1
Introdução

As mudanças ocorridas na sociedade nos últimos anos, devido principalmente a avanços tecnológicos e de comunicação, possibilitaram a criação e o armazenamento de grandes volumes de dados no formato digital. Entretanto, é sabido que o ser humano não consegue tratar de forma eficiente com grandes quantidades de dados. Surge então um novo desafio, que é lidar com tais volumes, conseguir extrair informação realmente útil e, por fim, adquirir novo conhecimento.

Empresas e organizações de todo mundo já fazem uso de técnicas avançadas de apoio à decisão, em particular as de Mineração de Dados [1] e *Data Warehousing* [2][3]. Ambas as abordagens caracterizam-se pela extração de informação proveitosa a partir de dados estruturados como planilhas, relatórios, gráficos e tabelas.

Entretanto, com o advento da digitalização e sucessivamente da Internet, criou-se um grande repositório *on-line* de textos e páginas. Assim, pesquisas recentes mostram que, pela naturalidade do formato, cerca de 80% do conteúdo *on-line* está em formato textual [4]. Da mesma forma, outra pesquisa aponta para que, nas empresas e organizações, o percentual de dados não-estruturados seja o mesmo [5]. Entretanto, até pouco tempo atrás, as informações ocultas nesse tipo de dado não eram usadas para a obtenção de algum tipo de vantagem competitiva ou suporte à tomada de decisão.

Conseqüentemente, nasce uma nova área de conhecimento que ainda está em processo de amadurecimento teórico: a **Mineração de Textos** (*Text Mining*), também conhecida por **Mineração de Dados Textuais** (*Text Data Mining*) ou **Descoberta de Conhecimento em Textos** (*Knowledge Discovery in Texts*).

O principal objetivo de se minerar textos é descobrir conhecimento novo e inovador a partir de massas de texto livre, isto é, na forma natural que conhecemos e lidamos diariamente, agregando valor comercial a empresas e organizações.

1.1. Objetivos da Dissertação

O objetivo desta dissertação é explorar a Mineração de Textos através de um estudo amplo e completo do que atualmente é considerado “estado da arte”. Por se tratar de uma área bastante recente, ainda há muitas divergências entre o que seria a **metodologia ideal** para a descoberta de conhecimento a partir de textos. Deste modo, é explorado a fundo um modelo sugerido em outros trabalhos, o qual é constituído das seguintes etapas: **Coleta**, **Pré-processamento**, **Indexação**, **Mineração** e **Análise**.

1.2. Organização da Dissertação

No capítulo 2, são discutidos os fundamentos da Mineração de Textos, com a apresentação de cada uma das principais áreas de conhecimento que a compõe, seguido de alguns exemplos de aplicações práticas em áreas como Medicina e Direito.

No capítulo 3, são descritas as etapas da metodologia estudada, composta por: coleta, pré-processamento, indexação, mineração e análise de resultados.

No capítulo 4, são apresentadas as principais tarefas da etapa de mineração, aonde cada uma destas atende a um objetivo específico.

No capítulo 5, é apresentado o desenvolvimento de um sistema para mineração de textos, seguido de aplicação prática em um estudo de caso.

Finalmente, no capítulo 6, são apresentadas as conclusões deste trabalho.