Extração de Informações na Web

Mário Henrique A. C. Adaniya[1], Mario Lemes Proença Jr[1]

Resumo: É observado um crescimento exponencial nas informações contidas na Web, e com todo este crescimento, por muitas razões deixamos de agregar valor ao nosso conhecimento ou utilizar informações pelo simples fato de não termos a capacidade de processar demasiado volume. Para tanto, estudos em áreas como Recuperação de Informação e Extração de Informação visam tratar os documentos em si e a informação contida nestes documentos. E a WebMining engloba as duas áreas entre outras, transformando tais informações em algo útil para nós.

1 Introdução

Cada vez mais encontramos toda e qualquer informação que precisamos disponíveis *online*. É uma tendência que grandes editoras com revistas e publicações impressas estão aderindo, mantendo os impressos tradicionais e publicando virtualmente os mesmos conteúdos e adicionando outros exclusivos na edição *online*. A *World Wide Web (Web)* é um meio de comunicação popular e interativo para disseminar informação atualmente [11]. Todo dia, mais e mais páginas são indexadas pelos motores de buscas. Blogs surgem aos milhares com pessoas expressando suas idéias, opiniões e experiências. Sites de relacionamento, fóruns, *Wikis* armazenam conteúdos imensos dos mais diversificados assuntos. E neste pandemônio, como encontrar o que estamos procurando? Como descobrir se a informação recuperada é confiável?

[1] Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 6001 – 86051-990 – Londrina – PR – Brasil
{mhadaniya}@dc.uel.br, {proenca}@uel.br

2 Recuperação de Informação

Recuperação de Informação (RI) é a tarefa de encontrar documentos relevantes a partir de um corpus ou conjunto de textos em resposta a uma necessidade de informação de um usuário [13]. Recuperação de Informação possui limites muito bem delimitados, e qualquer tarefa além de prover ao usuário os documentos, não é um sistema de recuperação de informação.

2.1 Recuperação de Informação na WEB

Muitas características devem ser levadas em conta quando estamos recuperando documentos na WEB [10]:

- **Tamanho da Internet** – O tamanho da Internet, segundo Zhang e seu grupo de pesquisa [14], estima-se que em Janeiro de 2008 a Internet continha 62400000 hostnames ativos. É observada que para a Internet a cada cinco anos ela dobra de tamanho;
- **Dinamismo da Internet** – As técnicas de Recuperação de Informação são geralmente estáticas, enquanto a Web está em constante metamorfose;
- **Duplicação** - 30% do conteúdo da Internet é uma cópia de algum conteúdo existente;
- **Comportamentos específicos** – É estimado que 85% dos usuários utilizam apenas a primeira página retornada das *search engines*, e 28% modificam sua consulta original;
- **Múltiplos tipos de usuário** – Possui muitos tipos de usuários e cada usuário utiliza a Internet para uma tarefa específica;
- **Idiomas** – Como a *Internet* se tornou algo mundial, as páginas são encontradas em mais de 100 idiomas;
- **Alta Linkagem (High Linkage)** – Cada página contém aproximadamente oito links para outras páginas;

Com estas características, podemos ter uma noção da dificuldade do campo de Recuperação de Informação na Web. E muitas vezes, o usuário não sabe expressar sua necessidade, tornando a tarefa muito mais penosa.

3 Extração de Informação

O contraste entre os objetivos dos sistemas de Extração da Informação e Recuperação de Informação podem ser descritos como seguem: Recuperação de Informação recupera

documentos relevantes de uma coleção, enquanto Extração de Informação extrai informações relevantes de documentos. Consequentemente, as duas técnicas se complementam, e usadas em combinação podem prover uma ferramenta poderosa [7].

3.1 Abordagens

Na Extração de Informação, observamos claramente a distinção de duas abordagens [2]: *Knowledge Engineering* e *Automatic Training*.

Em *Knowledge Engineering* o sistema é praticamente construído manualmente *pelo knowledge engineer*[2]. Sua construção se baseia no conhecimento que o engenheiro possui do cenário e domínio com o qual vai se trabalhar.

A abordagem de *automatic training* não necessita de um especialista, mas alguém que tenha o conhecimento suficiente do domínio da aplicação. Uma vez que os documentos de *corpus* foram anotados, um algoritmo de treino é executado, treinando o sistema para novos textos. Utilizam métodos estatísticos, e aprendem regras com a interação com o usuário.

Nenhuma das duas abordagens é superior a outra, pois a extração depende de muitas variáveis, e muitas vezes, variáveis externas, logo, não podemos apontar nenhuma abordagem como completa. Ambas utilizadas em conjunto caminha para um sistema ideal.

3.2 Tipos de Dado

A Extração ocorre em documentos, e eles são categorizados em três tipos [7]:

I. ***Documentos livre/sem estruturação***: Texto livre é basicamente o texto onde não encontramos nenhuma forma de estrutura, e é o tipo mais encontrado. Originalmente o objetivo de Extração de Informação era desenvolver sistemas capazes de extrair informações chaves de textos em linguagem natural.

II. ***Documentos semi-estruturados***: Não são textos totalmente livres de estrutura, mas também as estrutura existente não é tão severa, os textos semi-estruturados encontram-se no intermédio.

O pesquisador Sergel Abiteboul diferencia dentro do contexto de semiestruturados, em cinco categorias [1][5]: **Estrutura Irregular,Estrutura Implícita, Estrutura Parcial, Estrutura Indicativa e Estrutura Flexível**.

III. ***Documentos estruturados*** : Informações textuais contidas em banco de dados ou qualquer outro gênero de documento com uma estruturação rígida,

[2] É a pessoa mais familiarizada com o sistema de Extração de Informação, e conhece melhor o formalismo para expressar as regras para o sistema.

são a base de textos estruturados. Como seguem uma moldura sem grandes diferenças de um documento para outro, sua informação é facilmente extraída.

3.3 Avaliação

Os critérios de avaliação consistem em: quanta informação foi extraída (*recall*), quanto da informação extraída é correta (*precision*) e quanto da informação extraída é supérflua (*overgeneration*) [12]. Quando a *Cobertura* aumenta, a Precisão tende a diminuir e vice-versa, pois são inversamente proporcionais. *Precisão* e *Cobertura* estão sempre no intervalo de [0; 1], sendo 0 o pior resultado e 1 o melhor.

A *F-measure* mede considerando a precisão e a cobertura. O parâmetro β quantifica a preferência da cobertura sobre a precisão. Geralmente utilizamos β = 1, balanceando assim as duas medidas.

$$F - measure = \frac{(\beta^2 + 1) * Cobertura * Precisão}{\beta^2 * (Cobertura + Precisão)} \qquad (1)$$

Para esclarecer um pouco mais o conceito de Precisão e Cobertura, tomamos um total de 16 termos extraídos. Desses 16 termos, apenas 4 são nomes corretos e esperávamos no total 8 nomes, então nossa Precisão é de 50%. Resultando em uma precisão média. A Cobertura são os nomes extraídos corretamente sobre o total de termos que extraímos, resultando em apenas 25%. Isso significa que de toda informação extraída, apenas 25% é relevante para o domínio do sistema.

4 Web Mining

Web Mining é o uso das técnicas de Mineração de Dados para descobrir e extrair automaticamente a informação de documentos na Web [8]. A Mineração de Dados refere-se ao processo não trivial de identificação de padrões válidos, previamente desconhecidos e potencialmente úteis de dados [9]. Seguindo o conceito de Etzione, que utiliza da Descoberta do Conhecimento (*Knowledge Discovery Database*) como base, ele decompõe a Web Mining em quatro tarefas: *Resource finding* (Coleta de Documentos), *Information selection and pre-processing* (Pré-processamento), *Generalization* (Extração de Padrões) e *Analysis* (Análise).

4.1 Categorias de WEB Mining

Com o crescimento exponencial das fontes de informação disponíveis na Web ao nosso redor, cresce a necessidade de automatizar ferramentas que busquem as informações desejadas e corretamente. Ferramentas mais eficazes no rastreamento, tanto do lado dos

servidores como dos clientes, são comumente alvos de pesquisas e projetos na busca de uma mineração de dados. Do lado dos servidores, temos extensas listas de logs, registros de usuários ou perfil de usuário, entre outros itens que podem ser analisados [4].

4.1.1 Mineração de Conteúdo

A Mineração de Conteúdo e a Recuperação de Informação são muitas vezes utilizadas em conjunto. Enquanto uma realiza a mineração diretamente do conteúdo dos documentos a outra incrementa o poder de busca de outras ferramentas e serviços. Áudio, vídeo, dados simbólicos, metadados e vínculos de hipertexto fazem parte do conteúdo de documentos da Web atualmente, e como tal, na mineração de conteúdos também são analisados. Existem áreas de pesquisas destinadas a mineração de dados multimídias, entretanto, como uma enorme parte da Web é constituída de texto e hipertexto, permanecendo assim o foco em dados de texto.

Com o continuo crescimento da Web, as pesquisas voltadas para ferramentas mais eficazes, melhorias nas técnicas de mineração e extração de dados se desenvolveram. Podemos observar duas grandes abordagens quando tratamos de Mineração de Conteúdo: Baseado em Agente (*Agent-Based*) e Banco de Dados (*Database*).

Baseado em Agente (*Agent-Based*): Esta abordagem de mineração de dados trabalha diretamente com o campo de Inteligência Artificial, provendo um sistema autônomo ou semi-autônomo, que trabalha para a coleta de conhecimento e organização das informações na WEB delimitado pelo escopo do sistema.

Banco de Dados (*Database*): A abordagem de Banco de Dados, como o nome pressupõem, trabalha com a organização e integração dos documentos semi-estruturados para um documento estruturado, como em um banco de dados relacional, usando inclusive consultas e mecanismos de banco de dados para acesso e analise das informações. A área de mineração de textos está bem esclarecida, com muitas técnicas, uma das quais seria reestruturar o documento para uma linguagem entendida pela maquina. Uma mineração que vem ganhando destaque em pesquisas é a mineração em serviços da Web tais como grupo de noticias, grupos de e-mails, lista de discussão. Outro conceito é introduzido por estes pesquisadores, chamado de Web Intelligence, que promete transformar os serviços da Web em entidades inteligentes, de forma que elas possam interagir e se comunicar através de uma linguagem comum.

4.1.2 Mineração de Estrutura

Como o próprio nome descreve, nesta categoria de mineração estamos preocupados com a estrutura dos documentos Web e como estes estão ligados entre si. Os vínculos de ligação de hipertexto são os principais objetos de estudos nesta categoria. Podemos visualizar a Web como um grafo orientado, onde os nós representam páginas e as setas entre os pares de nós representam os vínculos entre as páginas. Como ocorre em citações bibliográficas quando um artigo é bastante citado indicando que provavelmente este artigo tem um peso

importante perante outros que abordam o mesmo tema, o mesmo pode ser observado entre os documentos Web. Podemos drasticamente comparar que se uma pagina contém muitas setas entrando, ela teria certa relevância quanto ao seu conteúdo ser confiável.

4.1.3 Mineração de Uso

A mineração de uso utiliza os dados secundários provindos de logs de servidores, 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. As aplicações da mineração de dados de uso são classificadas em duas categorias: aprendizado de perfil de usuário (modelagem em interfaces adaptativas) e aprendizado de padrões de navegação de usuário. Talvez umas das técnicas em mais utilização atualmente, devido ao grande número de E-Commerce, pois com isto podemos adaptar sites de acordo com o cliente, recomendar produtos de acordo com compras passadas ou baseadas nas similaridades entre perfis de usuários.

4.1.4 Web Semântica

A Web Semântica é uma extensão da web já existente, onde a informação ganha melhores significados, proporcionando aos humanos trabalhar melhor em conjunto com os computadores [3]. Acredita-se muito que Web Semântica será o próximo passo evolutivo da Web, pois possui uma linguagem semântica muito rica, e.g., *Web Ontology Language*[3].

Como somos expostos a muitas informações de diversas maneiras, não sabemos lidar com o que exatamente é correto ou útil para nós, resultando em uma "sobrecarga de informação". Observamos duas características importantes para este fenômeno: demasiado volume e a falta de uma definição semântica interpretável por programas e sistemas [6].

Algumas áreas estudadas na Inteligência Artificial casaram muito bem, pelo fato de serem mecanismos que captam a semântica do conteúdo e se ajustam de acordo com as necessidades. Uma das abordagens propostas era de dotar a Internet de inteligência própria, construindo páginas mais elaboradas e ricas semanticamente e onde agentes pudessem raciocinar sobre semântica, logo, modelando uma Web Semântica.

A semântica é obtida através de ontologias, que são modelos de dados representando o conhecimento adquirido sobre um mundo ou parte deste em um conjunto de conceitos existentes em um domínio e os relacionamentos entre estes. As ontologias descrevem geralmente: indivíduos, classes, atributos e relacionamentos.

Muitos problemas são enfrentados nesta área de estudo, mas grandes avanços são observados. Como fazer a ontologia chegar ao usuário comum sem ser tão complicada, como assegurar que o conteúdo será sempre preciso e claro, padrões ontológicos, entre outros são as discussões que direcionam as pesquisas na área.

[3] Web Ontology Language - http://www.w3.org/TR/owl-feature

5 Referências Bibliográficas

[1] ABITEBOUL, S. Querying semi-structured data. INTERNATIONAL CONFERENCE ON DATABASE THEORY (1997).

[2] APPELT, D. E., AND ISRAEL, D. J. Introduction to information extraction technology. Tutorial for IJCAI-99 (1999).

[3] BERNERS-LEE, T., HENDLER, J., AND LASSILA, O. The semantic web. Scientific American 279, 5 (2001), 34–43.

[4] COOLEY, R., MOBASHER, B., AND SRIVASTAVA, J. Web mining: Information and pattern discovery on the world wide web:, 1997.

[5] DA SILVEIRA, I. C. Extração semântica de dados semi-estruturados através de exemplos e ferramentas visuais. Master's thesis, Universidade Federal do Rio Grande do Sul, 2001.

[6] DE FREITAS, F. L. G. Anais do XXIII Congresso da Sociedade Brasileira de Computação., vol. Volume 8: Jornada de Mini-Cursos em Inteligência Artificial. SBC, 2003, ch. Ontologias e a Web Semântica, pp. 1–52.

[7] EIKVIL, L. Information extraction from world wide web - a survey.

[8] ETZIONE, O. The world wide web: quagmire or gold mine? Communications of the ACM 39 (1996), 65–68.

[9] FRAWLEY, W. J., PIATETSKY-SHAPIRO, G., AND MATHEUS, C. J. Knowledge discovery in databases: An overview. AI Magazine 13 (1992), 57–70.

[10] HUANG, L. A survey on web information retrieval technologies, 2000.

[11] KOSALA, R., AND BLOCKEEL, H. Web mining research: A survey. SIGKDD Explorations 2 (2000), 1–15.

[12] LEHNERT, W., AND SUNDHEIM, B. A performance evaluation of text-analysis technologies. AI Magazine 12 (1991), 81–94.

[13] SMEATON, A. Information retrieval: Still butting heads with natural language processing? Information Technology, M.T Pazienza ed., Springer-Verlag Lecture Notes in Computer Science 1299 (1997), 115–138.

[14] ZHANG, G.-Q., ZHANG, G.-Q., YANG, Q.-F., CHANG, S.-Q., AND ZHOU, T. Evolution of the internet and its core. New Journal of Physics 10 (December 2008), 1–11.