Mineração de Textos na Web

Luciano Antonio Digiampietri
Escola de Artes Ciências e Humanidades da Universidade de São Paulo
digiampietri@usp.br

Resumo: *Com o crescimento das informações disponíveis na Web e de ferramentas de trocas de mensagem digitais, blogs e wikis surge a necessidade do desenvolvimento de soluções automáticas para auxiliar os usuários a entender e organizar esse conteúdo. Este projeto pretende auxiliar nesse processo com o uso da Mineração de Textos. Enquanto na Mineração de Dados tradicional os dados estão pré-organizados (por exemplo, tabulados em uma planilha) e o processo de mineração (descoberta de conhecimento) só precisa tratar esses dados (por exemplo, com o uso de técnicas estatísticas), na Mineração de Textos, os dados encontram-se na forma de textos livres (escritos em língua natural) e, desta forma, existem maiores desafios para entender e extrair conhecimento desses dados. Como contribuições do desenvolvimento deste trabalho temos: (i) o estudo detalhado de técnicas de pré-processamento de texto e mineração de textos; (ii) a produção e disponibilização de ferramentas para auxiliar em cada uma das etapas do processo de Descoberta de Conhecimento; e (iii) desenvolvimento de ferramenta para visualização dos dados seguindo a organização/estruturação resultante da Mineração de Dados.*

1. Vínculo com o Programa Ensinar com Pesquisa

Este projeto contribuirá com o aprofundamento do entendimento do estudante nos seguintes assuntos ligados principalmente à área Inteligência Artificial: descoberta de conhecimento, recuperação de informação, mineração de textos, clusterização de dados (*data clustering*), classificação de dados e descoberta de associações. Estes assuntos são muito importantes tanto na indústria quanto na academia e tem ganhado destaque nos últimos anos devido ao crescimento na produção e disponibilização de dados na Web. Além de familiarizar o estudante com os conceitos gerais de mineração de texto, este projeto possibilitará um estudo teórico e prático de diversas técnicas de descoberta de conhecimento aplicadas em textos escritos em língua natural (especificamente em português). Como consequência, o aluno terminará este projeto apto a trabalhar ou pesquisar na área de mineração de texto (ou, de maneira mais abrangente, na área de descoberta de conhecimento). Além disso, o estudo realizado neste projeto poderá auxiliar o estudante na disciplina de Inteligência Artificial.

2. Introdução

Com o crescimento das informações disponíveis na Web e de ferramentas de trocas de mensagem digitais, além de blogs e wikis surge a necessidade do desenvolvimento de soluções automáticas para auxiliar os usuários a entender e organizar esse conteúdo.

A Mineração de Textos difere da Mineração de Dados, pois, nesta última, os dados estão pré-organizados (por exemplo, tabulados em uma planilha) e o processo de mineração (descoberta de conhecimento) só precisa tratar esses dados (por exemplo, com o uso de técnicas estatísticas). Já em Mineração de Textos, os dados encontram-se na forma de textos livres (escritos em língua natural) e, desta forma, existem maiores desafios para entender e extrair conhecimento desses dados.

Considerando sistemas web, sempre que um usuário tiver uma pergunta (ou desejar fazer uma busca sobre um dado conteúdo) são necessários sistemas cada vez mais inteligentes para responder a essa pergunta. Existem ao menos duas modalidades desses sistemas: sistemas de busca, os quais retornam um grande número de sítios Web e documentos relacionados a uma dada palavra ou expressão chave, e sistemas de resposta automática a questões (QAS - Question-Answering Systems) que retornam um ou mais resultados exatos a uma pergunta do usuário [1].

Responder automaticamente é a tarefa na qual um sistema de software responde a perguntas arbitrárias formuladas em língua natural. QAS são especialmente úteis quando um usuário necessita de informação específica e não quer, ou não tem tempo, para ler todos os documentos e sítios web relacionados com a pergunta [2]. Nestes casos, sistemas de questões frequentemente respondidas (FAQ - Frequent Asked Questions) são ferramentas extremamente úteis para aumentar a eficiência na resposta a perguntas específicas.

Uma direção para a solução para este tipo de sistemas de resposta automática é o trabalho existente em Recuperação de Informação (IR – Information Retrieve) que utilizam medidas precisas de similaridade para identificar possíveis relações entre novas perguntas e perguntas previamente respondidas. Entretanto, medidas de similaridade, que funcionam relativamente bem na comparação entre documentos, não costumam funcionar bem na comparação entre questões, que são textos relativamente pequenos; e, medidas tradicionais de similaridade para sentenças não funcionam bem quando há pouca repetição de palavras entre sentenças [3].

Por outro lado, sistemas de resposta automática podem se beneficiar de técnicas de descoberta de conhecimento (KDD - Knowledge Discovery Database) para agrupar, classificar ou encontrar associações entre perguntas e respostas ou mesmo para agrupar e organizar informações sobre um dado domínio. Além disso, se estivermos trabalhando em um domínio bem definido, é possível utilizar o conhecimento sobre o domínio (descrito na forma de uma ontologia de domínio, por exemplo) para aumentar a precisão do sistema.

Este projeto apresenta os primeiros passos para o desenvolvimento de um sistema de resposta automática. Nele, serão exploradas técnicas de Recuperação de Informação e Descoberta de Conhecimento para possibilitar a organização de informações disponíveis na web com o objetivo de desenvolver sistemas de busca mais inteligentes e sistemas de resposta automática a perguntas.

3. Sistema Proposto

A Figura 1 apresenta a arquitetura do sistema proposto neste projeto. Cada atividade deste sistema será discutida a seguir. Cada uma dessas atividades deverá ser implementada para possibilitar a mineração de textos proposta neste projeto.

Este sistema está contextualizado dentro de um projeto de pesquisa que está sendo desenvolvido pelo Grupo de Inteligência Artificial da EACH[1].

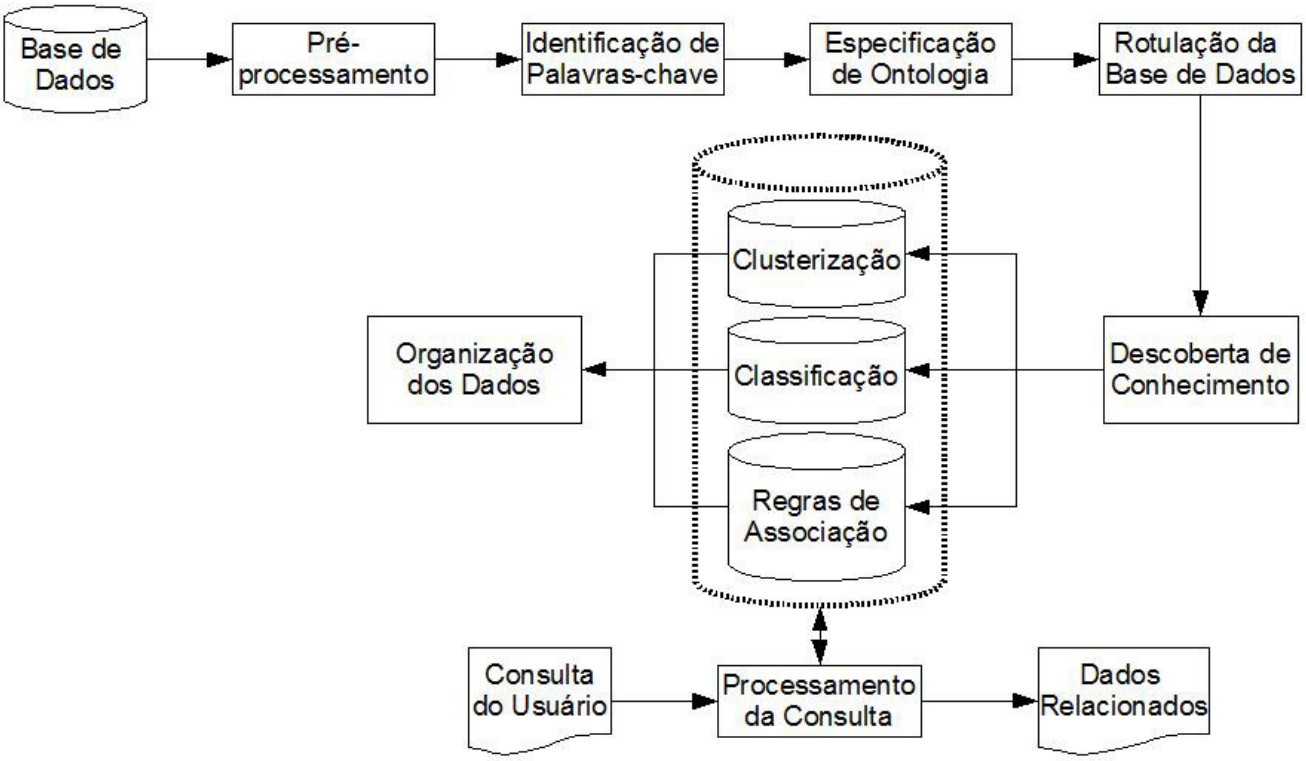

Figura 1 – Arquitetura Proposta

Pré-processamento: esta atividade executa diversos procedimentos sobre a base de dados, incluindo: a descoberta da raiz de cada palavra, identificação de sinônimos e remoção de palavras indesejadas (como artigos, por exemplo).

Identificação de Palavras-chave (e expressões-chave): esta atividade visa identificar as palavras ou expressões mais importantes no domínio da base de dados. Para isto, tipicamente é feita a combinação de duas estratégias: (i) identificação manual feita por um especialista no domínio e (ii) identificação de palavras relevantes a partir da comparação da frequência que essas palavras ocorrem na base de dados em relação a frequência esperada dessas palavras em um *corpus* maior de textos em língua portuguesa.

Especificação de Ontologia: esta atividade consiste na especificação de uma ontologia de domínio que será utilizada para a descrição dos conceitos e relacionamentos entre os conceitos presentes na base de dados. Embora existam diversas ferramentas para apoiar a especificação de ontologias, é fundamental a presença de um especialista de domínio para assegurar a qualidade da ontologia produzida.

[1] http://dgp.cnpq.br/buscaoperacional/detalhegrupo.jsp?grupo=0067103WZ9GQ1B

Rotulação da Base de Dados: esta atividade consiste na rotulação de cada registro da base de dados como pertencente a uma ou mais classes (estas classes podem ser, por exemplo, conceitos da ontologia). Esta rotulação inicial é feita manualmente. É importante notar que esta classificação manual, bem como a especificação da ontologia, não são atividades obrigatórias para um sistema de mineração de texto / resposta automática de perguntas, porém estas atividades são necessárias para a construção de um sistema robusto que utilize técnicas de aprendizado supervisionado e baseadas em semântica (que é a intenção deste projeto).

Descoberta de Conhecimento: esta atividade é responsável pela aquisição de conhecimento a partir da base de dados. Ela inclui a identificação de grupos de informações correlacionadas; classificação de novas consultas feitas pelo usuário e descoberta de associações entre os registros da base de dados.

Organização da Informação: esta atividade usa os resultados da Descoberta de Conhecimento para organizar informações de acordo com sua semântica, bem como para apresentá-la ao usuário, por exemplo, de uma maneira hierárquica ou sumarizada.

Processamento de Consultas: esta atividade recebe uma consulta do usuário escrita em linguagem natural, automaticamente processa essa consulta e apresenta para o usuário as informações relacionadas com a consulta feita.

4. Metodologia e Cronograma

Para o desenvolvimento deste projeto, existe um processo bem claro composto das seguintes atividades:

(i) Estudo das soluções existentes para cada uma das atividades da Figura 1;

(ii) Definição da base de dados a ser utilizada e aquisição dessa base de dados;

(iii) Desenvolvimento das atividades: *Pré-processamento* e *Identificação de Palavras-chave* utilizando os conhecimentos obtidos na etapa de estudo / revisão bibliográfica;

(iv) *Especificação de Ontologia* (caso não exista para o domínio da base de dados) e a *Rotulação da Base de Dados*;

(v) Estudo e desenvolvimento de técnicas de *Descoberta de Conhecimento* para o processamento da base de dados (incluindo técnicas de clusterização, classificação e descoberta de regras de associação);

(vi) Avaliação dos resultados obtidos: (a) de maneira automática, comparando a classificação gerada pelo sistema com a rotulação feita manualmente; e (b) de maneira manual verificando-se a semântica e a relevância dos *clusters* gerados e das regras de associação descobertas.

O cronograma a seguir relaciona as atividades listadas na metodologia com o prazo de execução dos projetos Ensinar com Pesquisa.

Tabela 1 – Cronograma da execução das atividades

	Atividade i	Atividade ii	Atividade iii	Atividade iv	Atividade v	Atividade vi
03/2010						
04/2010						
05/2010						
06/2010						
07/2010						
08/2010						
09/2010						
10/2010						
11/2010						
12/2010						
01/2011						
02/2011						

5. Conclusões

Este projeto pretende lidar com o problema de mineração de textos na Web. Este problema é complexo devido à flexibilidade da língua natural (na qual os textos são escritos) e, por outro lado, é um problema extremamente relevante à medida que mais informação é disponibilizada na Web e deseja-se extrair conhecimento dessa informação.

A solução almejada pretende ser genérica, podendo ser adaptada para diversos domínios de aplicação (bastando para isso a troca da base de dados e de sua rotulação, e o uso de uma ontologia compatível com essa base de dados). Além disso, o desenvolvimento deste projeto pretende aperfeiçoar os conhecimentos do aluno em todas as etapas envolvidas no processo de Descoberta de Conhecimento.

6. Referências

1. Song, W., Feng, M., Gu, N., Wenyin, L.: Question similarity calculation for faq answering. In: SKG '07: Proceedings of the Third International Conference on Semantics, Knowledge and Grid, Washington, DC, USA, IEEE Computer Society (2007) 298–301
2. Mollá, D., Vicedo, J.L.: Question answering in restricted domains: An overview. Comput. Linguist. 33(1) (2007) 41–61
3. Jeon, J., Croft, W.B., Lee, J.H.: Finding semantically similar questions based on their answers. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (2005) 617–618
4. Holsapple, C.W., Joshi, K.D.: A collaborative approach to ontology design. Commun. ACM 45(2) (2002) 42–47