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-
39 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

Figura 1 – Processo de Extração de Metadados 22
Figura 2 – Processo de Metodologia 27

Lista de tabelas

Tabela 1 – Seleção de artigos científicos para testes (ainda em desenvolvimento) . 29
Tabela 2 – Os metadados e seus pesos atribuídos 31
Tabela 3 – Resultados obtidos em cada metadado e sua precisão 31

Lista de abreviaturas e siglas

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

Sumário

1 INTRODUÇÃO 21
1.1 Delimitação do Problema 22
1.2 Objetivo Geral 22
1.2.1 Objetivos Específicos 23
1.3 Resultados Esperados 23
1.4 Limitações do Trabalho 23
1.5 Justificativa 24
1.6 Estrutura 24

2 REVISÃO DE LITERATURA 25

3 METODOLOGIA 27
3.1 Seleção das Técnicas 27
3.2 Seleção de Artigos 28
3.3 Infraestrutura Computacional 28
3.3.1 Testes In Loco 29
3.4 Quadro Comparativo 30
3.5 Metadados, Pesos e Resultados 30
3.5.1 Índice de Confiabilidade 31

4 TESTES 33
4.1 Ambiente de Testes 33
4.1.1 Servidores de Teste 33

5 RESULTADOS 35

6 CONCLUSÃO 37
6.1 Trabalhos Futuros 37
6.2 Considerações Finais 37

Referências 39

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.

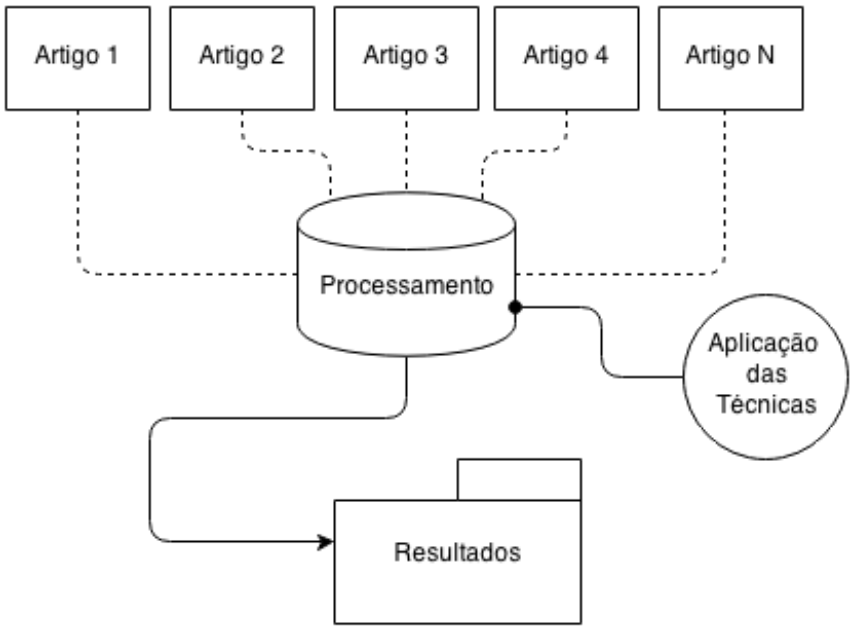
Sabemos da existência de bases de conhecimento de maneira global e nacional, porém quando estamos falando da contribuição social, em pequena escala, interpessoal, estamos falando que contribuições físicas, com envio de sugestões de artigos para nossos amigos pesquisadores. Este envio é feito de maneira 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

Figura 1 – Processo de Extração de Metadados



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 pressupõe 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.

Já na questão de testes de cada técnica de extração de metadados, as técnicas que serão selecionadas deverão ser livres, ou seja, ter seu uso liberado sem a necessidade de pagamento de licenças. Deste modo excluímos todas as técnicas que rodam exclusivamente em plataforma Windows, por exigir licenças de software e fugirem das previsões de teste deste projeto. Assim, os projetos deverão necessariamente utilizar de linguagens de programação livres (ou de código aberto) e que rodem em sistemas operacionais derivados do Unix, como o Linux, por exemplo.

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 metadados 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

Este trabalho tem como metodologia uma pesquisa de caráter não-experimental e quantitativa, por se tratar de coleta de informações e comparação de resultados de técnicas de extração de metadados em artigos científicos.

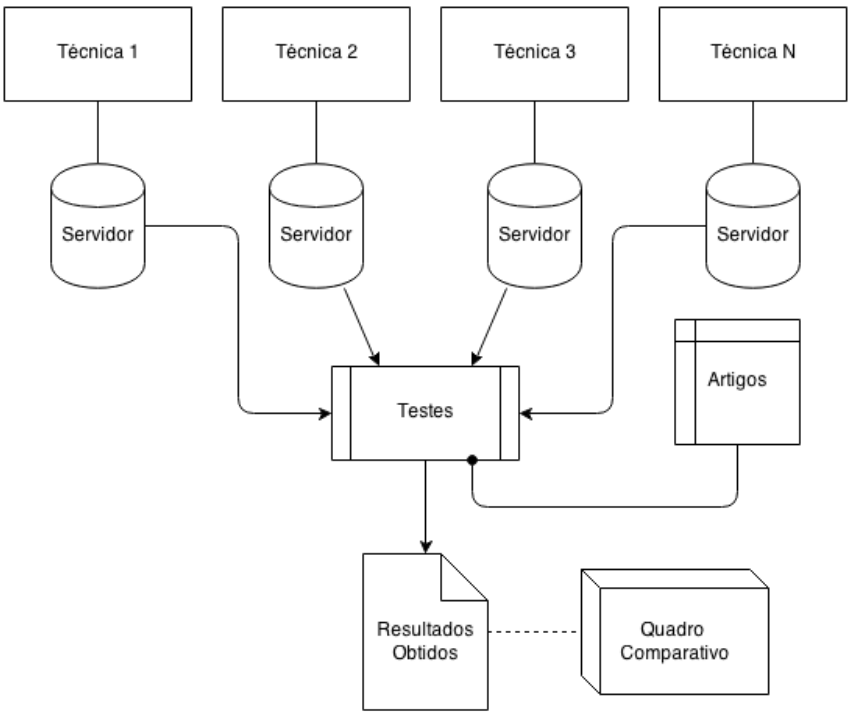
Desta maneira, a pesquisa de modo padrão não traz alteração nos ambientes pesquisados, apenas os analisa e os compara com base em padrões estabelecidos como sendo de resultado adequado. Assim, o projeto em si trata muito mais de pré-seleção de documentos e técnicas a serem testadas como também dos resultados já analisados e que são necessários para uma técnica ser considerada produtiva.

Primeiramente, são filtradas as técnicas encontradas a fim de analisar realmente as que são necessárias dentro do objetivo da pesquisa, fazendo do projeto o mais conciso possível. Desta forma, diversos elementos serão utilizados a fim de se obter os resultados desejados. Os elementos se relacionam entre si e estão identificados na figura 2.

3.1 Seleção das Técnicas

As técnicas que foram selecionadas dentro do capítulo de Revisão de Literatura compreendem um universo atuante e de caráter livre, independente da linguagem de

Figura 2 – Processo de Metodologia



programação que foi utilizada, exceto pelos detalhes explicados nas limitações do trabalho.

Assim sendo, todas as técnicas catalogadas e definidas como sendo importantes serão testadas de maneira completa e independente, ou seja, sem interferência de nenhuma outra técnica em seus resultados finais.

3.2 Seleção de Artigos

Visando provar a eficiência destas técnicas, desejamos ter informações de saída consideradas corretas para que seus resultados possam ser comparados e verificados com exatidão. Assim, foi selecionada uma série de artigos científicos das mais diversas áreas de pesquisa, de diversos eventos distintos, com padrões visuais totalmente diferentes e que são possível de ser analisados e coletados.

Desde modo a lista destes artigos compreende um total de 100 artigos variados, com seus metadados já extraídos manualmente e todos catalogados a fim de terem o resultado da aplicação das técnicas comparado com os resultados desejados. As áreas do conhecimento, seus respectivos eventos e o número de artigos de cada um estão identificados na Tabela 1.

Os artigos foram selecionados tomando como base a principal forma de análise das técnicas selecionadas para testes: o layout, ou seja, a disposição dos elementos nos artigos científicos. Esta seleção foi feite com base a abranger um maior número de representações, com disposições diferentes, tipografias diferentes e inclusive ordem diferentes nas exibições. Desta forma as técnicas poderão ser confrontadas e os resultados comparados com os resultados esperados.

Todos os artigos selecionados foram escritos na língua inglesa. Esta decisão foi tomada em virtude de além de ser a língua inglesa universal para disseminação de conhecimento, ela é a mais utilizada no meio acadêmico, de maneira a ter um universo muito maior de artigos escrito em inglês do que outras línguas.

Além disso a abrangência de outros idiomas entraria em um aspecto que não é objetivo deste trabalho abordar, visto a diversificação de culturas e símbolos, fazendo com que línguas orientais, como o mandarim ou japonês, tenham análises diferenciadas em função de suas diferenças na forma de representação.

3.3 Infraestrutura Computacional

Para que os testes sejam feitos de maneira adequada e independente, sem interferência de outras técnicas nos resultados, serão utilizados N servidores, sendo N o número total de técnicas a serem avaliadas.

Tabela 1 – Seleção de artigos científicos para testes (ainda em desenvolvimento)

Área de Conhecimento	Evento ou Conferência	Total de Artigos
Arquitetura e Urbanismo	Nome do Evento I	4
Arquitetura e Urbanismo	Nome do Evento II	3
Artes e Música	Nome do Evento I	3
Artes e Música	Nome do Evento II	2
Ciência da Computação	Nome do Evento I	2
Ciência da Computação	Nome do Evento II	4
Ciência da Informação	ENANCIB 2013	4
Ciência da Informação	Nome do Evento	3
Ciências Biológicas	Nome do Evento I	4
Direito	Nome do Evento I	3
Engenharia Civil	Nome do Evento I	2
Medicina	Nome do Evento I	2
Medicina	Nome do Evento II	3
Psicologia	Nome do Evento I	3
...	...	...
		100

Deste modo, para cada técnica avaliada será configurado um servidor com as linguagens de programação necessárias e todos os pré-requisitos que a técnica necessita para funcionar. Estes servidores serão definidos utilizando infraestrutura em **nuvem** (*Cloud Computing*), o que traz benefícios não somente de performance mas de flexibilidade quanto das tecnologias necessárias para o funcionamento de cada técnica.

Todos os servidores criados para os testes devem necessariamente rodar alguma versão do Linux e serão hospedados nos *data centers* da **Digital Ocean**[1], empresa de infraestrutura tecnológica de conhecimento mundial e referência em *Cloud Computing* no mercado computacional.

3.3.1 Testes In Loco

Algumas técnicas de extração de metadados disponibilizam acesso online a uma ferramenta gratuita para testes. Deste modo, para estes casos específicos não será necessária a criação de servidores e instalação dos pacotes, visto que o ambiente de testes poderá ser feito dentro da ferramente fornecida pelas desenvolvedoras das técnicas.

Este fato garante uma maior precisão nos resultados, inclusive pelo fato de o ambiente de testes estar 100% funcionando e ter sido disponibilizado pela mesma equipe de desenvolvedores do projeto, garantindo a eficácia nos resultados que essa técnicas fornece.

[1] Acesse o site em <http://digitalocean.com>

3.4 Quadro Comparativo

Visando uma comparação dos resultados eficaz todos os resultados serão inseridos em um documento formato planilha para serem comparados manualmente e as conclusões então obtidas. Para tal esta planilha de resultados a chamaremos de "Quadro Comparativo"e será exclusiva para comparação dos resultados das técnicas analisadas.

Com o objetivo de facilitar o acesso às informações este documento é disponibilizado como anexo deste projeto e ainda tem seu acesso liberado em formato online. Desta forma qualquer pessoa poderá ter acesso aos resultados obtidos pelos testes e comparar elas mesmas os dados contidos na planilha.

3.5 Metadados, Pesos e Resultados

A extração de metadados de artigos científicos engloba um processo onde os resultados obtidos, mais especificamente os metadados propriamente ditos, possuem características diferenciadas que podem influenciar em uma busca por artigos, feita por um pesquisador.

Desde modo atribuímos pesos para cada um dos metadados analisados, de maneira a identificar os mais importantes e que podem contribuir com um número maior de resultados de busca.

Alguns metadados são mais importantes que outros no que diz respeito à funcionalidade de pesquisa. Geralmente quando vamos buscar artigos, seja na Internet, ou em algum outro local, geralmente buscamos primeiro pelo título do artigo (quando procuramos por um artigo em específico) ou então pelo nome do autor (quanto procuramos artigos de um determinado autor).

Além disso, utilizamos também o título, juntamente com o resumo, para buscar de palavras chaves ou palavras que podem ser relevantes na pesquisa pelos documentos. Assim sendo alguns metadados devem ser mais considerados no resultado destas extrações, por serem mais importantes no ponto de vista da busca.

Assim sendo apresentamos a tabela 2, que demonstra como cada metadado teve sua importância interpretada e qual o peso que lhe foi atribuído, sendo utilizado o inteiro 1 para o peso mais baixo e o 5 para peso mais alto, sendo consequentemente o mais importantes.

Outro detalhe importante é a precisão de cada resultado para cada metadado analisado. Em alguns casos o título, por exemplo, não é extraído em 100% mas alguma variação dele.

Deste modo consideramos 3 (três) resultados possíveis para um resultado analisado:

Tabela 2 – Os metadados e seus pesos atribuídos

Metadado	**Relevância**	**Peso**
Título	Um dos termos mais buscados quando se pesquisa um artigo	5
Autor(es)	O segundo termo mais pesquisado	4
E-mail(s)	Pouco relevante no quesito pesquisa de artigos	1
Resumo	Importante por conter palavras chaves e o resumo propriamente dito	3
Referências	Muito importante e necessário, pois será utilizada na referência inversa de autores	5

Tabela 3 – Resultados obtidos em cada metadado e sua precisão

Resultado	**Precisão**
Preciso	1
Satisfatório	0.60
Inaceitável	0

1. **Preciso:** Quando um resultado atinge acima de 95% de precisão, ou seja, o campo foi extraído em 95% ou mais de sua totalidade.

2. **Satisfatório:** Quando um resultado atinge entre 90 e 94%, o que pode ser considerado satisfatório e a maioria do conteúdo consegue ser analisada sem maiores problemas.

3. **Inaceitável:** Quando o resultado atinge abaixo de 90%, ou seja, entre 0% e 89%. Este resultado no âmbito do presente projeto é considerado inaceitável.

Assim, temos o valor de cada resultado possível, que será também utilizado no processo de análise, conforme consta na tabela 3.

3.5.1 Índice de Confiabilidade

Considerando que cada metadado possui um peso diferente necessitamos calcular o índice de acertos a ser utilizado em cada resultado coletado para cada técnica aplicada. Assim sendo chegamos em uma fórmula matemática à qual chamamos "Índice de Confiabilidade", que calcula o resultado obtido através dos pesos que foram atribuídos.

Este índice utiliza os pesos anteriormente definidos e a precisão dos resultados obtida, de maneira a permitir chegar em um único resultado para cada técnica aplicada.

$$Fórmula a ser definida ainda$$

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.