Ciência e Tecnologia no caminho
da Cooperação Internacional

Extração automática de metadados

Oswaldo Luiz de Souza Ferreira, Fábio Duncan de Souza

IV Congresso Fluminense de Iniciação Científica e Tecnológica

17º Encontro de IC da UENF
9º Circuito de IC da IFF
5ª Jornada de IC da UFF

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

Ciência da Computação

RESUMO

As bibliotecas digitais vem sendo empregadas de forma crescente nas mais diversas áreas. Como consequência deste fato tem-se o aumento da demanda por ferramentas capazes de automatizar os processos que envolvem a submissão dos documentos para tais bibliotecas. Um exemplo de tarefa a ser realizada quando da submissão de um conteúdo para uma biblioteca digital é o cadastro dos metadados deste conteúdo. Isto se deve ao fato de que tais bibliotecas armazenam os documentos digitais e precisam de informações sobre estes para que consultas possam ser executadas e apresentadas aos usuários. Assim, torna-se fundamental a extração de metadados específicos do documento armazenado para que as informações sejam geradas de maneira a satisfazer os critérios estabelecidos. O objetivo deste projeto é criar uma ferramenta que possa extrair automaticamente tais metadados, buscando-os no corpo do texto dos documentos. Uma vez construída a ferramenta, esta será incorporada a Biblioteca Digital da EPCT (BD-EPCT), desenvolvida pelo Núcleo de Pesquisa em Sistemas de Informações do Instituto Federal Fluminense (NSI/IFF). Para a construção da primeira versão da ferramenta foi definida uma estratégia que se baseia na busca por padrões no corpo do texto de um conteúdo. Uma vez encontrado um padrão para o tipo de conteúdo avaliado, este padrão é modelado em um documento XML que serve como base para a ferramenta realizar a extração dos metadados. Esta primeira versão tem o objetivo de extrair metadados de tipos de conteúdos como monografias, dissertações e teses, considerados um dos mais relevantes para o universo da Biblioteca Digital da EPCT. Os resultados alcançados para os tipos de conteúdos avaliados foram bastante satisfatórios. O fato do tipo de conteúdo obedecer às normas da ABNT foi definitivo para o sucesso da ferramenta. No entanto, os demais tipos de conteúdos presentes na Biblioteca Digital da EPCT nem sempre obedecem a uma padronização específica, dificultando a ação da ferramenta desenvolvida. Como trabalhos futuros tem-se a análise e busca por padrões em artigos científicos. Caberá ainda investigar na literatura, estratégias para encontrar metadados em conteúdos onde a busca por padrões é dificultada, visando assim a extração automática de metadados dos demais tipos de conteúdos suportados pela BD-EPCT.

PALAVRAS CHAVE: extração, automática, metadados

ISSN 2177-6180