José Alberto Grossi Júnior

Análise Comparativa de Técnicas de Extração de Metadados em Artigos Científicos sob o Ponto de Vista do Resultado Comparativo Final

Belo Horizonte/MG, Brasil

2014, v-0.1.1

José Alberto Grossi Júnior

Análise Comparativa de Técnicas de Extração de Metadados em Artigos Científicos sob o Ponto de Vista do Resultado Comparativo Final

Dissertação de mestrado apresentada à coordenação do PPGCI/UFMG com o objetivo de obtenção de título de Mestre em Ciência da Informação

Universidade Federal de Minas Gerais – UFMG

Escola de Ciência da Informação

Programa de Pós-Graduação em Ciência da Informação

Orientador: Marcello Peixoto Bax

Belo Horizonte/MG, Brasil

2014, v-0.1.1

José Alberto Grossi Júnior

Análise Comparativa de Técnicas de Extração de Metadados em Artigos Científicos sob o Ponto de Vista do Resultado Comparativo Final/ José Alberto Grossi Júnior. – Belo Horizonte/MG, Brasil, 2014, v-0.1.1-

35 p. : il. (algumas color.) ; 30 cm.

Orientador: Marcello Peixoto Bax

Dissertação (Mestrado) – Universidade Federal de Minas Gerais – UFMG
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação, 2014, v-0.1.1.

1. Extração de informação 2. Metadados. I. Artigos científicos. II. Universidade Federal de Minas Gerais. III. Escola de Ciência da Informação. IV. Análise Comparativa de Técnicas de Extração de Metadados em Artigos Científicos sob o Ponto de Vista do Resultado Comparativo Final

CDU 02:141:005.7

José Alberto Grossi Júnior

Análise Comparativa de Técnicas de Extração de Metadados em Artigos Científicos sob o Ponto de Vista do Resultado Comparativo Final

Dissertação de mestrado apresentada à coordenação do PPGCI/UFMG com o objetivo de obtenção de título de Mestre em Ciência da Informação

Trabalho aprovado. Belo Horizonte/MG, Brasil, 20 de setembro de 2014:

Marcello Peixoto Bax
Orientador

Professor
Professor Convidado 1

Professor
Professor Convidado 2

Professor
Professor Convidado 3

Belo Horizonte/MG, Brasil
2014, v-0.1.1

Este trabalho é dedicado a todas as pessoas que desejam, de uma forma ou outra, superar seus objetivos pessoais.

Agradecimentos

A escrever.

Resumo

A necessidade de contribuição entre a comunidade acadêmica é evidente quando da necessidade de leituras específicas de artigos científicos de autores espalhados pelo mundo. Porém, esta contribuição se dá de maneira muito pessoal, com envios manuais de artigos quando da necessidade de certos nichos acadêmicos. A dificuldade apresentada geralmente é a centralização de artigos de maneira livre e compensatória, por meio de extração automática de metadados relevantes para o catálogo destes documentos, de maneira a permitir que qualquer pesquisador, devidamente reconhecido, possa compartilhar e obter estes documentos de maneira eficaz e anônima.

Este trabalho demonstra que as técnicas livres existentes para extração de metadados em artigos científicos não são suficientes para abranger os diversos formatos existentes de apresentação dos conteúdos, uma vez que são baseados em layout pré-definidos, sem possibilidade de expansão ou adaptação de acordo com a necessidade de certos grupos de pesquisa, cujo formato de apresentação deste tipo de documento se dá de maneira diferenciada, ou até mesmo, adaptada para seu universo de pesquisadores.

Palavras-chaves: artigos científicos. extração de metadados. extração de dados em artigos.

Abstract

The need of contribution existent in the academic community is focused based on the sharing of papers from authors around the world, when specific studies are needed. However, this contribution is made in a very basic and personal way, with papers sent by manual interactions from some specific research groups. The main goal is focused on the papers centralization in a free and compensatory format, by automatic relevant metadata extraction to the indexation of these documents, allowing any researcher to share and get these documents in a very effective manner.

This work shows how the existent metadata extraction techniques in scientific papers are not totally perfect to perform the different papers formats to present research works, once they are based on pre-defined layouts, without any change of customization according with some groups needs, because of a different presentation format, or even, adapted to your researchers' worlds.

Palavras-chaves: scientific papers. metadata extraction. data extraction on scientific papers.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

PDF	Portable Document Format
IEEE	Institute of Electrical and Electronics Engineers
RSL	Revisão Sistemática de Literatura

Sumário

1 Introdução

A necessidade de contribuição acontece de forma natural no ser humano. Os desejos em ajudar ao próximo e inclusive contribuir com alguma parte de sua formação é algo que desperta um desejo cada vez mais amplo do ponto de vista social.

Somos seres realizados pela satisfação do outro, e seu sucesso de uma forma ou outra acarreta em nosso sucesso, nossa satisfação pessoal e de certa forma profissional. Sentimos atraídos por contribuir e por compartilhar conhecimento, sendo ele umas das principais formas de realização como pessoa.

No âmbito acadêmico sempre contribuímos de uma forma ou outra com a formação de nossos colegas e parceiros de pesquisa. Esta contribuição pode ser feita com base em uma conversa informal ou até mesmo com uma ajuda em documentação ou sugestão de um texto para leitura. Esta sugestão de leitura geralmente possui um caráter muito técnico, e envolve na maioria dos casos a utilização de artigos acadêmicos.

Sabemos da existência de bases de conhecimento de maneira global e nacional, porém quando estamos falando da contribuição social, em pequena escala, interpessoal, estamos falando que contribuições físicas, com envio de sugestões de artigos para nossos amigos pesquisadores. Este envio é feito de maneira informal, e reduz tempo e aumenta consequentemente a praticidade do processo de pesquisa.

Sendo assim, esta experiência como objetivo global seria uma ferramenta poderosa de apoio à pesquisa, com pesquisadores compartilhando conhecimentos de maneira informal, anônima, e segura. Esta forma de disseminação de conhecimento traria um benefício muito grande socialmente falando, uma vez que pesquisadores iriam se unir, mesmo que virtualmente, na transmissão de conhecimento entre si próprios, fazendo do processo de pesquisa um processo mais focado e evitando o desperdício de tempo durante a fase de pesquisa e busca por conhecimento.

Para isso, a utilização de técnicas de extração de metadados deve ser utilizada de maneira eficaz, para que de maneira automática diversos artigos sejam analisados e catalogados em pequenos universos de pesquisa. Entende-se por metadados os campos básicos e necessários para que uma pesquisa por nome, por exemplo, seja feita com sucesso. Resume-se então que os metadados que esperam-se ser extraídos destes artigos são: o título do artigo, o nome e e-mail de seus autores, o resumo/abstract e as referências utilizadas.

Basicamente estes campos já permitem que uma pesquisa mais detalhada fosse feita e então o artigo localizado. Já as referências são necessárias para se fazer referências inversas de autores que publicam e são citados posteriormente, facilitando ainda mais aos

pesquisadores poder, por exemplo, encontrar artigos semelhantes de uma mesma área do conhecimento.

1.1 Delimitação do Problema

De modo geral, as técnicas livre existentes para que essa extração de metadados seja feita são focadas em layouts pré-definidos, geralmente de conferências e/ou congressos internacionais, que possuem um padrão visual parecido, como é o caso do IEEE por exemplo, que segue de referência para diversos outros eventos tomando seu layout como base.

Porém, existem diversos outros eventos que possuem layouts de artigos considerados fora do padrão e, portanto, necessitam de adaptações destas técnicas para que seus trabalhos possam ser analisados e catalogados de maneira eficaz. Esta customização promoveria uma série de tentativas para verificar o melhor layout para ser utilizado em cada caso, automaticamente.

1.2 Objetivo Geral

Este trabalho possui como objetivo geral provar que as técnicas livres de extração automática de metadados em artigos científicos ainda necessitam ajustes e principalmente flexibilidade para abranger um maior números de documentos e prover então uma contribuição maior perante a comunidade acadêmica.

A necessidade de customização é uma tendência natural de qualquer ramo de atividade, de maneira a promover possibilidades de ferramentas auto-suficientes capazes de suprir as necessidades de grupos específicos de pesquisas, de eventos ou conferências, que possui padrões de apresentação de artigos personalizados e que demandam de uma análise diferenciada para que possa ser indexada e então analisada por sistemas de informação.

1.2.1 Objetivos Específicos

Com base na diferenciação de formas de apresentação de artigos científicos este trabalho tem como Objetivos Específicos identificar pontos em que técnicas de extração de metadados necessitam de adaptações flexíveis por parte da comunidade em geral, permitindo que artigos sejam analisados de maneira diferente em virtude de especificações distintas e necessidades diferenciadas de grupos de pesquisa.

Os padrões existentes no mercado são de maneira geral insuficientes para suprir as necessidades dos mais diversos eventos e/ou conferências existentes, afunilando a apenas

uma pequena parcela de artigos, o que acaba gerando um desconforto e uma ineficácia das técnicas de extração de metadados existentes atualmente.

1.3 Resultados Esperados

As formas de extração de dados em artigos científicos são geralmente baseadas em layouts, ou seja, em pequenos pedaços onde certas informações devem ser informadas. Porém em virtude da grande diversidade de materiais produzidos e em função das adaptações realizadas por grupos e/ou eventos de pesquisa, este layout padrão não se mostra eficiente na abrangência total das necessidades do meio.

Assim sendo, espera-se que certos artigos científicos não tenham seus metadados analisados de maneira eficaz por todas as técnicas livres existentes de extração de dados, uma vez que adaptações são necessárias a fim de contribuir para uma globalização destas análises, permitindo a customização então de técnicas de extração com base em mercados ou culturas diferentes.

1.4 Limitações do Trabalho

Este trabalho limita-se aos artigos científicos difundidos na comunidade científica em formato PDF, excluindo aqueles em que seu conteúdo é disponibilizados através de imagens escaneadas de documentos físicos, o que impede, em um primeiro momento, de ter os textos analisados em sua forma original, sem necessidade de processamento extra a fim de obter todo o material textual contido em tais imagens.

Além disso o trabalho pressupões que a língua inglesa seja utilizada como padrão no meio, de maneira a permitir que através de um único idioma o conhecimento seja difundido e aplicado em diversas culturas, independente de especificidades e diferenças culturais, permitindo uma difusão do conhecimento em sua mais pura forma de apresentação.

1.5 Justificativa

De maneira geral, a necessidade de centralizar estes artigos científicos existe, e a contribuição seria uma forma de aumentar cada vez mais o acesso aos materiais de pesquisa. Sendo assim, esta forma de análise e extração de matadados traria benefícios para que este repositório fosse criado, tendo então milhões e milhões de documentos em suas bases de dados.

Este trabalho é feito justamente para prover esta visão do que ainda precisa ser melhorado e pensado para que estas técnicas abranjam diversos padrões encontrados no mercado, permitindo além que usuários possam contribuir com seus próprios padrões.

1.6 Estrutura

Esta pesquisa é estruturada iniciando com uma introdução sobre o tema, a definição do problema, os objetivos gerais e específicos e sua justificativa.

O segundo capítulo tem como base o referencial teórico feito através de uma RSL (Revisão Sistemática de Literatura), tendo como base (KITCHENHAM, 2004), que propõe um passo-a-passo para um revisão de literatura eficaz e atingindo os resultados desejáveis pela pesquisa.

No terceiro capítulo temos a metodologia para o desenvolvimento do trabalho, as técnicas que serão aplicadas e principalmente como serão feitas. Posteriormente, no capítulo quarto temos os testes propriamente ditos, como eles foram realizados, os ambientes de teste, a seleção de artigos para testes e no quinto capítulo os resultados obtidos.

No sexto capítulo temos a conclusão, trabalhos futuros e considerações finais sobre o trabalho apresentado.

2 Revisão de Literatura

3 Metodologia

4 Testes

4.1 Ambiente de Testes

4.1.1 Servidores de Teste

5 Resultados

6 Conclusão

6.1 Trabalhos Futuros

6.2 Considerações Finais

Referências

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. [S.l.], 2004. Citado na página 24.