6
Exemplos de Aplicações de Mineração de Textos

Nas subseções seguintes serão apresentadas três aplicações de nosso modelo para mineração de textos: Classificação automática, Extração de Informação e Interface em Linguagem Natural (ILN). Após estas aplicações, este capítulo mostra um pouco sobre a importância do pré-processamento de textos para o campo da WebSemântica.

6.1.1.
Classificação

O processo de classificação automática de um texto passa pelo processo de classificação dos lexemas (entidades e objetos reconhecidos) que o compõem. Cada lexema identificado incrementa a capacidade de discriminação semântica do classificador. Isto é, se soubermos, no limite, dizer o significado de cada objeto lingüístico no texto teremos informação mais precisa para a classificação de um ou mais textos (hipótese base de nossa metodologia).

Definir o significado de uma dada palavra é ainda uma questão filosófica muito complicada. No entanto, nos inspiram em um dos pensamentos de (Wittgenstein, L., 1979): "o importante não é saber o significado da palavra e sim jogar o jogo da linguagem". Fazemos uma ressalva de que para Wittgenstein o significado só existe no momento do uso, e desse modo, qualquer sistema computacional para uma dada linguagem seria inviável.

Para exemplificar essa aplicação, utilizamos o classificador Bayesiano (seção 2.1.3). Primeiramente o pré-processamento extrai os lexemas de todos os documentos. Em seguida, através da freqüência relativa, apenas os lexemas significantes são selecionados. Para cada lexema selecionado, um peso é associado de acordo com o cálculo da freqüência relativa.

Para contornar o problema do significado, o modelo de classificação utilizado contempla uma abordagem estatística para modelar os diversos sentidos de uma palavra. Isto é, para cada palavra existe uma distribuição de probabilidade que varia conforme a usabilidade da palavra. No Anexo I existem algumas

distribuições de probabilidade importantes que nos ajudam a visualizar esse modelo estatístico.

A cada termo novo, pré-processado e adicionado ao léxico aumentamos a potência de classificação dos documentos.

Na fase de treinamento, o algoritmo estima a probabilidade de o documento pertencer a cada categoria. Esse processo de treinamento se dará com um conjunto de textos alvo pré-classificados e uma base de documentos representativa do domínio em questão.

O conjunto de lexemas selecionado será chamado de "conjunto semântico" do contexto. Sendo assim, dado um texto, este pertencerá à categoria que contiver uma soma ponderada suficiente de lexemas do "conjunto semântico".

Esse procedimento não é o mesmo que a aplicação de filtros por palavra-chave, visto que a semântica de um contexto é baseada em um conjunto de objetos. Sua intensidade é baseada exatamente na associação destes objetos. Sob o ponto de vista estatístico, essa associação é justamente a estimativa da matriz de covariância de todos os termos em todos os documentos. A ocorrência de apenas um lexema não torna uma categoria significante. Isso implica que um lexema sozinho não carrega consigo nenhuma semântica. Dessa forma, cada um deles pode ser retirado individualmente sem alterar a semântica do conjunto. De outra maneira, podemos dizer que não é a totalidade do conjunto que caracteriza a semântica do contexto.

6.1.2. Extração de Informações

A extração de informações (IE) pode ser entendida como desde o reconhecimento e classificação de entidades mencionadas até a associação entre eles segundo o conteúdo do texto (Pérez, C. e Vieira, R., 2003).

Após o pré-processamento visto durante essa tese, o texto é etiquetado, o que permite a implementação de algoritmos estatísticos para determinar os relacionamentos significantes entre os termos extraídos. Tem-se utilizado também modelos simbólicos para determinar esses relacionamentos. Diversos trabalhos sobre a extração automática de relações em um corpus vêm sendo desenvolvidos com ênfase especial nas relações de hierarquia semântica (ex. "Banana é uma

fruta" ou "Caule é uma parte da planta") (Snow, R. et al, 2004); (Widdows, D., 2003); (Wu, S. et al, 2003).

Para reduzir a exploração das quase infinitas relações que podemos ter em uma base textual, alguns padrões sintáticos pré-definidos são utilizados para a extração de informações (Hearst, M. A., 1992).

A extração de informações trata de processos que seletivamente estruturam e combinam dados encontrados – de forma implícita ou explícita - nos textos (Cowie, J. e Wilks, Y., 2000); (Grishman, R., 1997). Cada processo é especificado de acordo com o objetivo pretendido de mineração que pode ser, por exemplo, associar bandas de rock com suas respectivas gravadoras. Tais processos podem, por um lado, resultar em algum tipo de base de dados e, por outro, se apoiar em bases de dados, como ontologias e taxonomias semânticas, para atingir os objetivos pretendidos (Phillips, W. e Riloff, E., 2002); (Snow, R. et al, 2004).

Dentre as bases de dados lexicais/semânticas mais populares encontram-se a WordNet (Fellbaum, C., 1998) e Cyc (Lenat, D., 1995) – uma base manualmente alimentada com esquema da representação do senso comum, elas contêm sinônimos, hiperônimos e definições de palavras da lingua inglesa. Porém, tais bases, embora sejam extremamente valiosas, tanto para consulta quanto para sua utilização por aplicações de mineração de textos, e ofereçam enormes vantagens na execução das tarefas de treinamento de algoritmos de aprendizado, também apresentam algumas limitações.

A primeira delas diz respeito à sua elaboração, que é feita manualmente, o que significa um trabalho lento e árduo. A conseqüência desse caráter manual é a dificuldade para atualização e extensão, que também dependeriam de trabalho manual. Tais limitações têm levado pesquisadores a se interessar por métodos mais autométicos de extração de informações (Caraballo, S. A., 2001); (Cederberg, S. e Widdows, D., 2003); (Girju, R. et al, 2003); (Snow, R. et al, 2004); (Widdows, D., 2003).

Além disso, para (Freitas, M. C., 2004), uma fraqueza de especial importância para a EI é a quantidade insuficiente de nomes próprios que constam na base. O fato de tais nomes constituírem uma classe ainda mais "aberta" que a dos substantivos comuns, uma vez que novos nomes, principalmente de empresas, podem ser criados a qualquer momento, deixa ainda mais evidente a necessidade

de atualização constante e, conseqüentemente, de metodologias capazes de extrair informações automaticamente.

Ainda em (Freitas, M. C., 2004), na recuperação de informações a importância do reconhecimento de nomes próprios aparece, por exemplo, quando o usuário busca por documentos que se refiram a uma entidade cujo nome é apenas parcialmente conhecido, como Gutierrez, ou que se refiram a uma determinada classificação semântica, como bancos ou cervejarias. Para a área de Extração de Informações, também pode ser útil ao usuário uma classificação de CityBank como um banco americano que, por sua vez, integra uma categoria superior "empresa".

Só para exemplificar, uma ontologia (seção 4.1.4) é um modelo relacional do mundo que se quer trabalhar. Em um hospital teríamos pessoas, lugares e horários. Dentro de pessoas podemos ter médico, enfermeira, paciente etc. Dentro de lugares podemos ter quarto, recepção, estacionamento etc.

Os padrões seriam formas sintáticas de se extrair o conhecimento dessa ontologia. Exemplo, "o médico Antônio Lopes atendeu a paciente Mariana Bastos": Antônio Lopes seria classificado como pessoa/médico, Mariana Bastos como pessoa/paciente e Mariana como paciente de Antônio.

6.1.3. Interface em Linguagem Natural

Usamos como exemplo de aplicação de Interface em Linguagem Natural (ILN), a integração do pré-processamento proposto com o sistema desenvolvido em (Nascimento, M. R., 1993).

Segundo (Nascimento, M. R., 1993), a aplicação de ILN já conta com vários trabalhos, a exemplo, para a língua inglesa, dos sistemas Lunar (Woods, 1972); Lifer (Hendrix, 1978); Planes (Waltz, 1978); Intellect (Harris, 1977); TQA (Demerau, 1985); Eufid (Templenton & Burger, 1983). E para a língua portuguesa (Coelho, 1981), (Assis, 1986), (Branco, 1987), (Árabe, 1989), (Martins, 1990).

A aplicação da dissertação de (Nascimento, M. R., 1993) foi de um Assistente Inteligente. Para isso, estudou e implementou uma funcionalidade de Respostas Automáticas a perguntas em linguagem natural. O trabalho de (Muradas, A., 1995), depois, ampliou o sistema de (Nascimento, M. R., 1993)

com o conceito de PLN e desenvolveu um sistema interpretador de perguntas capaz de converter uma consulta em linguagem natural para a linguagem de consulta SQL segundo um banco de dados relacional pré-especificado.

Segundo (Muradas, A., 1995), um sistema de respostas automáticas é constituído por uma sequência de perguntas, e respostas, que atendem às perguntas realizadas pelo usuário. A perguntas são feitas em linguagem natural e podem se apresentar de diversas formas. Através dos estudos dos trabalhos de (Amori, D. R., 1990); (Muradas, A., 1995); (Nascimento, M. R., 1993) selecionamos três tipos principais de interrogativas:

- **Interrogativa Fundamental**: apresenta um ponto de interrogação no final e requer apenas um SIM ou NÃO como resposta. (ex. "O João comeu todas as maçãs?")
- **Interrogativa Descritiva**: apresenta um ponto de interrogação no final e requer como resposta uma descrição ou uma lista. (ex. "Quantas maçãs João comeu?"; "Quais frutas o João comeu?")
- **Comando**: não contém ponto de interrogação, mas são precedidas por uma expressão pré-locutória. (ex. "Informe-me se João comeu todas as maçãs." ou "Gostaria de saber se João comeu todas as maçãs.")

Em (Muradas, A., 1995) e (Barros, F. A. e DeRoeck, A., 1994) vemos a importância de um bom pré-processamento. Esses trabalhos mostram que para converter as perguntas em consultas SQL é necessário resolver tarefas de PLN como anáforas e elipses.

Uma anáfora é a referência a uma entidade previamente definida. Essa referência pode se dar através de pronomes ou classes ontológicas hierarquicamente superiores, exemplo:

```
- Qual o título do artigo de Alex Garcia? (Entidade, seção 4.2.7)
- Qual o endereço dele? (Anáfora pronominal, seção 4.2.10.3)

- Qual o tamanho de um Beija-Flor? (Entidade, seção 4.2.7)
- Esse pássaro tem algum predador? (Ontologia, seção 4.1.4)
```

E resolução de elipses através do modelo de pré-processamento proposto se dá pela substituição da entidade mencionada na consulta SQL, exemplo:

```
- Quais os pesquisadores que moram no Grajaú ? (Entidade “Grajaú”)
- e na Tijuca? (Entidade “Tijuca”)
```

Todas essas técnicas funcionam muito bem em um ambiente controlado. Porém, quando deixamos o usuário livre para escrever um texto, normalmente ocorrem imprecisões de escrita. É fácil verificar isso em uma caixa de email. O conceito de imprecisão da informação escrita foi bem descrito por (Owei, V., 2002). Para essa aplicação é muito importante a utilização de um corretor ortográfico automático como visto na seção 4.2.10.5.

Após o pré-processamento das perguntas, estas serão convertida para a consulta SQL que deve seguir rigorosamente o banco de dados estruturado. Essa estrutura pode ser feita utilizando os modelos pré-definidos na seção 6.1.2 de Extração de Informações. Essa mesma técnica pode ser usada para inserir as informações da base textual nas tabelas.

Um exemplo do funcionamento de um implementação desse sistema em Java pode ser encontrado no Anexo II. As perguntas são livres e as respostas baseadas na ontologia pré-definida. Nesse exemplo, primeiro classificamos as entidades segundo a ontologia de locais, pessoas e empresas, depois extraímos a informação sob o modelo “Argentina é um país da América Latina” e alimentamos um banco estruturado. Finalmente convertemos a pergunta livre em uma consulta SQL para fornecer a resposta ao usuário.

6.2. Web Semântica

A Web Semântica atualmente se insere nas principais linhas de investigação que norteiam os desdobramentos da Internet atual. Tem por principal objetivo fornecer estruturas e dar significado semântico ao conteúdo das páginas da Internet, favorecendo um ambiente onde agentes de software e usuários possam trabalhar de forma cooperativa.

Neste novo contexto, a Internet será capaz de representar associações entre informações que, em princípio, poderiam não estar relacionadas. Para isso, são

necessários conjuntos estruturados de informações (dados e metadados) e um conjunto de regras de inferência que ajudem no processo de dedução automática. Estas regras são explicitadas através de um tratamento terminológico, que permite representar explicitamente a semântica/significado dos dados.

A gestão de conteúdos na Internet envolve uma série de procedimentos de uniformização e padronização de informações, baseado numa visão integradora, reunindo informações de setores diferentes a partir de princípios comuns que possibilitam a interoperabilidade dos diversos sistemas informacionais de uma instituição ou de várias.

No âmbito do conhecimento de gestão de conteúdo, dois conceitos são de fundamental importância: o de organização e o de comunicação. O conceito de organização pressupõe procedimentos classificatórios (seção 6.1.1). Tais procedimentos possibilitam o agrupamento e a recuperação de informações de um órgão ou mais de um. Estas informações podem estar em forma estruturada (bancos de dados, por exemplo) e não estruturadas (textos integrais, por exemplo). Esses procedimentos classificatórios, além de evidenciarem os contornos de atuação de um órgão, facilitam os processos de seleção e de tratamento de informações. O produto deste processo classificatório se apresenta como um mapa de conteúdos das atividades das organizações que as produzem.

Entretanto, para que este mapa de conteúdos possa verdadeiramente funcionar como um mecanismo que permita não somente a socialização, mas também a integração das informações, é necessário que exista interoperabilidade entre os diversos sistemas de uma instituição. Isto significa promover a capacidade dos sistemas potencializarem oportunidades de intercâmbio e re-utilização de informações, interna ou externamente.

Neste contexto, a comunicação deve ser entendida como uma série de procedimentos que permitem a transmissão de conteúdos informativos, a partir de uma visão integrada desses conteúdos. Insere-se, neste domínio, a importância de ações como definição de metadados (seção 4.2.5) e construção de terminologias padronizadas, além da possibilidade de visualização de processos inter-institucionais. Estas ações viabilizam o tratamento e a recuperação das informações e utilização de outros recursos oferecidos no domínio de diversas instituições.

Por outro lado, é necessário também um meio que viabilize a comunicação não somente entre os sistemas, mas entre o sistema e o usuário deste sistema. Este meio é uma linguagem padronizada, também denominado de terminologia, e mais recentemente, no âmbito da Web Semântica, de ontologias (seção 4.1.4).

Porém, depois de todas essas definições para o funcionamento da Web Semântica, esse modelo esbarra com a principal dificuldade que é a manutenção da coerência dos conteúdos com os metadados. Primeiramente a da inserção dos metadados nas páginas e depois a permanente verificação de coerência futura.

Nesse sentido, os procedimentos automáticos de pré-processamento ora propostos podem ajudar em muito a expansão e a viabilização da Web Semântica por meio da inserção automática de metadados nas páginas HTML[4] e verificação de coerência estatística dos mesmos.

[4]Essa abordagem já está sendo testada em projeto CNPq para o Portal Cidadão.