN, MA e ZHANG, 2003).

Ainda explorando a árvore DOM, Bajula (2006) adota uma técnica semissupervisionada baseada em segmentação visual, para selecionar a área desejada na página, reduzindo a entropia da árvore DOM e, alimenta um sistema de aprendizagem não supervisionado, estatístico, para entender a composição da árvore DOM local e recuperar a informação desejada (BAJULA, 2006).

As técnicas baseadas na análise da estrutura da página *web*, mediante o uso da árvore DOM ou não, são eficientes apenas em casos específicos. No entanto, são apenas aceitáveis em uso genérico (PASTERNACK; ROTH, 2009). A complexidade das páginas *web* acaba por forçar a criação de um grande número de regras, frequentemente baseadas no uso de expressões regulares, que torna o aprendizado

[47] Do inglês: modelo de documento em objetos.

de máquina difícil, suscetível a erros e de vida útil reduzida (PASTERNACK; ROTH, 2009).

Pesquisas recentes, Nicol, Wood *et al.* (2001), Pasternack e Roth (2009) e Kohlshütter, Fankhouser e Nejdl (2010), incorporam técnicas estatísticas e de mineração de textos ao processo tentando encontrar uma forma mais eficaz de separar as partes da página *web* utilizando principalmente as características do texto.

Pasternack e Roth (2009) dividiram o problema de extração em três partes: decidir que a página contém um artigo; identificar todos os caracteres entre a primeira e última palavra deste artigo e remover tudo que não é “artigo”.

Em seu estudo, Pasternack e Roth (2009), definem “artigo” como sendo um contínuo coerente de prosa sobre um único tópico, ou sobre tópicos relacionados, que compõem o conteúdo informacional da página *web*. Esta pesquisa adota esta definição. O estudo de Pasternack e Roth (2009) descreve uma técnica de redução a *tokens*[48] do conteúdo *web* e a aplicação de um algoritmo de aprimoramento de subsequência máxima para uma classificação local do valor de cada *token*, baseada em um algoritmo semissupervisionado do tipo Naive-Bayes treinado com um *corpus* pré-classificado.

A redução a *tokens*, realizada como uso do *Stemming*[49] de Porter (PORTER, 2006) torna esta técnica pouco interessante para o português como é falado e escrito no Brasil requerendo adaptações às características do idioma português (LACERDA; MALHEIROS, 2006). Pesquisas como as de Russell e Norvig (2004) e Orengo e Santos (2007) parecem indicar que, técnicas como o *Stemming* podem aumentar a entropia em vez de reduzi-la, ainda que o algoritmo proposto por Orengo e Santos (2007) atenda as necessidades específicas de redução da língua portuguesa.

Uma das técnicas estatísticas de detecção dos artigos em páginas *web*, segundo a definição de Pasternack e Roth (2009), mais interessantes para os propósitos desta pesquisa, o algoritmo *Boilerplate*, baseia-se apenas no número de palavras e na densidade de *links* em cada seção do documento *web*. Partindo do princípio que áreas com muitos *links* contém pouco conteúdo informacional (KOHLSCHÜTTER; FANKHAUSER; NEJDL, 2010).

[48] Em inglês a palavra *token* representa uma unidade semântica, uma característica de distinção, algo que distingue (HORNBY, 2010). Em mineração entende-se *token*, como a unidade textual que contém o significado (CHAKRABARTI, 2003).

[49] Do inglês: decorrentes.

Boilerplate é uma palavra em inglês usada para descrever um material jornalístico que pode ser usado repetidamente em edição de jornais que, usualmente, era distribuído em uma placa de metal parecida com as placas usadas para ferver água no século XIX (HORNBY, 2010). A expressão chegou ao século XXI significando uma parte de texto importante e reutilizável, em artigos de jornais, revistas e *sites web*.

Kohlschüstter, Frankhauser e Nekdl (2010) não justificam a escolha do termo, mas referem-se à extração de artigos como extração de *Boilerplate*, implicitamente atrelando os significados *Boilerplate* e artigo. A técnica proposta pelos autores consiste na seleção dos blocos de texto contidos nas *tags* HTML com características de bloco (<p>, , <td>, etc.), a geração dos *tokens* de texto destas áreas, o cálculo de médias de sentenças, *tokens* e linhas e, finalmente o cálculo da densidade de *links* e palavras em cada área. O cálculo da densidade de palavras é realizado com o uso de um comprimento arbitrário de caracteres aplicado a cada linha (oitenta no estudo original) (KOHLSCHÜTTER; FANKHAUSER; NEJDL, 2010), como pode ser visto na Equação 14:

$$p_{(b)} = \frac{\text{Número de tokens em b}}{\text{Número de linhas em b}} \qquad (14)$$

Onde, a densidade de *links* é determinada pela relação entre o número de *tokens* contidos em *tags* <a> e o número de *tokens* contido no bloco.

Os resultados obtidos por Kohlschüstter, Frankhauser e Nekdl (2010) apresentam percentuais de sucesso na ordem de 98,1% a um custo de processamento relativamente baixo quando comparado aos outros algoritmos de mineração utilizados para a mesma finalidade. O custo de processamento para a separação do artigo (ou *Boilerplate*), onde está o conteúdo informacional, da página resume-se a contagem de palavras e cálculos de densidade, sendo dispensável a utilização das técnicas de *Stemming*.

O estudo de Kohlschüstter, Frankhauser e Nekdl (2010) indica que os textos das páginas *web*, considerando apenas as densidades de palavras e textos, pode ser dividido em dois grandes grupos: textos curtos, que incluem menus e anúncios e textos longos que incluem os textos com conteúdo informacional.

2.6.1 Análise de opinião

A mineração de opinião é uma técnica de mineração de dados que pretende classificar textos de acordo com o conteúdo emocional, ou opinativo (LIU, 2010). Técnicas de análise de opinião já foram utilizadas para tarefas tão distintas quanto prever o valor de ações em bolsas de valores, utilizando a opinião de usuários da rede social Twitter® (BOLLEN; MAO; ZENG, 2011) ou fazer uma análise da competição setor automobilístico da China (XINZHOU; QIANG; ANQI, 2012).

Um dos primeiros estudos nesta área usou as palavras de referência: *excellent* e *poor*[50]. Retiradas do idioma inglês, para classificar um conjunto de textos com objetivos semânticos usando uma métrica estatística para comparar a distância entre os textos avaliados e as palavras de referência (TURNEY, 2002). Contudo, a maior parte dos estudiosos parece concordar que o estudo seminal nesta área seja o estudo de Bo Pang, Lillian Lee e Shivakumar Vaithyanathan (2002). Neste estudo os pesquisadores aplicaram técnicas de aprendizado de máquina à classificação de textos considerando que pode existir um número relevante de frases que possuem características negativas ou positivas ainda que não possuam nenhuma palavra com estas características.

O estudo de Pang, Lee e Vaithyanathan (2002) foi realizado no domínio da resenha de filmes, considerando frases e palavras contidas em revisões realizadas por profissionais e amadores. O mérito do estudo foi verificar se a análise de opinião poderia ser realizada como um caso especial das técnicas de classificação. Para tal, utilizaram os classificadores por algoritmos probabilísticos - Naive-Bayes e Máxima Entropia, e *Support Vector Machines*[51] (SVM). Obtendo resultados melhores com o uso do SVM. Ressalte-se que em todos os casos tenha sido usado o mesmo conjunto de treinamento. Todavia, o SVM é conhecido por ser especialmente caro para a classificação de textos que requeiram uma base de treinamento extensa (COLAS; BRAZDIL, 2006).

Em redes sociais virtuais onde os textos são notadamente curtos, e em alguns casos limitados, as técnicas de aprendizado de máquina probabilísticas apresentam resultados interessantes e promissores (BOLLEN; MAO; ZENG, 2011). Bibliotecas

[50] Do inglês: excelente e pobre.
[51] Do inglês: máquina de vetores de suporte.

disponíveis *online* como o OpinionMiner[52] permitem a classificação de textos em inglês de acordo com o conteúdo opinativo, de cada fragmento mediante uma técnica em duas fases. A primeira fase faz uso de uma seleção de sentenças baseadas em um classificador Naive-Bayes para a determinação da subjetividade ou não de uma sentença. Na segunda fase realiza-se uma classificação de sentenças baseada em um arquivo de treinamento construído sobre um léxico anotado da língua inglesa (WILSON; HOFFMANN *et al.*, 2005).

Este modelo, originalmente proposto por Pang e Lee (PANG e LEE, 2004), de divisão do algoritmo em duas fases, uma que separa e classifica as sentenças e outra onde as sentenças são avaliadas de acordo com seu conteúdo tem sido adotado por pesquisadores que estudam o uso de algoritmos diversos em cada fase (MARKERT; HOU; STRUBE, 2012) (TAN; LEE *et al.*, 2011).

Em língua portuguesa, está sendo desenvolvido um estudo de pesquisa pela equipe do Data Management and Information Retrieval[53] de Lisboa, Portugal. Neste grupo, Mário J. Silva, Paula Carvalho, Carlos Costa, Luís Sarmento (2010) desenvolveram um léxico, o SentiLex-PT 01[54], anotado em português com 25.406 formas flexionadas de 6.321 lemas da língua portuguesa. Este léxico contém anotações referentes a polaridade, se positiva, negativa ou neutra; gênero; a natureza do alvo da polaridade e a forma como esta polaridade foi atribuída, se manual ou automática e já se mostrou efetivo para análise de opinião (SILVA; CARVALHO *et al.*, 2010).

2.6.2 Técnicas de varredura na *web*

Um *web-crawler* é um programa que, a partir de uma determinada URL, grava todo o conteúdo da página *web* e, recursivamente recupera todos os *links* para outras páginas *web* repetindo o processo até que não existam mais páginas para serem analisadas ou que alguma condição limite seja atingida (CHANDRAMOULI; GAUCH; ENO, 2010).

[52] OPINIONMINER: Disponível em: http://www.cs.pitt.edu/mpqa/opinionfinder_1.html, acessado em 28 jul. 2012.
[53] DMIR. Disponível em: http://dmir.inesc-id.pt/project/Main_Page, acessado em 31 jul. 2012.
[54] SENTILEX-PT 01. Disponível em: http://dmir.inesc-id.pt/project/SentiLex-PT_01, acessado em 30 jul. 2012.

O termo em inglês *web-crawler* não tem tradução consensual em português, mas pode ser entendido como colheita, varredura ou recuperação. Em inglês, antes do sentido usado na tecnologia da informação, referia-se apenas ao movimento feito por tratores no campo durante a colheita, deriva do movimento das esteiras característico de alguns tratores e veículos militares (HORNBY, 2010).

A prática de referenciar documentos de acordo com o interesse pessoal de cada autor na *web* cria uma rede de interconexões que constitui um exemplo de grafo, já classificado como rede complexa (BARABÁSI; DEZS *et al.*, 2003). Sendo assim, a única forma efetiva de coletar informação na *web* é o uso de um sistema de *crawling*, que significa varrer cada documento disponível em busca de *links* para outros documentos e continuar este processo sucessivamente até que não existam documentos para serem lidos, ou que sejam recuperadas informações suficientes (CHAKRABARTI, 2003). O primeiro *web-crawler*, o *Wanderer* de Matthew Gray (GRAY, 1993), foi escrito na primavera de 1993, praticamente coincidindo com a criação do *Mosaic*, o primeiro navegador *web* (HEYDON; NAJORK, 1999). Os *sites* de busca que varrem a *web* necessitam de aplicativos de *crawling* que sejam, essencialmente, escaláveis e extensíveis (HEYDON; NAJORK, 1999).

A criação e a manutenção de um serviço de *web-crawler* são tarefas não triviais e apresentam dificuldades complexas, notadamente para serviços de indexação que dependam da atualidade das informações recolhidas como fator de qualidade, ou determinante de uso (HSIEH; GRIBBLE; LEVY, 2010). No entanto, segundo Manning, Raghavan e Schutze (2009) existem funcionalidades básicas que um *crawler* deve possuir:

a) robustez: a capacidade de evitar armadilhas de recursividade, quando uma página aponta para outra e esta aponta para a primeira, sejam estas armadilhas propositais ou não;

b) adequação: capacidade de entender os requisitos das páginas e serviços que são varridos;

c) distribuição: capacidade de varrer vários documentos de forma independente e paralela;

d) escalabilidade: arquitetura flexível que permita o aumento de capacidade de varredura, com a inclusão de novas máquinas, de forma transparente ao processo de varredura;

e) eficiência: melhor uso possível dos recursos disponíveis na arquitetura de hardware;

f) qualidade: a capacidade de reconhecer a qualidade das páginas seja mediante o uso de um dicionário histórico ou de avaliação preliminar para varrer primeiro as páginas de melhor qualidade;

g) atualidade: capacidade de permitir a atualização contínua da base armazenada;

h) extensibilidade: capacidade adaptação ao uso de novos protocolos, formatos de informação ou dispositivos, sem alteração no serviço.

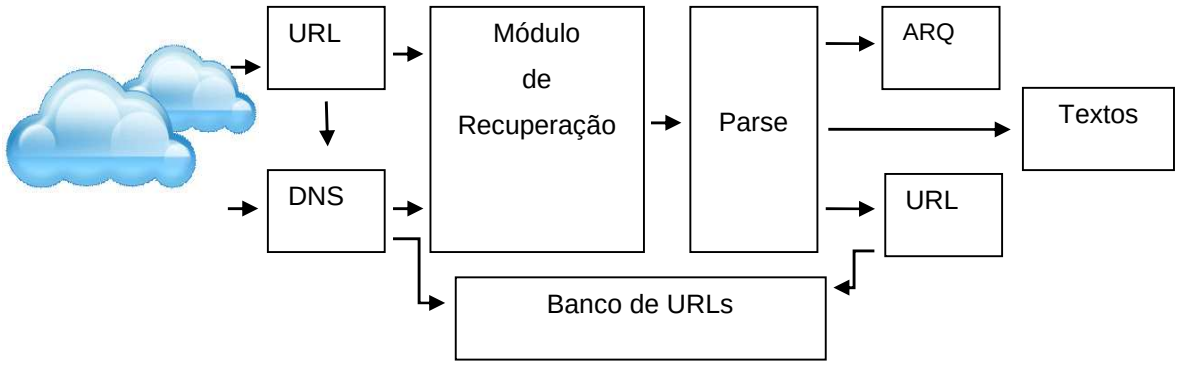
2.6.2.1 Arquitetura típica de um *web-crawler*

A arquitetura típica de um *web-crawler* deve incluir um módulo para a resolução de URLs, um módulo de recuperação e um módulo de *parse*[55] (CHAKRABARTI, 2003). Em inglês a palavra *parse* pode ser entendida como sendo a ação de quebrar uma sentença em seus componentes básicos para análise.

No caso de um *web-crawler* a função do *parse* é separar o texto e outras informações contidas em uma página *web* para avaliação posterior. A Figura 5 mostra um diagrama em blocos de um sistema *web-crawler* conforme proposto por Chakrabarti (2003).

[55] Do inglês: analisar.

FIGURA 9 - DIAGRAMA DE UM *WEB CRAWLER*



FONTE: tradução livre e adaptação de Hisieh, Gribble e Levy (2010).

O módulo de resolução de domínios em endereços IP atende as necessidades de velocidade do *web-crawler*. Uma vez que o serviço de *Domain Name Service*[56] (DNS) tenha sido acessado e o nome do domínio resolvido não é necessário repetir todo o processo de resolução para cada documento de um determinado domínio, ou *site*. Em contra partida, documentos que sejam referenciados em *sites* diferentes irão requer a atualização das informações de resolução IP (HSIEH; GRIBBLE; LEVY, 2010).

O banco de URL*s* armazena o IP das URL**s** já resolvidas e as URL**s** dos *links* que forem encontradas nos documentos recuperados. A principal função deste banco é minimizar o tempo de recuperação diminuindo os acessos repetidos a *Internet*.

O módulo de recuperação gerencia tanto os processo de resolução quanto os processos de repetição, em caso de erro, e de armazenamento na camada de persistência.

O módulo de *parse* deve ser capaz de entender as informações contidas em cada documento *web* recolhido e, quando for o caso, descartar os documentos que não contenham informações que atendam as especificações anteriormente estabelecidas.

As pesquisas recentes de varredura, ou *crawling*, na *web* incluem o Mercator, estudo de Heydon e Najork (1999); o Saltícius desenvolvido por Robin Burke (2001); o Wire que é uma pesquisa de Baeza-Yates e Castillo (2002) e ainda uma proposta de *crawler* em paralelo de Cho e Garcia-Molina (2002). Entretanto, nenhum deles é

[56] Do inglês: serviço de nomes de domínio.

tão interessante para este estudo quando a arquitetura proposta por Chakrabarti (2003) anteriormente descrita no item **Erro! Fonte de referência não encontrada.**.

2.6.2.2 Varrendo redes sociais virtuais

A análise de redes sociais, aplicada às conexões dos documentos *web*, ou as relações compartilhadas em redes sociais virtuais já foram usadas como fator de classificação das páginas recolhidas (CHANDRAMOULI; GAUCH; ENO, 2010).

Cada serviço de rede social virtual possui o seu conjunto particular de dados, sua estrutura própria de apresentação e um conjunto específico de informações (GJOKA; KURANT *et al.*, 2011). O *crawling* de uma rede social virtual, com o objetivo de coletar informações de identificação, e da produção, de cada usuário pode ser realizado mediante a criação de um gráfico social para esta rede (GJOKA; KURANT *et al.*, 2011). Mesmo assim, sendo esta uma rede complexa, o uso do gráfico não dispensa a tarefa de percorrer todo a rede (BARABÁSI; DEZS *et al.*, 2003).

A tarefa de fazer um *crawler* em uma rede social virtual não é trivial. A quantidade de informação produzida (vídeo, áudio, textos e *links*), demanda software e *hardware* específicos (YE; LANG; WU, 2010). Acrescente-se a isso o fato que muitas das redes sociais virtuais usam AJAX e *Dynamic Hypertext Markup Language*[57] (DHTML) para aumentar a interação com os usuários e facilitar o uso do *site*. Tecnologias como estas, que tornam a geração de páginas dinâmicas, dependente da ação dos usuários e independente de *links* fixos aumentando a complexidade do processo de varredura e coleta (YE; LANG; WU, 2010). A opção pelo uso do grafo social apresenta dificuldades adicionais.

O principal ativo de uma empresa que mantém um *site* prestando o serviço de rede social virtual é justamente o grafo social (YE; LANG; WU, 2010). Cada serviço limita o acesso às informações dos usuários com o intuito de proteger seus ativos financeiros e garantir a segurança dos dados dos usuários (GJOKA; KURANT *et al.*, 2011).

Da mesma forma que um *web-crawler* tradicional pode acabar em uma página sem nenhum *link* externo, por falta de intenção do autor, ou pela existência de um sistema de geração de páginas dinâmico (HSIEH; GRIBBLE; LEVY, 2010). Os *social-*

[57] Do inglês: HTML dinâmico. Trata-se de um termo para englobar um conjunto de tecnologias que adicionam movimentos e interação com usuários a páginas *web*.

crawlers, utilizando o grafo social, podem encontrar usuários que não compartilham informações na rede e, graças a essa falta de relacionamento, ficam privados das informações necessárias a continuidade da varredura (YE; LANG; WU, 2010). Os pesquisadores Ye e Wu (2010) listaram as características mínimas necessárias para definir a eficiência de um *social-crawler*:

a) eficiência: a velocidade com que os atores, nós da rede, são varridos;

b) sensibilidade: a como a estrutura da rede social e usuários que não compartilham informações afetam a coleta;

c) viés: indica a diferença estatística entre as propriedades estatísticas do grafo amostrado e do grafo global.

A literatura científica classifica os algoritmos de amostragem da rede social em dois grandes grupos: grafos transversos e caminhadas randômicas. Nas técnicas transversas os atores são visitados sem reposição. Uma vez que um ator é visitado não é amostrado novamente. Neste caso, estão os algoritmos *Breadth-Search-First*[58]**,** *Depth-First Search*[59], *Forest Fire*[60] e o *Snowball Sampling*[61] (GJOKA; KURANT *et al.*, 2011). Entre estes o se destaca o *Breadth-Search-First* por apresentar a característica de amostrar todos os atores e conexões de uma determinada parte do grafo social (NAJORK; WIENER, 2001).

As técnicas de caminhadas randômicas incluem os algoritmos *Metropolis-Hasting Random Walk*[62] e *Re-Weighted Random Walk*[63] (GJOKA; KURANT *et al.*, 2011). Ambos possuem a vantagem de não dependerem do provedor do serviço de rede social virtual para o fornecimento de amostragens uniformes e não precisam de um número excessivo de atores para qualificar a rede (YE; LANG; WU, 2010). Esta independência é fundamental para os objetivos desta pesquisa e, desta forma,

O *Metropolis-Hasting Random Walk* foi o algoritmo escolhido para recuperação de interações sociais, graças a sua capacidade de gerar amostras estatisticamente representativas e de gerar uma imagem estática da rede social virtual. A amostra

[58] Do inglês: primeira largura de busca.
[59] Do inglês: primeira busca profunda.
[60] Do inglês: floresta de fogo.
[61] Do inglês: amostragem em bola de neve.
[62] Do inglês: caminhada randômica Metropolis-Hastings
[63] Do inglês: reponderação de caminhada randômica.

obtida parece possuir significância estatística como ferramenta de análise comportamental ou de classificação de *links* (GJOKA; KURANT *et al.*, 2011).

A caminhada randômica baseada no algoritmo *Metropolis-hashing* foi desenvolvida com o objetivo de gerar amostras, estatisticamente significantes, de uma rede de forma a suprir qualquer viés na seleção dos nós, durante a amostragem (GJOKA; KURANT *et al.*, 2011).

2.6.3 Classificação de páginas *web*

Os *web-crawling* utilizam a estrutura de conexões, *links*, entre documentos na *web* para encontrar as páginas, recolher as informações, localizar nestas páginas as palavras chave, ou *keywords* inglês, e, em um segundo momento, fornecer estes *links,* endereços e palavras chaves para os algoritmos de classificação de informação..

Entre as políticas de classificação destaca-se o Pagerank de Page, Brin e Rajeev (1999) utilizado pelo Google e por outros pesquisadores como indicador de qualidade das páginas recolhidas (NAJORK; WIENER, 2001).

2.6.3.1 *Keywords*

Keywords[64], ou palavras chaves em português, são estudadas como termos capazes de indicar uma linguagem, ou conhecimento, específico (WILLIAMS, 1983). Desde os primeiros anos da *web sites* de busca usam as *keywords* como indicadores de autoridade e classificação de páginas *web (CHAKRABARTI, 2003)*. Mesmo *sites* como o Google® e o Bing® que usam algoritmos de classificação diferentes, continuam usando as *keywords* em seus algoritmos (TEEVAN; ALVARADO *et al.*, 2004) (BAEZA-YATES; RIBEIRO-NETO, 1999). Ainda que muitos destes algoritmos sejam protegidos por uma capa de sigilo, é possível inferir a importância das *keywords* para a ordem de classificação por meio da divulgação de alterações nos algoritmos de busca e classificação usados por estes *sites* (SEOMOZ, 2011).

[64] Do inglês: palavra chave.

3 PROCEDIMENTOS METODOLÓGICOS

Destaca-se aqui, a característica metodológica da pesquisa, levando-se em consideração as necessidades específicas do problema. Desta forma, a pesquisa está descrita em termos de métodos e técnicas com vistas a conferir o caráter austero necessário e indispensável à pesquisa científica.

O presente estudo pode ser classificado como experimental, exploratório, descritivo e metodológico.

A natureza experimental se manifesta na criação de um conjunto de ferramentas técnico matemáticas para recuperar, classificar e avaliar informações recolhidas na *web* e em redes sociais virtuais por esta suportada.

Para a obtenção de conhecimento referente às áreas estudadas, foi realizado um estudo exploratório de caráter bibliográfico segundo os conceitos de Cervo e Bervian (1983) com atenção a história e a busca dos princípios básicos de cada área pesquisada.

O caráter descritivo (TRIVIÑOS, 1987) da pesquisa aparece na caracterização das redes sociais, virtuais ou não, na descrição da estrutura da *web*, na descrição dos métodos de mineração na *web* e de análise de opinião.

Já as características metodológicas (DEMO, 1985) aparecem manifestas nos métodos e técnicas desenvolvidos para ajuste dos fatores de adequação dos algoritmos de classificação. Segundo Demo (1985, pg. 13):

> há pesquisa metodológica, dedicada a indagar por instrumentos, por caminhos, por modos de se fazer ciência, ou a produzir técnicas de tratamento da realidade, ou a discutir abordagens teórico-práticas (DEMO, 1985).

Sendo replicável esta pesquisa atende esta classificação por meio do desenvolvimento de instrumentos de coleta de dados, com o intuito de permitir a criação de um ambiente de definição para um constructo que se deseja medir, criar ferramentas além de testar estas mesmas ferramentas em relação à precisão e confiabilidade (DEMO, 1985).

3.1 CONTEXTUALIZAÇÃO METODOLÓGICA

A escolha das ferramentas para a criação do ambiente de testes levou em consideração o seu uso em pesquisas científicas e as facilidades de acesso e custo. Para a criação do ambiente de testes foi desenvolvido um sistema formado por um conjunto diverso de ferramentas disponíveis gratuitamente *online*, entre as quais é possível destacar o Google AdWords®[65], o buscador Microsoft Bing® e o algoritmo *Boilerplate* de Kohlshüster, Fankhauser e Nejdl (2010), diversas bibliotecas de software, *script*s desenvolvidos em software livre e de código aberto, o léxico SentiLex-PT 01 e *script*s especificamente desenvolvidos para programar as funcionalidades necessárias.

3.1.1 Especificação do domínio

O domínio escolhido para a prova de conceito deve ser representado pelo termo *notebook.* Um termo que representa um domínio ligado a tecnologia, sem apelo de gênero e que permeia toda a sociedade contendo um conjunto de informações diverso. Para evitar qualquer viés na determinação deste domínio, foi apresentado um questionário *online* para os usuários da rede social Twitter, seguidores da conta @depijama, administrada pelo autor. O questionário apresentado se encontra no Apêndice A desta pesquisa.

3.1.2 Questionário para especificação do domínio

Para selecionar quais informações são mais utilizadas para a tomada de decisão de compra de *notebooks* foi realizada uma pesquisa *online* utilizando-se para isso os seguidores da conta @depijama do Twitter que se identificaram como brasileiros junto a esta rede social virtual, um total de 1567 entre os 17017 seguidores no momento da realização do questionário. Destes 1567 inquiridos, 234 responderam o questionário, representando 15% do universo.

[65] Disponível na internet em: http:adwords.google.com/keywords

O universo foi definido mediante apenas um recorte realizado a partir do fator geográfico. Os usuários selecionados foram aqueles que, anteriormente e de forma espontânea e independente, haviam preenchido seus perfis no Twitter como estando localizado em uma cidade ou estado do Brasil ou, simplesmente com a palavra Brasil. Como esta informação é opcional nesta rede social virtual e a maior parte das pessoas, 65% do total de seguidores da conta @depijama, não preenche este campo existem apenas 5955 seguidores com esta informação preenchida e entre estes existem apenas 1567 que atendem os requisitos geográficos.

Optou-se pela utilização de um questionário com cinco perguntas diretas, de múltipla escolha, e uma pergunta aberta atendendo os requisitos definidos por Sue e Ritter (2007) para questionários *online*. Visando a simplicidade de uso pelos respondentes sem deixar de abranger as três das etapas de processo de compra definidos por Samara e Morcsh (2005): busca da informação, avaliação das alternativas do produto e avaliação das alternativas de compra. Que atendem a proposta de utilização do sistema desenvolvido para esta pesquisa fornecendo dados referentes à tomada de decisão. Foi aplicada uma e somente uma questão dependente de uma resposta anterior e apenas uma questão teve característica aberta. Ao fim do período de disponibilidade as respostas foram recolhidas e exportadas para o aplicativo Microsoft Excel® para análise e a geração dos gráficos.

3.1.3 Operacionalização do questionário

O questionário envolveu dois processos independentes: a recuperação da lista de todos os seguidores da conta @depijama e a aplicação do questionário propriamente dito.

Para a recuperação da lista de seguidores foi desenvolvido um *script* em *Hypertext Preprocessor*[66] (PHP), rodando no servidor Apache®[67] que enviou as requisições necessárias por meio da API do Twitter® e recuperou os dados referentes a todos os seguidores da conta @depijama que atenderam o recorte estabelecido. Uma vez que a lista de seguidores foi obtida, enviou para cada um deles uma

[66] Do inglês: preprocessador de hipertexto.
[67] O Apache é um servidor HTTP capaz de responder a requisições neste protocolo, esta categoria de aplicativos é popularmente conhecida como servidor *web*.

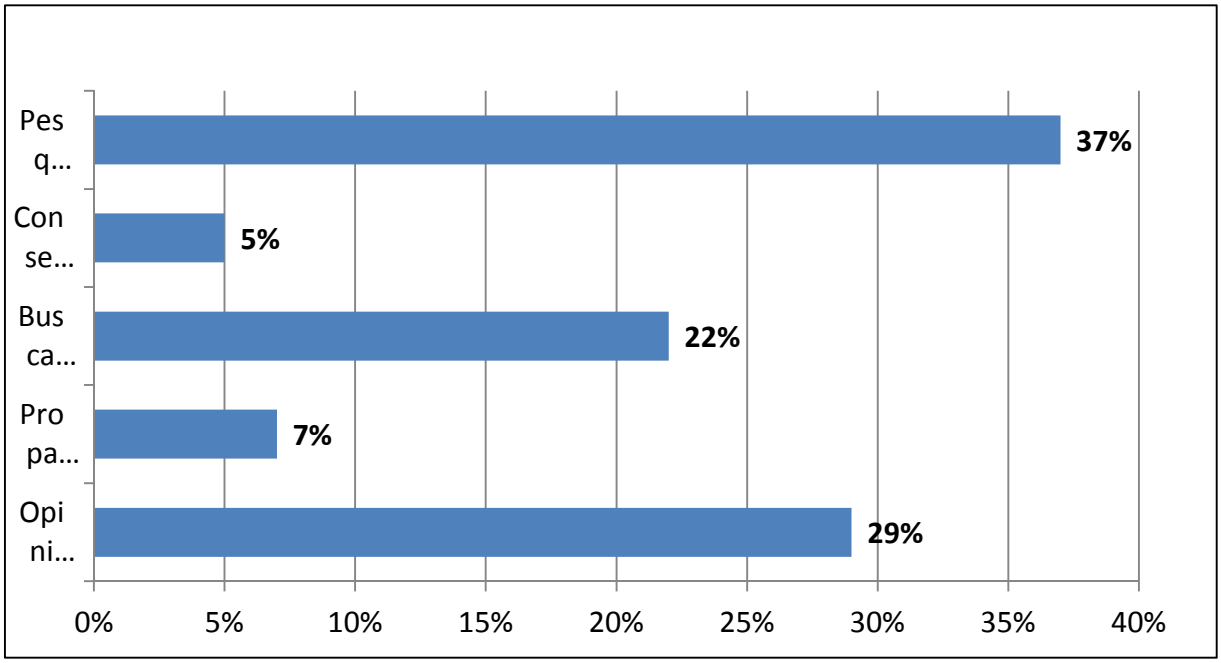
mensagem pessoal contendo uma solicitação de participação na pesquisa e um *link* para o *site* da pesquisa.

Para a realização do questionário foi usado o software livre LimeSurvey[68] também rodando em PHP no servidor Apache®. O questionário não exigiu nenhum tipo de registro por parte dos usuários, garantindo o anonimato, e esteve disponível entre os dias 10 de maio de 2012 e 11 de junho de 2012.

3.1.4 Descrição e análise do questionário

O resultado do questionário está apresentado na forma de gráficos para facilitar a visualização das diferenças percentuais entre as respostas.

GRÁFICO 1 - PARA COMPRAR UM NOTEBOOK QUAL DESTAS OPÇÕES É MAIS IMPORTANTE?



FONTE: o autor (2013)

É possível observar, no Gráfico 1 que as opções pesquisa de preço, opinião do amigo e busca na *Internet* se destacam. Com uma vantagem para pesquisa de preço, parecendo indicar que o preço é o fator decisivo na decisão de compra. Observe, no entanto, que a diferença percentual entre estes três fatores é igual ao valor percentual atribuído a propaganda.

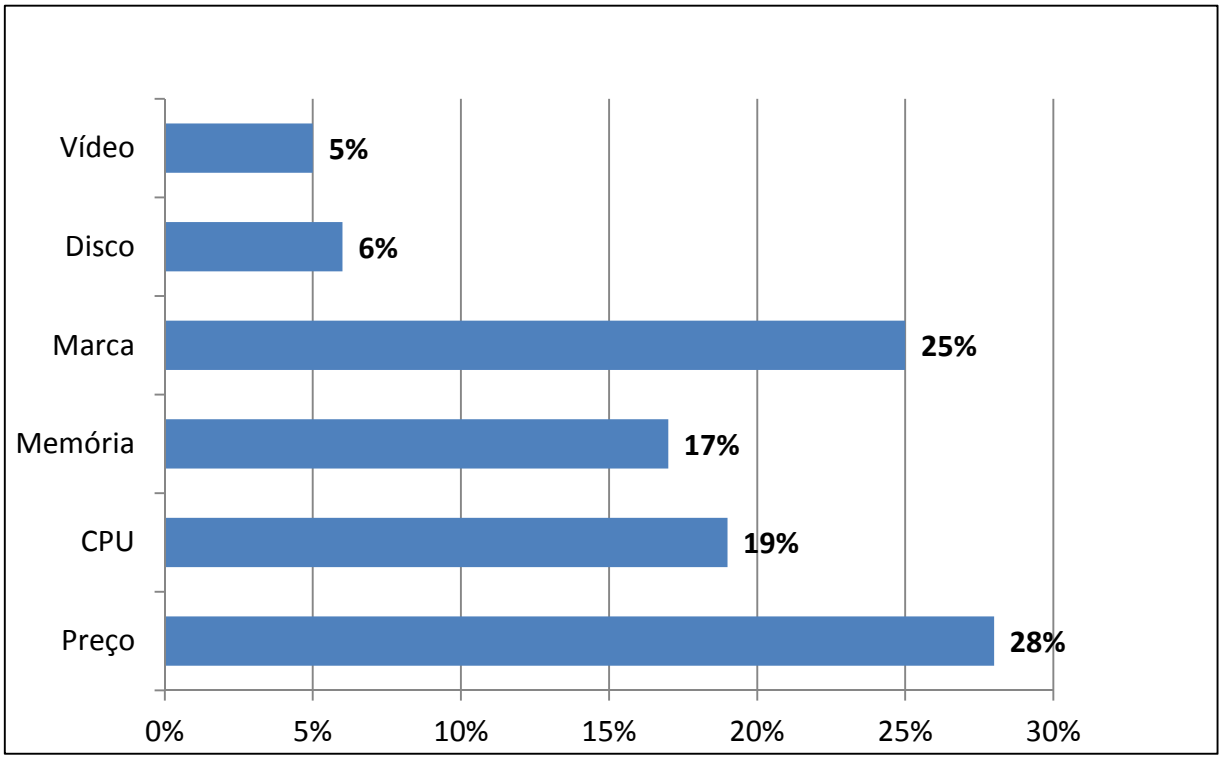
[68] LIMESURVEY. Disponível em: http://www.limesurvey.org/pt/, acessado em 01 ago. 2012.

No Gráfico 2 é possível observar o resultado percentual referente aos fatores considerados importantes para a tomada de decisão de compra segundo os respondentes.

As opções oferecidas foram: vídeo, disco, marca, memória, CPU e preço. Observa-se que existe uma pequena predominância dos fatores preço e marca e que as características relacionadas ao vídeo parece ser o fator menos importante para a decisão de compra.

As respostas à segunda questão parecem corroborar o resultado apresentado na questão anterior indicando que preço é um fator importante para a tomada de decisão de compra.

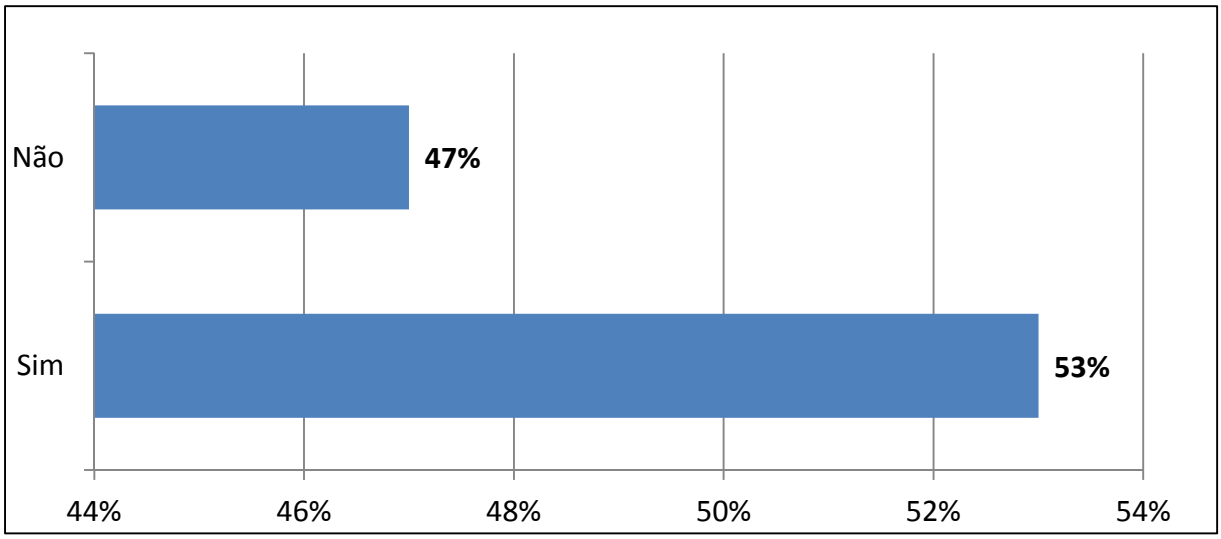
GRÁFICO 2 - QUANDO VOCÊ VAI COMPRAR UM NOTEBOOK, QUAL O FATOR MAIS IMPORTANTE PARA SUA ESCOLHA?



FONTE: o autor (2013)

No Gráfico 3 é possível observar o resultado da pergunta três. Esta pergunta foi criada na intenção de identificar o efeito da marca na decisão de compra.

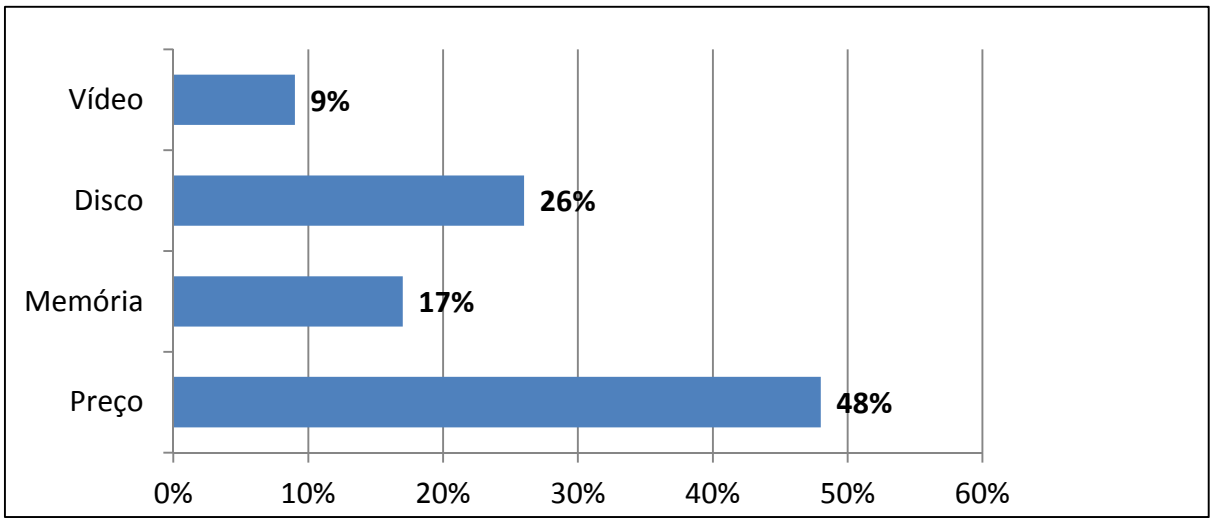
GRÁFICO 3 - CONSIDERANDO DOIS NOTEBOOKS COM O MESMO PREÇO E A MESMA CPU, VOCÊ TROCARIA DE MARCA?



FONTE: o autor (2013)

Todos os entrevistados que escolheram a opção não na pergunta de número 3, tiveram a opção de responder a pergunta quatro. Nesta pergunta, o objetivo foi tentar indicar que informação poderia provocar uma escolha de uma marca diferente, descartando-se a CPU[69]. O Gráfico 4 apresenta o resultado desta questão:

GRÁFICO 4 - SE VOCÊ NÃO TROCA DE MARCA DEVIDO A CPU O QUE O FARIA TROCAR DE MARCA?

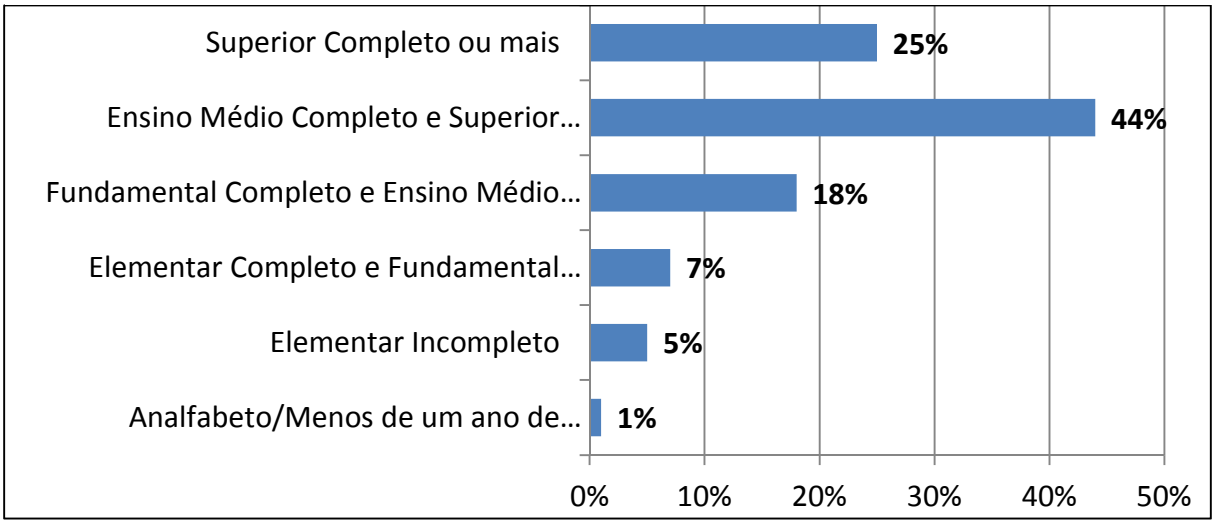


FONTE: o autor (2013)

[69] Abreviatura em inglês de Central Processing Unit ou unidade central de processamento

Observa-se que a indicação de preço como fator decisivo fica mais consistente. A próxima questão caracteriza os entrevistados de acordo com a educação em seis níveis. O Gráfico 5 apresenta o resultado obtido com esta questão:

GRÁFICO 5 - QUAL O SEU GRAU DE ESCOLARIDADE?



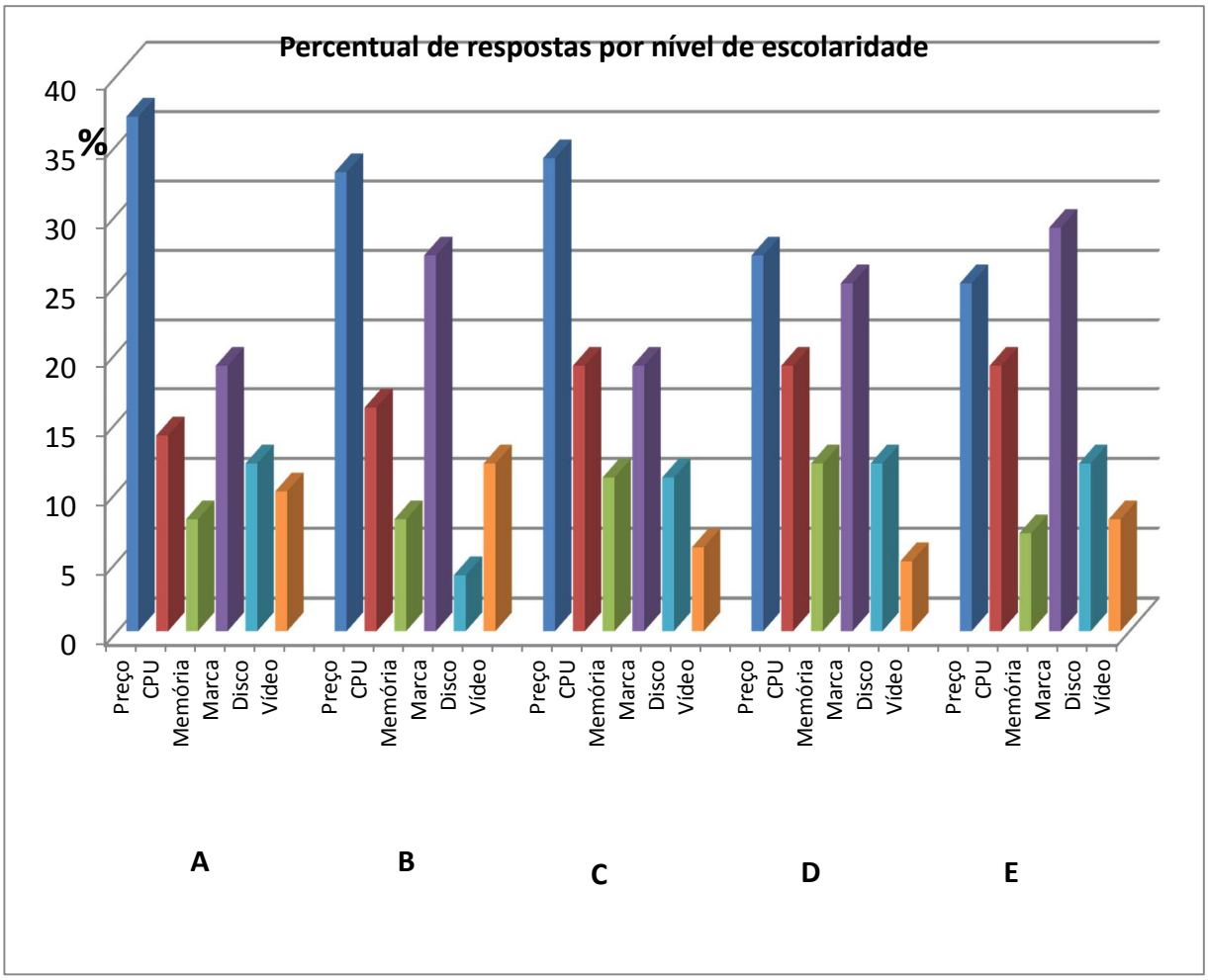
FONTE: o autor (2013)

3.1.4.1 Observações adicionais sobre o questionário

O nível de escolaridade serve com referência da forma em que a informação que pode ser ofertada. O levantamento de escolaridade serve como referência da capacidade de interpretação, ou letramento, da audiência (BRANDÃO; SPINILLO, 1998; SOARES, 1996) e permite o desenvolvimento de uma interface adequada a cada usuário melhorando a qualidade do conhecimento que pode ser adquirido (BRANDÃO; SPINILLO, 1998; SOARES, 1996).

Para observar o efeito do nível de escolaridade na tomada de decisão foram reavaliadas as respostas às questões anteriores considerando como amostra o nível escolar. No entanto, foi removido o primeiro nível (Analfabeto/menos de um ano de escolaridade) por causa da quantidade de respondentes que se classificou nesta faixa (menos de 1%), a distribuição encontrada está apresentada no Gráfico 6.

GRÁFICO 6 - RESPOSTAS A QUESTÃO 2 POR NÍVEL DE ESCOLARIDADE



FONTE: o autor (2013)

No Gráfico 6 as letras A, B, C, D, E representam:

A: elementar incompleto;
B: fundamental completo e ensino médio incompleto;
C: elementar completo e fundamental incompleto;
D: ensino médio completo e superior incompleto;
E: superior completo ou mais.

É possível observar que o grau de importância atribuído a cada uma das opções varia levemente de acordo com o nível de escolaridade. Destaca-se a variação dos fatores preço e marca.

O questionário parece indicar que fatores como preço e marca são importantes para a caracterização de um domínio relacionado a compra de *notebook* e como tal devem ser destacados e, possivelmente utilizados como balizadores. Fatores relacionados a capacidade de processamento e *hardware*, parecem ocupar um lugar secundário nesta decisão. Talvez agindo apenas como complemento ou, devido a

fatores sociais ou culturais (KOTLER; KELLER, 2006) que não foram explorados nesta pesquisa.

3.1.5 Seleção das palavras chave

A interação mais importante com o usuário se dá na criação do léxico de definição do domínio. Este léxico, um conjunto de palavras chaves (ou *keywords*, em inglês) representa a forma com que o usuário expressa suas ideias relacionadas ao domínio pesquisado.

Além dos termos explicitados pelo próprio questionário foram utilizados os termos citados na última pergunta do questionário, a pergunta número seis, à saber: “Em poucas palavras, explique por que você compraria um *notebook”*.

Os textos fornecidos pelos respondentes foram adicionados sequencialmente em um único arquivo de texto e sofreu um processamento específico. Inicialmente forma removidos os caracteres em maiúsculo e depois foram removidos os termos comumente usados em português.

A remoção dos termos comuns foi possível com a utilização de um léxico disponível *online* e usado com sucesso em algoritmos de mineração de textos (SNOWBALL TEAM, 2009).

Por fim, o texto contendo todas as respostas da pergunta seis foi transformado em uma distribuição de frequências e esta, por sua vez em uma nuvem de palavras com o uso da Linguagem de programação R. O resultado deste processamento pode ser visto na Figura 10.

Deve-se observar que a análise destes termos extrapola o escopo desta pesquisa. Estes termos servem apenas para complementar o conjunto de palavras chaves que é utilizado para a caracterização do domínio ampliando o léxico usado para busca e recuperação de páginas *web* e interações sociais.

FIGURA 10 - NUVEM DE PALAVRAS RESULTANTE DAS RESPOSTAS DA QUESTÃO 6.

FONTE: o autor (2013)

3.2 MODELO OPERACIONAL DO SISTEMA

O domínio de exploração foi definido através do uso do questionário. Este domínio representa extensivamente a informação necessária para a recuperação de páginas *web* e para a criação do gráfico social. A forma escolhida para caracterizar o domínio é o uso de *keywords*. Este léxico é ampliado com o uso da ferramenta de *keywords* do serviço Google AdWords® e fornecido ao sistema para alimentar os *scripts* de recuperação.

Uma vez que tenha acesso ao léxico, o sistema é responsável pelo *crawler* na *web* e nas redes sociais virtuais e pela criação da camada de persistência para armazenamento dos dados recolhidos.

Na camada de persistência estarão armazenados os dados colhidos diretamente da *web* e das redes sociais virtuais, o léxico de termos relativos ao domínio e os gráficos sociais virtuais resultantes da aplicação de cada termo do léxico nas redes sociais virtuais.

Cada informação recolhida sofrerá a aplicação de quatro fatores de classificação:

a) o primeiro fator de classificação dos dados recolhidos é determinado pela colocação das páginas *web* nas páginas de resposta das ferramentas de busca. Para este estudo foi utilizado o Microsoft Bing®;

b) o segundo fator de classificação é atribuído considerando-se o uso percebido da página *web* e a participação, ou não, de usuários na criação de conteúdo nesta página ou nas páginas do mesmo site. Informações oriundas de sites com comentários (fóruns, blogs e rede sociais virtuais) têm um peso maior que as informações retiradas de páginas estáticas ou de publicidade;

c) o terceiro fator de classificação é atribuído de acordo com as características de cada gráfico social virtual criado. Links encontrados em interações sociais que contenham as *keywords* têm um peso relacionado a existência ou não de um gráfico social virtual e a sua topologia dentro deste gráfico;

d) o quarto fator de classificação é determinado pela análise da opinião explicitada no texto recuperado através de algoritmos específicos para este fim.

Os fatores de classificação compõem o fator α de adequação do algoritmo Pagerank (PAGE; BRIN *et al.*, 1999).

Com relação ao gráfico social, as métricas de avaliação de nós e *links*, relativas ao grafo formado para cada *keyword* são utilizados como fatores de classificação e, novamente, comporão o *fator α* de adequação do mesmo Pagerank, desta vez aplicado apenas aos gráficos sociais virtuais criados para cada *keyword*.

O *crawler* é executado em três passos: páginas simples como *blogs*, de estrutura conhecida e repetitiva, são varridos de forma automática a partir da existência do arquivo *sitemap.xml* (GOOGLE, 2011) e sofrerão o processo de separação da informação com o uso do algoritmo *Boilerplate* (KOHLSCHÜTTER; FANKHAUSER; NEJDL, 2010); Páginas complexas, com alto grau de interação, como fóruns são varridos de acordo com o *sitemap.xml* acrescido de um modelo de varredura determinado de forma manual e característico de cada *site*. Por fim, as redes sociais *online* são varridas de acordo com as funcionalidades disponíveis mediante as APIs particulares de cada serviço.

Uma exceção deste processo de *crawler são os sites* de fabricantes e/ou venda de produtos. Estes *sites* são tratados de forma individual com a criação de um modelo de varredura e recolhimento supervisionado. Minimizando a existência de eventuais erros de interpretação pelos algoritmos. O *script* é personalizado para cada *site* e usará um conjunto de regras de varredura da árvore DOM baseadas em expressões regulares.

O resultado do processo é um conjunto de informações relacionadas com as *keywords* que são formatadas de acordo com as regras determinadas em um arquivo específico, em formato JSON (CROCKFORD, 2006), que deve especificar o dado, sua característica, e o formato de exibição que, para os fins desta pesquisa, se limitou a determinar a forma como o titulo da página classificada aparecerá para o usuário.

3.3 ARQUITETURA DO SISTEMA

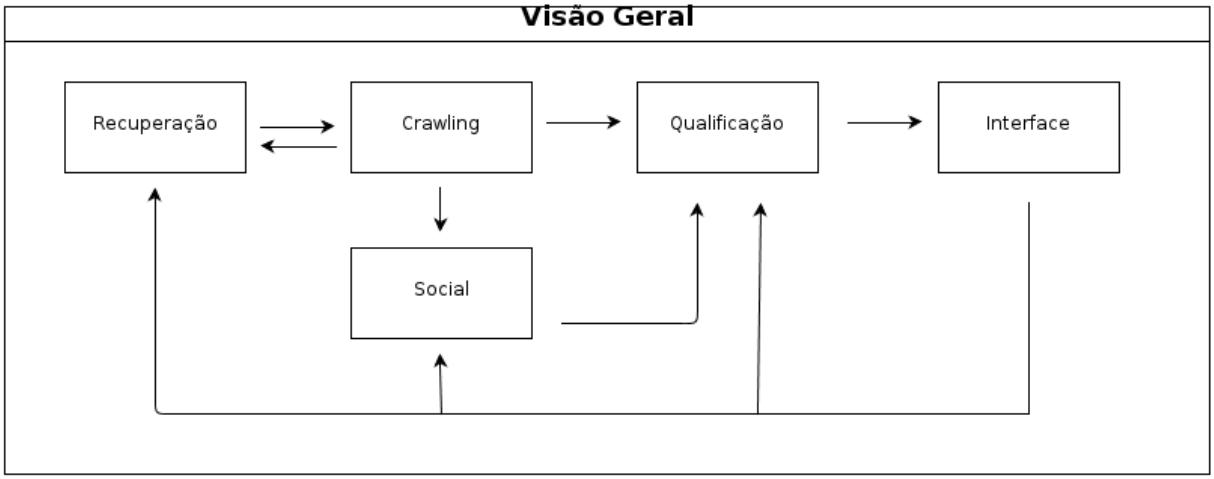
A arquitetura de software utilizada, por força das bibliotecas e componentes utilizados baseia-se, não rigidamente, na arquitetura *Model View Controller*[70] (MVC), desenvolvida com técnicas de orientação a objetos. Entretanto, não está restrita a nenhuma linguagem de programação ou arquitetura de software. De fato, na construção do sistema foram utilizadas bibliotecas de funções e *script*s em R, PHP, Python, C e Java de acordo com a necessidade pontual de cada uma. Além de ferramentas prontas e disponíveis em software livre ou de forma gratuita na *internet* em arquiteturas MVC, cliente servidor e orientada a eventos.

As bibliotecas de função, conjunto de ferramentas de desenvolvimento, aplicativos ou componentes utilizados durante a pesquisa foram escolhidas por atenderem apenas duas condições: serem livres e gratuitas e terem sido utilizadas anteriormente em estudos de pesquisa científica.

O sistema consiste de cinco módulos: recuperação, *crawling*, classificação, social e relatório. Como pode ser visto na Figura 11:

[70] Do inglês: modelo, visualização e controlador.

FIGURA 11 - VISÃO GERAL DO SISTEMA



FONTE: o autor (2013)

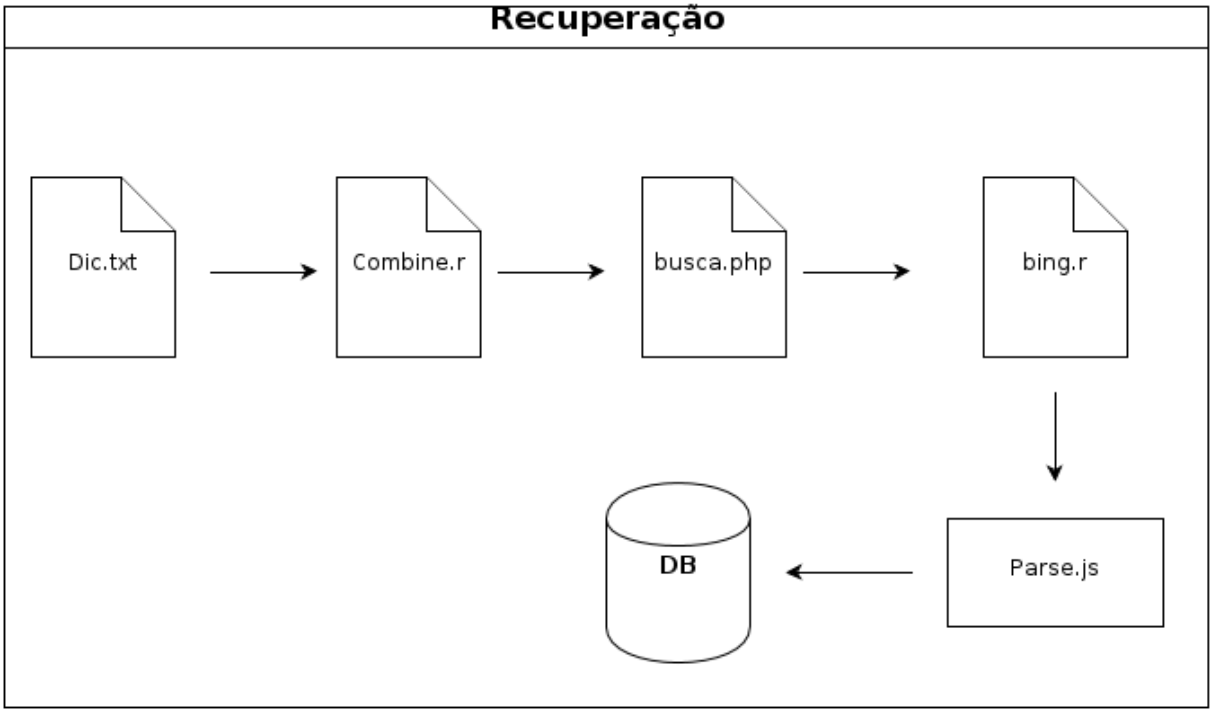
Cada módulo foi dividido em um conjunto de *script*s, ou programas interligados em uma forma particular e específica. Como não foi adotada nenhuma arquitetura padrão para o desenvolvimento de software, não existe um protocolo único de comunicação entre os módulos ou entre os componentes internos de cada módulo. A troca de informações entre os módulos se dá por meio de chamadas diretas a procedimentos ou mediante o uso da camada de persistência.

3.3.1 Módulo de recuperação

Consiste dos *script*s e programas necessários para a seleção de URL*s* relevantes à recuperação das informações desejadas. Este módulo utiliza um arquivo de texto gerado pelo usuário contendo as *keywords*, denominado *dic.txt*; Um *script* responsável pela geração de combinações para os termos de busca a partir da consulta ao serviço *Google AdWords Keywords Tool* ®[71]; Um arquivo para a interface com o Microsoft Bing®; e um arquivo de *parse*. O diagrama em blocos deste módulo pode ser visto na Figura 12:

[71] Disponível na Internet em: https://adwords.google.com/o/keywordtool

FIGURA 12 - MÓDULO DE RECUPERAÇÃO



FONTE: O autor (2013)

O *Google AdWords Keyword Tool*® permite a geração de uma lista com os termos de busca mais utilizados no Google®, para cada conjunto de *keywords* submetido. O conjunto de termos submetido ao *Google AdWords Keywords Tool*® originou-se do questionário aplicado *online*.

A opção pelo serviço do Google® foi realizada levando-se em consideração a gratuidade, a facilidade de uso e, acima de tudo, a relevância dos termos gerados em relação às *keywords* fornecidas. Este serviço, por si só, já fornece os termos mais utilizados pelos usuários da *Internet* para um determinado domínio de busca (TUSAR, 2009).

O serviço de buscas da Microsoft® foi selecionado graças ao limite de buscas mensais (5000) que podem ser realizadas de forma gratuita. Os *scripts* deste módulo são:

a) o arquivo dic.txt: o arquivo dic.txt usa o padrão de caracteres UTF-8 e contém um argumento de busca por linha, separados pelos códigos ASCII de quebra de linha e retorno de carro (CR+LF, 0x0D 0x0A). Cada linha deste arquivo contém um argumento de busca que foi originado do questionário *online*;

b) o *script* combine.r: consiste de um *script* criado na linguagem R (R DEVELOPMENT CORE TEAM, 2010), que lê o arquivo dic.txt e gera todas as combinações possíveis de termos em grupos de 1, 2 e 3 termos criando a primeira versão do vetor de *keywords*. O vetor de *keywords* é utilizado contra o serviço *Google AdWords Keywords Tool*® e o resultado é armazenado na segunda geração do vetor de *keywords*, sem repetição. A geração final do vetor de *keywords*, para o domínio *notebook*, contém 4547 itens com até três termos cada um. Cada um das linhas deste vetor é utilizada como argumento para o *script* busca.php;

c) o *script* busca.php: este *script*, desenvolvido na linguagem PHP (PHP TEAM, 2009), faz uma chamada a biblioteca Bing API PHP©[72] para cada item armazenando no vetor de *keywords*. A opção pelo uso da biblioteca Bing API PHP© e a consequente chamada direta ao serviço de buscas da Microsoft® permitiu a simplificação do processo de *parse*. A API do Microsoft Bing© devolve um arquivo XML com as páginas já ordenadas de acordo com a classificação do próprio serviço. Para cada argumento constante no vetor de busca são recuperadas 100 páginas representando um total de 454.700 URLs representando 327.406 URLs únicas que foram passadas ao *script parse*.js através de uma área comum de armazenamento na camada de persistência;

d) o *script parse*.js: Este *script* considera apenas as URLs de páginas *web*. De forma a evitar problemas com eventuais erros na formação de uma URL, só foram consideradas as URLs que apontavam para recursos HTTP do tipo HTML e HTM reduzindo a base e evitando arquivos mais complexos como o PDF, DOC, PPT ou DOCX. A identificação do tipo de documento apontado pela URL é realizada uma análise da resposta HTTP do servidor hospedeiro observando-se o tipo *Multipurpose Internet Mail Extensions*[73] (MIME) devolvido, o que permite identificar inequivocamente o tipo de arquivo apontado pela URL e mantendo-se apenas os tipos HTML e HTTP. Cada URL recuperada e validada tem seu o seu domínio DNS, extraído e armazenado e relacionado com o item do vetor de keyword que o gerou.

[72] BING API PHP. Disponível: http://bingapiphp.codeplex.com/, acessado em 04 ago. 2012.
[73] Do inglês: extensões multi propósito para correspondência na Internet

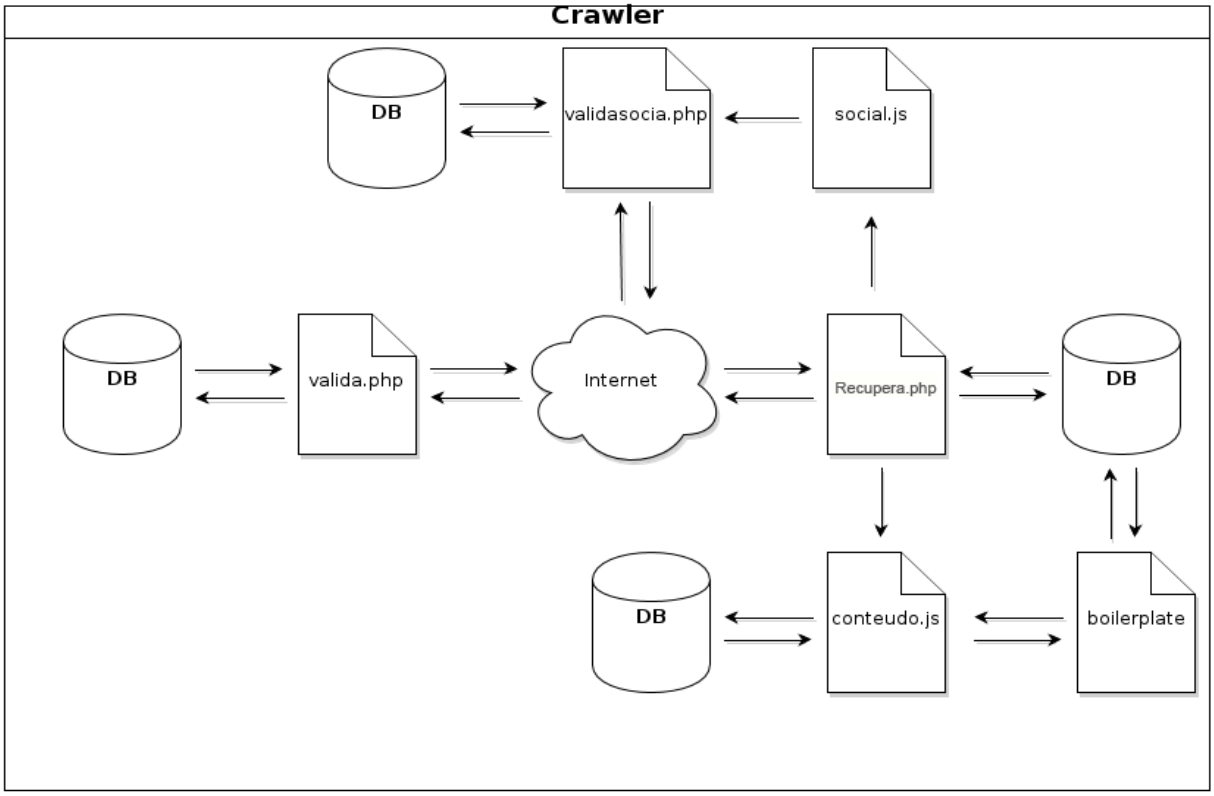
3.3.2 Módulo de *crawler*

Este módulo é responsável pela recuperação de páginas *web*, e pela identificação de usuários de redes sociais que estejam relacionados com o domínio de informação pesquisado, mediante o uso de uma ou mais *keywords* em suas interações sociais.

Este módulo também é responsável pela recuperação de páginas cuja URL tenha sido citada em interações sociais, comparando estas referências com os dados já armazenados na camada de abstração.

A Figura 13 apresenta o diagrama em blocos do módulo de *crawler*.

FIGURA 13 - MÓDULO DE *CRAWLER*



FONTE: o autor (2013)

Os *scripts* deste módulo possuem as seguintes funções:

a) o *script* valida.php: provê as rotinas de validação de URLs. Valida a existência da página mediante a observação das mensagens de erro enviadas pelo servidor hospedeiro e em seguida providencia a resolução DNS. Em caso de erro HTTP, as URLs são verificadas novamente, em períodos aleatórios de

tempo, até três vezes antes de serem excluídas da camada de persistência. A comprovação da existência, e disponibilidade, de uma URL é feita por meio da análise das mensagens de status devolvido pelo servidor hospedeiro em resposta a requisição HTTP correspondente. A resolução DNS é feita com o uso das funções intrínsecas do PHP para este fim.

b) o *script* recupera.php: a função do *script* recupera.php é recolher o conteúdo das páginas *web* e armazenar este conteúdo na camada de persistência fornecendo este mesmo conteúdo para os *script*s de *parse* específico o social.js e o conteúdo.js. O status da recuperação é guardado na camada de persistência, relacionado à URL.

c) Este resultado é positivo quando o *script* recupera.php encontrar uma tag <\body> no texto recuperado. Este *script* começa a recuperação da URL na tag <body>, onde existe maior probabilidade de encontrar conteúdo informacional como foi definido por Pasternack e Roth (2009), excluindo todo o conteúdo que exista entre as tags. <head></head>, <*script*></*script*>,<style></style>, onde a probabilidade de existência de texto e informação é praticamente nula (PASTERNACK; ROTH, 2009).

d) o *script* conteudo.js: o *script* conteudo.js verifica a existência de um esquema de recuperação de dados referente à página recuperada, na camada de abstração e, caso exista, lê este esquema, aplica as regras de validação e armazena o resultado novamente na camada de abstração. O esquema de recuperação é definido em JSON, de forma manual, para sites de grande volume de participação de usuários, como fóruns, para sites de e-commerce, para os sites dos fabricantes e para sites de venda.

   Este esquema de recuperação especifica as classes dos elementos HTML/CSS que contém a informação necessária e a expressão regular necessária para a extração desta informação. O conteúdo recuperado é enviado ao *script* Boilerplate que implementa o algoritmo de mesmo nome para a recuperação do conteúdo informacional;

e) o *script* social.js: este *script* recebe o conteúdo recolhido pelo *script* recupera.php e recupera referências às redes sociais virtuais. Dados relativos

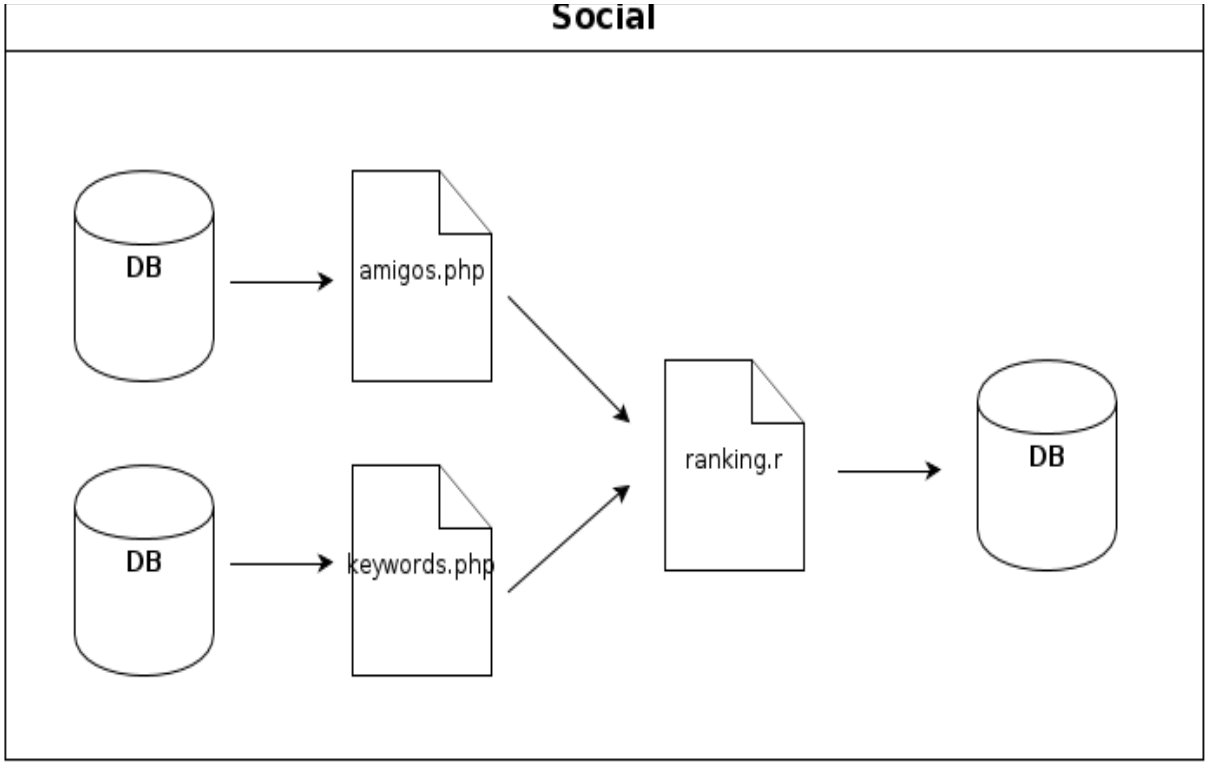
a página no Facebook®, usuário do Twitter® ou perfil no Linkedin® são recuperados, armazenados na camada de persistência e encaminhados para o *script* validasocial.php. Referências sociais posicionadas no rodapé, ou em barras laterais nas páginas são marcadas com o status de secundárias e são penalizadas durante a geração do Pagerank e do gráfico social virtual;

f) O *script* validasocial.php: o *script* validasocial.php é responsável pela validação dos indicadores de redes sociais recuperados nas páginas *web* mediante o uso das APIs de cada rede social virtual e tem o objetivo de, analisando os dados disponíveis em cada indicador, inferir se o indicador aponta para uma pessoa ou para um sistema automático de interação, esta distinção é realizada por meio da detecção de interações, considerando a aleatoriedade ou não destas interações e a existência ou não de interações múltiplas entre o usuário de rede social indicado e outros usuários. Referências sociais identificadas com inativas são excluídas da camada de persistência e não são utilizadas novamente.

3.3.3 Módulo social

Este módulo, apresentado na Figura 14, é responsável pela criação da estrutura do gráfico social virtual e é responsável pela descoberta das relações entre os atores recuperados das diversas redes sociais que tenham interações relacionadas ao domínio de informação pesquisado e, posteriormente pela aplicação do algoritmo Pagerank (PAGE; BRIN *et al.*, 1999) a este gráfico social virtual. A aplicação do Pagerank diretamente ao gráfico social virtual atribui a cada ator uma classificação anterior a aplicação das métricas da análise de rede social.

FIGURA 14 - MÓDULO SOCIAL



FONTE: o autor (2013)

O gráfico social virtual criado por este *script* não obedece aos relacionamentos sociais explícitos em cada serviço de rede social virtual. Em vez disso, este gráfico é criado observando-se apenas as interações e sub interações entre os indivíduos da rede, notadamente sua resposta a um determinado estímulo social. Desta forma, terá mais peso uma repostagem de uma determinada interação por um usuário já identificado no sistema, mas que não faça parte da rede de amigos do postador original, que as relações declaradas pelo postador original.

Este módulo é composto dos seguintes *scripts:*

a) o *script* amigos.php: o *script* amigos.php, de posse das referências sociais inicia um processo de crawler nas redes sociais, usando as APIs disponíveis, para recuperar as interações realizadas nestas redes relacionadas com a informação desejada e armazena estas interações na camada de persistência. Tomando o cuidado de tornar anônimos os usuários pesquisados atribuindo individualmente um código específico do sistema. São recuperadas todas as interações realizadas pelos usuários da rede social, relativos as *keywords* geradas pelo *script* combine.r. O gráfico social virtual criado que relaciona

apenas os usuários de redes sociais virtuais que foram encontrados em páginas *web* que contém as *keywords* do arquivo dic.txt, ou geradas pelo *script* combine.r, ou ainda nas redes sociais particulares destes usuários e que produziram interações sociais contendo estas mesmas *keywords*. Esta rede social virtual é utilizada no processo de classificação da informação. No caso do gráfico social virtual criado os nós podem ser um usuário de rede social ou um documento *web* e as ligações entre estes nós são determinadas levando-se em consideração o número de interações encontradas para cada keyword, a existência ou não de links *web* nestas interações, o comprimento da interação social, a densidade de *keywords* média nas interações recuperadas. O valor de cada nó, ou ligação, é calculado de acordo com as métricas utilizadas na análise de redes sócias. Cada um destes valores é armazenado na camada de persistência para cada keyword e para cada rede social virtual temporária criada;

b) o *script keywords*.php: percorre a estrutura social armazenada pelo *script* amigos.php e procura as interações realizadas que contenham as *keywords* especificadas pelo arquivo dic.txt ou geradas pelo *script* combine.r. As eventuais interações localizadas são armazenadas na camada de persistência para o uso pelo módulo de crawler. Este *script* também descartou interações que não atendessem critérios de confiabilidade conforme apresentado no estudo de Castillo, Mendoza e Poblete (2011) que incluem comprimento, a existência ou não de URL e a citação de outros usuários;

c) o *script* ranking.r: este *script*, ranking.r, é responsável pela aplicação do algoritmo Pagerank no gráfico social virtual produzindo outro fator de classificação. O fator α de adequação do Pagerank é determinado pelos fatores levantados pelo *script* amigos.php; no cálculo da cadeia de Markov, indispensável ao Pagerank, é considerada a relação entre o número de interações retransmitidas por um determinado usuário e o número de interações suas que foram retransmitidas por outros usuários reproduzindo a metodologia por Weng, Lim *et al.* (2010) e criando um índice de classificação para os nós do gráfico social virtual.

3.3.4 Módulo de classificação

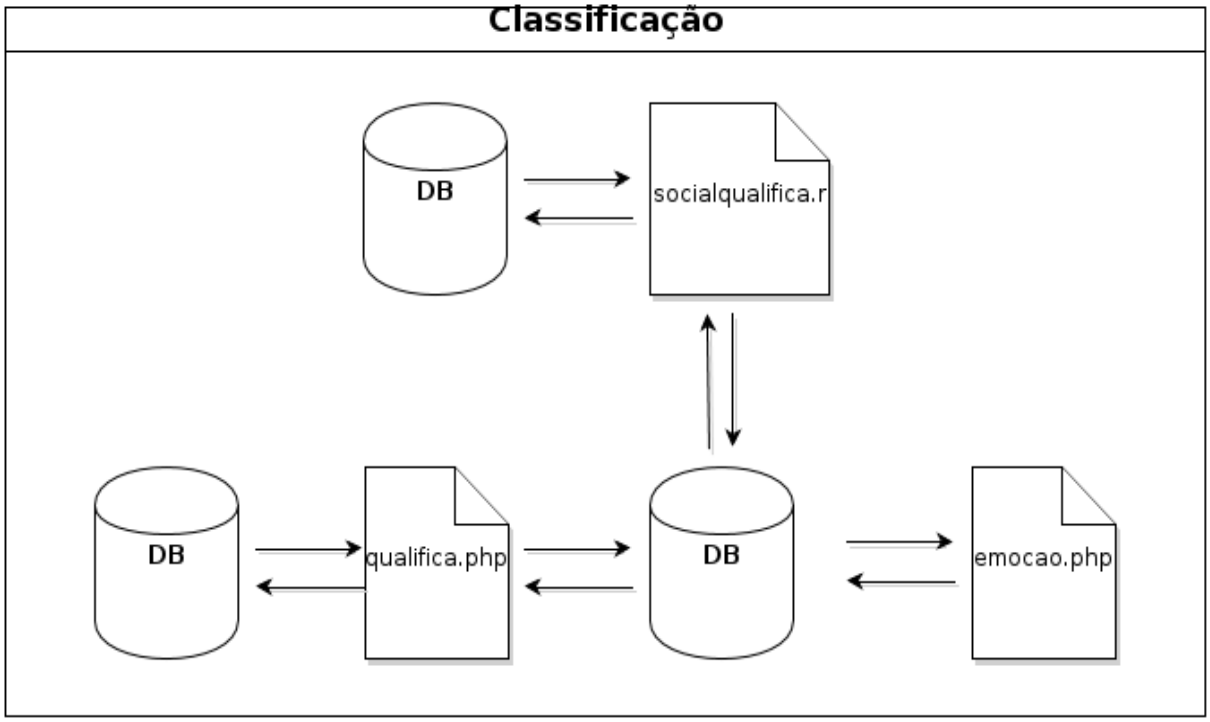
Este módulo é responsável por fazer a análise de opinião do conteúdo recolhido nas páginas *web*, nas redes sociais virtuais e calcular a classificação de cada informação, utilizando os diversos fatores de classificação armazenados na camada de persistência acrescidos do valor positivo ou negativo correspondente a opinião expressa no conteúdo recolhido.

O valor da opinião determina um fator de classificação que pode ser positivo ou negativo e varia entre os valores de -1 a +1 de acordo com uma normalização que é realizada de forma automática levando em consideração que o maior resultado positivo é equivalente ao valor +1 e o menor, ou mais negativo, é equivalente ao valor -1.

Os outros fatores de classificação derivam do posicionamento da página no Microsoft Bing®, do resultado da aplicação do Pagerank e quantidade de interação social, normalizada em um processo semelhante ao utilizado para o conteúdo emocional.

A Figura 15 apresenta o diagrama de blocos deste módulo:

FIGURA 15 - MÓDULO DE CLASSIFICAÇÃO



FONTE: o autor (2013)

Este módulo é composto pelos seguintes *scripts:*

a) o *Script* socialqualifica.r: Este *script*, socialqualifica.r, aplica os índices obtidos pelo algoritmo Pagerank às informações provenientes das redes sociais, ou que tenham sido originadas de uma referência social contida em um link disponível em uma interação social;

b) o *script* qualifica.php: O *script* qualifica.php é responsável pela aplicação dos índices obtidos pelo algoritmo Pagerank às páginas obtidas pelo *web*-crawler, originadas de interações sociais, ou não, e aos nós da rede social virtual, para a realização deste cálculo, os diversos índices armazenados na camada de persistência são transformados em índices de uma média aritmética ponderada. Os pesos aplicados a cada índice são definidos de forma supervisionada e podem variar de acordo com o domínio da informação desejada. O resultado desta média indica o fator de classificação da informação recolhida e, consequentemente da página *web* original ou do usuário de rede social que criou a interação;

c) o *script* emocao.php: O *script* emocao.php é responsável pela determinação do valor opinativo contido em cada fragmento de informação armazenado na camada de persistência. O valor relativo à opinião é calculado mediante a soma entre a densidade de tokens com valor positivo e a densidade dos tokens com valor negativo, encontrados em cada fragmento de informação. O valor opinativo, quando positivo aumenta o peso da URL ou da interação social, quando negativo diminui este peso. A identificação do token, como positivo ou negativo, é feita por meio da comparação deste token, com o léxico padrão SentiLex-PT 01. A densidade por sua vez é calculada considerando a média de caracteres contida em cada fragmento de informação como o comprimento padrão para o fragmento em questão. Desta forma a densidade, d, pode ser calculada pela equação 12.

$$\mathrm{d} = \frac{\text{Número de tokens no fragmento}}{\text{Número de caracteres médio}} \qquad (12)$$

b) O comprimento padrão é determinado por meio do cálculo da média do comprimento de cada conteúdo informacional armazenado em três categorias:

livre, que inclui todas as páginas *web*; limitado, que incluí o Twitter® e variável que inclui interações em sites como o Facebook® e o Linkedin®. O valor opinativo atribuído a cada informação é normalizado considerando o maior e o menor valor encontrado em todo o conjunto de informações recolhidas

3.3.5 Módulo de interface

O módulo de interface é responsável pela interação com o usuário, e a geração da página de resultados. Este módulo organiza a informação qualificada e apresenta o resultado a cada consulta.

Este módulo interage com os outros módulos permitindo que o usuário possa entrar com o dicionário de classificação do domínio de informação, *dic.txt*, as regras de recolhimento de informação para os *sites* específicos e os pesos para os cálculos da classificação. Este módulo consiste de um único *script* em PHP, denominado de interface.php, que contém as classes necessárias para o relacionamento com o usuário e gera as páginas *web* necessárias a este relacionamento.

4 ANÁLISE DE RESULTADOS

Entre os dias 10 de maio de 2012 e 20 de dezembro de 2012, de forma esporádica e aleatória, foram recolhidas 327.406 URLs únicas das quais apenas 12.456 páginas *web* foram relacionadas com algum tipo de interação social e 16.777.200 interações sociais das quais 335.684 foram classificadas como interações sociais válidas totalizando 348.140 registros diferentes na camada de persistência, sujeitos a classificação, como pode ser visto no Quadro 2.

QUADRO 2 - INFORMAÇÕES RECOLHIDAS

Origem	**Recuperadas**	**Validadas**
Páginas *web*	327.406	12.456
Interações sociais	16.777.200	335.684

FONTE: o autor (2013)

As informações contidas nestes registros foram classificadas de acordo com a opinião expressa em três grupos: 55.702 informações positivas e 192.400 informações negativas e 100.038 informações neutras como pode ser visto no Quadro 3.

QUADRO 3 - CLASSIFICAÇÃO SEGUNDO A OPINIÃO EXPRESSA NAS INFORMAÇÕES RECOLHIDAS

Classificação	**Valor Absoluto**	**Valor Percentual**
Positivas	55.702	15,99%
Neutras	100.038	55,26%
Negativas	192.400	28,73%
Total	348.140	100,00%

FONTE: o autor (2013)

As informações recolhidas permitiram também a identificação de 244 gráficos sociais virtuais entre estes foram encontrados 34 com dez ou mais nós.

Os atores destes Gráficos Sociais Virtuais (GSV) foram responsáveis pela criação, ou divulgação de aproximadamente 3,47% (12.080) do total de informações

recolhidas conforme pode ser visto no Quadro 4 onde o total de atores no GSV é representado por tn.

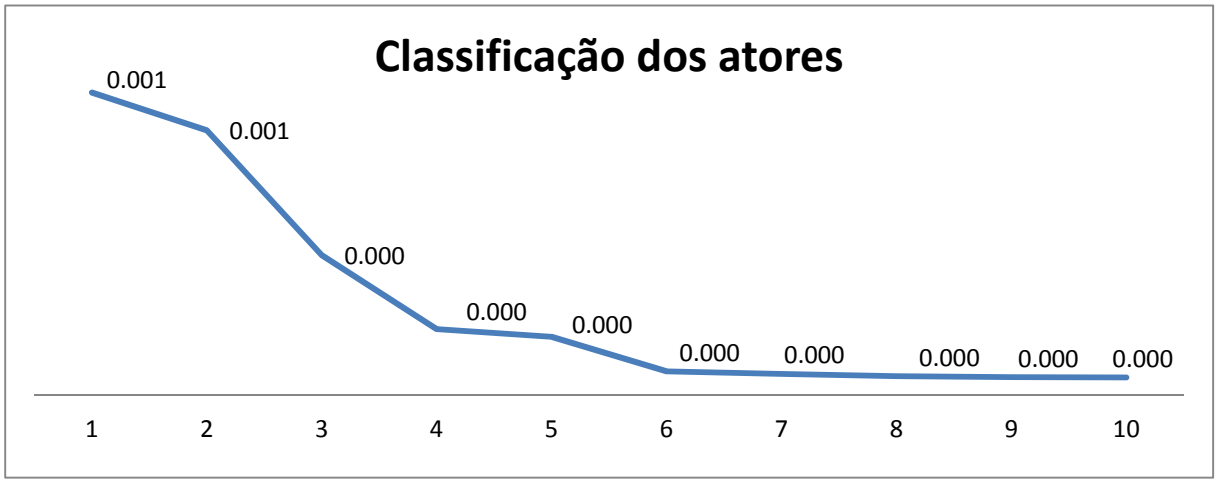
QUADRO 3 - INFORMAÇÕES DEVIDO A GRÁFICOS SOCIAIS VIRTUAIS

		Interações		
Faixas	**GSV**	**Pg.** *web*	**Sociais**	**Total**
$\mathrm{tn} = 2$	815	3268	5298	8566
$2 < tn < 7$	123	685	807	1492
$7 \leq \mathrm{tn} \leq 10$	87	568	984	1552
$\mathrm{tn} > 10$	34	218	252	470
Total	**1059**			**12080**

FONTE: o autor (2013)

As faixas escolhidas para uso nos Quadro 4 e 5 foram selecionadas de forma a destacar os menores e os maiores GSVs em termos de números de nós. As faixas centrais foram escolhidas apenas para dividir os GSVs de acordo com o diâmetro do grafo, de forma equitativa. Os atores destes gráficos sociais virtuais foram classificados de acordo com sua *betweenness*, centralidade e número de relações e o *ranking*[74]. No Gráfico 7 é possível observar o valor, normalizado, definido pelos algoritmos e atribuído a cada um dos dez atores melhor classificados:

GRÁFICO 7 - CLASSIFICAÇÃO NORMALIZADA DOS ATORES



FONTE: o autor (2013)

[74] Ranking: este valor contém o resultado do algoritmo pagerank para cada ator.

A distribuição apresentada no Gráfico 7 parece indicar que a rede social virtual formada pelos gráficos sociais virtuais recolhidos pode ser classificada como rede complexa.

As interações sociais na *web*, sejam elas através de *sites* ou em redes sociais virtuais parecem apresentar um alto nível de ruído. Este ruído pode ser percebido no fato que apesar do número de interações sociais recolhidas (16.777.200), pouco mais de 2% (335.684) puderam ser consideradas válidas. Este percentual é decorrente dos fatores de descarte utilizados para a validação. Notadamente por que os algoritmos utilizados para validação não consideram símbolos de pontuação como capazes de conter informação. Consequentemente, as interações contendo *emoticons*[75] tais como ":)" e ":(" foram ignoradas.

Observando apenas as informações relacionadas de forma direta aos gráficos sociais virtuais descobertos é possível perceber uma discrepância quanto aos valores opinativos quando comparamos com o conjunto de todas as informações recuperadas como pode ser observado no Quadro 5.

QUADRO 4 - COMPARATIVO DA ANÁLISE DE OPINIÃO NO GRÁFICO SOCIAL VIRTUAL

Faixas	**GSV**	**Info.**	**Positivas**	**Neutras**	**Negativas**
$\mathrm{tn} = 2$	815	8566	1798	6451	317
$2 < tn < 7$	123	1492	313	1125	54
$7 \leq \mathrm{tn} \leq 10$	87	1552	324	1171	57
$\mathrm{tn} > 10$	34	470	98	355	17
Total	**1059**	**12080**	**2533**	**9102**	**445**
Apenas GSV			**20,96%**	**75,34%**	**3,68%**
Todo Conjunto			**15,99%**	**55,26%**	**28,73%**

FONTE: o autor (2013)

A remoção do conteúdo informacional das páginas recolhidas apresentou poucas dificuldades. No conjunto de páginas tratadas o *script boilerplate* se mostrou eficiente e aproximadamente 0,2% (25 páginas) das 12.456 páginas únicas recuperadas necessitaram de *scripts* personalizados.

[75] Do inglês: ícones de emoção

No Quadro 6 está uma comparação entre os *sites* classificados nas dez primeiras posições de acordo com sua atividade principal, comparados com os resultados encontrados nas ferramentas de busca do Google® e da Microsoft®. Este quadro apresenta apenas as 10 primeiras páginas de cada serviço de busca, especificamente configurados para apresentar apenas páginas em português como falado no Brasil, sem considerar a localização do usuário que faz a busca. A busca foi realizada no dia 01 de Janeiro de 2013 as 20:30h.

QUADRO 6 - COMPARAÇÃO ENTRE OS SITES RETORNADOS COM OS SERVIÇOS COMERCIAIS

Google	Bing	Pesquisa
Vendas - Extra	Vendas - Buscapé	Fórum - Clube do Hardware
Vendas - Mag. Luiza	Imagens - Bing	Revista - Info Abril
Vendas - Lj. Americanas	Vendas - UOL	Fabricante - HP
Vendas - UOL	Vendas - Mag. Luíza	Blog - Tecmundo
Vendas - Ricardo Eletro	Vendas - Ponto Frio	Vendas - Buscapé
Vendas - Buscapé	Vendas - BondFaro	Vendas – Mag. Luiza
Vendas - Walmart	Vendas - Extra	Blog- Vidaestilo
Vendas - Casas Bahia	Vendas - Lj. Americanas	Blog - Infowester
Vendas - Mercado Livre	Vendas - Zura	Vendas - Extra
Vendas - ShopTime	Vendas - Toda Oferta	Vendas – Lj. Americanas

FONTE: o autor (2013)

No Apêndice B estão três listas contendo as dez primeiras URLs retornadas em sua integralidade, tanto pelo sistema desenvolvido para esta pesquisa quanto pelos dois serviços de busca anteriormente referidos. A diferença entre os resultados apresentados pelos *sites* de busca pode ser creditada as diferenças entre seus algoritmos de classificação.

A diferença nos resultados apresentado pelo sistema criado e os *sites* comerciais parece dever-se aos fatores de classificação baseados nos gráficos sociais virtuais e na mineração de opinião. Da mesma forma, a diferença encontrada na quantidade de informações consideradas positiva, neutras ou negativas, quando se observa o conjunto completo ou o conjunto derivado das interações sociais parece indicar uma tendência a opinião negativa quando o processo de troca de informação se dá de forma social.

5 CONSIDERAÇÕES FINAIS

O conceito central que norteou esta pesquisa foi a busca e classificação de informações na *web* utilizando como ferramenta de classificação as próprias interações sociais disponíveis na *web.* Este conceito suscitou um conjunto de objetivos, previamente descritos, e alcançados como descrito nesta conclusão.

A pesquisa bibliográfica permitiu a caracterização das redes sociais, notadamente destacando as características de complexidade da *web*; e a identificação das métricas necessárias à caracterização de nós em uma rede social, satisfazendo o objetivo específico (a). Métricas como a centralidade, adjacência e centralidade *betweenness* se mostraram interessantes para a avaliação de nós e foram utilizadas para a classificação das informações contidas nas interações sociais e nas páginas *web* recolhidas.

O estudo da arquitetura típica das páginas *web* permitiu geração do conhecimento necessário a criação de *scripts* personalizados para a extração de conteúdo informacional de sites tão diversos quanto o Clube do Hardware (um fórum) e o Dell (um *site* de fabricante), fornecendo o conhecimento necessário a satisfação do objetivo específico (b).

A integração entre pesquisa bibliográfica e metodológica permitiu identificar algumas técnicas de mineração de opinião e escolher a técnica desenvolvida por Silva e Carvalho (2010) para uso no sistema, satisfazendo o objetivo específico (c).

O conceito de classificação, por si só, induz a conceitos de valor e ordem. Esta pesquisa argumentou em favor do uso de um gráfico social virtual, inferido a partir das interações explicitadas, como fator de valoração da informação e, usou como suporte a valoração a opinião expressa pelos próprios autores da informação.

Evitando os processos de busca e recuperação de informações genéricas, durante esta pesquisa foi criado um léxico específico para a caracterização do domínio representado pela palavra *notebook* e este léxico foi utilizado como argumento de busca inicial. Busca esta que resultou em um conjunto de páginas classificadas com base nos gráficos sociais virtuais e nas opiniões expressas pelos autores da informação. Atendendo de forma plena o objetivo geral traçado no início desta pesquisa.

A comparação entre os resultados obtidos para um domínio específico (*notebook*) parece indicar que o ruído existente na informação *online* pode ser contornado por ferramentas de classificação suportadas por interações sociais. Os resultados parecem indicar que os atores sociais trocam informações que vão além das informações ofertadas pelos serviços comerciais de busca na *web*. Esta troca, induzida pela necessidade, prenuncia que os serviços de busca atualmente disponíveis podem estar pecando justamente na qualidade da informação que ofertam aos seus usuários.

A contribuição científica derivada desta pesquisa pode ser encontrada tanto na validação de técnicas de mineração de opinião e classificação de atores em redes sociais quanto na proposta, de uso de um gráfico social virtual como ferramenta de classificação de informações disponíveis na *web*.

A análise dos resultados obtidos sugere várias opções para a continuidade da pesquisa, entre estes se destacam:

a) a adequação dos algoritmos de mineração de opinião aos costumes atuais. É premente a necessidade da inclusão de *emoticons*, imagens e *memes*[76] no processo de avaliação da opinião contida na informação;

b) um estudo de análise semântica dos termos utilizados para a definição de um domínio de informação e, a caracterização deste domínio de acordo com seu significado;

c) a criação de um algoritmo de classificação de informação semântico;

d) a criação de um sistema de recuperação contínuo, em tempo real, que permita a avaliação temporal do valor da informação recolhida.

A análise dos resultados obtidos também parece indicar a necessidade de novas pesquisas na caracterização do domínio para a busca de informações, indicando a necessidade de uma caracterização mais acurada do usuário.

[76] Trata-se de um elemento cultural que pode ser compartilhado na internet, na forma de imagens, animações ou áudio.

REFERÊNCIAS

AGUIARI, V. Facebook supera Orkut no Brasil. **Info**, 2011. Disponivel em: <http://info.abril.com.br/noticias/internet/facebook-supera-orkut-no-brasil-diz-site-27042011-4.shl>. Acesso em: 1 Ago. 2011.

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. The diameter of the world wide web. **Nature**, London, Reino Unido, v. 401, p. 130-131, Set. 1999.

ALEXA. Top Sites in Brazil. **Alexa**, 2011. Disponivel em: <http://www.alexa.com/topsites/countries/BR>. Acesso em: 01 Ago. 2011.

ANDERSON, F. JOHN GUARE. **The theatre database**, 2002. Disponivel em: <http://www.theatredatabase.com/20th_century/john_guare_001.html>. Acesso em: 12 Jan. 2011.

BACKSTROM, L. et al. Four degrees of separation. **The Cornell University Library**, 2012. ISSN arXiv:1111.4570v3. Disponivel em: <http://arxiv.org/abs/1111.4570v3>. Acesso em: 10 Jan. 2012.

BAEZA-YATES, R.; CASTILLO, C. Balancing volume, quality and freshness in web crawling. **Soft computing systems - Design, management and applications**, Santiago, Chile, 22 Ago. 2002. 565--572.

BAEZA-YATES, R.; CASTILLO, C. Crawling the infinite Web: five levels are enough. **Lecture notes in computer science**, Roma, Itália, v. 3243, p. 156--167, Out. 2004.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York, NY, USA: ACM Press, 1999.

BAJULA, S. **Browsing on small screens:** recasting web-page segmentation into an efficient machine learning framework. Proceedings of the 15th international conference on World Wide Web. New York, NY, USA: ACM. 2006. p. 33--42.

BAKHSHANDEH, R. et al. **Degrees of separation in social networks**. Proceedings, The fourth international symposium on combinatorial search. Barcelona, Espanha: AAAI Publications. 2011.

BALAKRISHNAM, V. K. **Theory and problems of graph theory**. Orono, MN, USA: McGraw Hill, v. , 1997.

BARABÁSI, A.-L. **Linked:** The new science of networks. 1ª. ed. Cambridge, Reino Unido: Perseus publishing, v. I, 2002.

BARABÁSI, A.-L. et al. **Scale-free and hierarchical structures in complex networks**. AIP Conference Proceedings. Melville, NY, USA: [s.n.]. 2003. p. 114 -130.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in randon networks. **Science**, New York, NY, USA, v. 286, n. 5439, p. 509-512, Outubro 1999.

BARNES, J. A. Class and committeess in a norwegian island parish. **Human relations**, London, Reino Unido, v. I, n. 1ª, p. 39-58, Mai. 1954.

BARR, A. Google+ attracts 25 million visitors: comScore. **Reuters**, 2011. Disponivel em: <http://www.reuters.com/article/2011/08/02/us-google-idUSTRE7716WL20110802>. Acesso em: 02 Ago. 2011.

BAR-YOSSEF, Z.; RAJAGOPALAN, S. **Template detection via data mining and its applications**. Proceedings of the 11th international conference on World Wide Web. [S.l.]: ACM. 2002. p. 580--591.

BERGAMIM JR., G. Em protesto, mães promovem "mamaço" em São Paulo. **Folha de São Paulo**, 2011. Disponivel em: <http://www1.folha.uol.com.br/cotidiano/915030-em-protesto-maes-promovem-mamaco-em-sao-paulo.shtml>. Acesso em: 20 Jul. 2011.

BERLOW, E. L. Strong effects of weak interactions in ecological communities. **Nature**, London, Reino Unido, v. 398, n. 1ª, p. 330-334, Mar. 1999. ISSN ISSN : 0028-0836.

BERNERS-LEE, T. HTML. **World Wide Web Consortium - History**, 1992. Disponivel em: <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/MarkUp/MarkUp.html#4>. Acesso em: 12 Mai. 2012.

BERNERS-LEE, T. World Wide Web Consortium - History. **SGML**, 1992. Disponivel em: <http://www.w3.org/History/19921103-hypertext/hypertext/WWW/MarkUp/SGML.html>. Acesso em: 12 Mai. 2012.

BERNERS-LEE, T. et al. Hypertext Transfer Protocol -- HTTP/1.1. **Ietf**, 1994. Disponivel em: <http://www.ietf.org/rfc/rfc1738.txt>. Acesso em: 12 Out. 2011.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. RFC 3986. **The Internet Engineering Task Force (IETF)**, 2005. Disponivel em: <http://www.ietf.org/rfc/rfc3986.txt>. Acesso em: 22 Mai. 2012.

BERNERS-LEE, T.; FISCHETTI, M. **Weaving the web - The original design and ultimate destiny of the world wide web by its inventor**. 1ª. ed. San Francisco, CA, USA: Harper, v. I, 2002. ISBN ISBN-10: 006251587X.

BERNERS-LEE, T.; MASINTER, L.; MCCAHILL, M. Uniform Resource Locators (URL). **Internet Ingeneering Task Force**, 1994. Disponivel em: <http://www.ietf.org/rfc/rfc1738.txt>. Acesso em: 10 Jan. 2012.

BLACK, K. **Business statistics - contemporary decision making**. 4ª. ed. Danvers, MA, USA: John Wiley & Sons, Inc., 2010.

BLACKBURN, B. Japan Earthquake and Tsunami: Social Media Spreads News, Raises Relief Funds. **Abc News**, 2011. Disponivel em: <http://abcnews.go.com/Technology/japan-earthquake-tsunami-drive-social-media-dialogue/story?id=13117677>. Acesso em: 20 Jul. 2011.

BOCCALETTI, S. et al. **Complex networks:** Structure and dynamics. 1ª. ed. [S.l.]: Elsevier B.V., v. 1, 2006. 175–308 p.

BOLLEN, J.; MAO, H.; ZENG, X.-J. Twitter mood predicts the stock market. **Journal of computational science**, New York, NY, USA, 21 Set. 2011. 115-121.

BOYD, D. M.; ELLISON, N. B. Social Network Sites: Definition, History, and Scholarship. **Journal of Computer-Mediated Communication**, New Jersey, NY, USA, 13 Out. 2007. 210–230.

BRANDÃO, A. C. P.; SPINILLO, A. G. Aspectos gerais e específicos na compreensã de textos. **Psicología reflexao e crítica**, Porto Alegre, RS, Brasil, v. 11, n. 2, p. 253-272, 1998. ISSN ISSN 0102-7972.

BURKE, R. **Salticus:** Guided crawling for personal digital libraries. 1st ACM/IEEE-CS joint conference on Digital libraries. [S.l.]: [s.n.]. 2001. p. 88--89.

CASTILLO, C.; MENDOZA, M.; POBLETE, B. **Information Credibility on Twitter**. 20th International Conference on World Wide Web. New York, NY, USA: [s.n.]. 2011. p. 675–684.

CERVO, A. L.; BERVIAN, P. A. **Metodologia científica para uso de estudantes universitários**. São Paulo, SP, Brasil: McGraw-Hill, 1983.

CHAKRABARTI, S. **Mining the web:** Discovering knowledge from hypertext data. New York, NY, USA: Morgan Kaufmann Pub, v. 1, 2003.

CHANDRAMOULI, A.; GAUCH, S.; ENO, J. **A popularity-based URL ordering algorithm for crawlers**. 3rd Conference on human system interactions (HSI). [S.l.]: [s.n.]. 2010. p. 556--562.

CHEN, Y.; MA, W.-Y.; ZHANG, H.-J. **Detecting web page structure for adaptive viewing on small form factor devices**. Proceedings of the 12th international conference on World Wide Web. New York, NY, USA: ACM. 2003. p. 225 - 233.

CHO, J.; GARCIA-MOLINA, H. **Parallel Crawlers**. Proceedings of the 11th international conference on World Wide Web. [S.l.]: [s.n.]. 2002. p. Parallel Crawlers.

CHRISTLEY, R. M. et al. Infection in social networks: Using network analysis to identify high-risk individuals. **American Journal of Epidemiology**, Oxford, Reino Unido, v. 162, n. 10, p. 50-58, 21 setembro 2001.

COLAS, F.; BRAZDIL, P. Comparison of SVM and some olderclassification algorithms in text. **Artificial Intelligence in Theory and Practice**, Boston, MA, USA, 21-24 Ago. 2006. 169-178.

CROCKFORD, D. **JSON:** The fat-free alternative to XML. Proceedings of XML. Boston, MA, USA: [s.n.]. 2006.

CSERMELY, P. **Weak links:** Stabilizers of complex systems from proteins to social networks. 1ª. ed. Berlim, Alemanha: Springer-Verlag, v. I, 2006. ISBN ISBN-10 3-540-31151-3.

CUNNINGHAM, W. Correspondence on the Etymology of Wiki. **Cunningham & Cunningham, Inc.**, 2004. Disponivel em: <http://c2.com/doc/etymology.html>. Acesso em: 12 março 2012.

CURBERA, F. et al. Unraveling the web services web an introduction to SOAP, WSDL,and UDDI. **Internet computing, IEEE**, New York, NY, Usa, v. 6, n. 2ª, p. 86--93, 22 Mai. 2002.

DELL'AMORE, C. Prehistoric Dice Boards Found—Oldest Games in Americas? **National Geographic News**, 10 dez. 2010. Disponivel em: <http://news.nationalgeographic.com/news/2010/12/101210-dice-gaming-gambling-native-american-indian-casinos-science/>. Acesso em: 15 Mar. 2011.

DEMO, P. **Introdução à metodologia científica**. São Paulo: Atlas, 1985.

DIESTEL, R. **Graph theory**. New York, NY, USA: Springer-Verlag, 2000.

DRUCKER, P. Drucker on management: The economy's Power Shift. **Wall Street journal**, New York, NY, USA, 24 Set. 1992. 16.

DRUCKER, P. The next information revolution. **Forbes ASAP**, New York, NY, USA, 24 Ago. 1998. 47-56.

EBIZMBA. Top 15 Most Popular Social Networking Sites - August 2011. **ebizmba.com**, 2011. Disponivel em: <http://www.ebizmba.com/articles/social-networking-websites>. Acesso em: 6 Ago. 2011.

ECMA. **ECMAScript Language Specification -ECMA-262**. ECMA International. Genebra, Suíça, p. 258. 2011.

FACEBOOK. Statistics. **Facebook**, 2011. Disponivel em: <http://www.facebook.com/press/info.php?statistics>. Acesso em: 02 Ago. 2011.

FENSEL, D. et al. **Enabling semantic web services:** The web service modeling ontology. 1ª. ed. Berlim, Alemanha: Springer-Verlag, v. I, 2007.

FETTE, I.; MELNIKOV, A. The WebSocket Protocol. **IETF HyBi Working Group**, 2011. Disponivel em: <http://people.csail.mit.edu/emax/papers/www2012-webbox.pdf>. Acesso em: 10 Jun. 2012.

FIELDING, R. et al. Hypertext Transfer Protocol -- HTTP/1.1. **World Wide Web Consortium**, 1999. Disponivel em: <http://www.w3.org/Protocols/rfc2616/rfc2616.html>. Acesso em: 22 Mai. 2012.

FIELDING, R. T. Architectural styles and thedesign of network-based software architectures. **University of California, Irvine**, http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm, p. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm, 1 jan. 2000. ISSN ISBN:0-599-87118-0. Disponivel em: <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>. Acesso em: 12 Ago. 2011.

FRAGOSO, S. D. Eu odeio quem odeia. Considerações sobre o comportamento dos usuários brasileiros na ‘tomada’ do Orkut. **XXIX Congresso Brasileiro de Ciências da Comunicação**, Brasília, 06 Set. 2006. 255-274.

FRAGOSO, S. D. Eu odeio quem odeia. Considerações sobre o comportamento dos usuários brasileiros na ‘tomada’ do Orkut. **XXIX Congresso Brasileiro de Ciências da Comunicação**, Brasília, 6 Setembro 2006. 255-274.

FRISBY, D. **George Simmel (key sociologists)**. 2ª. ed. London, Reino Unido: Taylor & Francis Group, v. 1, 2002. ISBN ISBN: 0-203-52018-1.

FURCHE, T. et al. **Opal:** Automated form understanding for the deep web. World Wide Web conference 2012 - web engineering. Lyon, França: [s.n.]. 2012. p. 10.

FURCHE, T.; GOTTLOB, G.; SCHALLHART, C. Diadem: Domains to databases. In: W.LIDDLE, S.; SCHEWE, K.; XIAOFANG ZHOU **Database and expert systems applications**. Berlim, Alemanha: Springer Berlin Heidelberg, v. 7446, 2012. Cap. 2, p. 1-8.

GAMASUTRA. Most Popular Facebook Games: From FarmVille to King.com's Sagas. **Gamasutra**, 2012. Disponivel em:

<http://gamasutra.com/view/news/180569/Most_Popular_Facebook_Games_From_FarmVille_to_Kingcoms_Sagas.php#.UQkdHGduZ9g>. Acesso em: 30 Jan. 2013.

GARTON, L.; HAYTHORNTHWAITE, C.; WELLMAN, B. Studying online social networks. **Willey OnLine Library**, Indianapolis, 26 Junho 1997. ISSN DOI: 10.1111/j.1083-6101.1997.tb00062.x. Disponivel em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00062.x/abstract>. Acesso em: 30 Jun. 2011.

GJOKA, M. et al. Practical recommendations on crawling online social networks. **IEEE Journal on selected areas in communications**, Los Angeles, CA, USA, v. 29, n. 9ª, p. 1872--1892, Nov. 2011.

GOOGLE. About sitemaps. **Webmasters tools**, 2011. Disponivel em: <http://support.google.com/webmasters/bin/answer.py?hl=en&answer=156184>. Acesso em: 07 jan. 2012.

GOOGLE. Introducing the Google+ project: Real-life sharing, rethought for the web. **The Official Google Blog**, 2011. Disponivel em: <http://googleblog.blogspot.com/2011/06/introducing-google-project-real-life.html>. Acesso em: 12 Jul. 2011.

GRANOVETTER, M. S. The strench of weak ties. **American Journal of Sociology**, Chicago, IL, USA, v. 78, p. 1360-1380, Jun. 1973.

GRAY, M. Credits and Background. **MIT - Old Home Page of Matthew K. Gray**, 1993. Disponivel em: <http://www.mit.edu/people/mkgray/net/background.html>. Acesso em: 13 março 2012.

HEYDON, A.; NAJORK, M. Mercator: A scalable, extensible web crawler. **World Wide Web Magazine**, Los Angeles, CA, USA, v. 2, n. 4, p. 219--229, Ago. 1999.

HORNBY, A. S. **Oxford Advanced Learner's Dictionary**. Oxford, Reino Unido: Oxford University Press, 2010.

HSIEH, J. M.; GRIBBLE, S. D.; LEVY, H. M. **The architecture and implementation of an extensible web crawler**. Proceedings of the 7th USENIX conference on Networked systems design and implementation. San Jose, CA, USA: USENIX Association. 2010. p. 22-36.

INTERNATIONAL TELECOMMUNICATION UNION. **Open systems interconnection Model - X200**. International Telecomunication Unioun - ITU. Genebra, Suíça, p. 63. 1994.

JACKSON, M. H. Assessing the structure of communication on the world wide web. **Journal of Computer-Mediated Communication**, Malden, MA, USA, v. 3, n. 1, p. 25-43, Set. 1997. ISSN ISSN: 1083-6101.

KALLAS, P. Top 10 Social Networking Sites by Market Share of Visits. **DreamGrow Social Média**, 2011. Disponivel em: <http://www.dreamgrow.com/tag/social-networking-market-share-2011/>. Acesso em: 02 Ago. 2011.

KASHTAN, N. et al. Eficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. **Bioinformatics**, Los Angeles, CA, USA, 8 Jan. 2004. 1746-1758.

KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. **Journal of the ACM**, New York, NY, USA , 12 Set. 1999. 604 - 632.

KOHLSCHÜTTER, C.; FANKHAUSER, P.; NEJDL, W. **Boilerplate detection using shallow text features**. Proceedings of the third ACM international conference on web search and data mining. New York, NY, USA: ACM. 2010. p. 441--450.

KOTLER, P.; KELLER, K. L. **Marketing management**. 12ª. ed. Upper Siddle River, NJ, USA: Pearson Pretince Hall, v. 1, 2006. ISBN ISBN: 0-13-145757-8.

KUNDER, M. D. The size of the World Wide Web (The Internet). **WorldWebSize.com**, 2012. Disponivel em: <http://www.worldwidewebsize.com/>. Acesso em: 12 Jan. 2012.

LACERDA, M. A.; MALHEIROS, M. D. G. Automatic extraction of keywords for the portuguese language. **Computational Processing of the Portuguese Language**, New York, NY, USA, 13 Ago. 2006. 204--207.

LEINER, B. M. et al. A Brief History of the Internet. **Cornell University Library**, 1999. Disponivel em: <http://arxiv.org/abs/cs/9901011v1>. Acesso em: 12 Jan. 2011.

LESKOVEC, J.; HORVITZ, E. **Planetary-scale views on an instant-messaging network**. World wide web conference series: Proceedings of the 16th international conference. Beijing, China: [s.n.]. 2008. p. 915-924.

LIU, B. Sentiment analysis and subjectivity. In: LIU, B. **HandBook of natural language processing**. Miami, FL, USA: Chapman and Hall, 2010. p. 627 -666.

LUGANO, G. Egypt – history's first "Facebook revolution"? **Helsink Times**, Helsink, p. 10, 3 Março 2011. Disponivel em: <Egypt – history's first "Facebook revolution"?>. Acesso em: 4 Ago. 2011.

MANNING, C. D.; RAGHAVAN, P.; SCHUTZE, H. **An introduction to information retrieval**. Cambridge,Reino Unido: Cambridge University Press, v. 1, 2009.

MARKERT, K.; HOU, Y.; STRUBE, M. **Collective classification for fine-grained information status**. Proceedings of the 50th annual meeting of the association for computational linguistics. Jeju Island, Korea: [s.n.]. 2012. p. 1-14.

MASSEY, D. S. Presidential Address: A brief history of human society:The origin and role of emotion in social life. **American Sociological Review**, Philadelphia, PA, USA, 01 Fev. 2002. 1-29.

MCCORMICK, T. H.; SALGANIK, M. J.; ZHENG, T. How Many People Do You Know?: Efficiently Estimating Personal Network Size. **Journal of the American Statistical Association**, Washington, 01 Mar. 2010. 59-70.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a Feather: Homophily in Social Networks. **Annual Review of Sociology**, Palo Alto, CA, USA, v. 27, p. 415-444, Ago. 2001.

MERIAN-ERBEN, V. Königsberg. **Preussen**, 1652. Disponivel em: <http://www.preussen-chronik.de/bild_jsp/key=bild_kathe2.html>. Acesso em: 22 Julho 2012.

METZ, J. et al. **Redes complexas: Conceitos e aplicações**. Instituto de Ciências Matemáticas e de Computação. São Carlos, SP, Brasil, p. 45. 2007.

MILGRAM, S. The small-world problem. **Psychology Today**, New York, NY, USA, v. I, n. 1ª, p. 61-67, Mai. 1967.

MISLOVE, A. et al. **Measurement and analysis of online social networks**. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. San Diego, CA, USA: ACM. 2007. p. 29--42.

MOODY, D.; WALSH, P. **Measuring the value of information:** An asset valuation approach. 7th European conference on information systems (ECIS'99). Frederiksberg, Dinamarca: Copenhagen bussiness school. 1999. p. 23-25.

MORENO, Y.; NEKOVEE, M.; PACHECO, A. F. Dynamics of rumor spreading in complex networks. **Physical Review E**, v. 69, n. 6, junho 2006.

NAJORK, M.; WIENER, J. L. **Breadth-first search crawling yields high-quality pages**. Proceedings of the 10th international conference on World Wide Web. Hong Kong, China: [s.n.]. 2001. p. 114--118.

NELSON, T. H. Parallel Documents, Deep Links to Content, Deep Versioning and Deep re-use. **ACM Computing Surveys (CSUR)**, New York, NY, USA, v. 31, n. 4es, 1999.

NEUMANN, E.; PRUSAK, L. Knowledge networks in the semantic web. **Briefings in bioinformatics**, Oxford, Reino Unido, v. 8, n. 3ª, p. 141-149, 04 Mai. 2007.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, Oxford, Reino Unido, v. 45, p. 167--256, Mar. 2003.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The structure and dynamics of networks**. 1ª. ed. New Jersey, NJ, USA: Princeton University Press, v. I, 2006. ISBN ISBN: 978-0-691-11356-2.

NEWMAN, M.; BARABASI, A.-L.; WATTS, D. J. **The Structure and Dynamics of Networks**. 1ª Edição. ed. New Jersey: Princeton University Press, v. I, 2006. ISBN ISBN: 978-0-691-11356-2.

NEWMAN, M.; BARABÁSI, A.-L.; WATTS, D. J. **The structure and dynamics of networks**. Princeton, NJ, USA: Princeton University Press, 2006.

NEWSWIRE. Today's Global Youth Would Give Up Their Sense of Smell to Keep Their Technology. **PR Newswire**, 2011. Disponivel em: <http://www.prnewswire.com/news-releases/todays-global-youth-would-give-up-their-sense-of-smell-to-keep-their-technology-122605643.html>. Acesso em: 01 Ago. 2011.

NICOL, G. et al. **Document object model (DOM) level 3 core specification**. World Wide Web Consortium. Paris, França, p. 146. 2001.

NIELSEN ONLINE. The Social Media Report Q32011. **Nielsen OnLine**, 2011. Disponivel em: <http://cn.nielsen.com/documents/Nielsen-Social-Media-Report_FINAL_090911.pdf>. Acesso em: 23 Fev. 2012.

NURSEITOV, N. et al. **Comparison of JSON and XML Data Interchange Formats:** A Case Study. 2011 Third International Conference on Communications and Mobile Computing. Louisville, KY, USA: [s.n.]. 2009. p. 157-62.

ORENGO, V. M.; SANTOS, D. D. Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. In: SANTOS, D. **Radicalizadores versus analisadores morfológicos:** Sobre a participação do Removedor de Sufixos da Língua Portuguesa nas Morfolimpíadas. Lisboa, Portugual: IST Press, 2007.

PAGE, L. et al. **The PageRank citation ranking: Bringing order to the web**. Stanford InfoLab. Stanford, CA, USA, p. 17. 1999.

PALLA, G. et al. Uncovering the overlapping community structure of complex networks in nature and society. **Nature**, London, Reino Unido, v. 435, n. 7043, p. 814--818, junho 2005.

PALLIS, G.; ZEINALIPOUR-YAZTI, D.; DIKAIAKOS, M. D. Online Social Networks: Status and Trends. **New Directions in Web Data Management**, Berlin, Alemanha, v. 1, n. 331, p. 213–234., Ago. 2011.

PANG, B.; LEE, L. **A sentimental education:** Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the Association for Computational Linguistics. Barcelona, Espanha: [s.n.]. 2004. p. 271–278.

PANG, O.; LEE, L.; VAITHYANATHAN, S. **Thumbs up? Sentiment classification using machine learning techniques**. Proceedings of empirical methods in natural language processing. Philadelphia, PA, USA: [s.n.]. 2002. p. 79--86.

PASTERNACK, J.; ROTH, D. **Extracting article text from the web with maximum subsequence segmentation**. Proceedings of the 18th international conference on world wide web. New York, NY, USA: ACM. 2009. p. 971-980.

PHP TEAM. PHP Language Reference. **PHP**, 2009. Disponivel em: <http://php.net/manual/en/langref.php>. Acesso em: 12 Ago. 2012.

PICARD, A. The history of Twitter, 140 characters at a time. **The Globe and Mail**, 2011. Disponivel em: <http://www.theglobeandmail.com/news/technology/tech-news/the-history-of-twitter-140-characters-at-a-time/article1949299/>. Acesso em: 12 Julho 2011.

POOL, I. D. S.; KOCHEN, M. Contacts and influence. **Social networks**, Lausanne, Holanda, v. 1, n. 1, p. 5-51, 1978.

PORTER, M. The porter stemming algorithm. **The porter stemming algorithm**, 2006. Disponivel em: <http://tartarus.org/~martin/PorterStemmer/>. Acesso em: 22 Ago. 2012.

PORTER, M. E.; MILLAR, V. E. How information gives you competitive advantage. **Harvard Business Review**, Boston, Massachusetts, USA, v. 63, n. 4, p. 149-153, Jul. 1985.

R DEVELOPMENT CORE TEAM. **R:** A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2010. ISBN ISBN 3-900051-07-0.

RAPOPORT, A.; HORRATH, L. J. A study of a large sociogram. **Systems research and behavioral science**, New York, NY, USA, v. 6, n. 4, p. 279-292, Jan. 2007. ISSN DOI: 10.1002/bs.3830060402.

RUSSELL, S.; NORVIG, P. **Inteligência artificial**. Rio de Janeiro, RJ, Brasil: Elsevier, 2004.

SALLET, J. et al. Social network size affects neural circuits in macaques. **Science**, Washington, DC, USA, v. 334, p. 697-700, Nov. 2011.

SAMARA, B. S.; MORSCH, M. A. **Comportamento do Consumidor, coceitos e casos**. São Paulo, SP, Brasil: Prentice Hall, 2005.

SCHMIDT, K. **Góbekli Tepe, southeastern Turkey a preliminary report on the 1995-1999 excavations**. Paléorient. Paris: Persée - Ministère de l'Enseignement supérieur et de la Recherche. 2000. p. 45-54.

SEOMOZ. Google algorithm change history. **SeoMoz**, 2011. Disponivel em: <http://www.seomoz.org/google-algorithm-change>. Acesso em: 12 Ago. 2012.

SHAPIRO, C.; VARIAN, H. R. **Information rules:** a strategic guide to the network economy. 1ª. ed. Boston, MS, USA: Harvard Business School Press, v. 1, 1999.

SHAPIRO, L. **Something from the oven**. 1ª. ed. New York, NY, USA: Penguin Group, v. I, 2004.

SHEPELYANSKY, D. L. CheiRank versus PageRank. **Quantware**, 2011. Disponivel em: <http://www.quantware.ups-tlse.fr/QWART/cheirank/>. Acesso em: 11 Jan. 2012.

SILVA, M. J. et al. **Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis**. Faculdade de Ciências da Universidade de Lisboa. Lisboa, Portugal, p. 14. 2010. (doi: 10455/6694).

SNOWBALL TEAM. A Portuguese stop word list. **Snowball**, 2009. Disponivel em: <http://snowball.tartarus.org/algorithms/portuguese/stop.txt>. Acesso em: 02 Ago. 2012.

SOARES, M. Letramento: um tema em três gêneros. **Presença Pedagógica**, v. 2, n. 10, julho/agosto 1996.

SUE, V. M.; RITTER, L. A. **Conducting Online Surveys**. Thousand Oaks, CA, USA: Sage Publications, 2007.

TAN, C. et al. **User-level sentiment analysis incorporating social networks**. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, CA, USA: [s.n.]. 2011. p. 1397-1405.

TEEVAN, J. et al. **The perfect search engine is not enough:** A study of orienteering behavior in directed search. Proceedings of the SIGCHI conference on Human factors in computing systems. New York, NY, USA: [s.n.]. 2004. p. 415-422.

THOMSEN, J.; BRABRAND, C. WebSelF: A web scraping framework. **Journal of Web Engineering**, Paramus, NJ, Usa, 22 Mar. 2012. 347--361.

TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. **Sociometry**, Boston, MA, USA, 01 Dez. 1968. 425-443.

TRIVIÑOS, A. N. S. **Introdução à pesquisa em ciências sociais:** A pesquisa qualitativa em educação. São Paulo, SP, Brasil: Atlas, 1987.

TURNEY, P. D. **Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews**. Proceedings of the 40th Annual Meeting of the Association forComputational Linguistics (ACL). Philadelphia, MS, USA: [s.n.]. 2002. p. 417-424.

TUSAR, F. K. **Google's keyword analysis tool and compare its various methods for finding the most popular search terms on the web**. Departament of computer science and engineering - Brac University. Dahka, Bangladesh, India, p. 33. 2009.

VAIL, E. Knowledg mapping: getting started with knowledge management. **Information systems management**, NY, New York, USA, 1999. 16 -23.

WAI, Y. L.; LAU, F. C. M. A context-aware decision engine for content adaptation. **Pervasive computing, IEEE**, Los Alamitos, CA, USA, v. 1, n. 3, p. 41 - 49, Set. 2002.

WASSERMAN, S.; FAUST, K. **Social networks analysis:** Methods and applications. 1ª. ed. New York, NY, USA: Cambridge University Press, v. I, 1994. ISBN ISBN 0-521-38269-6.

WATTS, D. J.; STROGATZ, S. H. Colletive dynamics of small-world. **Nature**, London, Reino Unido, v. 393, p. 440–442, Jan. 1998.

WENG, J. et al. **Twitterrank:** finding topic-sensitive influential twitterers. Procedings. of the third ACM international conference on Web. New York, NY, USA: [s.n.]. 2010. p. 261–270.

WIJNIA, E. **Understanding weblogs:** A communicative perspective. Blog talks 2.0: The european conference on weblogs. Krems, Austria: Donau-Universität Krems Kulturwissenschaft. 2004. p. 38-82.

WILLIAMS, R. **Keywords:** a vocabulary of culture and society. New York, NY, USA: Oxford University Press, 1983.

WILSON, P. et al. **A system for subjectivity analysis. Demonstration and Descprition**. Conference on Empirical Methods in Natural Language Processing. Vancouver, Canadá: [s.n.]. 2005. p. 32-36.

WILSON, S. A Brief History of LinkedIn: The Rise of Online Business Networking. **Knol**, 2010. Disponivel em: <http://knol.google.com/k/stephen-wilson/a-brief-history-of-linkedin/qkgetcb5gnql/16#>. Acesso em: 03 Ago. 2011.

WINER, D. XML-RPC Specification. **XML-RPC**, 199. Disponivel em: <http://xmlrpc.scripting.com/spec.html>. Acesso em: 13 maio 2012.

WORLD WIDE WEB CONSORTIUM. SOAP Version 1.2 Part 0: Primer (Second Edition). **World Wide Web Consortium**, 2007. Disponivel em: <http://www.w3.org/TR/2007/REC-soap12-part0-20070427/>. Acesso em: 11 Fev. 2012.

WORLD WIDE WEB CONSORTIUM. Extensible Markup Language (XML) 1.0 (Fifth Edition). **World Wide Web Consortium**, 2008. Disponivel em: <http://www.w3.org/TR/REC-xml/>. Acesso em: 30 Jan. 2013.

WORLD WIDE WEB CONSORTIUM. Cascading Style Sheet. **World Wide Web Consortium**, 2012. Disponivel em: <http://www.w3.org/Style/CSS/>. Acesso em: 22 Abr. 2012.

WORLD WIDE WEB CONSORTIUM -W3C. Web services description language (WSDL) 1.1. **World Wide Web Consortium - W3C**, 2001. Disponivel em: <http://www.w3.org/TR/wsdl>. Acesso em: 03 Mar. 2012.

XINZHOU, X.; QIANG, W.; ANQI, C. Analysis of competition in chinese automobile industry based on an opinion and sentiment mining system. **Journal of intelligence studies in business**, Halmstad, Suécia, 23 Jun. 2012. 41-50.

YE, S.; LANG, J.; WU, F. **Crawling online social graphs**. 12th International Asia-Pacific Web Conference. [S.l.]: [s.n.]. 2010. p. 1-7.

YEE, N. Motivations for play in online games-. **CyberPsychology \& Behavior**, New York, NY, USA, v. 9, n. 6, p. 772--775, Jun. 2006.

APÊNDICE A - QUESTIONÁRIO DE ESPECIFICAÇÃO DO DOMÍNIO

O questionário foi composto de cinco questões diretas de múltipla escolha e uma questão aberta. A questão de número quatro só foi apresentada para os entrevistados que responderam não a questão de número três. A intenção da pergunta número quatro foi definir qual dos fatores é mais importante na decisão de compra, a marca ou o preço à saber:

1. Para comprar um *notebook* qual destas opções é mais importante?

    a) Opinião de amigo;

    b) Propaganda de marca;

    c) Busca na *Internet*;

    d) Conselho de técnico;

    e) Pesquisa de preço.

2. Quando você vai comprar um *notebook*, qual o fator mais importante para sua escolha?

    a) Preço;

    b) CPU;

    c) Memória;

    d) Disco;

    e) Vídeo;

    f) Marca.

3. Considerando dois *notebooks* com o mesmo preço e CPUs diferentes, você trocaria de marca?

    a) Sim;

    b) Não.

4. Se você não troca de marca devido a CPU o que o faria trocar de marca?

    a) Preço;

    b) Memória;

    c) Disco;

    d) Vídeo.

5. Qual o seu grau de escolaridade?

    a) Analfabeto/Menos de um ano de instrução;

    b) Elementar Incompleto;

    c) Fundamental Completo e Ensino Médio Incompleto;

    d) Elementar Completo e Fundamental Incompleto;

    e) Ensino Médio Completo e Superior Incompleto;

    f) Superior Completo ou mais.

6. Em poucas palavras, explique por que você compraria um *notebook*:

APÊNDICE B - TABELAS DE URLS RECUPERADAS

Neste apêndice estão as listas das URLs completas, na forma como devolvidas pelos serviços de busca Google® e Bing® no dia doze de dezembro de 2012 às 22:30h. É importante destacar que estes resultados são variáveis e dependem de condições fora do controle, ou escopo, desta pesquisa.

Google
http://www.extra.com.br/Informatica/*Notebook*/?Filtro=C56_C57
http://www.magazineluiza.com.br/*notebook*/informatica/s/in/note/
http://www.americanas.com.br/linha/267868/informatica/*notebooks*-netbooks-e-ultrabooks
http://shopping.uol.com.br/*notebook*.html#rmcl
http://www.ricardoeletro.com.br/Loja/Informatica/*Notebook*/49-82
http://www.buscape.com.br/*notebook*.html
http://www.walmart.com.br/categoria/informatica/*notebooks*/?fq=C:247/254/
http://www.casasbahia.com.br/dep/Informatica?Filtro=C56
http://*notebooks*.mercadolivre.com.br/*notebooks*-laptops/*notebook*
http://www.shoptime.com.br/sublinha/320381/informatica/laptops/*notebooks*

É possível observar à predominância absoluta de *sites* voltados a venda de produtos. Sejam eles de venda direta ao consumidor ou para a pesquisa de preços.

Bing
http://www.buscape.com.br/*notebook*.html
http://br.bing.com/images/search?q=*notebook*&qpvt=*notebook*&FORM=IGRE
http://shopping.uol.com.br/*notebook*.html#rmcl
http://www.magazineluiza.com.br/*notebook*/informatica/s/in/note/
http://www.pontofrio.com.br/Informatica/*Notebook*/?Filtro=C56_C57
http://www.bondfaro.com.br/*notebook*.html
http://www.extra.com.br/Informatica/*Notebook*/?Filtro=C56_C57
http://www.americanas.com.br/linha/267868/informatica/*notebooks*-netbooks-e-ultrabooks
http://informatica.zura.com.br/preco/*notebook*.html
http://todaoferta.uol.com.br/*notebook*#rmcl

No resultado apresentado pelo Bing® também existe uma predominância de sites voltados a venda de produtos.

Em nenhum dos dois serviços de buscas *online* utilizados como referência para esta pesquisa foi possível encontrar nenhum *blog*, fórum ou site de fabricante entre

os 40 primeiros *links* retornados. O primeiro blog, http://www.tecmundo.com.br, apareceu no Bing® na posição 63.

Pesquisa
http://www.clubedohardware.com.br/duvidas/46
http://info.abril.com.br/reviews/*notebooks*/
http://www8.hp.com/br/pt/home.html
http://www.tecmundo.com.br/*notebook*
http://www.buscape.com.br/*notebook*.html
http://www.magazineluiza.com.br/*notebook*/informatica/s/in/note/
http://vidaeestilo.terra.com.br/homem/como-limpar-a-tela-do-seu-*notebook*,d008518b2b137310VgnCLD100000bbcceb0aRCRD.html
http://www.infowester.com/dicascnotebook.php
http://www.extra.com.br/Informatica/*Notebook*/?Filtro=C56_C57
http://www.americanas.com.br/linha/267868/informatica/*notebooks*-netbooks-e-ultrabooks

Nota-se o aparecimento de *blogs,* fóruns e revistas entre os dez primeiros resultados, além de um *site* de fabricante e vários *sites* dedicados à venda de equipamentos eletrônicos.