1
Introdução

1.1
Caracterização do problema

A apresentação de informação sob a forma eletrônica vem crescendo de modo acelerado nas últimas décadas, principalmente por conta da internet. O grande alcance desse meio de comunicação e a facilidade de disponibilização de conteúdo são os principais fatores que incentivam a disposição de textos na internet.

Dessa forma é necessário criar maneiras de acesso a esses textos de forma rápida e objetiva.

Quando se trata de notícias de jornal, a necessidade de obtenção de informação de forma rápida e objetiva é mais crítica, pois o usuário deseja estar sempre atualizado e informado sobre as notícias publicadas nas suas fontes de preferência.

Com os mecanismos atuais de aquisição de informação na internet, como por exemplo, a ferramenta de busca mais popular Google, nem sempre os resultados são de acordo com a necessidade do usuário, retornando muita informação irrelevante ao contexto da busca. Assim sendo, só resta ao usuário realizar uma busca manual e exaustiva no conjunto de respostas retornado pela busca. Um dos principais fatores para isso é que freqüentemente o usuário não sabe expressar sua busca por palavras chaves, muitas vezes expressando sua consulta com poucas palavras e com palavras que possam gerar ambigüidade [Baeza-Yates et al., 2004]. Além disso, não existe nenhuma noção de semântica nesse tipo de busca, retornando ao usuário textos irrelevantes que simplesmente possuem as palavras chaves expressas na sua consulta [Baeza-Yates & Ribeiro-Neto, 1999].

Uma solução para o problema de organização e aquisição de informação é a classificação automática de textos. Com a classificação automática de textos o usuário pode especificar categorias de interesse, por exemplo, *economia do Brasil,*

e política externa, e o sistema buscar na internet, ou em fontes especificadas pelo usuário, retornando ao usuário documentos referentes aos tópicos desejados. Para isso, é necessária a criação de um conhecimento de domínio sobre os tópicos especificados. Até os anos 80, a técnica mais popular para extração desse conhecimento era através da construção manual de regras realizada por engenheiros do conhecimento, como por exemplo, o sistema CONSTRUE [Hayes & Weinstein, 1990], criado pelo grupo Carnegie para a agência de notícias Reuters. A grande desvantagem desses sistemas é que exigia um esforço humano considerável para a criação manual das regras e caso o conhecimento de domínio fosse atualizado, seria necessária novamente uma intervenção dos engenheiros do conhecimento.

A partir de então, foram propostas abordagens que se baseiam em técnicas de aprendizado de máquina. Nessas abordagens, através de um processo indutivo de aprendizado, é criado automaticamente um classificador, dado um conjunto de relações entre documentos e rótulos associados. Uma vez criado o classificador, dado um documento não conhecido, o algoritmo classifica o documento nas categorias aprendidas na fase de treinamento.

No entanto, muitas das abordagens propostas resolveram o problema de classificação binária de textos não considerando o caso mais geral que é a possibilidade de um documento estar associado a mais de um rótulo. Dentre as propostas que trataram o problema de classificação multirótulo, a maioria transforma este problema em subproblemas de classificação binária, considerando que existe independência entre as categorias. Além disso, utilizam limiares, que são muito específicos para o conjunto de treinamento utilizado, não possuindo grande capacidade de generalização na aprendizagem.

1.2 Objetivo

Esta dissertação propõe dois algoritmos de classificação automática de textos baseados no algoritmo multinomial naive Bayes e sua aplicação em um ambiente on-line de classificação automática de notícias com realimentação de relevância pelo usuário, combinando técnicas de aprendizado de máquina e mineração de textos.

O sistema proposto é constituído de três etapas: na primeira o usuário especifica um conjunto de fontes, um conjunto de categorias de interesse, e o sistema busca documentos nas fontes de informação. Em seguida o usuário associa os documentos às categorias podendo classificar um documento em mais de uma categoria e também criar novas categorias. Por último, o sistema busca novos documentos e os classifica de acordo com o treinamento realizado anteriormente. O processo de classificação e treinamento se repete tantas vezes quanto o usuário queira.

Muitas são as vantagens de se utilizar tal técnica: A semântica das categorias é subjetiva, possibilitando ao usuário uma melhor especificação de sua necessidade de informação. Outra vantagem é que a possibilidade de realimentação do usuário e a adaptação do sistema às correções de falsos positivos e falsos negativos implicam em uma grande flexibilidade do sistema e convergência para uma alta acurácia. Além disso, a categorização de um documento em mais de uma classe é bem mais realista, uma vez que é comum a sobreposição de categorias.

Para testar a eficiência dos algoritmos propostos, foram realizados testes na base de dados da Reuters e na base de dados Ohsumed.

1.3 Trabalhos relacionados

A classificação de textos é uma área muito rica no que se refere à quantidade de trabalhos publicados e pesquisas. A tarefa de classificação de textos data da década de 60, porém com a criação da internet se tornou cada vez mais necessária, uma vez que a quantidade de informação disponível vem crescendo de forma acelerada.

Desta forma, muitos classificadores foram propostos, entre eles o classificador naive Bayes, utilizado nesse trabalho. Tal classificador assume que existe independência entre as palavras de um texto (por isso é chamado de "naive"), o que na maioria das tarefas é uma suposição falsa. Porém, Domingos & Pazzani, [1997] mostraram teoricamente que a suposição de independência de palavras na maioria dos casos não prejudica a eficiência do classificador.

Basicamente, existem dois tipos de modelos estatísticos para os classificadores naive Bayes. O modelo multinomial, usado nessa dissertação, que representa um documento como um vetor de freqüências das palavras no texto e o modelo binário, que representa um documento como um vetor binário de palavras, considerando apenas a ocorrência das palavras no texto. O modelo binário foi apresentado em [Robertson & Sparck Jones, 1976; Koller & Sahami, 1997]. Já o modelo multinomial foi apresentado em [Joachims, 1998; McCallum & Nigam, 1998]. Lewis, [1998] apresenta uma comparação teórica entre o modelo binário e o modelo multinomial. Já McCallum & Nigam, [1998], através de experimentos, comparam o modelo multinomial com o modelo binário e concluem que o modelo multinomial apresenta melhores resultados.

Outras técnicas foram propostas além do classificador naive Bayes como, por exemplo, support vector machines [Vapnik 1998], que atualmente tem-se mostrado muito eficiente [Joachims, 1998], K-nearest neighbors [Yang & Liu, 1999], árvores de decisão [Breiman et al., 1984], Boosting [Schapire & Singer, 2000] e redes neurais [Shavlik & Eliassi-Rad, 1998].

Grande parte dos trabalhos resolve o problema da classificação binária e para resolver o problema multirótulo, propõe a transformação do problema em subproblemas binários. A técnica mais conhecida, chamada de "one-vs-all" cria para cada classe um classificador binário, capaz de distingui-la das demais classes do domínio. Assim, um exemplo associado a uma categoria c_i é um exemplo positivo para c_i, e um exemplo negativo para todas as outras classes as quais o documento não pertence.

Entretanto, McCallum [1999] propõe um algoritmo bayesiano que soluciona o problema de classificação multirótulo através de um modelo misto e "Expectation Maximization".

1.4 Organização da dissertação

Os capítulos a seguir estão organizados da seguinte forma. O capítulo 2 apresenta os fundamentos teóricos de máquina de aprendizado, a tarefa de classificação automática de textos, duas principais abordagens do algoritmo naive Bayes e as principais medidas para apresentar a eficiência de classificadores de

texto. No capítulo 3 são apresentados os dois algoritmos de classificação propostos nessa dissertação, assim como sua aplicação em um ambiente online de filtragem de documentos com realimentação de relevância do usuário. O capítulo 4 apresenta os resultados obtidos em testes realizados na base de notícias da Reutes e na base de documentos médicos Ohsumed e analisa os resultados obtidos. O capitulo 5 apresenta os principais trabalhos relacionados. Finalmente, o capitulo 6 apresenta as conclusões e sugestões para trabalhos futuros.