5
Conclusão

Nesta dissertação foram propostos dois algoritmos de classificação automática de textos multirótulo.

Além disso, foi proposta a utilização de tais algoritmos em um ambiente completo de mineração de textos, onde inicialmente o usuário especifica um conjunto de categorias de interesse e um conjunto de fontes de notícias da internet. O sistema, então, retorna ao usuário um conjunto de notícias, possibilitando ao usuário selecionar para cada notícia um conjunto de categorias associadas. A partir do treinamento do usuário, um novo conjunto de notícias é retornado, onde cada notícia está associada a um conjunto de categorias. Então, o usuário pode corrigir os possíveis erros cometidos pelo classificador.

Com o propósito de apresentar o desempenho de tais algoritmos, realizaram-se experimentos em duas bases de documentos freqüentemente utilizadas na literatura, a base de notícias Reuters 21578 e a base de documentos médicos Ohsumed.

A fim de que os resultados obtidos fossem comparáveis com outros trabalhos, foram realizados experimentos utilizando partições pré-definidas, no caso da Reuters, a partição ModApté e no caso da base Ohsumed, a partição dos 20.000 primeiros documentos.

Comparando-se os resultados obtidos com outros trabalhos, pode-se concluir que os dois algoritmos possuem uma eficiência satisfatória com relação às medidas Micro Recall, Micro Precision e Micro F1.

Porém, nos experimentos realizados na base da Reuters R(90), os dois algoritmos apresentaram a medida Macro F1 consideravelmente inferior aos trabalhos anteriores. Uma vez que a base de notícias Reuters R(90) possui uma distribuição não uniforme entre os documentos de treinamento e o conjunto de categorias, pode-se concluir que tal desempenho se deve ao fato de tais algoritmos serem muito sensíveis à distribuição entre documentos e categorias. Isso pode ser uma desvantagem, uma vez que tais algoritmos aplicados ao sistema de mineração

de textos proposto necessitam de uma maior intervenção do usuário, já que deve ser apresentada uma maior quantidade de exemplos de treinamento para cada categoria.

Outra desvantagem é que as complexidades teóricas da fase de classificação do algoritmo pseudo-multirótulo e da fase de treinamento do algoritmo multirótulo são exponenciais na quantidade de categorias. Entretanto, considerando que, na prática, os documentos são associados a combinações de categorias relativamente pequenas, isso não é um problema.

As vantagens de se utilizar tais algoritmos propostos são:

- Não é necessária a manutenção dos documentos de treinamento. Uma vez realizado o treinamento, os documentos podem ser descartados.

- Os algoritmos são perfeitamente adaptáveis ao contexto de classificação on-line, uma vez que dado um novo documento de treinamento, não é necessário retreinar toda a base para contemplar o novo conhecimento.

- Os dois algoritmos não utilizam limiares, que são muito específicos para o conjunto de treinamento utilizado, não possuindo grande capacidade de generalização na aprendizagem. Desta forma, não é necessária uma fase de validação para se encontrar valores ideais para tais parâmetros.

- Os dois algoritmos consideram que existe dependência entre categorias, que, na prática, é o caso mais comum.

5.1 Contribuições

As principais contribuições deste trabalho são:

- Criação de dois algoritmos de classificação automática de textos de fácil implementação, que possibilitam resolver o problema mais genérico de classificação (multirótulo), sem a necessidade de calcular limiares e considerando que existe dependência entre as categorias.

- Proposta de criação de um sistema de mineração de textos aplicado ao contexto de notícias de jornal, com a possibilidade de realimentação de relevância por parte do usuário e treinamento on-line.

- Possibilidade de aplicação dos algoritmos propostos no sistema de mineração de textos.

5.2 Trabalhos futuros

- Implementação do sistema de mineração de textos e realização de testes para verificar a eficiência de uma aplicação prática dos algoritmos propostos neste trabalho.

- Inclusão de relações entre categorias, ou seja, permitir que o usuário apresente ao sistema relações entre categorias e o sistema agregue tais informações ao aprendizado.

- Possibilitar que o aprendizado seja semi-supervisionado, reduzindo a necessidade de o usuário associar para cada documento um conjunto de categorias, tarefa que pode ser bastante trabalhosa e suscetível a erros.

- Realização de melhorias no algoritmo multirótulo, verificando novas condições de parada na fase de classificação.

- Realização de testes com outras técnicas de amortização além da técnica utilizada neste trabalho.