Mineração na Web

LEANDRO BALBY MARINHO
ROSARIO GIRARDI

UFMA – UNIVERSIDADE FEDERAL DO MARANHÃO
DEINF – DEPARTAMENTO DE INFORMÁTICA
GESEC/DEINF, Avenida dos Portugueses,
s/n, Campus Universitário do Bacanga,
São Luís - Maranhão - Brasil,
CEP 65080-040

lbalby@uol.com.br
rgirardi@deinf.ufma.br

Resumo: A Web é hoje a maior fonte de informação eletrônica que dispomos. Entretanto, por causa da sua natureza dinâmica, a tarefa de achar informações relevantes se torna muitas vezes uma experiência frustrante. Muitos esforços de pesquisa têm sido feitos no sentido de sanar esse problema. Um deles é a utilização de técnicas de mineração de dados para a descoberta de informações na Web. Este tutorial apresenta uma visão geral da mineração na Web, as fases do processo e as categorias em que se divide.
Palavras chaves: WWW, Mineração de dados, Recuperação de informação, Descoberta de conhecimento.

1. Introdução

Somos testemunhas do enorme aumento de informações e recursos na Web nos últimos anos. Mais de um bilhão de páginas são indexadas pelos motores de busca [Pal, 2000] e achar a informação desejada pode algumas vezes se tornar uma tarefa penosa. Essa abundância de informações e recursos instigou a necessidade do desenvolvimento de ferramentas automáticas de mineração e descoberta de informações na Web.

De forma geral, a mineração na Web pode ser conceituada como a descoberta e análise inteligente de informações úteis da Web [Cooley, 1997]. Pode-se estar interessado, por exemplo, na informação contida dentro dos documentos da Web – mineração de conteúdo - na informação contida entre os documentos da Web – mineração de estrutura – ou na informação contida na utilização ou interação com a Web – mineração de uso. Essas são as três categorias em que se divide a mineração na Web, de acordo com a parte da Web a ser minerada. Para cada classificação são desenvolvidas técnicas e metodologias distintas, muitas delas herdadas de outras áreas disciplinares como Aprendizagem de máquina, Bancos de dados, Estatística, Recuperação de informação, Inteligência artificial e Redes sociais.

A mineração na Web já não é tão recente, ela vem sendo citada e estudada desde meados de 1996, mas tem realmente ganhado importância nestes últimos anos. Podemos apontar dois fatores principais que contribuíram para isso:

- *Aumento das transações comerciais na Web*, que motivaram o desenvolvimento de técnicas para a mineração de uso, pois através delas os sites de venda puderam aprender acerca dos perfis dos compradores para montarem melhores estratégias de venda e marketing;
- *O desenvolvimento da Web semântica* [Decker, 2000] *e da tecnologia dos agentes da informação* [Sycara, 1996], onde as técnicas de mineração na Web são utilizadas. A Web semântica poderá entre outras coisas estender a inteligência dos agentes e não apenas o seu conhecimento. Dessa forma, os serviços da Web poderão eles próprios tornar-se entidades dotadas de comportamento autônomo, que poderão entre outras coisas, comunicar-se através de uma linguagem comum. A mineração na Web será uma ferramenta crucial a ser utilizada pelos agentes e serviços nessa visão da Web, pois ela os ajudará em várias tarefas, dentre as quais estão busca por informações, personalização e talvez até como mecanismo de aprendizado.

Entretanto, têm-se muitos desafios e problemas a serem contornados antes que a Web possa realmente se transformar num meio mais rico, amigável e inteligente na qual todos possamos explorar e compartilhar.

O objetivo deste tutorial é apresentar uma visão geral sobre a mineração na Web, as fases do processo e as categorias em que se divide. A seção 2 descreve os principais conceitos da mineração na Web e a sua divisão segundo o processo geral de descoberta de conhecimento em bases de dados, falando um pouco sobre cada fase do processo. A seção 3 apresenta as categorias em que se divide a mineração na Web e o contexto em que cada uma delas se aplica.

2. Mineração na Web

A Web é uma vasta coleção de documentos heterogêneos. Possui natureza dinâmica e milhões de páginas surgem e desaparecem todos os dias. Por isso sente-se um anseio para que a Web realmente alcance todo o seu potencial e se torne uma ferramenta mais utilizável, eficaz e compreensível. Nesse contexto a mineração de dados aparece como uma possibilidade óbvia a ser explorada. Em parte pelo seu grande sucesso quando aplicada a bancos de dados tradicionais, e em parte porque a Web parece ser uma área fértil em potencial para a aplicação de suas técnicas.

A mineração de dados refere-se ao processo não trivial de identificação de padrões válidos, previamente desconhecidos e potencialmente úteis dos dados [Frawley, 1992]. Entretanto, utilizar e compreender os dados disponíveis na Web não é uma tarefa simples, pois esses dados são muito mais sofisticados e dinâmicos do que os sistemas de armazenamento de bancos de dados tradicionais. Enquanto estes últimos utilizam estruturas de armazenamento bem definidas e estruturadas, a Web não possui qualquer controle sobre a estrutura ou o tipo dos documentos que virtualmente armazena. Outro aspecto que diferencia a mineração de dados tradicional da mineração na Web é a existência de vínculos de hipertexto entre os seus documentos. Os vínculos de hipertexto são uma rica fonte de informações a ser explorada, pois dentre outras coisas, ajudam no processo de ranqueamento de páginas pelos motores de busca e na identificação de micro-comunidades na Web.

Apesar das diferenças e particularidades entre as duas abordagens - mineração em dados tradicionais e mineração de dados da Web -, a metodologia utilizada para a mineração na Web, segue os mesmos passos utilizados no processo geral de descoberta de conhecimento em bases de dados (KDD – Knowledge Database Discovery). O processo de mineração na Web é dividido em 4 sub-tarefas [Etizione, 1996], que na verdade são análogas às fases do processo KDD.

Com base nas quatro fases descritas a seguir e na representação da Figura 1, a mineração na Web pode ser vista como a utilização de técnicas de mineração de dados para a recuperação automática, extração e avaliação de informação para a descoberta de conhecimento em documentos e serviços da Web. Aqui avaliação inclui tanto ‘ generalização’ quanto ‘análise’ [Pal, 2000].

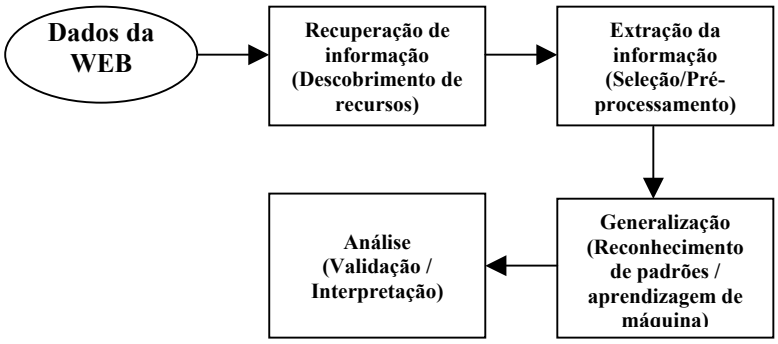

Figura 1 – Sub-tarefas da mineração na Web

2.1. Recuperação de informação (descobrimento de recursos)

A recuperação de informação ou descobrimento de recursos trata da automatização do processo de recuperação de documentos relevantes, que inclui, principalmente, representação, indexação e busca por documentos.

Um índice é basicamente uma coleção de termos retirados dos documentos com ponteiros para os lugares onde as informações sobre os documentos podem ser encontradas [Pal, 2000]. A indexação de páginas Web, para facilitar o processo de recuperação, é bem mais complexa que o processo de indexação utilizado em bancos de dados tradicionais. A enorme quantidade de páginas na Web, seu dinamismo e atualizações freqüentes fazem da indexação uma tarefa aparentemente impossível. E, na verdade, esse é um dos grandes desafios dos serviços de busca atuais: indexar toda a Web. Os serviços de busca - programas destinados a consultar e recuperar informações armazenadas tanto em bancos de dados, páginas HTML ou texto - ainda estão bem longe disso e

isso influi na recuperação das informações desejadas, pois algumas vezes os usuários estão atrás de uma informação que está justamente na porção da Web que ainda não foi indexada.

A indexação de documentos na Web pode ser humana ou manual ou automática [Pal, 2000], e está baseada nos modelos tradicionais de recuperação de informação: espaço vetorial, estatístico e lingüístico [Baeza-Yates, 1999] [Girardi, 1998] [Salton, 1983].

A indexação é essencialmente um processo de classificação onde é realizada uma análise conceitual do documento ou elemento de informação. Por exemplo, nas técnicas baseadas no modelo do espaço vetorial, a indexação envolve a atribuição de elementos de informação a certas classes, onde uma classe é o conjunto de todos os elementos de informação para o qual um termo de indexação (ou palavra-chave), em particular, tem sido atribuído. Os elementos de informação podem fazer parte de várias classes. Algumas técnicas atribuem pesos aos termos de indexação de um elemento de informação de forma a refletir sua relativa relevância [Girardi, 1998]. Nas técnicas baseadas no modelo estatístico, os termos de indexação são extraídos a partir de uma análise de freqüência das palavras ou frases em cada documento e em toda a fonte de informação. Nas técnicas lingüísticas, os termos de indexação são extraídos utilizando técnicas de processamento da linguagem natural, por exemplo, análise morfológica, lexical, sintática e semântica [Girardi, 1995].

2.2. Seleção da informação / Extração e pré-processamento

Uma vez tendo sido os documentos recuperados, o próximo passo é transformar ou pré-processar esses documentos de forma que algoritmos de mineração de dados e aprendizagem de máquina possam ser aplicados de forma efetiva.

O campo disciplinar conhecido como extração de informação presta um grande serviço à mineração da Web, no que diz respeito à fase de extração e pré-processamento da informação. Denomina-se extração de informação à tarefa de identificar fragmentos específicos que constituem o núcleo semântico de um documento em particular e construir modelos de representação da informação (conhecimento) a partir dele [Pal 2002]. Os métodos geralmente envolvem a escrita de código específico, popularmente chamados de *wrappers* responsáveis pelo mapeamento do documento para algum modelo de representação do conhecimento. O problema é que para cada documento da Web temos que escrever um código específico, tornando o trabalho manual. Como os documentos da Web não possuem uma semântica agregada às informações que contém, e nem mesmo um padrão de como apresentar essas informações ao usuário, temos que aprender acerca da estrutura individual de cada documento e escrever código para essa estrutura em particular. Daí a dificuldade de estendermos ou generalizarmos esse código para outros documentos.

Vários métodos foram desenvolvidos para a extração de informação tanto em documentos desestruturados quanto em semi-estruturados. [Kushmerick, 1997] por exemplo, descreve vários aspectos e técnicas da extração de informação, [Freitag, 1998] fala sobre a aplicação de algoritmos de aprendizagem de máquina para a extração de informação de documentos HTML e [Soderland, 1999] fala sobre a aprendizagem de regras para a extração de informação de documentos semi-estruturados e texto comum.

É importante salientar a diferencia entre as fases de recuperação e extração de informação. As técnicas de extração de informação buscam derivar conhecimento de documentos recuperados segundo a forma como um documento está estruturado e representado enquanto as técnicas de recuperação de informação visualizam o documento apenas como um conjunto de palavras [Pal 2002].

2.3 Generalização

Após as informações terem sido extraídas e algum modelo de representação das informações ter sido construído, são utilizadas técnicas de mineração de dados e aprendizagem de máquina para descobrir novo conhecimento a partir do que já existe. É nessa fase que os algoritmos de mineração vão descobrir novo conhecimento em cima do que já existe. Um exemplo que nos daria uma idéia de como seria uma saída de um desses algoritmos é dado abaixo.

a) 70% das pessoas que acessam a seção sobre natação também acessam a seção sobre artes marciais;
b) 80% dos sites que abordam o tema Fórmula-1 possuem links apontando para sites que falam da vida de Ayrton Senna;

No primeiro exemplo, a saída poderia dar uma indicação ao mantenedor da loja virtual sobre as preferências e perfis de seus clientes, de forma a montar estratégias de vendas que possam impelir o usuário a comprar mais. No outro exemplo, a saída descobre uma relação interessante entre os sites podendo dar novos caminhos aos usuários interessados nesses tópicos.

O maior problema em aprender ou descobrir novos conhecimentos da Web é a falta de marcação semântica das informações. Muitos algoritmos de mineração de dados requerem como entrada exemplos positivos ou negativos de algum conceito. Se, por exemplo, tivéssemos um conjunto de páginas da Web marcadas como exemplos positivos e negativos do conceito *portal,* seria fácil modelar um algoritmo classificatório para a classificação automática de novas páginas como portais ou não portais. Embora a Web atual dificulte o

processamento das suas informações por parte das máquinas, a Web Semântica [Lee, 2001] fornece uma solução a este problema.

Agrupamento ou *clustering* é uma técnica de classificação que não requer entradas com marcação semântica, e por isso tem sido aplicada com sucesso em grandes conjuntos de documentos HTML [Cutting, 1992]. No clustering, documentos são agrupados de acordo com a sua similaridade, portanto, um novo documento é classificado de acordo com a sua similaridade com algum conjunto de documentos existente. Uma boa referência para clustering no contexto da mineração na Web pode ser encontrada em [Lingras,2002].

As Regras de associação também podem ser utilizadas nessa fase do processo. Regras de associação são basicamente expressões do tipo $X \Rightarrow Y$ onde X e Y são conjuntos de itens. $X \Rightarrow Y$ expressa que toda vez que uma transação T contiver X então ela provavelmente também conterá Y. A probabilidade ou confiança da regra é a porcentagem de transações contendo Y junto a X comparado ao total de transações contendo X. A idéia de minerar regras de associação se origina nos dados de super-mercados e afins onde regras como "O cliente que compra o produto x também comprará o produto y com probabilidade (confiança) de c%" [Pal, 2000].

A Web em sua concepção foi construída de forma a atender as necessidades de visualização e consumo de seres humanos, onde os textos são quase sempre escritos em linguagem natural sem nenhuma semântica que facilite seu processamento. Isso instigou o desenvolvimento de um novo conceito para a Web, chamada de "Web semântica" onde além de outras coisas promete escrever os documentos da Web com uma semântica agregada às informações de forma que as máquinas possam compreendê-los e processá-los. Um bom trabalho para aprofundamento sobre Web Semântica está em [Lee, 2001]

2.4 Análise

Uma vez os padrões tendo sido descobertos os analistas precisam de técnicas e ferramentas apropriadas de modo a entender, visualizar, interpretar e validar esses padrões. O sistema WEB-MINER [Mobasher, 1997] , por exemplo, propõe uma linguagem de consulta estruturada para a consulta do conhecimento descoberto (na forma de regras de associação e padrões seqüenciais). Outros sistemas utilizam técnicas de OLAP (On-line Analytical Processing) [Girardi, 1998] R. Girardi. "Main Approaches to Software Classification and Retrieval". Em: Ingeniería del Software y reutilización: Aspectos Dinámicos y Generación Automática. Editores J. L. Barros y A. Domínguez. (Universidad de Vigo – Ourense, del 6 al 10 de julio de 1998). Julio, 1998.

[Girardi, 1995] R. Girardi,. *"Classification and Retrieval of Software through their Descriptions in Natural Language",* Ph.D. dissertation, No. 2782, University of Geneva, December 1995.

[Han, 2000] com o propósito de simplificar a análise de estatísticas de uso em logs de acesso.

3. Categorias da Mineração na Web

Nesta seção é apresentada uma visão geral das categorias em que se divide a mineração na Web, assim como algumas das técnicas utilizadas em cada uma delas. A mineração na Web se divide em três categorias de acordo com a parte da Web a ser minerada: mineração de conteúdo, mineração de estrutura e mineração de uso.

A mineração de conteúdo aborda a mineração dos dados contidos dentro dos documentos da Web. A grande quantidade de formatos que os dados podem assumir (textos comuns, páginas HTML, imagens, áudio, vídeo, etc.) acabam dirigindo as técnicas de mineração a serem utilizadas.

A mineração de estrutura por outro lado, aborda a mineração das informações contidas entre os documentos da Web. Os documentos da Web se relacionam basicamente através de vínculos de hipertexto, e esses vínculos escondem informações valiosas e interessantes não só sobre a topologia da Web, mas também sobre como os documentos se relacionam.

A mineração de uso por sua vez, aborda a mineração das informações de uso da Web, que em outras palavras, são as informações sobre como o usuário interage com a Web. Nessa categoria são tratadas questões como personalização, interfaces adaptativas e aprendizado de perfis de usuários.

3.1 Mineração de conteúdo

A mineração de conteúdo trata do descobrimento de informações úteis do conteúdo, dados, documentos e serviços da Web [Pal, 2000]. Vale salientar que o conteúdo da Web não se constitui apenas de texto ou hipertexto, mas abrange uma ampla variação de tipos de dados, tais como áudio, vídeo, dados simbólicos, metadados e vínculos de hipertexto. Apesar de já existir uma área de pesquisa destinada ao estudo da mineração de dados multimídia, o foco ainda são os dados de texto e hipertexto, que na verdade são os que constituem o grosso da Web. Uma boa referência para pesquisa sobre mineração de dados multimídia é [Zaiane, 1998].

Os dados de texto da Web podem ser de três tipos: desestruturados, tais como textos comuns, semi-estruturados, tais como documentos HTML, e estruturados, tais como as tabelas de bancos de dados. No tratamento de dados desestruturados utiliza-se KDT (Knowledge Discovery in Texts) ou mineração de dados em

textos. A mineração em textos é uma área bem amadurecida e a sua cobertura em detalhes está além do escopo deste tutorial, mas uma boa referência é [Mladenic, 1998].

A extração de conhecimento da Web e a sua modelagem em uma representação simbólica para a aplicação de técnicas de mineração de dados é descrito em [Ghani, 2000].

Algumas outras abordagens que tratam da mineração de dados em texto sugerem reestruturar os documentos de forma que eles se tornem legíveis para as máquinas, ou seja, técnicas para a inserção de marcas (tags) semânticas nas informações [Pal, 2000].

A mineração em hipertexto envolve a mineração de páginas HTML, as quais além de texto contém vínculos hipertexto. Um excelente tutorial apresentando esse assunto é descrito em [Chakrabarti, 2000].

A mineração em serviços da Web tais como grupos de notícia, grupos de e-mail, lista de discussão e bibliotecas digitais também é uma área que tem cada vez mais chamado a atenção dos pesquisadores, principalmente pesquisadores envolvidos na área de WI ("Web Intelligence") [Zhong, 2002]. A WI promete, dentre outras coisas, transformar os serviços da Web em entidades inteligentes, de forma que elas possam interagir e se comunicar através de uma linguagem comum, elevando assim, a Web a um outro nível de tecnologia da informação. Em [Levy, 2000] são discutidos os sistemas de Internet inteligentes em geral, abordando temas como modelagem de usuários, descobrimento e análise em fontes de informações remotas, integração da informação e gerenciamento de sites da Web.

Há uma fina linha separando a mineração de conteúdo e a recuperação de informação na Web. Não há um consenso sobre a relação entre as duas, alguns afirmam que a recuperação da informação na Web pode ser vista como uma instância da mineração de conteúdo, e outros associam a mineração de conteúdo com recuperação inteligente de informação. Isso acontece porque algumas vezes as duas acabam trabalhando juntas para alcançar determinado objetivo e uma acaba por complementar a outra.

Há basicamente duas estratégias para a mineração de conteúdo: uma realiza a mineração diretamente do conteúdo dos documentos e a outra incrementa o poder de busca de outras ferramentas e serviços. Na primeira estratégia, os documentos pretendidos já foram recuperados e já estão prontos para serem minerados. Na segunda estratégia, a mineração de conteúdo presta um grande "favor" às ferramentas e serviços de recuperação de informação, pois ajuda a realizar o processo de indexação e categorização dos documentos. Dessa forma, percebemos que quando a mineração de conteúdo utiliza a segunda estratégia, ela complementa o processo de recuperação de informação, sendo utilizada como uma ferramenta pelos motores e serviços de busca, daí nesse caso ser descrito por alguns como recuperação inteligente de informação.

A mineração de conteúdo pode seguir duas abordagens: baseada em agentes ou baseada em bancos de dados. A abordagem baseada em agentes envolve o desenvolvimento de sistemas de inteligência artificial que podem agir de forma autônoma ou semi-autônoma para a descoberta e organização de informações da Web de acordo com os interesses de um usuário em particular [Kosala, 2000]. Geralmente, a abordagem baseada em agentes pode ser dividida em três categorias: agentes de busca inteligentes, agentes de filtragem e/ou categorização da informação e agentes de interface [Cooley, 1997]. A abordagem de bancos de dados focaliza-se nas técnicas para transformar os dados semi-estruturados ou desestruturados da Web em modelos de dados estruturados onde mecanismos de consulta, como, por exemplo, a linguagem SQL, possam ser utilizados, assim como técnicas de mineração de dados para a análise.

3.2 Mineração de estrutura

Enquanto que na mineração de conteúdo da Web estamos interessados no que há dentro dos documentos, na mineração de estrutura o interesse está nas informações que existem de forma implícita entre os documentos. Esta categoria envolve a mineração da estrutura que há por trás da interligação entre os documentos da Web. O que liga esses documentos são os vínculos de hipertexto, os quais são os principais objetos de estudo nesta categoria.

A Web pode ser visualizada como um grafo orientado, onde os nós representam páginas, e as setas entre pares de nós representam vínculos entre as páginas. Essa representação da Web em forma de grafo apresenta uma forte semelhança com as chamadas redes sociais [Kumar, 2002] que, juntamente com a análise de citações, inspirou a pesquisa dessa categoria de mineração.

A teoria moderna de redes sociais foi desenvolvida a partir do trabalho de Stanley Milgram [Kumar, 2002]. Em 1967, Milgram conduziu experimentos onde ele pedia que diversas pessoas residentes em Omaha, Nebraska, conduzissem uma carta para um associado seu que morava em Boston. As pessoas só podiam enviar a carta para outra pessoa que elas conhecessem pelo primeiro nome, e essas pessoas por sua vez só podiam retransmitir a carta para uma pessoa que elas também conhecessem pelo primeiro nome. O objetivo era de que a carta chegasse ao seu associado no menor número de "passos" possíveis. Milgram descobriu que o número médio de "passos" ao longo do caminho das cartas que conseguiam chegar com sucesso era seis, criando o folclore de que quaisquer duas pessoas residentes nos Estados Unidos estavam ligadas em uma rede social com "seis graus de separação".

Os pesquisadores têm explorado continuamente as similaridades entre a Web e as redes sociais, desenvolvendo técnicas que incrementam o poder dos motores de busca e dos sistemas de gerenciamento do conhecimento.

Nas citações bibliográficas quando um artigo é bastante citado isso indica que provavelmente este é um artigo importante e de maior autoridade perante outros que abordam o mesmo tema. Acontece o mesmo com as páginas e documentos da Web. Os vínculos de hipertexto dão indicações interessantes de como as páginas se relacionam entre si, links apontando para uma página, por exemplo, podem indicar a sua importância, enquanto links "saindo" de uma página podem indicar entre outras coisas a continuação ou complemento dos tópicos abordados por ela.

Alguns algoritmos foram propostos para a modelagem da topologia da Web tais como o HITS ("Hyperlinked Induced Topic Search") [Kleinberg, 1998] e o PageRank [Brin, 1998]. Esses modelos são aplicados principalmente para calcular a qualidade ou relevância das páginas da Web. Uma das regras utilizadas é que quanto mais páginas estiverem apontando para uma determinada página, mais relevante ela será. Várias medidas são tomadas para garantir que as páginas que apontam tenham credibilidade. Alguns exemplos são o sistema Clever [Chakrabarti, 1999] e o site de busca Google [Brin, 1998]. Algumas outras aplicações destes modelos são a categorização de páginas Web e a descoberta de micro-comunidades na Web [Kumar, 1999].

3.3 Mineração de uso

A mineração de uso da Web focaliza-se em técnicas que possam prever o comportamento do usuário enquanto ele interage com a Web [Kosala, 2000]. Enquanto a mineração de conteúdo e a mineração de estrutura utilizam os dados reais ou primários da Web, a mineração de uso lida com os dados secundários provenientes da interação do usuário com a Web. Os dados de uso da Web incluem dados provenientes de logs de servidores web, logs de servidores proxy, logs de browsers, perfis de usuário, cookies, seções ou transações de usuários, pasta favoritos, consultas do usuário, cliques de mouse e qualquer outro dado gerado pela interação do usuário com a Web.

O processo de mineração de uso da Web pode ser classificado segundo duas abordagens [Borges, 1998]. A primeira mapeia os dados de uso do servidor Web em tabelas relacionais antes das técnicas adaptadas de mineração de dados serem aplicadas. A segunda utiliza os dados de logs diretamente utilizando técnicas especiais de pré-processamento. Assim como no KDD, a limpeza e pré-processamento dos dados, aqui, é uma parte crucial do processo, pois a qualidade desses dados vai determinar a eficiência dos algoritmos de mineração. Uma boa referência para a descrição e comparação de métodos de pré-processamento para a mineração de uso da Web pode ser encontrada em [Cooley, 1999].

As aplicações da mineração de uso da Web podem ser classificadas em duas categorias principais: aprendizado de perfil de usuário ou modelagem em interfaces adaptativas (personalização) e aprendizado de padrões de navegação de usuário. A mineração de uso da Web despertou interesse especial no comércio eletrônico, principalmente pela sua necessidade de aprender acerca do comportamento dos clientes, perfis de compra, preferências e padrões de navegação. Alguns sites populares de comércio eletrônico já utilizam estas técnicas não só para a adaptação do site de acordo com o perfil do usuário, mas como para fazer recomendações de produtos de acordo com compras anteriores, ou baseadas na similaridade entre perfis de usuários. Um bon trabalho para o aprofundamento no tema são [Srivastava, 2000], [Salton, 1983] G. Salton, "An Introduction to Modern Information Retrieval". New York: McGraw-Hill, 1983.

[Spilopoulon, 1999] e a tese de doutorado de Robert Coolley [Cooley, 2000] que trata da mineração de padrões de usuários.

4. Conclusão

Este tutorial apresentou uma visão geral da mineração de dados da Web. Foram analisadas as diferentes fases do processo de mineração: descobrimento de recursos, extração de informação, generalização e análise. Também foi apresentada uma classificação das diferentes modalidades de mineração: mineração de conteúdo, mineração de estrutura e mineração de uso.

O que dificulta o estudo da mineração da Web é a delimitação de seu escopo por se tratar de uma área multidisciplinar vinculada a outras áreas, como a recuperação de informação, a aprendizagem de máquina e os agentes da informação.

Este trabalho procurou, de uma forma concisa, caracterizar e definir o alcance da mineração da Web; estabelecer seu relacionamento com outras disciplinas e oferecer indicações para o aprofundamento no estudo dos diferentes tópicos da área.

5. Referências bibliográficas

[Baeza-Yates, 1999] R. Baeza-Yates, B., Ribeiro-Neto, **"**Modern Information Retrieval". New York: ACM Press Series/Addison Wesley, 1999.

[Borges, 1998] J.Borges & M.Levene, “Mining association rules in hypertext databases”. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98). New York City, New York, USA, 1998.

[Brin, 1998] S. Brin & L. Page, “The anatomy of a large scale Web Search Engine”. In Seventh International World Wide Web Conference, Brisbane, Australia, 1998.

[Chakrabarti, 1999] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, & A. Tomkins, “Mining the link structure of the World Wide Web“, 1999.

[Chakrabarti, 2000] S. Chakrabarti, “Data mining for hypertext”. ACM SIGKDD Explorations, 2000.

[Cooley, 1997] R.Cooley, B.Mobasher & J. Srivastava, “Web mining: information and pattern Discovery on the World Wide Web”. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.

[Cooley, 1999] R.Cooley, B.Mobasher & J. Srivastava, “Data preparation for mining world wide web browsing patterns”. Knowledge and Information Systems, 1999.

[Cooley, 2000] R. W. Cooley, “Web usage mining: Discovery and application of Interesting Patterns from Web data”. PhD thesis, Dept. of Computer Science, University of Minesota, 2000.

[Cutting, 1992] D.D. Cutting, J. Karger, J. Pederson & J. Scatter, “A cluster based approach to browsing large document collections”. Proceedings of the Fifteenth International Conference on Research and Development in Information Retrieaval, 1992.

[Decker, 2000] S. Decker, S. Melnick, F. V. Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, I.Horrocks, “THE SEMANTIC WEB:*The Roles of XML and RDF*”. IEEE Internet Computing, 2000.

[Etzione, 1996] O. Etzione, “The World Wide Web Quagmire or gold mine ” Communications of the ACM, vol.39, no.11, pp. 65-68, 1996.

[Frawley, 1992] W.J. Frawley, G. Piatetsky-Shapiro, and C.J. Matheus. Knowledge discovery in databases: an overview. In G. Piatetsky-Shapiro & W.J. Frawley, editors, “*Knowledge Discovery in Databases”.* AAAI / MIT Press, 1991.

[Freitag, 1998] D. Freitag, “Information Extraction from HTML: Application of a General Machine Learning Approach”. American association for Artificial Intelligence, 1998.

[Ghani, 2000] R. Ghani, R. Jones, D. Mladenic, K. Nigam, & S. Slattery, “Data mining on symbolic knowledge extracted from the Web”. In Proceedings of the Sixth Internaional Conference on knowledge Discovery and Data mining (KDD-2000), 2000.

[Girardi, 1998] R. Girardi. “Main Approaches to Software Classification and Retrieval”. Em: Ingeniería del Software y reutilización: Aspectos Dinámicos y Generación Automática. Editores J. L. Barros y A. Domínguez. (Universidad de Vigo – Ourense, del 6 al 10 de julio de 1998). Julio, 1998.

[Girardi, 1995] R. Girardi,. *"Classification and Retrieval of Software through their Descriptions in Natural Language",* Ph.D. dissertation, No. 2782, University of Geneva, December 1995.

[Han, 2000] J .Han, “OLAP Mining: An integration of OLAP with Data Mining”. School of Computing Science, Simon Fraser University, British Columbia, Canada, 2000.

[Kleinberg, 1998] J.M. Kleinberg, “Authoritative Sources in a Hyper-linked Enviroment”. In Proc. ACM-SIAM Symposyum on Discrete Alghoritms, 1998.

[Kosala, 2000] R. Kosala & H. Blockeel, “Web mining research: a survey”. *SIG KDD Explorations,* vol.2, pp. 1-15, 2000.

[Kumar, 1999] S.R. Kumar, P. Raghavan, S. Rajagopalan & A. Tomkins, “Trawling the web for emerging cybercommunities”. In Proceedings of the Eighth WWW Conference, 1999.

[Kumar, 2002] R. Kumar, P. Raghavan, S. Rajagopalan & A. Tomkins, “The Web and Social Networks”, IEEE Computer, vol.35, no.11, 2002, pp.32-36.

[Kushmerick, 1997] N. Kushmerick, “Wrapper Induction for Information Extraction”. Doctoral thesis. University of Washington, Department of Computer Science and Engineering, 1997.

[Lee, 2001] BERNERS – LEE, Tim, HENDLER, James, LASSILA, Ora. The Semantic Web**.** Scientifc American, May 2001.

[Levy, 2000] A. Levy & D. Weld, “Intelligent Internet Systems”. Artificial Intelligence, vol.118, no.1-2, 2000.

[Lingras, 2002] P. Lingras, “Rough Set Clustering for Web Mining”. Saint Mary´s University, 2002.

[Mladenic, 1998] M. Mladenic & M. Globelnik, “Efficient text categorization”. In Proceedings of Text Mining Workshop on the 10th European Conference on Machine Learning, 1998.

[Mobasher, 1997] B. Mobasher, N. Jain, E.H. Han & J. Srivastava, “Web Mining: Patterns from WWW transactions”. Tech. Rep. TR96-050, Dept. of Computer Science, University of Minesota, 1997.

[Pal, 2000] Sankar K. Pal, Varum Talwar, Pabitra Mitra, “Web Mining in Soft Computing Framework: Relevant, State of the Art and Future Directions”, 2000.

[Salton, 1983] G. Salton, “An Introduction to Modern Information Retrieval”. New York: McGraw-Hill, 1983.

[Spilopoulon, 1999] M. Spilopoulon, “Data mining for the Web”. In Principles of data mining and knowledge discovery, Second European Symposium, 1999.

[Srivastava, 2000] J. Srivastava, R.Cooley, M. Deshpande & P.N.Tan., “Web usage mining: Discovery and applications of usage patterns from Web data”. SIG KDD Explorations, 2000.

[Soderland, 1999] S. Soderland, “Learning Information Extraction Rules for Semi-structured and Free Text”. Machine Learning 1-44. ©Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 1999.

[Sycara, 1996] K. Sycara, K. Decker, A. Pannu, M. Williamson & D. Zeng, “Distributed Intelligent Agents”. The robotics institute, Carnegie Mellon University, 1996.

[Zaiane, 1998] O. R. Zaiane, J. Han, Z. –N. Li, S.H. Chee, & J.Chiang, “Multimidia data miner: a system prototype for multimedia data miner”. In Proc. ACM SIGMOD Intl. Conf. on Management of Data, pages 581-583, 1998”.

[Zhong, 2002] N. Zhong, J. Liu, Y. Yao, “In Search of the Wisdom Web”. IEEE Computer, vol.35, no.11, 2002, pp.27-31.