Área: CV ( ) CHSA ( ) ECET (x)

INCORPORAÇÃO NO SCIENTIA.NET DOS ARTIGOS DOS PRINCIPAIS REPOSITÓRIOS ON-LINE

Camila Vilarinho de Sousa (ICV/UFPI), Vinicius Ponte Machado (Orientador, Departamento de Informática e Estatística – UFPI)

Introdução

O Scientia.Net é uma rede social on-line que visa disponibilizar de forma automática aos seus usuários informações relevantes relacionadas ao seu perfil. Além disso, o Scientia.net é uma agregador de conteúdo contido em diversos serviços da Internet (fóruns, repositórios de artigos, sites, blogs e demais redes sociais). A ferramenta permitirá a interação de seus usuários (estudantes, professores e pesquisadores) com base nos seus interesses em comum.

Os repositórios de artigos hoje constituem uma referencia para professores, pesquisadores, alunos e para a população em geral, permitindo coleta, integração, preservação e o compartilhamento de conhecimento, contribuindo assim para o desenvolvimento da educação, da cultura e da ciência.

Este trabalho trata do estudo e implementação de mecanismos no Scientia.Net que permitem a incorporação automática de artigos das principais bases científicas *on-line* de acordo com o perfil do usuário, disponibilizando automaticamente estes artigos de acordo com a área de interesse do usuário

Metodologia

Para o desenvolvimento deste projeto foi analisada a viabilidade da recuperação de artigos das principais bases cientificas on-line (SciELO, ACM , Portal CAPES, IEEExplorer4e CiteSeer). Para a incorporação no Scientia.Net foram estudadas aplicações que coletam dados desses sites.

Foi selecionada para os primeiros testes de implementação a base de artigos CiteSeer, por ser de livre acesso para o publico em geral. Dentre os mecanismos estudados para a incorporação dos artigos no Scientia.Net estão *Heritrix*, *HTML Parser* e *Jsoup*, sendo este último o escolhido para utilização por ser mais fácil de manipular e mais intuitivo.

Resultados e Discussão

Com o *Jsoup* foi possível capturar uma página através de sua URL e depois efetuar o parser em tags de conteúdo especifico. Usou-se o método *connect* passando a URL como parâmetro. Esta URL foi modificada para incluir a busca pela área de interesse do usuário cadastrado, coletando as seguintes informações dos artigos: título, autor, link e *abstract*.

Os dados coletados foram salvos em um banco de dados, e depois incorporados ao Scientia.Net para serem usados como indicações de artigos no perfil dos usuários.

Conclusão

Este trabalho propôs a incorporação no Scientia.Net dos Artigos dos principais repositórios online, para isso foram estudados os repositórios SciELO, ACM, Portal CAPES, IEEExplorer e CiteSeer. Verificou-se a possibilidade de se trabalhar somente com o CiteSeer no protótipo desenvolvido, pois este repositório tem acesso livre, ou seja, não necessita de assinatura ou convênio com alguma universidade.

Dentre os mecanismos para a incorporação estudados, o escolhido foi o Parser HTML *Jsoup*, com ele foi feita a extração dos dados de interesse do CiteSeer, e posteriormente salvo em um banco de dados para serem então incorporados ao Scientia.Net.

Apoio

Universidade Federal do Piauí

Referências

[1] jsoup: Java HTML Parser.<http://jsoup.org/>. Acesso em 1 de março de 2012.

[2] HTML Parser < http://htmlparser.sourceforge.net/ >acesso em 1 de março de 2012.

[3] Heritrix <https://webarchive.jira.com/wiki/display/Heritrix/Heritrix> acesso em 1 de março de 2012.

[4] CiteSeerX Data. Disponível em:< http://csxstatic.ist.psu.edu/about/data >Acesso em 10 de Setembro de 2011.

[5] SciELO < http://www.scielo.org > Acesso em 10 de Setembro de 2011.

[6] ACM Digital Library < http://dl.acm.org/ > Acesso em 10 de Setembro de 2011.

[7] Portal de periódicos da capes. <http://www.periodicos.capes.gov.br.ez17.periodicos.capes.gov.br/> Acesso em 10 de Setembro de 2011.

[8] IEEExplorer Digital Library. <http://ieeexplore.ieee.org/>. Acesso em 10 de Setembro de 2011.