8
Conclusão e Trabalhos Futuros

O trabalho consiste em mostrar a importância de um bom pré-processamento de dados para uma mineração de textos eficaz. Vimos que todo texto de linguagem natural está atrelado a uma língua e toda língua tem particularidades específicas que não podem ser desprezadas em nenhum algoritmo de mineração de textos, nem nos mais simples.

Para extrair informações mais ricas dos textos devemos processá-los previamente de acordo com o conhecimento específico da língua. Vimos ainda que ao fazer um bom pré-processamento, tornamos mais simples a mineração de textos e suas aplicações.

Esse trabalho apresentou uma pesquisa em que é proposto um sistema de pré-processamento automático para mineração de textos em português utilizando técnicas de inteligência computacional baseada em conceitos existentes na literatura, como redes neurais, sistemas dinâmicos, e estatística multidimensional

O objetivo da tese de doutorado foi, portanto, contribuir com a área de mineração de textos na fase de pré-processamento. Ao se propor um modelo autônomo e automático de enriquecimento de dados textuais, estendeu-se o tradicional modelo de *bag-of-words*, de ênfase mais estatística, transformando-o no modelo do tipo *bag-of-lexems* com maior aproveitamento do conteúdo lingüístico do texto numa abordagem mais computacional.

A existência de um modelo automático de pré-processamento poupa esforços na parte mais custosa do processo de mineração de textos, e ainda, como visto no texto, melhora a semântica dos resultados.

Foram mostrados ainda alguns exemplos de aplicação de mineração de textos como: classificação de documentos, extração de informações e interface de linguagem natural (ILN) além da importância para o âmbito da WebSemântica. Os exemplos de classificação de textos foram utilizados para iluminar a diferença dos procedimentos de mineração de textos com ou sem o modelo ora proposto.

Entendemos que um bom pré-processamento economiza espaço de armazenamento devido a uma seleção mais enxuta dos termos, o que vimos ser crítico nos modelos Espaço-Vetorial. Esta seleção reduzida também aumenta a

precisão dos índices em modelos de Redes Neurais, aumenta a interpretabilidade em modelos discriminantes e melhora a condição inicial de modelos baseados em transformação.

O trabalho foi ainda complementado com o desenvolvimento e a implementação de uma instância do sistema de pré-processamento de textos. Como visto nos resultados, o sistema melhora os índices de acerto, parsimônia, tempo e generalização, além da maior interpretabilidade dos resultados.

Os resultados nos mostram que essa abordagem consegue estar entre as melhores do mundo para a língua portuguesa pela avaliação HAREM efetuada em 2005 e promovida pela instituição Linguateca em Portugal.

Durante esse período, foi feita uma vasta revisão da literatura recente sobre todas as etapas do processo de mineração de textos podendo ser aproveitado como guia para estudos futuros mais profundos. No entanto, a principal dificuldade encontrada nesta pesquisa foi a pouca quantidade de trabalhos sobre a etapa de pré-processamento para a língua portuguesa tanto no Brasil quanto no exterior.

Esse trabalho contribui, então, em grande parte, para viabilização de trabalhos futuros em mineração de textos para a língua portuguesa utilizando o módulo de reconhecimento de entidades lingüísticas e armazenamento em XML.