

MINERAÇÃO DE OPINIÕES DE USUÁRIOS NA BUSCA DE CONHECIMENTO

Rafael Guedes*

Derkian Afonso**

Lúcia Helena de Magalhães***

RESUMO

Este trabalho visa utilizar as técnicas de *Web Mining* para mineração de opiniões de usuários da Web, com o objetivo de descobrir informações relevantes com base em opiniões postadas em fóruns de discussão. O objetivo é extrair características que permitam analisar os objetos em estudo de acordo com as opiniões manifestadas. Para validação dos resultados, utilizou-se o software Weka para processamento e contagem das palavras-chave. Desse modo, com base nos histogramas dos elementos em análise, o trabalho procurou identificar e comparar características dos produtos em estudo.

PALAVRAS-CHAVE: RECUPERAÇÃO DA INFORMAÇÃO. WEB MINING. MINERAÇÃO DE OPINIÃO. CONHECIMENTO.

* Graduado pelo curso de Sistemas de Internet das Faculdades Integradas Vianna Júnior

** Graduado pelo curso de Sistemas de Internet das Faculdades Integradas Vianna Júnior

*** Professora das Faculdades Integradas Vianna Júnior

INTRODUÇÃO

Com o aumento de popularidade da WEB, um grande volume de dados e informação foi gerado e publicado em inúmeras páginas web, nas quais se encontram dados das mais diversas áreas do conhecimento. Isto permitiu o surgimento de um subconjunto altamente dinâmico da World Wide Web, tais como blogs, fóruns de discussão, etc. Desta forma, usuários tendem a expressar opiniões e postam seus comentários sobre as pessoas, organizações ou produtos em páginas da internet. Por exemplo, um usuário que está planejando comprar um laptop de uma determinada marca pode querer sondar a opinião dos outros usuários na internet sobre as características da marca que se pretende adquirir.

Desta forma, este volume de informações disponível na Internet e as taxas diárias de crescimento tornam cada vez mais necessários mecanismos eficientes e eficazes para recuperação e mineração de conhecimentos úteis da web, pois a busca de informação relevante e necessária em tempo útil torna-se cada vez mais uma necessidade crítica.

Como "O que os outros pensam" sempre foi uma importante fonte de informação para a maioria dos tomadores de decisão, esta pode ter como suporte o arcabouço de informações que podem ser filtradas e analisadas no ambiente Web. São muitas as pessoas que procuram amigos para, por exemplo, recomendar o serviço de um pedreiro, um modelo de automóvel, ou, até mesmo, um candidato às eleições locais. Quando se trata de emprego, há aqueles que solicitam cartas de referência ou de recomendação. A internet, atualmente, tem sido um canal de referência, onde é possível encontrar relatos de opiniões e experiências, considerando os mais diversos assuntos, de forma que, num processo cada vez mais ascendente, as pessoas manifestam suas opiniões através da Web.

Edição Especial

A presente pesquisa visa descrever o processo de Recuperação da Informação, método utilizado para encontrar as opiniões necessárias para esta pesquisa, além de apresentar os conceitos de Web Mining, descrever suas subáreas, que são *Web Structure Mining*, *Web Usage Mining* e *Web Content Mining*, com destaque especial em *Web Mining Opinion*, que é o objetivo principal desta pesquisa.

Além disso, este trabalho propõe uma abordagem estatística baseada em frequência de palavras. Apesar da sua aparente simplicidade, fornece resultados precisos sem tanto esforço, se comparado com o trabalho de ler documentos imensos e dispersos na web sobre uma pesquisa de interesse.

Estima-se que este estudo possa ser considerado relevante na área de mineração de dados na Web, na medida em que os trabalhos relacionados à extração de elementos que qualificam opiniões de usuários na Web ainda são restritos.

1 MINERAÇÃO DE OPINIÃO NA WEB

1.1 Recuperação da Informação

Formulada de amplas e profundas informações, a internet nos dias atuais é uma fonte de conhecimento. Porém, encontrar a informação certa é uma tarefa de pesquisa desafiadora.

Para tal, é crucial a utilização da Recuperação da Informação, que lida com a representação, armazenamento, organização e o acesso a elementos de informação.

Edição Especial

Recuperação de Informação é, portanto, a área da computação que trabalha com o armazenamento de documentos e a recuperação automática de informações que a eles são associadas. Tal como uma ciência, mas no campo da pesquisa, cobre a busca por informações em documentos, e na busca pelos próprios documentos (YATES e NETO, 1999).

A representação e organização dos elementos de informação devem fornecer um acesso fácil à informação, na qual se está a procura.

Sua extensão cobre também a busca por metadados¹ que descrevam documentos e a busca em banco de dados, quer sejam relacionais e isolados ou banco de dados interligados em uma rede de hipermídia, tais como a World Wide Web. Tal informação pode estar disponível sob forma de textos, de sons, de imagens ou de dados. Todavia, há muita confusão entre os termos e conceitos "recuperação de dados", "recuperação de documentos", "recuperação de informações" e "recuperação de textos".

No presente artigo, o contexto de um sistema de recuperação de informação foi baseado, principalmente, em determinar quais os documentos de informações de uma determinada coleção dos mesmos contêm as palavras-chave que serão utilizadas para a mineração e futura análise.

Para um aproveitamento eficaz dos resultados, foca-se nos mecanismos de busca e diretórios de pesquisa, por serem recursos essenciais na recuperação da informação na Web, tal como, informação textual e online. É possível também para a análise, o uso de diferentes sites especializados sobre mecanismos de busca como, por exemplo, o Google e diretórios de pesquisa disponíveis na Web.

Segundo Yamaoka (2003), um diretório de pesquisa na Web é uma árvore de assuntos ou um diretório de assuntos, isto é, uma lista de assuntos distribuída em categorias, tais como, diretórios acadêmicos ou comerciais, por exemplo.

¹ Metadados são “dados sobre dados”. Metadados referem-se a estrutura descritiva da informação sobre outro dado, o qual é usado para ajudar na identificação, descrição, localização e gerenciamento de recursos da web. (MAGALHÃES, 2008, p.7)

Edição Especial

Para toda e qualquer pesquisa, o êxito de encontrar quais documentos são valiosos e significantes dependerá em sua boa parte da linguagem. É essencial a escolha de palavras e termos que designem a menção ou relaciona-se aos documentos.

Caso o tema da pesquisa seja próprio de línguas estrangeiras, utilizar de sites próprios ao linguajar de origem como mecanismos de pesquisa e diretórios de assuntos, permitem uma maior chance de sucesso.

São inúmeras as formas de se conseguir uma pesquisa eficaz: utilizar de opções tais como, por exemplo, uma pesquisa somente em um determinado idioma, frase exata entre aspas duplas ou frase exata com formulário padrão de escolha.

Segundo Yates (1999), os modelos clássicos de recuperação da informação consideraram que cada documento é descrito por um conjunto de palavras-chave representantes, chamados “termos de índice”. Um termo índice pode ser simplesmente uma palavra, que sua semântica nos ajudará a encontrar os principais documentos com o tema. Assim, termos de índice são usados para indexar e resumir o conteúdo do documento. Em geral, os índices são, em boa parte, desempenhados por substantivos que possuem significado próprio, e também, porque sua semântica é mais fácil de identificar e compreender. Adjetivos, advérbios e conectivos são menos úteis como termos de índice, principalmente porque eles trabalham como um complemento, entretanto eles serão essencialmente úteis na tarefa de mineração.

Dependendo da pesquisa, pode ser interessante considerar todas as palavras distintas em uma coleção de documentos como termos de índice.

Por exemplo, em um dado conjunto de termos de indexação para um documento, percebe-se que nem todos os termos são igualmente úteis e eficientes para descrever o conteúdo do documento. Na verdade, existem termos que são simplesmente e termos que são contraditórios a pesquisa. Decidir sobre a importância de um termo para resumir o conteúdo de um documento não é uma questão trivial.

Edição Especial

De toda forma, vale salientar que nenhum diretório de busca tem todas as informações existentes na Internet categorizadas e, periodicamente e gradativamente surgem novas formas de compilação e indexação das informações em áreas especializadas do conhecimento que são direcionados a um determinado público-alvo.

A recuperação da informação, aplicada a mecanismos de busca, exige-se conhecer os diferentes aspectos da Internet como os diretórios e os termos utilizados e é essencial para que se possa identificar quais os recursos a serem utilizados na recuperação de uma determinada informação.

Utilizar-se de mecanismos e diretórios de busca permitem ao usuário uma forma de trabalho mais rápida e eficaz. Sabendo que ao formular sua pesquisa, pode-se recuperar diferentes tipos de informação.

1.2 Web Mining

As técnicas de mineração de dados na web surgiram das técnicas de Data Mining. A mineração de dados teve início nos anos 80 quando profissionais da área começaram a se preocupar com o crescente armazenamento dos dados inutilizados nos computadores das empresas. Mineração de dados é um conceito que se assemelha ao termo inglês KDD (Knowledge discovery in Database), ou seja, descoberta de conhecimento nos bancos de dados. (AMO, 2010).

O procedimento consiste em algumas etapas: (AMO, 2010)

- a) Limpeza dos dados: Nessa etapa eliminam-se os ruídos e os dados inconsistentes.
- b) Integração dos dados: Etapa onde diferentes bancos de dados podem se combinar produzindo uma única base de dados.

Edição Especial

- c) Transformação dos dados: Nesse ponto os dados podem ser transformados para um formato mais apropriado à aplicação dos algoritmos de mineração.
- d) Mineração: Consiste na aplicação de técnicas inteligentes a fim de definir a extração dos padrões de interesse.
- e) Avaliação ou pós processamento: Identificam-se os padrões interessantes de acordo com os critérios do usuário.
- f) Visualização dos resultados: Consiste em apresentar ao usuário através de técnicas de representação os dados minerados.

Com o crescente uso da web para diversas finalidades, cresceu também a quantidade de informações armazenadas nos servidores e a frequência que esses dados são pesquisados remotamente cresceu proporcionalmente ao número de usuários da rede. Esses dados estão armazenados em diversas partes do mundo. Portanto, para que essas constantes pesquisas sejam possíveis, é extremamente necessário que se faça a mineração de dados nas paginas da web.

A Mineração na Web é uma área de pesquisa que visa integrar as tecnologias Web e a Mineração de Dados, focalizando o desenvolvimento de novas ferramentas e métodos para análise e descoberta de conhecimento de dados na Web. Ela pode ser definida como a descoberta e análise de informações úteis, novas e interessantes da Web, onde, a partir das informações descobertas, seja possível demonstrar características, comportamentos, tendências e padrões de navegação do usuário na Web (COOLEY et al., 1997).

Web Mining, de acordo com Kosala e Blockeel (2000), pode ser dividida em três subáreas: *Web Structure Mining*, *Web Usage Mining* e *Web Content Mining*. A seguir, será apresentada uma breve descrição sobre cada uma dessas categorias, com maior ênfase em *Web Content Mining*, que é o objetivo desta pesquisa.

a) Mineração de estruturas: A mineração de estruturas procura descobrir um modelo na estrutura de links na web. Baseia-se na topologia de hiperlinks. Pode ser útil ao categorizar paginas da web e na geração de informações similares encontradas em diferentes sites.

Edição Especial

b) Mineração de logs: Entende-se por mineração de logs ou em inglês web usage mining, o estudo do comportamento dos usuários com relação ao acesso de um determinado site. O servidor que armazena o site em questão faz o registro dos acessos em logs que podem ser estudados e analisados para se obter informações importantes com relação ao uso do site, assim pode-se formular o site para um melhor proveito por parte do usuário, aumentando assim a frequência com que esse usuário vai acessá-lo e até mesmo aumentar o número de usuários do site.

c) Mineração de conteúdo: Essa técnica de mineração consiste em pesquisar o conteúdo como texto, imagens e outros componentes das páginas.

A mineração de conteúdo nos permite achar mais facilmente o conteúdo localizado nas páginas (LIU, 2007). Pode ser muito útil na categorização das e indexação do que é publicado nas páginas da web. Um exemplo simples desse tipo de mineração é quando se pesquisa na ferramenta “Google search” um nome de alguém (artista, jogador de futebol, etc.) ou mesmo qualquer outra coisa, e selecionamos no canto esquerdo superior o link “imagens”, o motor vai minerar o conteúdo dos sites e encontrar todas as imagens que possuem nomes iguais ou parecidos com a pesquisa que foi feita.

Com o uso dessas três categorias de descoberta de informações relevantes na Web, podemos transformar a rede em um ambiente amigável a todos os níveis de conhecimento e de experiência de uso, tornando assim a web cada vez mais popular e atingindo mais pessoas simultaneamente ao redor do mundo.

Portanto, Mineração de Opinião de Usuários da Web é um tipo de mineração de conteúdo, no qual as informações são opiniões de usuários, sobre determinados produtos, que são extraídas de fórum de discussão.

1.3 Procedimentos Metodológicos

Quando se discute mineração de opiniões de usuários na Web, um dos métodos mais utilizados nas pesquisas sobre Mineração de Opinião tem sido a Recuperação da Informação. Assim, o trabalho concentra-se em três pontos relevantes no processo de Mineração de Opinião na Web: a Recuperação da Informação, o processamento das informações recuperadas e análise dos resultados obtidos a partir da mineração das opiniões selecionadas.

Para este estudo, o primeiro passo é recuperar páginas com opiniões relevantes sobre o tema em questão. Com as técnicas desenvolvidas para recuperação da informação na web, percebe-se que várias ferramentas podem ser aplicadas para buscar opiniões de acordo com suas características. Optou-se, nesta pesquisa, pela utilização do *Google* por ser uma ferramenta conceituada por pesquisadores, devido ao seu desempenho na recuperação de páginas web. Máquinas de busca, como o *Google*, são ferramentas essenciais para recuperar informação da Web.

A quantidade de informações na Internet é tão grande e diversificada que é praticamente impossível encontrar tudo o que se precisa sem o uso de um mecanismo de busca. Assim, o uso desta tecnologia foi de fundamental importância no processo de recuperação das opiniões de usuário na web.

Com o uso desta ferramenta, foram pesquisados mais ou menos 30 sites com opiniões sobre sete marcas de notebooks, um número que foi considerado razoável em relação à quantidade de opiniões colhidas e que seria suficiente para a análise.

Optou-se por marca de notebooks, uma escolha aleatória, mas que, com certeza, encontrar opiniões sobre este produto não seria uma dificuldade desta pesquisa e se o resultado fosse positivo, esta metodologia poderia ser aplicada em qualquer outro produto.

Edição Especial

Após recuperação das opiniões, usou-se o software *Websphinx*. Este software permite salvar páginas web para um disco local para que as mesmas possam ser folheadas off-line. Além de permitir concatenar páginas de modo que elas possam ser visualizadas ou impressas como um único documento. Assim, esta ferramenta permitiu fazer um arquivo único com as 30 páginas recuperadas, o que facilitou a separação das opiniões por marca.

Com as opiniões organizadas, a principal preocupação é extrair elementos textuais, que podem ser palavras ou pequenas expressões, consideradas importantes por caracterizar o produto em estudo. Os termos relevantes são escolhidos, ou seja, palavras-chave, e são posteriormente armazenados sob uma forma estruturada, permitindo, assim, um rápido e eficiente acesso à informação. Palavras-chave são termos que expressam sentimentos positivos ou negativos capazes de qualificar um produto ou objeto em análise.

Estes termos são definidos por tomadores de decisões e de acordo com o objeto a ser avaliado, uma vez que o objetivo da metodologia é não se restringir à língua ou a um objeto específico.

Após seleção de todas as características no documento de opinião, o passo seguinte é preparar os dados para o processamento, ou seja, gerar o arquivo, na extensão do software escolhido para a mineração das opiniões. Optou-se pelo software *Weka* pelo fato de o mesmo melhor se adequar à pesquisa e apresentar melhores resultados em termos de comparação. Além do mais, o *weka* é um software livre que permite, através de métodos simples, de reconhecimento de padrão ou algoritmo de aprendizado de máquina, extrair as informações de textos (MARKOV, Z.; LAROSE, D. T., 2007).

Após carregamento do arquivo no formato arff, formato reconhecido pelo *Weka*, são selecionadas as palavras que interessam para a análise do objeto em si e o sistema examina todas as características e gera os resultados.

Edição Especial

Com o processamento das opiniões no *Weka*, é possível ver algumas estatísticas interessantes sobre os termos, onde se podem fazer as devidas comparações através dos diagramas gerados.

2 RESULTADOS E DISCUSSÃO

A tabela 1 ilustra os resultados fornecidos pelo software *Weka* após o processo de mineração das opiniões de usuários sobre marcas de notebooks.

No.	id	\$acabamento_perfeito Numeric	\$adorno Numeric	\$boa Numeric	\$bom Numeric	\$bonito Numeric	\$compatível Numeric	\$confiável Numeric	\$desempenho Numeric	\$elegante Numeric	\$excelente Numeric	\$fino Numeric	\$gosto Numeric	\$leve Numeric	\$mais_vendido Numeric	\$melhor Numeric	\$muito_bom Numeric	\$recomendo Numeric	\$super Numeric	\$ótimo Numeric
1	...	1.0	3.0	6.0	6.0	1.0	4.0	2.0	6.0	3.0	2.0	5.0	7.0	4.0	1.0	18.0	1.0	3.0	3.0	1.0
2	...	2.0	2.0	13.0	21.0	0.0	2.0	2.0	19.0	3.0	6.0	3.0	3.0	3.0	2.0	12.0	2.0	3.0	8.0	2.0
3	...	2.0	3.0	4.0	7.0	0.0	2.0	3.0	7.0	2.0	9.0	2.0	2.0	2.0	2.0	4.0	1.0	2.0	10.0	2.0
4	...	1.0	1.0	13.0	9.0	0.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	10.0	1.0	1.0	1.0	1.0
5	...	1.0	2.0	11.0	16.0	2.0	3.0	1.0	8.0	2.0	9.0	1.0	2.0	6.0	1.0	9.0	1.0	3.0	3.0	1.0
6	...	1.0	1.0	5.0	12.0	0.0	1.0	3.0	1.0	1.0	4.0	1.0	5.0	2.0	1.0	12.0	3.0	7.0	2.0	1.0
7	...	2.0	2.0	5.0	4.0	0.0	2.0	2.0	4.0	2.0	2.0	2.0	2.0	2.0	2.0	5.0	2.0	4.0	2.0	2.0

Figura 1: Exemplo de freqüência de palavras que caracterizam as marcas de notebooks

A figura acima explana as frequências das palavras que qualificam positivamente as marcas dos notebooks em análise. A maneira mais simples de usar um termo como característico em um documento é verificar se o termo aparece ou não no documento. O valor de uma característica em um documento é considerado pelo número de ocorrências desta no documento. As frequências com que os termos ocorrem no documento são os parâmetros utilizados para representar as características dos objetos em análise.

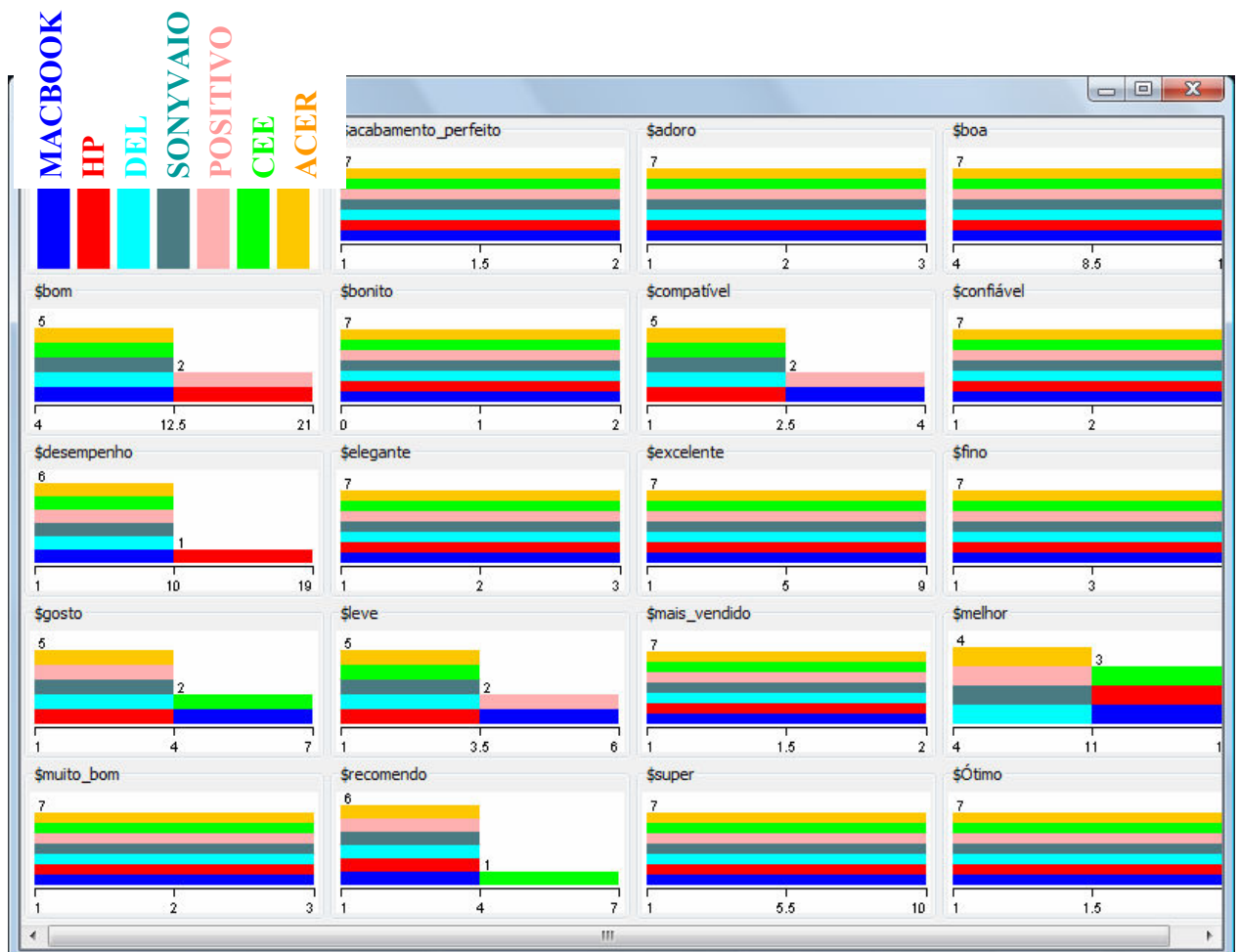
Assim, com os resultados obtidos, conforme apresentado na tabela 1, é possível fazer a análise comparativa das variáveis capazes de qualificar as marcas do produto em estudo, de acordo com a opinião dos usuários. Na figura, observa-se que foram selecionadas sete marcas de notebooks, ou seja, sete classes que estão representadas na figura 1, numeradas de um a sete, que representam

Edição Especial

respectivamente as marcas MACBOOK, HP, DEL, SONYVAIO, POSITIVO, CEE e ACER.

Ao analisar os valores mais acentuados, verifica-se que o termo “bom” apareceu com destaque nas marcas HP e Positivo, com frequência de 21 vezes na marca HP e 16 vezes na marca Positivo. A marca HP, destacou-se também no atributo “desempenho”, com 19 ocorrências desta característica. A marca MacBook ganhou destaque no atributo “melhor”, com 18 ocorrências. Em relação a esta características, também tiveram destaques as marcas HP, SonyVaio e CEE.

A figura 2 ilustra os histogramas gerados pelo processamento no *Weka*. Os histogramas mostram algumas estatísticas interessantes sobre os termos, onde se podem fazer as devidas comparações.



Edição Especial

Figura 2: Histogramas das características gerados pelo Weka

A figura 2 apresenta os histogramas comparativos entre as opiniões que caracterizam as marcas de notebooks em análise. Examinando os diagramas é possível encontrar os termos mais específicos para cada documento. Por exemplo, ao analisar o diagrama que representa o termo “recomendo”, observa-se que a marca CEE foi a mais recomendada.

CONCLUSÃO

Atualmente, a busca por informação na web tem crescentemente se tornado importante e popular. Isso se deve à riqueza de informação presente na web e, como se vive uma época de constantes mudanças, buscar novos conhecimentos e permanecer sempre atualizado é de extrema importância. Portanto, buscar informações qualitativas sobre produtos, eventos, etc, em opiniões de usuários na Web e descobrir informações relevantes com base nestas opiniões, tornou-se de suma importância.

Como se pode verificar, o acesso às informações relevantes torna-se imprescindível para que as pessoas possam decidir a melhor maneira de direcionar seus projetos e tomar decisões importantes. Como obter essas informações normalmente não é fácil, esta pesquisa propôs uma metodologia para mineração de opiniões na Web, que objetiva extrair características importantes contidas em documentos de opiniões de usuários e facilitar a tomada de decisão por parte dos dirigentes.

Nesta pesquisa, conclui-se que os resultados gerados pelo *Weka*, após execução em uma base de textos já preparada para processamento, facilitaram a visualização do conhecimento, de forma organizada. Embora tenha trabalhado com um universo de informações bastante subjetivas, as opiniões de usuários, os

Edição Especial

resultados ilustrados em forma de tabela e gráficos trouxeram facilidades na interpretação dos resultados obtidos.

Pode-se concluir também que, com base no processamento destas opiniões, a metodologia pode ser aplicada para a análise de outros objetos, sejam eles produtos ou serviços, tais como analisar opiniões de usuários sobre hotéis antes de fazer uma reserva, ou verificar opiniões de usuários em relação a uma universidade, antes de prestar o vestibular.

O que dificultou o estudo sobre Mineração de opiniões da Web é a demarcação de seu escopo, por se tratar de uma área multidisciplinar vinculada a outras áreas, como a recuperação de informação, estatística e extração da informação, isto exige um estudo que abrange mais de uma área do saber.

Alem disso, no Brasil, parece que os projetos ainda estão bastante tímidos e a maioria das referencias bibliográficas encontradas estão em inglês.

Também, varias dificuldades encontradas provem da escolha do tema em questão, pois por se tratar de opiniões de usuários, depara-se com vários problemas, alguns usuários postam em um fórum de opinião sobre marca de um produto, outros assuntos totalmente diferentes do tema relacionado, além de digitar várias palavras erradas, o que acaba ocasionando uma distorção no resultado.

Outro fator que prejudica o trabalho é o fato das opiniões serem totalmente desestruturadas, exigindo um esforço manual na organização destes arquivos para que eles possam ser reconhecidos pelo software escolhido para a mineração das opiniões em busca de informações relevantes.

MINING THE OPINIONS OF USERS IN SEARCH OF KNOWLEDGE**ABSTRACT**

This work aims to use the techniques of Web Mining for mining reviews of Web users, in order to discover relevant information based on reviews posted on discussion forums. The goal is to extract features for analyzing the objects under study in accordance with the views expressed. To validate the results, we used the Weka software for processing and counting of keywords. Thus, based on histograms of the elements under review, the study sought to identify and compare characteristics of the products under study.

KEYWORDS: INFORMATION RETRIEVAL, WEB MINING, OPINION MINING, KNOWLEDGE.

REFERÊNCIAS

AMO, S. **Técnicas de Mineração de Dados**. Universidade Federal de Uberlândia. Disponível em: <http://inf.cp.utfpr.edu.br/ligia/papers/jai-cap5.pdf>. Acesso em 24 fev. 2010.

COOLEY, R. W. et al. **Web mining**: information and Pattern Discovery on the World Wide Web. In: Proceedings of International Conference on Tools with Artificial Intelligence (ICTAI), 1997, 10p.

YAMAOKA, E. J. Recuperação de Informação na Web. Disponível em: http://www.ct.ufrj.br/bib/bibliotecaonline/pesqapoio/Recuperacao_informacao_web.pdf. Acesso em 01 jun. 2010.

Edição Especial

YATES, R. B., NETO, B. R. Modern information retrieval. Disponível <http://people.ischool.berkeley.edu/~hearst/irbook/>. Acesso em 01 març. 2010

KOSALA, R.; BLOCKEEL H. “**Web Mining Research: a survey**”. In: *SIGKDD Explorations*, vol. 2 -1, June 2000, pp. 1-15.

LIU, B. **Web Data Mining**. Exploring Hiperlinks, Contents, and Usage Data. Chigago. Springer, 2007.

MAGALHÃES, L. H.,. Uma análise de ferramentas para mineração de conteúdo de páginas Web. (Dissertação de mestrado) Universidade Federal do Rio de Janeiro. 2008.

MARKOV, Z.; LAROSE, D. T. **Data Mining the web: Uncovering Patterns in Web Content, Structure, and Usage**. New Britain: Wiley, 2007.