Um sistema Multi-agente para Busca e Filtragem de Informações de Domínio na Internet

Marcelo Nascimento Costa
mcosta@cos.ufrj.br

Orientadores: Cláudia M. L. Werner e Marta L. Q. Mattoso

COPPE/UFRJ - Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro
Caixa Postal 68511 – CEP. 21945-970
Rio de Janeiro – Brasil

Abstract

Domain Engineering (DE) aims at providing a group of reusable components within a specific domain. Application Engineering (AE) aims at constructing applications by reusing resulting DE components. This work presents a multi-agent system, which enables intelligent information retrieval, by providing domain engineers with a broad and relevant amount of information to support them in the accomplishment of DE/AE processes. The system adopts techniques such as user modeling, machine learning, and domain ontologies. The proposed approach allows information search on several sources of heterogeneous and/or distributed data throughout the Web. One important objective is to improve the algorithms that are already used by agents, involved in machine learning and information filtering. The techniques currently adopted for user modeling are being revised and some new characteristics will be added to the user profile. Another issue considered by the approach is the documentation of reusable components. XML, a standard defined by W3C, is used for this purpose and further publication of components at the Web. XML provides more semantics to documents and helps the search for components at the Web.

Palavras-Chave: Busca e Filtragem de Informações, Engenharia de Domínio, Agentes Inteligentes, Modelagem de Usuário, Ontologias, Aprendizagem de Máquina, XML**.**

1. Introdução

Uma das técnicas mais importantes, atualmente, na área de Reutilização de Software, é a Engenharia de Domínio (ED). Para a realização de um processo de ED de forma abrangente, é importante, dentre outros aspectos, prover acesso a diversos repositórios de informações, com o intuito de ampliar o conhecimento do engenheiro sobre o domínio em análise. Esses repositórios podem armazenar informações provenientes de diversos processos de ED realizados para o mesmo domínio, ou para algum domínio relacionado. Atualmente, a Web é considerada uma fonte de informação importante para ser consultada, mas devido ao seu grande volume de informações disponíveis, é preciso aplicar processos de filtragem para que somente as informações mais relevantes sejam apresentadas ao usuário.

A Engenharia de Aplicação (EA) trata da construção de uma aplicação a partir da reutilização de componentes[1] provenientes de um ou mais domínios que tenham sido previamente analisados por um processo de ED.

Este trabalho procura prover uma solução para a questão da busca e filtragem de informações sobre um domínio disponíveis na Web, no contexto de ED e/ou EA. A solução implica na extensão de um sistema de software multi-agente inteligente (envolvendo agentes de navegação, agentes de filtragem e de recuperação de informações) existente, descrito em [1]. Esta extensão consiste em adicionar ao sistema a capacidade de realizar a busca e filtragem de informações em repositórios distribuídos e heterogêneos, incluindo a Web, através de agentes. Este trabalho está sendo desenvolvido no contexto da infra-estrutura Odyssey[2][2], provendo uma ferramenta de apoio ao usuário na busca de informações do domínio durante a execução da ED e/ou EA.

Uma contribuição importante da dissertação diz respeito à publicação da documentação de componentes reutilizáveis de forma a facilitar sua busca através da Web. Esses componentes são publicados em XML[3], que provê uma maior semântica aos documentos, facilitando sua consulta por usuários remotos.

O trabalho tem um caráter interdisciplinar, envolvendo diversas áreas de conhecimento da computação: Engenharia de Software, envolvendo a documentação e recuperação de componentes; Banco de Dados com o aspecto da busca e recuperação de informação; e Inteligência Artificial através dos agentes inteligentes, modelagem de usuário, filtragem de informações, algoritmos de aprendizagem de máquina e serviços de ontologias.

2. Trabalhos Relacionados

O problema da recuperação de componentes vem sendo objeto de várias pesquisas, consistindo diversas abordagens descritas, encontradas na literatura, como OntoSeek [4] e esquemas de facetas[5]. Entretanto, todas estas limitam o processo de classificação, armazenamento e recuperação a componentes de código de programa, sendo estes armazenados em repositórios locais e homogêneos. Estes trabalhos não têm como premissa básica a reutilização de outros tipos de componentes, tais como especificação de requisitos e demais produtos do ciclo de desenvolvimento de software.

Em [3], Guerrieri trata da documentação de componentes gerados durante o processo de desenvolvimento de software, utilizando XML. Entretanto, somente a descrição de componentes Java é prevista, utilizando a biblioteca JavaXMLdoc. Nesta dissertação, considera-se a documentação em XML de todos os componentes gerados ao longo do processo de desenvolvimento. Atualmente no Odyssey, a documentação é gerada, em HTML, através da ferramenta Framedoc [2]. Esta ferramenta será estendida para incluir a geração em XML. Com a disponibilização em XML, será mais fácil a consulta a componentes Odyssey por usuários remotos.

---

[1] Componentes são considerados todos os produtos resultantes do ciclo de desenvolvimento de software em qualquer nível de abstração conceitual, arquitetural e implementacional.

[2] Infra-estrutura de apoio a reutilização, em desenvolvimento na COPPE/UFRJ, que tem como objetivo apoiar as atividades dos processos de ED e EA.

Existem na literatura alguns trabalhos relacionados com as técnicas de Inteligência Artificial utilizadas nesta dissertação, tais como o SAIRE [6], que utiliza os conceitos de *perfil de usuário* e agrupamento de perfis de usuários em *estereótipos*. O WebWatcher [7] e o Letizia [8] são sistemas baseados em agentes, que adotam algoritmos de aprendizagem de máquina e de filtragem de informações para tratar dos problemas de apoio inteligente à navegação na Web e de filtragem de informações relevantes ao usuário. Nesta dissertação, as abordagens de aprendizagem de máquina e de modelagem de usuário adotadas são similares às citadas acima, sendo adaptadas ao contexto do Projeto Odyssey, onde ao invés de *links* entre páginas, utiliza-se o conceito de *links* entre os itens[3] de modelos relacionados.

3. Metodologia

Inicialmente, foi construído o protótipo de um sistema multi-agente que apoia o usuário na navegação entre os diversos diagramas disponíveis na infra-estrutura Odyssey [1] (figura 1). Este protótipo utiliza técnicas como: i) modelagem de usuário, que engloba a criação de um perfil de usuário e de agrupamento de usuários com perfis semelhantes (estereótipo); ii) aprendizagem de máquina, onde o sistema identifica o comportamento do usuário em relação às sugestões feitas por ele (*feedback do usuário*), sem que este o perceba, atualizando sua base de conhecimento sobre o estereótipo correspondente.

A partir da construção deste protótipo, foi possível visualizar oportunidades de extensão da arquitetura para ampliar a capacidade de recuperação inteligente de informações do domínio (figura1). Essas extensões são descritas a seguir.

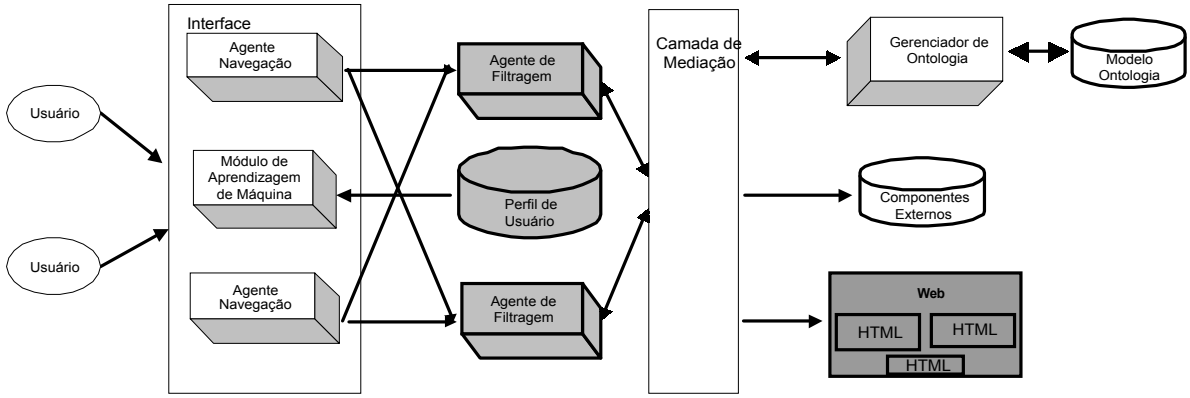

Figura 1. Arquitetura inicial e Extensão proposta (em negrito)

Um primeira extensão está relacionada a busca na Web, o agente deverá poder executar dois tipos de busca: uma previamente validada pelo engenheiro do domínio e a outra sem esta validação (i.e., livre).

No primeiro caso, o analista do domínio deverá criar um repositório de *links* para cada domínio, que apontam para páginas Web onde componentes ou informações do domínio consideradas relevantes para um usuário Odyssey estão disponíveis. Neste caso, também são definidos conceitos (i.e., termos do domínio), extraídos da ontologia do domínio[4], que descrevem essas páginas, associando pesos que indicam o grau de relevância de cada um deles para facilitar a consulta do usuário.

A segunda forma (i.e., livre) implica que o agente deverá acessar algum mecanismo de metabusca existente na WWW, enviando as palavras-chave da consulta do usuário, e retornar para o usuário uma página resposta com um conjunto de *links* que podem ser acessados livremente. O agente deverá apoiar o usuário através da verificação dos títulos e textos descritivos de cada página Web retornada pela mecanismo de busca, e comparando-os com o

[3] Item é qualquer componente do Odyssey.
[4] Ontologia nesse contexto pode ser definida como um vocabulário de termos e relacionamentos entre eles.

grafo de hipernônimos, hiponônimos e sinônimos dos termos da ontologia do domínio e com as informações pertencentes ao perfil do usuário. Como resultado é gerada uma classificação decrescente de *links*, de acordo com o grau de sua importância para o usuário.

Uma segunda extensão da arquitetura trata do refinamento das estratégias (algoritmos) utilizadas nos módulos de aprendizagem de máquina e de filtragem de informação. Os algoritmos serão modificados para a solução de alguns possíveis problemas identificados em [1]. Um deles se refere a questão de *feedback loop,* onde os usuários se deixam influenciar pela posição dos componentes, escolhendo sempre o primeiro na classificação de componentes fornecidos pelo agente. Serão ainda adicionadas algumas novas características ao perfil do usuário, como por exemplo, o usuário poderá atribuir um peso para indicar o grau de relevância de cada termo armazenado no perfil. Desse modo, os algoritmos de navegação e de filtragem de informação poderão se tornar mais precisos.

Uma terceira extensão identificada diz respeito à adaptação da atual documentação de componentes utilizada no Odyssey [2], provendo uma maior semântica aos documentos através da utilização do padrão XML, de forma a facilitar sua recuperação por usuários remotos.

Atualmente, arquitetura proposta está sendo detalhada através do refinamento dos diagramas de classes e de seqüência envolvidos. Estão sendo estudados na literatura problemas relacionados às técnicas de filtragem de informação e aprendizagem de máquina. Em seguida, dar-se-a início a implementação.

4. Considerações Finais

Este trabalho apresenta uma ferramenta importante para a recuperação e filtragem de informações do domínio, visando a ampliação do conhecimento do engenheiro do domínio/aplicação sobre o domínio, através da disponibilização de informação na Web. Para facilitar a recuperação de componentes por usuários remotos, a documentação de componentes utilizará o padrão XML para publicação na Web.

É importante ressaltar que a ferramenta, embora voltada para o Odyssey, pode ser instanciada, com algumas adaptações, em qualquer ambiente de apoio a Engenharia de domínio. A validação do sistema multi-agente deverá ser realizada pelos próprios engenheiros de domínio/aplicação através de sua utilização em diversas situações de engenharia de domínio/aplicação, estando prevista uma primeira utilização no domínio legislativo.

5. Referências

[1] Braga, R.; Costa, M.; Werner, C.; Mattoso; "A Multi-Agent System for Domain Information and Discovery", XIV Simpósio Brasileiro de Engenharia de Software, João Pessoa, Outubro 2000 (aceito para publicação).

[2] Werner, C.; Braga, R.; Mattoso, M.; Murta, L.; Costa, M.; Pinheiro, R.; Oliveira, A.; "Infra-estrutura Odyssey: estágio atual", XIV Simpósio Brasileiro de Engenharia de Software, Caderno de Ferramentas, João Pessoa, Outubro 2000 (aceito para publicação).

[3] Guerrieri E., "Software Document Reuse with XML". Proceedings of the Fifth International Conference on Software Reuse, BC, Canada,1998.

[4] Guarino, N., "OntoSeek: Content-Based Access to the Web", In: IEEE Inteligent Systems, May/Jun, 1999.

[5] Prieto-Diaz, R, "Classification of Reusable Modules", In: IEEE Software, 4 (1): 6-16,1987.

[6] Odubiyi J., Kocur D., Weinstein S., Wakim N., Srivastava S., Gokey C., SAIRE – A Scalable Agent-Based Information Retrieval Endgine", Proceedings of the first International Conference on Autonomous Agents, Fevereiro 5-8,1997, California.

[7] Joachims T., Freitag D., Mitchell T., "WebWatcher: A Tour Guide for the World Wide Web"; Proceedings of IJCAI97, Agosto 1997.

[8] Lieberman, H. "An Automated channel-surfing interface agent for the Web". Artificial Intelligence-based Tools to Help W3 Users Workshop, the Fifth International World Wide Web Conference, Paris, France, Maio 1996.