10
Trabalhos Relacionados

A linha de pesquisa de classificação de documentos é bastante extensa, já havendo diversos trabalhos relacionados. Quase a totalidade dos trabalhos encontrados destina-se à classificação de documentos na Internet.

Em [61], é apresentado um agente para descobrir e classificar documentos HTML na Internet baseado em um perfil de usuário. O algoritmo adotado prevê a descoberta das classes envolvidas através do treinamento com um grupo de documentos cuja classe é previamente conhecida. Através da análise das palavras contidas nesse grupo de documentos, principalmente a freqüência com que ocorrem, o agente gera *queries* em um formato definido. Essas *queries*, se executadas em uma base de dados de documentos, selecionam documentos cuja classe mais se assemelha à qual os documentos pertencentes ao grupo de aprendizado estão inseridos. Tais *queries* podem ser traduzidas para o formato dos motores de busca mais conhecidos, como Alta Vista [62], Hot Bot [63] e Yahoo [64], retornando os documentos desejados. De forma semelhante, em [65] e [66] são apresentados dois agentes de software que também se utilizam de algoritmos de aprendizado a partir de um conjunto inicial de documentos, para posteriormente buscarem novas referências na Internet. Já em [67], [68] e [69], são apresentados agentes de software que baseiam sua classificação em regras pré-estabelecidas, que são baseadas em informações do domínio dos documentos analisados. Não existe, portanto, uma fase de aprendizado. São, dessa forma, menos cognitivos que os apresentados anteriormente.

Saindo da esfera acadêmica para a esfera comercial, existem algumas iniciativas para a realização de *webclipping* que, conforme foi apresentado anteriormente, pode ser encarada como uma classificação de documentos. Nos EUA, podemos citar a empresa WebClipping.doc [70], que se utiliza de um software proprietário para realizar pesquisas em diversos veículos de comunicação *online* e *newsletters*. No Brasil, apesar de existirem empresas de *clipping* que

realizam *webclipping*, este é realizado de forma artesanal, sem a utilização de nenhum software de automatização.

A grande diferença entre este trabalho e os demais é a da criação de um *framework*, ao contrário de uma aplicação pronta. A principal vantagem é a possibilidade de deixar flexíveis áreas como o algoritmo de classificação e os documentos categorizados, o que traz como principal benefício:

- Na esfera acadêmica, transforma o software em uma espécie de laboratório, onde diferentes algoritmos de classificação podem ser testados em diferentes tipos de documentos.
- Na esfera comercial, deixa o software extremamente flexível e de fácil evolução, que são duas qualidades que o protegem da obsolescência, um problema sério se considerarmos o atual cenário de constante surgimento de novas tecnologias.