JOÃO ALFREDO PINTO DE MAGALHÃES

Um Framework Multi-Agentes para Busca e Flexibilização de Algoritmos de Classificação de Documentos

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Pontifícia Universidade Católica do Rio de Janeiro

Rio de Janeiro
Setembro de 2002

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

João Alfredo Pinto de Magalhães

Um *Framework* Multi-Agentes para Busca e Flexibilização de Algoritmos de Classificação de Documentos

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática da PUC-Rio.

Orientador: Prof. Carlos J. P. de Lucena

Rio de Janeiro

Setembro de 2002

João Alfredo Pinto de Magalhães

João Alfredo Pinto de Magalhães é engenheiro de Computação formado pela Pontifícia Universidade Católica do Rio de Janeiro, da turma de 1996, tendo se formado em julho de 2000. Vem realizando pesquisas na área de engenharia de software desde 1999, quando se associou ao Laboratório de Engenharia de Software do Departamento de Informática da PUC-Rio. É sócio da Knewledge Technology, empresa de base tecnológica que criou com mais três amigos, dois dos quais foram colegas de turma, cujo foco é o desenvolvimento de software de missão crítica e tempo real.

Ficha Catalográfica

Magalhães, João Alfredo Pinto de

Um *Framework* multi-agentes para busca e flexibilização de algoritmos de classificação de documentos / João Alfredo Pinto de Magalhães; orientador: Carlos J. P. de Lucena. – Rio de Janeiro : PUC, Departamento de Informática, 2002.

[12], 104 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática.

Inclui referências bibliográficas.

1. Informática – Teses. 2. Framework. 3. Sistemas multi-agentes. 4. Classificação de documentos. 5. Separação de responsabilidades. 6. Web semântica. I. Lucena, Carlos J. P. de. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Este trabalho é dedicado a Deus e aos meus queridos pais, Mariza da Silva e Alfredo Magalhães, por tudo o que eles representam em minha vida.

Agradecimentos

A execução e finalização deste trabalho de pesquisa de dois anos de duração não se dariam sem a participação de algumas pessoas para as quais gostaria de mostrar aqui meu agradecimento sincero.

A Deus, por estar sempre orientando meus passos, encorajando-me a enfrentar todos os desafios e chegar até aqui.

Aos meus pais, Mariza e Alfredo, e à minha tia Mábia, pela compreensão, amor e carinho de sempre.

Ao meu orientador, professor Carlos J. P. de Lucena, por diversas vezes clarear dúvidas em minha mente e minimizar os erros durante todo o trabalho de pesquisa. As conversas de amigo e os inúmeros "bom momentos" foram especiais e caracterizam seu estilo *leve* e *sério* de orientar.

Aos professores, Ruy Milidiú e Simone D. J. Barbosa, pela atenção, pelas idéias e pelo apoio na definição deste trabalho.

A todos os funcionários do Departamento de Informática e da Fundação Padre Leonel Franca que, de alguma forma, contribuíram para a realização desse trabalho, em especial à Vera A. S. Menezes e ao Luís Fernando, que muito me ajudaram nesses dois anos.

Aos meus dois grandes amigos, Matheus e Daniel, cuja amizade eu tenho como um dos melhores resultados de minha passagem pela PUC. Obrigado por todo o apoio e companheirismo nas incontáveis madrugadas que passamos no ICQ!

Ao Luiz Mário e Leonardo, meus amigos, que muito me apoiaram com idéias e incentivos durante esta caminhada.

À Lívia, pelo apoio irrestrito, por toda a compreensão, cumplicidade e amor de sempre.

Aos colegas do laboratório TecComm, pela amizade e ajuda, em especial ao Otavio “Pirulito”, Francisco “Chicão”, Akeo, Alessandro “Véio”, Gustavo “Guga”, Leandro “Daflon”, Cristiano “CriCri”, Lucimar e Viviane. Obrigado pelos ensinamentos e incentivo que muito colaboraram para a realização deste trabalho!

Aos amigos da PUC, Adriana, Alésio, Flávio “Sufflair” Rodrigo, Marcel, Elton, Juliana, Viviane Braconi, Luciana Lima, Maíra, Taciana, Heron, Fábio e Fábio “Fabiomar” Marcos, pela alegria, companheirismo e pelos momentos de descontração, combustíveis para inspiração e criatividade.

A todos os professores do Departamento de Informática que, durante a minha graduação, compartilharam comigo seus conhecimentos, sem os quais teria sido impossível realizar esse trabalho. Em especial, ao Marcus Poggi, pela amizade, ajuda e pelo incentivo recebido para dar início a esta caminhada.

À CAPES, à FAPERJ e ao Departamento de Informática pelo apoio financeiro que possibilitou a realização deste trabalho.

Resumo

Magalhães, João Alfredo Pinto de; Lucena, Carlos José Pereira de. **Um Framework Multi-Agentes para Busca e Flexibilização de Algoritmos de Classificação de Documentos**. Rio de Janeiro, 2002. 116p. Dissertação de Mestrado - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Vivemos na era da informação, onde o conhecimento é criado numa velocidade nunca antes vista. Esse aumento de velocidade teve como principal razão a Internet, que alterou os paradigmas até então existentes de troca de informações entre as pessoas. Através da rede, trabalhos inteiros podem ser publicados, atingindo um público alvo impossível de ser alcançado através dos meios existentes anteriormente. Porém, o excesso de informação também pode agir no sentido contrário: muita informação pode ser igual a nenhuma informação. Nosso trabalho foi o de produzir um sistema multi-agentes para busca e classificação de documentos textuais de um domínio específico. Foi construída uma infra-estrutura que separa as questões referentes à busca e seleção dos documentos (plataforma) das referentes ao algoritmo de classificação utilizado (uma aplicação do conceito de *separation of concerns*). Dessa forma, é possível não só acoplar algoritmos já existentes, mas também **gerar novos algoritmos levando em consideração características específicas do domínio de documentos abordado.** Foram geradas quatro instâncias a partir do *framework*, uma aplicação de *webclipping*, um componente para auxílio a *knowledge management*, um motor de busca para *websites* e uma aplicação para a *web semântica.*

Palavras-chave

framework, sistemas multi-agentes, classificação de documentos, separação de responsabilidades, web semântica

Abstract

Magalhães, João Alfredo Pinto de; Lucena, Carlos José Pereira de. **Thesis Title in English**. A multi-agent framework for search and flexibilization of document classification algorithms. Rio de Janeiro, 2002. 116p. MSc. Dissertation - Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro

We are living in the information age, where knowledge is constantly being created in a rate that was never seen before. This is mainly due to Internet, that changed all the information exchange paradigms between people. Through the net, it is possible to publish or exchange whole works, reaching an audience impossible to be reached through other means. However, excess of information can be harmful: having too much information can be equal to having no information at all. Our work was to build a multi-agent framework for search and flexibilization of textual document classification algorithms of a specific domain. We have built an infra-structure that separates the concerns of document search and selection (platform) from the concerns of document classification (an application of the separation of concerns concept). It is possible not only to use existing algorithms, but also to generate new ones that consider domain-specific characteristics of documents. We generated four instances of the framework, a webclipping application, a knowledge management component, a search engine for websites and an application for the semantic web.

Keywords

framework, multi-agent systems, document classification, separation of concerns, semantic web

Sumário

1 – INTRODUÇÃO 12

2 – GESTÃO DE CONHECIMENTO (*KNOWLEDGE MANAGEMENT*) 16

3 – A WEB SEMÂNTICA 21

3.1 Interoperabilidade Semântica na WWW 23

3.2 Linguagens para Representação de Ontologias para a Web 26

3.2.1 – Simple HTML Ontology Extension (SHOE) *26*

3.2.2 – Ontology Inference Layer (OIL) *27*

3.2.3 – Darpa Agent Markup Language (DAML) *29*

4 – AGENTES DE SOFTWARE 32

5 – ALGORITMOS DE GERAÇÃO DE CATEGORIAS 34

6 – O *FRAMEWORK* AVESTRUZ 36

7 – O *FRAMEWORK* E SEU SISTEMA MULTI-AGENTES 39

7.1 O Agente de Software 40

7.1.1. Questões de implementação *40*

7.1.1.1 Subsistema Processador de Documentos 43

7.1.1.2 Subsistema Armazenador de Documentos a Visitar 46

7.1.1.3 Subsistema de Memória 49

7.1.1.4 Subsistema de Pesquisa 52

7.1.1.5 Subsistema Gerador de Relatórios 54

7.1.1.6 Subsistema de Gerência 56

7.1.2 .Questões de Comunicação e Coordenação *58*

7.2 O Sistema Multi-Agentes 59

7.2.1 Questões de Implementação *59*

8 – INSTANCIANDO O *FRAMEWORK* 61

8.1 Classes a serem instanciadas 61

8.1.1 Questões de Plataforma *63*

8.1.2 Questões de Classificação *67*

8.2 Arquiteturas de instanciação 68

8.2.1 Distribuição Intra-processo *69*

8.2.2 Distribuição Inter-processo *74*
8.2.2.1 Arquitetura Espelho 76
8.2.2.2 Arquitetura Geradora 78

9 – AS INSTANCIAÇÕES DO *FRAMEWORK* 80

9.1 – O *Webclipper* 80
9.1.1 Questões de Implementação *82*
9.1.1.1 Subsistema Processador de Documentos 83
9.1.1.2 Subsistema Armazenador de Documentos a Visitar 85
9.1.1.3 Subsistema de Memória 85
9.1.1.4 Subsistema de Pesquisa 86
9.1.1.5 Subsistema Gerador de Relatórios 89
9.1.2 Questões de Configuração *90*
9.1.3 Utilização do Webclipper *90*

9.2 – O KM *Probe* 91
9.2.1 Questões de Implementação *93*
9.2.1.1 Subsistema Processador de Documentos 93
9.2.1.2 Subsistema Armazenador de Documentos a Visitar 94
9.2.1.3 Subsistema de Memória 94
9.2.1.4 Subsistema de Pesquisa 95
9.2.1.5 Subsistema Gerador de Relatórios 97
9.2.2 Questões de Configuração *97*
9.2.3 Utilização do KM Probe *97*

9.3 – O *Site Seeker* 98
9.3.1 Questões de Implementação *98*
9.3.1.1 Subsistema Processador de Documentos 99
9.3.1.2 Subsistema Armazenador de Documentos a Visitar 99
9.3.1.3 Subsistema de Memória 99
9.3.1.4 Subsistema de Pesquisa 99
9.3.1.5 Subsistema Gerador de Relatórios 100
9.3.2 Questões de Configuração *100*
9.3.3 Utilização do Site Seeker *100*

9.4 – O *Semantic Probe* 100
9.4.1 Questões de Implementação *101*
9.4.1.1 Subsistema Processador de Documentos 101
9.4.1.2 Subsistema Armazenador de Documentos a Visitar 102
9.4.1.3 Subsistema de Memória 102
9.4.1.4 Subsistema de Pesquisa 102
9.4.1.5 Subsistema Gerador de Relatórios 103
9.4.2 Questões de Configuração *103*
9.4.3 Utilização da Semantic Probe *103*

10 – TRABALHOS RELACIONADOS 104

11 – CONCLUSÕES E TRABALHOS FUTUROS 106

11.1 Conclusões 106

11.2 Trabalhos Futuros 109

12 – REFERÊNCIAS 111