Fabrício Jailson Barth

Recuperação de documentos e pessoas em ambientes empresariais através de árvores de decisão

São Paulo
2009

Fabrício Jailson Barth

Recuperação de documentos e pessoas em ambientes empresariais através de árvores de decisão

Tese apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do Título de Doutor em Engenharia Elétrica.

Área de concentração:
Sistemas Digitais

Orientador:
Prof. Dr. Edson Satoshi Gomi

São Paulo
2009

Este exemplar foi revisado e alterado em relação à versão original, sob responsabilidade única do autor e com anuência de seu orientador.

São Paulo, de junho de 2009.

Assinatura do autor

Assinatura do orientador

Ficha Catalográfica

Barth, Fabrício Jailson

Recuperação de documentos e pessoas em ambientes empresariais através de árvores de decisão. / F. J. Barth – ed.rev. – São Paulo, 2009.

87 p.

Tese (Doutorado) — Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1. Recuperação da Informação. 2. Aprendizado Computacional. 3. Gestão da Informação I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais. II. t.

Dedicatória

À Fernanda G. Moreira, minha esposa, com amor, admiração e gratidão por sua compreensão, carinho, presença e incansável apoio ao longo do período de elaboração deste trabalho.

Agradecimentos

Ao Prof. Dr. Edson Satoshi Gomi, pela atenção, orientação e apoio durante a realização deste trabalho.

Aos professores, amigos, colegas e alunos, que, nos anos de convivência, muito me ensinaram, contribuindo para meu crescimento pessoal, científico e intelectual.

Ao Laboratório de Engenharia de Conhecimento (KNOMA), por colocar a disposição os equipamentos para a realização dos experimentos.

Resumo

Este trabalho avalia o desempenho do uso de árvores de decisão como função de ordenação para documentos e pessoas em ambientes empresariais. Para tanto, identificou-se atributos relevantes das entidades a serem recuperadas a partir da análise de: (*i*) dinâmica de produção e consumo de informações em um ambiente empresarial; (*ii*) algoritmos existentes na literatura para a recuperação de documentos e pessoas; e (*iii*) conceitos utilizados em funções de ordenação para domínios genéricos. Montou-se um ambiente de avaliação, utilizando a coleção de referência CERC, para avaliar a aplicabilidade do algoritmo C4.5 na obtenção de funções de ordenação para o domínio empresarial. O uso do algoritmo C4.5 para a construção de funções de ordenação mostrou-se parcialmente efetivo. Para a tarefa de recuperação de documentos não trouxe resultados bons. Porém, constatou-se que é possível controlar a forma de construção da função de ordenação a fim de otimizar a precisão nas primeiras posições do *ranking* ou otimizar a média das precisões (MAP). Para a tarefa de recuperação de pessoas o algoritmo C4.5 obteve uma árvore de decisão que consegue resultados melhores que todas as outras funções de ordenação avaliadas. O MAP obtido pela árvore de decisão foi $0,83$, enquanto que a média do MAP das outras funções de ordenação foi de $0,74$. Percebeu-se que a árvore de decisão utilizada para representar a função de ordenação contribui para a compreensão da composição dos diversos atributos utilizados na caracterização dos documentos e pessoas. A partir da análise da árvore de decisão utilizada como função de ordenação para pessoas foi possível entender que uma pessoa é considerada especialista em algum tópico se ela aparecer em muitos documentos, aparecer muitas vezes nos documentos e os documentos onde aparece têm uma relevância alta para a consulta.

Palavras-chave: Aprendizagem de Funções de Ordenação. Recuperação da Informação. Aprendizado Computacional. Gestão da Informação.

Abstract

This work evaluates the performance of using decision trees as ranking functions for documents and people in enterprises. It was identified relevant attributes of the entities to be retrieved from the analysis of: (*i*) the production and consumption of information behavior in an enterprise, (*ii*) algorithms for documents and people retrieval at literature, and (*iii*) the concepts used in ranking functions for generic domains. It was set up an evaluation environment, using the CERC collection, to evaluate the applicability of the C4.5 algorithm to obtain a ranking function for the enterprise domain. The use of C4.5 algorithm for the construction of ranking function was proved to be partially effective. In the case of documents retrieval the C4.5 has not found good results. However, it was found that is possible to control the way of building the ranking function in order to optimize the precision in the first positions of the ranking or optimize the mean average precision (MAP). For the task of people retrieval the C4.5 algorithm developed a ranking function that obtain better results than all other ranking functions assessed. The value of MAP obtained by decision tree was $0,83$, while the average MAP of other ranking functions was $0,74$. The decision tree used to represent the ranking function contributes to understanding the attributes composition used in the characterization of documents and people. Through the analysis of the decision tree used as ranking function for people, we could realise that a person is considered expert in any topic if he/she appear in many documents, appear many times in same documents and documents where he/she appears have a high relevance to the query.

Keywords: Learning to Rank. Information Retrieval. Machine Learning. Information Management.

Sumário