1
Introdução

1.1.
Motivação

Informações podem ser armazenadas das mais variadas maneiras. O modo mais clássico de armazenamento de informação é através da palavra escrita, impressa. O acesso à informação estocada desta forma é lento, difícil, e de pouco rendimento. E, nestes casos, para todas as etapas da manipulação da informação é necessária a presença do ser humano, que com suas limitações na capacidade de aquisição de conhecimento e processamento de grande volume de informação, constituiu o principal gargalo do processo (MANDEL, SIMON, & DELYRA, 1997).

A modernização dos últimos anos tornou as tecnologias de informação uma realidade inerente às vidas de todos nós. Das grandes multinacionais às pequenas empresas, das instituições públicas ao ensino e na nossa própria casa, termos como informática, computador, Internet e multimídia, entre tantos outros, passaram a fazer parte das tarefas do dia-a-dia, transformando-se em instrumentos fundamentais do trabalho. Todo este avanço tecnológico proporcionou meios muito mais eficientes para o armazenamento e disseminação de informação, esta, no formato digital, uma condição necessária para o amplo uso dos computadores no seu processamento.

Hoje, a *World Wide Web*, ou simplesmente *Web*, já é o maior repositório de dados do mundo (CHAKRABARTI, 2003). Principal serviço oferecido pela Internet, a *Web*, passa por uma verdadeira explosão de conteúdo, pois deixou de ser palco apenas das grandes empresas e passou a ser utilizada por indivíduos comuns. Segundo dados do **Internic**[1], órgão do Departamento de Comércio do Governo dos Estados Unidos responsável pelo registro de domínios pessoais e

[1] Acrônimo de "Internet's Network Information Center". Disponível em http://www.internic.net

comerciais deste país, há, **atualmente**[2], mais de 500 milhões de *sites* ativos, somente na América do Norte. A Figura 1 ilustra o crescimento do número de **domínios**[3] registrados nos primeiros anos da Internet e nos últimos oito anos.

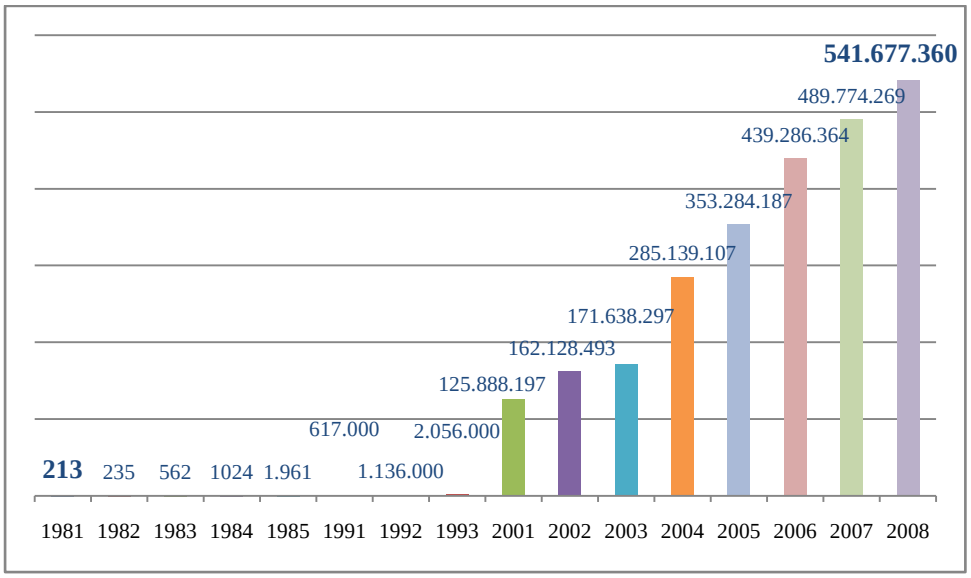

Figura 1 – Número de servidores de internet registrados por ano

Porém, o massivo crescimento de dados disponíveis na *Web* traz consigo grandes desafios em termos de disponibilização de informação. Embora usados muitas vezes como sinônimos, os termos dados e informação possuem significados distintos:

- Dados: um fenômeno qualquer, desprovido de significado e contexto (LAUDON & LAUDON, 2002).
- Informação: resultado do processamento, manipulação e organização de dados que passam a ter significados e, portanto, podem ser contextualizados, interpretados e compreendidos (GOLDSCHMIDT & PASSOS, 2005).

Os dados estão por toda parte. A maioria das organizações não sofre falta de dados, mas, sim, de uma abundância de dados redundantes e inconsistentes (SINGH, 2001). A informação desejada encontra-se entre os bits e bytes armazenados, por exemplo, em um disco rígido. Esses dados, após uma série de

[2] Estatísticas referentes à JUN/2008.

[3] Nome utilizado para localizar e identificar um computador na Internet.

processamentos que envolvem, por exemplo, operações lógicas, serão transformados em informação. E, finalmente, quando essa informação é recuperada e interpretada, chega-se ao conhecimento. É o que está resumido no diagrama da Figura 2.

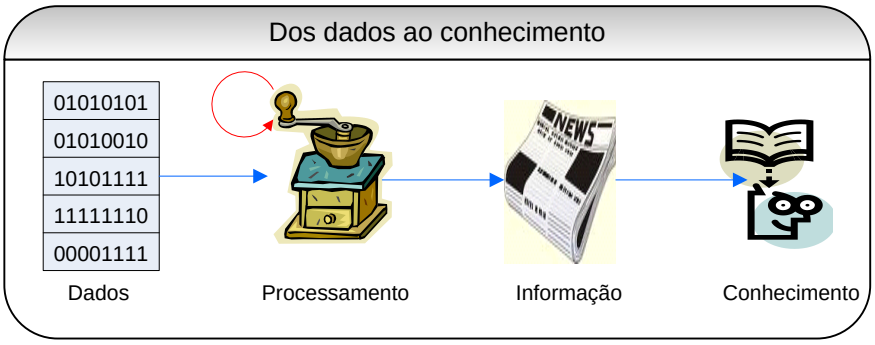

Figura 2 - Processo de obtenção de conhecimento

A *Web* é um ambiente em que reina a cultura liberal e informal de propagação de conteúdo. Toda esta atmosfera desprovida de regras e censura favoreceu a sua expansão vertiginosa de volume e diversidade de conteúdo, porém, também é responsável pela grande redundância e inconsistência de dados presentes neste ambiente.

Atualmente, a volumosa quantidade de dados na *Web*, muitas vezes, torna difícil encontrar a informação desejada. Uma típica busca por informação na *Web* começa em uma **máquina de busca**[4]**.** Para isto, é necessário selecionar algumas palavras relacionadas ao assunto desejado e submetê-las ao sistema de busca que retornará como resultado todos os documentos que contêm os termos desejados. Embora, muitas vezes o objetivo de encontrar a informação requisitada seja atendido, esta abordagem deixa muito a desejar:

- Consultas com termos amplos retornam **milhões de resultados**[5]**.**
- Dificuldades em determinar o prestígio ou a confiabilidade da informação de um documento.

[4] Do termo em inglês, *search engines*.

[5] Consulta realizada pelo termo "casa", no Google, retornou 533 milhões de documentos (01/07/2008).

- Existência de termos, muitas vezes, repletos de ambigüidade, como por exemplo, a procura pelo presidente Lula ou pelo molusco Lula. Em (GOMES, 2008), o assunto de ambigüidade é trado sobre o enfoque de Mineração de Textos.

Outro grande problema recorrente desta abordagem é a dificuldade encontrada pelos usuários em expressar a necessidade de informação por meio de palavras chaves. Conforme (SPINK, WOLFRAM, JANSEN, & SARACEVIC, 2001), aproximadamente cinqüenta de dois por cento das intenções por busca de informação em máquinas de busca são reformuladas. Em média, o número médio de consultas por necessidade de informação é de 4,86. Deste total de consultas reformuladas, segundo o estudo de (SPINK, WOLFRAM, JANSEN, & SARACEVIC, 2001), aproximadamente um terço das consultas modificadas sofreram modificações nos termos submetidos, mas permaneceram inalteradas quanto ao número total de termos. Quanto ao restante, mais de quarenta por cento incluíram termos diferentes dos submetidos na primeira consulta e apenas vinte e cinco por cento excluíram termos utilizados na primeira consulta.

Sendo a maior fonte de dados do mundo, todo usuário com necessidade de informação recorre a *Web*, por saber que a resposta necessária está lá. A grande dificuldade, porém, é saber onde e como obter.

Métodos de Recuperação de Informação sempre foram utilizados para o armazenamento de documentos e a recuperação automática de informação associada a eles (BAEZA-YATES & BERTIER, 1999) (MANNING, RAGHAVAN, & SCHÜTZE, 2007). Antes desta explosão de informação, tarefas de recuperação de informação eram restritas a bibliotecas, nas quais com a ajuda de um bibliotecário, qualquer assunto poderia ser encontrado. Entretanto, com a enorme massa de dados disponíveis, atualmente, a relevância das informações retornadas nem sempre atende às necessidades. Aliado a isto, surge o fato da possível existência de conhecimento, até então, desconhecido, presente nestes dados.

Em virtude do crescimento contínuo do volume de dados eletrônicos disponíveis, técnicas de extração de conhecimento automáticas tornam-se cada vez mais necessárias para manipular essa gigantesca massa de dados. **Mineração**

de Textos[6] ou **Descoberta de Conhecimento em Textos**[7] surge, neste contexto, como uma abordagem a obtenção de informação de qualidade a partir de bases de dados em formato textual. O principal objetivo das técnicas de Mineração de Textos é a manipulação de documentos em formato textual com o objetivo da obtenção do conhecimento implícito presente nestes (ARANHA & PASSOS, 2006).

Assim, dentre as motivações para o desenvolvimento deste trabalho duas necessidades estão presentes: a de prover um mecanismo para a recuperação de informação na *Web* diferente da abordagem que utiliza palavras-chaves e um conjunto de técnicas que seja capaz de extrair, analisar e interpretar todo o conhecimento implícito desta imensa base de dados disponível pela Internet.

1.2. Objetivos do Trabalho

O principal objetivo deste trabalho é pesquisar, propor, implementar e avaliar um método inteligente de recuperação de informação na *Web.* Para isto, é proposta uma metodologia de coleta de dados na *Web,* que com auxílio de técnicas de Mineração de Textos, permite que a necessidade de informação do usuário seja expressa, não por palavras, mas, sim, por um documento de exemplo. Além disso, este processo de coleta inteligente é capaz de analisar as opções de rastreamento disponíveis (*hyperlinks*) e selecionar aquelas que, possivelmente, venham a apresentar conteúdo de maior relevância, ignorando as opções de pouca ou nenhuma relevância.

1.3. Descrição do Trabalho

Coletar dados na *Web* não é uma tarefa trivial. Inteligente ou não, todo processo de coleta de dados na *Web* consiste na utilização de *web crawlers* (HEATON, 2002). *Web crawlers* são *softwares* que, uma vez alimentados por um

[6] Do termo em inglês, *Text Mining*.

[7] Do termo em inglês, *Knowledge Discovery in Texts – KDT*.

conjunto inicial de *URLs* (sementes), iniciam o procedimento metódico de visitar um *site*, armazená-lo em disco e extrair deste os *hyperlinks* que serão utilizados para as próximas visitas. Muitas são as técnicas envolvidas na construção de um *crawler*. Estudos sobre técnicas de recuperação de informação (capítulo "4"), tanto na área de Recuperação de Informação Clássica, como na de Recuperação de Informação na Internet, permitiram a obtenção do embasamento teórico necessário para o desenvolvimento do *crawler* utilizado neste trabalho.

Um processo de coleta inteligente de dados na *Web* é específico: mais do que coletar e armazenar qualquer documento *web* acessível, analisa as opções de *crawling* disponíveis para encontrar *links* que, provavelmente, fornecerão conteúdo de alta relevância a um tópico ou objetivo definido *a priori*. Na abordagem proposta neste trabalho, tópicos são definidos, não por palavras chaves, mas, pelo uso de documentos textuais como exemplos.

Para tanto, faz-se necessário analisar e extrair informação em documentos textuais, principal objetivo de Mineração de Textos. O estudo sobre esta área (capítulos "1" e "2") envolveu, tanto a abordagem sintática, baseada somente em dados estatísticos, como a abordagem semântica, em que a linguagem natural assume papel central no processo de Mineração. Dentre as técnicas baseadas na abordagem semântica, destacam-se o Processamento de Linguagem Natural (seção "3.2.3") e a construção e uso de um dicionário *thesaurus* (seção "3.3.2"). A abordagem ao processo de Mineração de Textos (capítulo "2") é baseada na metodologia proposta por (ARANHA C. N., 2007) que segmenta o processo de Mineração de Textos em cinco etapas, encadeadas nesta ordem: coleta, pré-processamento, indexação, mineração e análise.

Na abordagem de coleta de dados proposta neste estudo, uma vez apresentado e analisado o documento exemplo, o *web crawler* será guiado em busca do seu objetivo: recuperar informação relevante sobre o documento. Realizando uma consulta nas máquinas de buscas disponíveis pelos termos mais representativos do documento exemplo ou baseado em sementes fornecidas pelo usuário, o processo de rastreamento é iniciado. A todo documento recuperado da *Web* é atribuído um grau de relevância em relação ao documento exemplo. Este grau de relevância juntamente com a análise individual de cada *hyperlink* do documento irão determinar o quão importante é um endereço de *URL*. Endereços

URLs importantes serão visitados primeiro, o que eleva a possibilidade de encontrar informação relevante no início do processo.

Embora inteligente, um processo de coleta de dados na *Web* pode encontrar muitos documentos relevantes. Nestes casos, ao final do processo de coleta de dados, objetivando selecionar os documentos mais representativos de um conjunto de documentos, outra técnica de Mineração de Textos é aplicada: a Clusterização de Documentos.

1.4. Organização da Dissertação

Capítulo "2" − A área de Mineração de Textos é abordada com a apresentação das definições utilizadas, atualmente; dos principais elementos de um processo de Mineração de Textos; da estrutura de um documento em linguagem natural; das características representativas de um documento; das abordagens ao processo de Mineração de Textos; e das áreas correlatas.

Capítulo "3" − São descritas as etapas da metodologia estudada, composta por: coleta, pré-processamento, indexação, mineração e análise de resultados. As principais métricas utilizadas na análise de resultados também são comentadas.

Capítulo "4" − São apresentados os fundamentos da área de Recuperação de Informação utilizados em Mineração de Textos, como por exemplo, os modelos de representação de documentos e as principais operações envolvidas nestes processos. A área de Recuperação de Informação na Internet também é abordada, com foco exclusivo para a *Web.*

Capítulo "5" − As técnicas e heurísticas envolvidas no processo de coleta específica de dados na *Web* baseado em *crawler* focado são descritas e comentadas.

Capítulo "6" − São apresentados a metodologia proposta e o desenvolvimento de um sistema para a coleta de dados inteligente na *Web*, seguido de aplicação prática em um estudo de caso.

Capítulo "7" − Finalmente, são apresentadas as conclusões deste trabalho, bem como, trabalhos futuros a serem realizados.