José Alberto Grossi Júnior

Análise Comparativa de Ferramentas de Extração de Metadados em Artigos Científicos

Belo Horizonte/MG, Brasil

Out 2015, v-0.8.0

José Alberto Grossi Júnior

Análise Comparativa de Ferramentas de Extração de Metadados em Artigos Científicos

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

Coorientador: Renato Rocha Souza

Belo Horizonte/MG, Brasil

Out 2015, v-0.8.0

José Alberto Grossi Júnior
Análise Comparativa de Ferramentas de Extração de Metadados em Artigos Científicos/ José Alberto Grossi Júnior. – Belo Horizonte/MG, Brasil, Out 2015, v-0.8.0-
40 p. : il. (algumas color.) ; 30 cm.

Orientador: Marcello Peixoto Bax
Coorientador: Renato Rocha Souza

Dissertação (Mestrado) – Universidade Federal de Minas Gerais – UFMG
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação, Out 2015, v-0.8.0.

1. Extração de metadados 2. Metadados. I. Artigos científicos. II. Extração de informação. III. Ciência da Informação. IV. Análise Comparativa de Ferramentas de Extração de Metadados em Artigos Científicos

CDU 02:141:005.7

José Alberto Grossi Júnior

Análise Comparativa de Ferramentas de Extração de Metadados em Artigos Científicos

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, 7 de outubro de 2015:

Marcello Peixoto Bax
Orientador

Renato Rocha Souza
Coorientador

Beatriz Valadares Cendón
Professora Convidada - ECI/UFMG

Max Cirino de Mattos
Professor Convidado - UNA

Renata M. Abrantes Baracho Porto
Professora Suplente - ECI/UFMG

Belo Horizonte/MG, Brasil
Out 2015, v-0.8.0

Resumo

São inúmeras as ferramentas para extração de metadados em artigos científicos, tendo cada uma sua particularidade, tecnologia e técnicas utilizadas. Porém, com a crescente produção científica e a grande variedade de editoras, eventos e congressos, um grande número de artigos permanece sem uma extração de metadados eficaz, o que dificulta a disseminação de conhecimento e principalmente a pesquisa eletrônica destes documentos.

Este trabalho realiza um teste com algumas ferramentas pré-selecionadas com um conjunto pré-determinado de artigos, que abrange diversas áreas do conhecimento, diversos eventos e formatos visuais diferentes. Estes testes são realizados em ambientes pré-configurados de acordo com a necessidade tecnológica de cada ferramenta, permitindo que todos os artigos tenham seus metadados extraídos por cada uma delas e seus resultados comparados individualmente.

Desta forma, com base nos resultados apresentados, pode-se identificar o comportamento de cada uma das ferramentas perante à extração de metadados, suas falhas, onde são necessários ajustes e onde se obtém um maior sucesso na extração. Além disso, é apresentado também o Índice de Confiabilidade, onde cada ferramenta recebe uma nota com base nos resultados obtidos na extração de metadados pela seleção de artigos realizada.

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

Abstract

Currently we can find numerous tools to extract metadata from scientific papers, each one with your particularity, technology and techniques. However, with the crescent scientific production and the numerous publishers, events and conferences, a large part of papers still remain without an effective metadata extraction, hindering the knowledge dissemination e mainly the electronic search for these documents.

The present work makes tests with pre selected tools with a set of scientific papers, covering different areas of knowledge, different events and layouts. These tests were made inside custom environments according the technologies each tool needs, allowing all papers to be tested and their metadata extracted, comparing results one by one.

Thereby, according the presented results, we can identify the behavior of each tool related to the metadata extraction, where it failed, where adjusts are needed and where it has success on the extraction. Moreover, we also present the Reliability Index, a grade received by each tool based on the metadata extraction results using the selected scientific papers.

Palavras-chaves: scientific papers, metadata extraction, data extraction in 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
ACM	Association for Computing Machinery
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
XML	eXtensible Markup Language
SVM	Support Vector Machines
HMM	Hidden Markov Models
CRF	Conditional Random Fields
URL	Uniform Resource Locators
HTML	HyperText Markup Language
DCMI	Dublin Core Metadata Initiative
SVM	Support Vector Machines
DOI	Digital Object Identifier
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
UFMG	Universidade Federal de Minas Gerais

Sumário

1 Introdução

Em virtude da grande produção científica existente nos dias atuais, ferramentas automatizadas de extração de metadados de artigos científicos são cada vez mais úteis. Elas contribuem para uma melhor organização dos artigos científicos e facilitam buscas mais rápidos e eficientes.

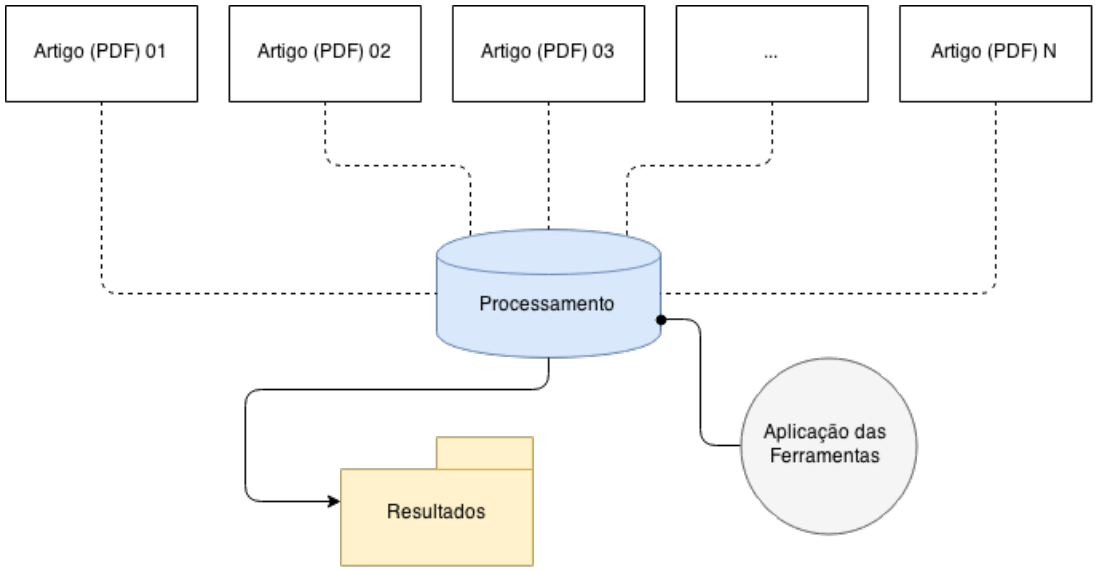
A pesquisa aqui realizada situa-se no campo da extração de metadados segundo a abordagem *machine learning*. O trabalho considera as ferramentas mais populares atualmente. Diversas ferramentas e técnicas para extração de metadados em artigos podem ser encontradas na literatura científica da área.

Algumas ferramentas são propriedades de universidades ou instituições privadas, o que dificulta a análise. Outras não permitem que testes automatizados sejam feitos, visto que não há acesso ao código fonte ou não podem ser utilizadas via linha de comando.

De modo geral, as ferramentas de extração são focadas em leiautes pré-definidos, geralmente seguindo modelos (ou *templates*) de revistas e encontros científicos, que possuem um padrão visual já estabelecido. Esse é o caso do IEEE (*Institute of Electrical and Electronics Engineers*), por exemplo, que serve de referência para diversos outros eventos da área da Ciência da Computação.

Porém, existem diversos outros eventos e revistas que empregam *templates* específicos. A extração nesses artigos exige adaptações das ferramentas.

Figura 1 – Processo de Extração de Metadados



Fonte: O próprio autor

Algumas ferramentas são aparentemente muito eficazes para um certo grupo de artigos, já seguindo um padrão visual pré-determinado. Porém, para alguns *templates* pouco comuns, de áreas de conhecimento diversas, elas não são tão eficazes. A eficácia varia de acordo com a tecnologia utilizada e, principalmente, de acordo com o princípio teórico utilizado.

Como definido por (MOHRI; ROSTAMIZADEH; TALWALKAR, 2012), *machine learning* permite uma forma de aprendizado com base em experiências passadas, através da utilização de dados coletados, que são analisados posteriormente seguindo padrões definidos.

A área é muito ampla e sua aplicabilidade é diversificada, podendo ser usada na classificação, processamento de linguagem natural, reconhecimento de fala, detecção de fraudes, diagnósticos médicos e sistemas de recomendações, além de mecanismos de buscas e extração de informação. Essa última é a aplicação foco neste trabalho.

Claro que as técnicas de extração existentes hoje são, de maneira geral, insuficientes para tratar todos os leiautes de artigos existentes, limitando-se a apenas uma parcela destes, que usam padrões visuais comuns. Espera-se que certos artigos científicos não tenham seus metadados extraídos com total exatidão.

Com base na diferenciação dos leiautes de artigos científicos, o objetivo da pesquisa é comparar o desempenho de ferramentas na tarefa de extração de metadados. Isso será feito com conjunto de documentos pré-selecionados para testes, dos mais diversos padrões e de diversas áreas do conhecimento.

Espera-se com isso poder identificar o desempenho de tais ferramentas, suas limitações e melhores aplicações: quais ferramentas apresentam melhores resultados para cada padrão visual? Que ferramenta é melhor aplicada para determinado tipo distinto de metadado?

O documento é estruturado iniciando com essa breve introdução e motivação sobre o tema.

O segundo capítulo traz o referencial teórico, onde são apresentados alguns conceitos básicos, além das técnicas mais utilizadas e as ferramentas mais comuns encontradas atualmente.

O terceiro capítulo apresenta a metodologia usada no trabalho, citando as ferramentas que serão testadas e principalmente o método usado nesta pesquisa para a realização dos testes.

Posteriormente, no capítulo quarto, faz-se a análise e apresentação dos resultados, explicando como os testes foram realizados, os ambientes de teste criados e os resultados coletados.

No quinto capítulo temos a discussão final e a exposição de algumas conclusões mais relevantes, além dos trabalhos futuros e considerações finais sobre o trabalho apresentado.

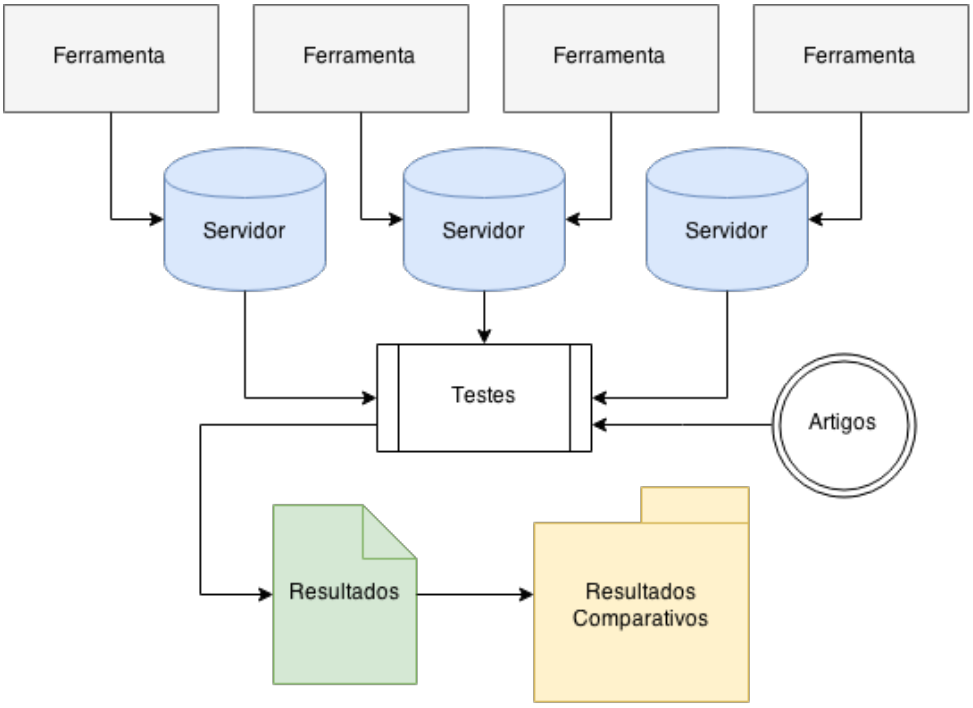
2 Metodologia

Este trabalho tem como metodologia uma pesquisa de caráter não-experimental e quantitativa, por se tratar de extração automática de metadados por ferramentas previamente selecionadas, tendo os resultados comparados com a extração manual do mesmo conjunto de artigos científicos.

Primeiramente, são filtradas as ferramentas encontradas a fim de analisar realmente as que possuem viabilidade técnica de testes dentro do objetivo da pesquisa. Desta forma, diversos elementos serão utilizados a fim de obter os resultados desejados.

De modo geral o procedimento de testes deste trabalho será realizado através da instalação e execução de cada ferramenta selecionada, permitindo que cada uma tenha seu conjunto necessário de tecnologias para seu correto funcionamento. Assim, os artigos selecionados para testes serão utilizados em cada uma destas ferramentas e seus resultados analisados, centralizados e consolidados a fim de se obter uma conclusão para cada análise encontrada. O fluxo de passos necessários para a realização destes testes pode ser melhor visualizado na Figura 2.

Figura 2 – Processo de Metodologia utilizado



Fonte: O próprio autor

Tabela 1 – Áreas do Conhecimento (CNPq)

Áreas do Conhecimento	**Subáreas**
Ciências Agrárias	157
Ciências Biológicas	104
Ciências da Saúde	76
Ciências Exatas e da Terra	243
Ciências Humanas	163
Ciências Sociais Aplicadas	185
Engenharias	305
Lingüística, Letras e Artes	53
Outros	4
Total	**1288**

Fonte: <http://www.memoria.cnpq.br/areasconhecimento/index.htm>

2.1 Escolha do Corpus

Visando provar a eficiência das ferramentas - juntamente com a implementação das técnicas por elas utilizadas -, desejamos ter resultados exatos da extração de metadados, de maneira que possam ser comparados e verificados com os metadados manualmente extraídos. Deste modo, foi selecionada uma série de artigos científicos das mais diversas áreas de pesquisa, com padrões visuais distintos.

Em virtude da necessidade de realizar testes com as ferramentas selecionadas em um ambiente real e representativo, foi realizada uma pesquisa no site do CNPq <http://www.cnpq.br/> a fim de obter a relação das áreas e subáreas do conhecimento reconhecidas oficialmente. Deste modo foi constatada a existência de 9 (nove) áreas do conhecimento, totalizando 1.288 (mil duzentas e oitenta e oito) subáreas, conforme pode ser verificado na Tabela 1.

Com base no grande número de subdivisões de cada área do conhecimento (vide Tabela 1), a seleção das subáreas será limitada a 2 (duas), sendo a escolha feita com base na existência de curso de graduação e/ou departamento na Universidade Federal de Minas Gerais (UFMG), o que facilitaria o contato com professores e coordenadores dos respectivos cursos. Desta forma, excluindo-se a área "Outros", seriam analisadas 16 (dezesseis) subáreas, sendo 2 (duas) para cada uma das 8 (oito) áreas do conhecimento.

Com as subáreas selecionadas, foi realizada uma entrevista com professores e/ou coordenadores de cada curso ou departamento correspondente na UFMG, obtendo então as bases de dados e/ou revistas mais utilizadas e relevantes para cada subárea do conhecimento, construindo um *corpus* realmente significativo. A relação dos professores entrevistados e as subáreas do conhecimento selecionadas pode ser vista na Tabela 2.

Para cada uma das subáreas do conhecimento selecionadas foram coletados 7 (sete)

Tabela 2 – Professores entrevistados para cada subárea do conhecimento.

Subárea do Conhecimento	**Professor(a)**
Arquitetura e Urbanismo	Eleonora Sad de Assis
Ciência da Computação	Adriano Alonso Veloso Alberto Henrique Frade Laender Arnaldo de Albuquerque Araújo
Ciência da Informação	Beatriz Valadares Cendón
Ciências Biológicas (Genética)	Mônica Bucciarelli Rodriguez
Ciências Biológicas (Zoologia)	Alfredo Hannemann Wieloch
Enfermagem	Eline Lima Borges Tania Couto Machado Chianca Daclé Vilma Carvalho
Engenharia Civil	Antônio Neves de Carvalho Júnior
Engenharia Mecânica	Antônio Eustáquio de Melo Pertence Alexandre Mendes Abrao
Fonoaudiologia	Sirley Alves da Silva Carvalho
Geologia	Adolf Heinrich Horn
História	Adriana Romeiro
Letras	Adriana Silvina Pagano
Medicina Veterinária	João Paulo Amaral Haddad Jenner Karlisson Pimenta dos Reis
Música	João Pedro Paiva de Oliveira
Psicologia	Carmen Elvira Flores-Mendoza Prado Livia de Oliveira Borges
Zootecnia	Ângela Maria Quintão Lana

artigos científicos. Em virtude da diversidade de quantidade de bases de dados informadas pelos professores de cada subárea do conhecimento (Tabela 3), a pesquisa foi limitada a 2 (duas) bases, na ordem apresentada pelos próprios pesquisadores, contemplando 4 (quatro) artigos para a primeira base e 3 (três) para a segunda. Para a subárea "Ciências Biológicas (Genética)" somente uma base de dados será utilizada, portanto dela serão retirados os 7 artigos necessários. No total foram selecionados 112 (cento e doze) artigos, contemplando 16 subáreas do conhecimento e 32 bases de dados, formando então o *corpus* utilizado como base desta pesquisa.

A seleção dos artigos em cada base de dados foi feita de maneira arbitrária, levando em consideração diferenças de leiautes e posicionamento dos elementos, permitindo que uma maior variedade de documentos seja analisada.

Todos os artigos selecionados foram escritos na língua inglesa. Esta decisão foi tomada em virtude de, além de ser a língua inglesa a universal para disseminação de conhecimento, ela é a mais utilizada no meio acadêmico, possuindo um universo muito maior e mais rico de artigos escritos no idioma. Além disso, algumas das ferramentas e

Tabela 3 – Bases de Dados informadas pelos professores entrevistados, por subárea do conhecimento.

Subárea do Conhecimento	**Bases de Dados**
Arquitetura e Urbanismo	Scielo, Web of Science, Scopus
Ciência da Computação	DBLP, ACM Digital Library, IEEE Xplore
Ciência da Informação	LISA, ISTA, LISTA
Ciências Biológicas (Genética)	PubMed
Ciências Biológicas (Zoologia)	Zoological Records, Biological Abstracts
Enfermagem	MedLine, Lilacs, CINAHL, EBSCO, IBECS, BDENF
Engenharia Civil	Construction and Building Materials (ELSEVIER), Cement and Concrete Composites (ELSEVIER), Composites Science and Technology (ScienceDirect), Cement and Concrete Research (ELSEVIER), Materials Research (Scielo)
Engenharia Mecânica	Scopus, ScienceDirect, Web of Science, SpringerLink, Elsevier, Research Gate
Fonoaudiologia	Pubmed, Bireme
Geologia	Springer, Scielo, Portal CAPES
História	Scielo, Jstor, Redalyc
Letras	Delta (Scielo), Periódicos Letras UFMG, Periódicos UFSC
Medicina Veterinária	PubMed, Scielo
Música	RISM, RILM, JSTOR, Grove Dictionary of Music
Psicologia	Scopus, PsycInfo, Scopus, Psicodoc
Zootecnia	Dairy Science, Animal, Poutry Science

respectivas técnicas utilizadas nos testes utilizam de "processamento de linguagem natural" para extração dos metadados, tendo por padrão a utilização do inglês na análise dos textos dos documentos.

Em virtude dessas colocações 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 por exemplo, tenham análises diferenciadas em função de suas diferenças nas formas de representação e leitura, necessitando de outras técnicas e/ou ferramentas mais direcionadas a fim de obter os resultados esperados.

No que tange a escolha das ferramentas para testes foi utilizado apenas um ponto na seleção: a sua utilização por linha de comando (*command line*). Embora algumas ferramentas possuem código aberto a extração de metadados faz parte de um contexto específico da aplicação, dificultando a utilização de apenas este recurso. Assim, foram

selecionadas para testes apenas as ferramentas que permitem o uso de sua funcionalidade de extração de metadados de maneira individual, independente da linguagem de programação ou tecnologia apresentada.

Deste modo, dentre as ferramentas apresentadas neste trabalho, presentes na **??**, as ferramentas selecionadas para teste foram: Cermine, CiteSeer, CrossRef e ParsCit.

2.2 Desenho do Experimento

Tendo selecionadas as ferramentas e os artigos que serão utilizados para os testes, parte-se para a instalação adequada de cada ferramenta, juntamente com as tecnologias necessárias e as linguagens de programação utilizadas pelos seus desenvolvedores.

Cada ferramenta será testada em separado, observando suas características particulares. Assim, cada artigo selecionado será testado para aquela ferramenta, anotando os resultados obtidos na extração. Estes resultados serão separados por metadados, o que permitirá calcular qual a porcentagem de acerto que cada ferramenta teve na extração de cada metadado analisado.

Assim, o processo será repetido para cada ferramenta e o resultado registrado, permitindo calcular sua porcentagem total de acertos de maneira simplificada. Para isso será criado um "Quadro Comparativo", no qual serão inseridos os resultados dos testes de cada ferramenta.

2.2.1 Metadados, Pesos e Resultados

Em se tratando de pesquisa por artigos científicos, pequenos detalhes podem fazer diferença. Dessa forma, uma extração de metadados não muito eficaz pode prejudicar direta ou indiretamente os resultados da busca. Por outro lado, alguns metadados tendem a ser mais utilizados na pesquisa que outros, o que implica em uma responsabilidade maior na eficiência de sua extração.

Geralmente quando vamos buscar artigos, procuramos primeiro pelo título (quando procuramos por um documento específico) ou então pelo nome do autor (quanto procuramos por artigos de um determinado pesquisador). Assim serão atribuídos pesos para cada um dos metadados, de maneira a valorizar essas informações que influenciam diretamente os resultados de busca.

A Tabela 4 mostra como cada metadado será interpretado e qual o peso que lhe será atribuído, sendo utilizado o inteiro 1 (um) para o peso mais baixo e o 5 (cinco) para o peso mais alto, sendo consequentemente o(s) metadado(s) mais importante(s). Os pesos utilizados, assim como a ordem de importância escolhida se fundamentam apenas na experiência do autor.

Tabela 4 – Os metadados e seus pesos atribuídos

Metadado	**Relevância**	**Peso**
Título	Um dos termos mais buscados quando se pesquisa um artigo	5
Autor(es)	Outro termo muito utilizado na busca por artigos	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	4

Como a extração de um metadado nem sempre ocorre de maneira 100% eficaz, visando uma avaliação mais detalhada de cada ferramenta, será calculada a precisão do resultado da extração de cada metadado, feita com base na porcentagem de sucesso obtida para aquele conjunto de caracteres. Este cálculo será feito com o uso da função `similar_text` da linguagem de programação PHP <http://php.net/similar_text>, que calcula a porcentagem de similaridade entre dois textos de acordo com o algoritmo proposto por Oliver (OLIVER, 1993). Assim, serão comparados:

1. O dado correto, retirado manualmente dos artigos;
2. O dado extraído, obtido por cada ferramenta.

Esta taxa de acerto será referenciada posteriormente, como por exemplo, por $P_{título}$ (porcentagem de acerto para o metadado título). Segundo a documentação da função `similar_text` temos:

> *"This calculates the similarity between two strings as described in Programming Classics: Implementing the World's Best Algorithms by Oliver (ISBN 0-131-00413-1). Note that this implementation does not use a stack as in Oliver's pseudo code, but recursive calls which may or may not speed up the whole process. Note also that the complexity of this algorithm is O(N**3) where N is the length of the longest string."*

Esta função recebe três parâmetros: o primeiro texto, o segundo texto e uma variável onde será armazenada a porcentagem de acerto. Como retorno tem-se um inteiro representando o número de caracteres em comum entre os dois textos comparados. Sua estrutura de utilização é a seguinte:

```
int similar_text ( string $first , string $second [, float &$percent ] )
```

Como cada ferramenta será testada em separado, os resultados da extração de cada artigo serão registrados, tendo o total da precisão calculado de acordo com a média aritmética dos resultados obtidos para aquele metadado. Por exemplo, para a Ferramenta "A" serão analisados 100 (cem) artigos. A precisão na extração do título de cada artigo ($P_{título1}$, $P_{título2}$, ..., $P_{tituloN}$), por exemplo, será somada e o resultado dividido pelo número de artigos - no caso 100. Assim tem-se a precisão geral para o metadado "Título" da Ferramenta "A" ($P_{título}$):

$$P_{título} = (P_{título1} + P_{título2} + P_{título3}... + P_{título100})/100$$

De posse dos resultados para cada metadado extraído podemos comparar as ferramentas a fim de conhecer a mais adequada para cada tipo de metadado. Podemos inferir, portanto, que a ferramenta "X" apresenta melhores resultados do que "Y" na extração do nome dos autores, por exemplo.

2.2.2 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 ferramenta testada. Assim chegamos a uma fórmula matemática à qual chamaremos "Índice de Confiabilidade", que calcula o resultado obtido através dos pesos que foram atribuídos a cada metadado, para cada ferramenta. Este índice é a nota final de cada ferramenta, levando em consideração todos os resultados obtidos por ela para o conjunto de artigos testado neste trabalho.

Este índice utiliza os pesos anteriormente definidos e a precisão dos resultados obtida, de maneira a permitir chegar a uma única nota final para cada ferramenta testada.

Esta fórmula é a média ponderada dos resultados alcançados na extração de cada metadado dos artigos, seguindo os pesos apresentados na Tabela 4. Cada peso é atribuído ao resultado encontrado em cada ferramenta.

A título de exemplo, após o teste de uma ferramenta, supondo que ela conseguiu extrair 87% dos títulos de todos os artigos com sucesso, sua precisão com relação ao título será 87 ($P_{título} = 87$), que será multiplicada pelo peso correspondente, neste caso, o inteiro 5. Isso ocorre para todos os metadados extraídos, seguindo seus respectivos pesos. A descrição de cada variável no Índice de Confiabilidade é apresentada na Tabela 5.

$$IC_{FerramentaX} = (5 * P_{título} + 4 * P_{autor} + 1 * P_{email} + 3 * P_{resumo} + 4 * P_{referências})/17$$

Assim, de posse do Índice de Confiabilidade de cada ferramenta podemos classificar cada uma com base nos seus resultados. Obviamente, esta classificação não tem por objetivo qualquer favorecimento de ferramentas, mas sim classificar cada uma delas com base nos

Tabela 5 – Descrição de cada variável no Índice de Confiabilidade

Variável	**Descrição**
$P_{título}$	Precisão na obtenção do título
P_{autor}	Precisão na obtenção do(s) autor(es)
P_{email}	Precisão na obtenção dos e-mails dos autores
P_{resumo}	Precisão na obtenção do resumo
$P_{referências}$	Precisão na obtenção das referências

resultados obtidos e critérios adotados neste trabalho. Desta forma, iremos classificar cada ferramenta seguindo as categorias abaixo:

1. **Precisa (P):** Quando o Índice de Confiabilidade é maior ou igual a 80 ($IC \geq 80$).
2. **Satisfatória (S):** Quando o Índice de Confiabilidade é maior ou igual a 60 e menor que 80 ($60 \leq IC < 80$).
3. **Insatisfatória (I):** Quando o Índice de Confiabilidade é menor que 60 ($IC < 60$).

2.3 Ambiente Tecnológico

As ferramentas testadas, por utilizarem das mesmas linguagens de programação ou por terem seu conjunto tecnológico muito semelhante, serão instaladas em um único servidor, permitindo que recursos computacionais sejam compartilhados, simplificando o trabalho de configuração em função de suas necessidades parecidas.

Este servidor será criado através de máquina virtual, o que traz benefícios não somente de performance mas de flexibilidade quanto às tecnologias necessárias para o funcionamento de cada ferramenta, permitindo que os testes possam ser feitos em sistemas operacionais distintos mas utilizando dos mesmos recursos computacionais da máquina de origem.

3 Análise e Apresentação de Resultados

Com o Corpus totalmente definido e as ferramentas devidamente instaladas no ambiente de testes foram realizados diversos experimentos para que os resultados pudessem ser analisados e apresentados numericamente.

Durante a extração dos metadados algumas observações puderam ser feitas tanto pela análise manual de cada resultado individual como também dos resultados em conjunto, tendo em vista os números apresentados pelas ferramentas utilizadas.

A ferramenta Cermine demonstrou-se de bem simples execução. Por se tratar de um arquivo em formato `.jar` (Java) em forma de executável, a extração ocorreu sem problemas, tendo os dados de saída da ferramenta gravados em arquivos isolados para posterior análise. Além disso, os resultados apresentados são os mais completos, com utilização de diversas tags XML que permitem que os dados sejam manipulados facilmente, da maneira que desejar e com uma grande riqueza de detalhes. O processo de extração dos metadados para cada artigo científico foi o mais lento das 4 (quatro) ferramentas testadas, demorando entre 15 e 20 segundos para uma completa análise de cada documento.

Já a ferramenta CiteSeer foi a que mais exigiu conhecimentos técnicos específicos para que pudesse ser testada. Sua execução dependeu da instalação de diversos outros componentes e serviços de terceiros, o que contribuiu para o aumento da complexidade de seu uso. Um fato interessante é que a ferramenta utiliza de outras ferramentas para alguns processos específicos de extração, como é o caso da sessão de referências, onde também utiliza a ferramenta ParsCit para realização do processo, porém com uma forma de entrada de dados um pouco diferenciada do que quando utilizada da sua maneira original e pura.

No caso da ferramenta CrossRef algumas particularidades devem ser mencionadas. Seus resultados de extração são apresentados de maneira muito básica, com campos muito genéricos e resultados pouco precisos, dificultando um pós-processamento dos dados, visando melhores resultados. Os metadados “autores”, “e-mails” e “resumo” não puderam ser extraídos. A versão atual de desenvolvimento da ferramenta não permite uma separação de dados muito específica, agrupando diversas informações em tags chamadas “sections”. Estas tags possuem informações textuais de modo geral, não sendo possível serem filtrados com a utilização da própria ferramenta. Portanto, para a ferramenta CrossRef somente os metadados “título” e “referências” foram extraídos. Os resultados para a extração das referências também merecem considerações, por serem apresentados de maneira muito genérica, em uma única tag, sendo impossível separar título e autor dentro do texto.

A ferramenta ParsCit também foi utilizada sem maiores dificuldades. Em virtude de sua particularidade de processar apenas arquivos de entrada em formato texto ou

XML, conforme sugerido pelos desenvolvedores, foi utilizada a ferramenta de linha de comando `pdftotext` (disponível em ambiente Linux) para conversão dos arquivos `.pdf` em arquivos `.txt`, permitindo que a ferramenta fosse utilizada conforme recomendações. Esta conversão foi feita em tempo de execução e os resultados coletados e gravados com sucesso.

De modo geral, exceto pela ferramente CrossRef as demais ferramentas tiveram um processo de extração bem eficaz visualmente e dentro do esperado, em virtude da grande diferenciação visual testada com o Corpus selecionado.

No que diz respeito à comparação dos resultados foi necessária uma padronização dos dados para que as quatro ferramentas pudessem ser testadas de maneira uniforme. Em virtude da apresentação dos resultados bastante detalhados, a ferramenta Cermine permitiu que os autores das referências fossem retornados seguindo a forma "primeiro nome" e em seguida "sobrenome". Já as demais ferramentas não apresentaram os resultados com tantos detalhes, variando em alguns momentos a ordem e disposição do nome dos autores. Assim, foi necessário um pré-processamento computacional a fim de manter, quando possível, o primeiro nome do autor antes do sobrenome, tornando a comparação a mais padronizada possível. Este pré-processamento foi realizado para todas as ferramentas testadas.

Já para a extração do metadado "e-mails", algumas ferramentas extraíram mais informações em conjunto, como foi o caso de algumas poucas extrações realizadas pela ferramenta Cermine. Em um destes casos a ferramenta retornou como e-mail o seguinte conteúdo: `Email: mvpein@yahoo.com`. Assim, sempre visando a comparação justa entre as ferramentas foi realizada uma análise em todas as comparações deste metadado para que somente pudessem ser comparados endereços de e-mails, o que tornou o processo bem simplificado. Os endereços de e-mail foram filtrados destas extrações com a utilização de expressão regular (**??**) de maneira a conseguir um conjunto homogêneo de dados comparados.

Os demais metadados foram comparados sem problemas. O metadado "título" foi comparado sem sua pontuação final, retirando, antes da comparação, qualquer caractere passível de erros como: asteriscos, pontos finais e espaços em branco. O resultado das extrações dos título foi feito seguindo a lógica anteriormente apresentada, comparando a similaridade entre os dois conteúdos através da função `similar_text` da linguagem de programação PHP, que apresenta como resultado um valor numérico representando o percentual de similaridade. Esta mesma lógica descrita foi aplicada para o metadado "resumo".

Os nomes dos autores foram comparados seguindo a mesma lógica do metadado "título", porém levando em consideração a ordem de apresentação e extração dos mesmos. Sendo assim, além de verificar a similaridade entre os nomes os testes levaram em consi-

deração a ordem de apresentação dada pelas ferramentas, requisito necessário para uma extração de sucesso.

Dentro do corpus escolhido diversos nomes de autores eram escritos com acentos, possuindo caracteres característicos de seu idioma de origem, como é o caso da autora polonesa "Anna Białk-Bielińska". Em virtude desta questão, as ferramentas se comportaram de maneiras bem distintas. Algumas conseguiram extrair os nomes de acordo com o artigo original, porém, outras substituíram caracteres como "ń" por apenas "n", ou ainda "n´", ainda que algumas ferramentas simplesmente desprezaram estes caracteres.

No caso específico do metadado "e-mail" a comparação foi realizada com base na identificação correta ou não do endereço eletrônico. Neste caso não foi considerada porcentagem de similaridade entre os resultados, ou seja, ou o endereço foi corretamente identificado ou não. Para estes resultados foram utilizados os inteiros 0 (zero) para a extração ineficiente e 100 (cem) para a extração eficiente.

Uma grande parte dos artigos utilizados no Corpus deste trabalho não possuía informações de e-mail dos autores. Desta forma, as extrações destes documentos foram desconsideradas, permitindo que as ferramentas tivessem seus resultados avaliados apenas para as extrações realmente computadas, valorizando ainda mais o trabalho de cada uma.

Já para a comparação das referências foram utilizadas duas informações: o título e o nome dos autores. Para o caso do título, a lógica utilizada foi a mesma utilizada no metadado "título", utilizando-se de um valor percentual para representação da similaridade entre os dados comparados. Já para o nome dos autores, a lógica seguiu a mesma do metadado "autores", onde é levado em consideração tanto a similaridade textual como também a ordem de apresentação. Deste modo, a extração de cada referência levou em consideração um peso de 60% do resultado para o título e 40% para os nomes dos autores, chegando em um numero final que representa o resultado da extração de cada referência identificada.

Com os dados de cada extração armazenados a comparação foi feita de maneira automática levando em consideração todos os pontos apresentados acima. Para cada subárea do conhecimento foi realizada uma comparação, registrando o resultado consolidado para cada artigo extraído, bem como a média aritmética dos resultados daquela subárea em específico. Portanto, para cada ferramenta e para cada subárea foi registrado um valor médio dos resultados.

Posteriormente foi feita a coleta destes dados separados por subáreas, porém consolidando-os para cada ferramenta. Assim foi calculada a média aritmética dos resultados de cada ferramenta para todas as subáreas analisadas, chegando então a uma nota final para cada ferramenta em cada metadado extraído, possibilitando então o cálculo do "Índice de Confiabilidade" (subseção 2.2.2).

Tabela 6 – Resultados da Cermine por subárea do conhecimento.

Subárea do Conhecimento	**Tit.**	**Aut.**	**Ema.**	**Res.**	**Ref.**
Arquitetura e Urbanismo	100	58.75	16.67	99.01	82.67
Ciência da Computação	88.27	71.87	21.43	98.83	77.25
Ciência da Informação	76.55	61.90	28.57	78.02	53.81
Ciências Biológicas (Genética)	91.58	81.00	50.00	84.72	96.11
Ciências Biológicas (Zoologia)	99.78	73.16	42.86	84.74	72.28
Enfermagem	99.77	39.38	16.67	98.09	81.69
Engenharia Civil	71.43	76.34	37.50	94.18	56.23
Engenharia Mecânica	99.45	75.97	58.33	77.97	82.87
Fonoaudiologia	100	77.75	71.43	98.13	80.05
Geologia	99.54	100	66.67	53.66	64.03
História	99.20	89.29	50.00	65.59	53.40
Letras	88.01	99.50	42.86	82.10	86.74
Medicina Veterinária	85.71	91.11	85.71	98.77	80.05
Música	99.03	90.61	66.67	95.47	68.50
Psicologia	88.46	63.96	47.62	92.53	63.25
Zootecnia	49.95	70.82	42.86	87.40	81.99
Média Geral	89.80	76.34	46.62	86.83	73.81

3.1 Resultados

Conforme esperado os resultados foram coletados de maneira individual, para cada artigo, e consolidados de maneira geral para cada ferramenta e para cada metadado. Os resultados apresentados por área do conhecimento estão presentes em 4 (quatro) tabelas, separadas por cada uma das ferramentas. Os resultados da ferramenta Cermine estão presentes na Tabela 6. Os resultados da CiteSeer estão na Tabela 7. Os resultado da ferramenta CrossRef na Tabela 8 e da ParsCit na Tabela 9. Todas as tabelas mostram o percentual de acerto separados por subárea do conhecimento e por metadados, representados pelas colunas Tit. (Título), Aut. (Autores), Ema. (E-mails), Res. (Resumo) e Ref. (Referências).

Para que os resultados pudessem ser melhores interpretados foi calculado o "Índice de Confiabilidade" para cada ferramenta, detalhado no capítulo de Metodologia (subseção 2.2.2). Para calcular este índice foram utilizadas as média dos resultados de extração de todas as subáreas, tomando os devidos pesos para cada metadado, obtendo-se então uma nota geral para cada ferramenta. Os resultados calculados para este índice estão presentes na Tabela 10.

De posse do "Índice de Confiabilidade" de cada ferramenta, conforme previsto na subseção 2.2.2, cada ferramenta foi classificada, de acordo com seus resultados de extração. Estes resultados e suas respectivas classificações estão presentes na Tabela 11.

Tabela 7 – Resultados da CiteSeer por subárea do conhecimento.

Subárea do Conhecimento	**Tit.**	**Aut.**	**Ema.**	**Res.**	**Ref.**
Arquitetura e Urbanismo	100	96.89	0	97.43	70.95
Ciência da Computação	100	83.75	23.81	99.81	71.79
Ciência da Informação	84.44	99.50	0	74.12	55.56
Ciências Biológicas (Genética)	80.92	83.15	28.57	60.63	25.15
Ciências Biológicas (Zoologia)	57.14	64.12	0	71.14	70.10
Enfermagem	71.43	52.82	0	70.31	34.67
Engenharia Civil	97.81	62.42	0	71.18	35.61
Engenharia Mecânica	71.11	46.00	0	71.36	63.18
Fonoaudiologia	100	61.14	0	94.68	61.85
Geologia	73.77	34.69	0	42.79	57.13
História	99.53	71.09	0	65.26	63.81
Letras	99.57	85.73	0	75.78	58.82
Medicina Veterinária	85.71	86.38	0	98.88	63.53
Música	49.02	56.87	0	54.28	54.55
Psicologia	94.93	83.85	14.29	88.87	66.75
Zootecnia	71.43	82.41	0	76.39	22.59
Média Geral	83.55	71.93	4.17	75.81	54.75

Tabela 8 – Resultados da CrossRef por subárea do conhecimento.

Subárea do Conhecimento	**Tit.**	**Aut.**	**Ema.**	**Res.**	**Ref.**
Arquitetura e Urbanismo	72.68	0	0	0	22.79
Ciência da Computação	64.19	0	0	0	14.64
Ciência da Informação	32.32	0	0	0	8.14
Ciências Biológicas (Genética)	47.05	0	0	0	14.62
Ciências Biológicas (Zoologia)	70.70	0	0	0	32.72
Enfermagem	55.96	0	0	0	10.29
Engenharia Civil	74.70	0	0	0	12.21
Engenharia Mecânica	89.27	0	0	0	27.50
Fonoaudiologia	71.43	0	0	0	13.92
Geologia	97.62	0	0	0	15.72
História	64.08	0	0	0	16.11
Letras	75.66	0	0	0	32.58
Medicina Veterinária	49.66	0	0	0	23.09
Música	84.63	0	0	0	28.16
Psicologia	82.92	0	0	0	23.19
Zootecnia	32.05	0	0	0	25.21
Média Geral	66.56	0	0	0	20.06

Tabela 9 – Resultados da ParsCit por subárea do conhecimento.

Subárea do Conhecimento	**Tit.**	**Aut.**	**Ema.**	**Res.**	**Ref.**
Arquitetura e Urbanismo	0	17.14	0	73.23	51.62
Ciência da Computação	37.54	58.36	47.62	74.82	69.72
Ciência da Informação	32.30	31.51	28.57	59.60	50.09
Ciências Biológicas (Genética)	8.97	1.17	0	40.91	39.76
Ciências Biológicas (Zoologia)	0	0	0	69.98	54.61
Enfermagem	11.06	14.29	0	65.65	37.24
Engenharia Civil	11.48	15.64	37.50	56.61	37.97
Engenharia Mecânica	14.29	23.23	22.22	55.93	55.34
Fonoaudiologia	5.70	0.89	0	27.74	62.52
Geologia	14.29	14.29	16.67	68.42	55.15
História	5.62	14.29	0	59.33	60.17
Letras	24.42	42.21	21.43	68.53	54.00
Medicina Veterinária	14.29	13.82	14.29	36.57	53.31
Música	34.24	42.86	0	81.39	55.73
Psicologia	14.29	14.29	14.29	71.14	68.84
Zootecnia	14.29	5.88	0	78.60	51.21
Média Geral	15.17	19.37	12.66	61.78	53.58

Tabela 10 – Índice de Confiabilidade de cada ferramenta

Ferramenta	**Resultado**
Cermine	79.81
CiteSeer	68.00
CrossRef	24.30
ParsCit	33.27

Tabela 11 – Classificação de cada ferramenta.

Ferramenta	**Índice de Confiabilidade**	**Classificação**
Cermine	79.81	Satisfatória
CiteSeer	68.00	Satisfatória
CrossRef	24.30	Insatisfatória
ParsCit	33.27	Insatisfatória

3.2 Ambiente de Testes

Para a realização das extrações e das comparações foi criado um ambiente de testes contendo todas as tecnologias necessárias para que as ferramentas pudessem ser executadas dentro do esperado. Desta maneira, foi utilizado um servidor virtual com a seguinte configuração:

- Sistema Operacional Linux Ubuntu 14.04 64 Bits
- 2GB de Memória RAM
- 20GB de Espaço em Disco

As tecnologias utilizadas foram instaladas de acordo com as recomendações de cada ferramenta, com suas dependências e necessidades de cada linguagem de programação. Foram instaladas as seguintes linguagens/bibliotecas, separadas de acordo com cada ferramenta:

- **Cermine:** Java OpenJDK Runtime Environment 1.7.0_79
- **CiteSeer:** Python 2.7.6, GROBID <https://github.com/kermitt2/grobid>, PDFBox <http://pdfbox.apache.org/>, PDF Classifier Jar, Java SE Environment (Maven).
- **CrossRef:** Ruby 2.1.2p95, RubyGem `pdf-extract` 0.0.1 e `pdf-reader` 1.3.2.
- **ParsCit:** Perl 5.18.2, G++ Compiler e CRF++ 0.51. Diversas outras dependências da linguagem Perl foram também instaladas: Class::Struct, Getopt::Long, Getopt::Std, File::Basename, File::Spec, FindBin, HTML::Entities, IO::File, POSIX, XML::Parser, XML::Twig, XML::Writer e XML::Writer::String.

4 Discussão / Trabalhos Futuros

Após todo o processo de pesquisa, de extração dos metadados pelas ferramentas analisadas e coleta de seus respectivos resultados, algumas considerações podem ser feitas, relativas aos objetivos propostos no início do trabalho.

Os resultados apresentados, de modo geral, foram inferiores às expectativas iniciais da pesquisa. As extrações não foram tão precisas quanto se imaginava. A grande diferença no leiaute dos elementos, presente no Corpus escolhido, realmente teve alto impacto nos resultados, principalmente no que diz respeito à extração dos autores e das referências.

Por outro lado, as ferramentas Cermine e CiteSeer obtiveram resultados para a extração do metadado "título" bem positivos, atingindo entre 83 e 89% de precisão. Já a ferramenta CrossRef ficou bem abaixo do esperado, com 66.56% de precisão apenas, porém acima da última colocada, a ferramenta ParsCit, que conseguiu extrair com sucesso apenas 15.17% dos resultados dos "títulos", muito abaixo do esperado.

Para o metadado "autores" a ferramenta com maior precisão foi a Cermine, que atingiu 76.34%, resultado próximo da segunda colocada, a CiteSeer, com 71.93%. Já as demais ferramentas não obtiveram êxito na extração dos nomes dos autores, ficando abaixo dos 20% de acerto.

Para a extração dos e-mails dos autores o resultado obtido, de modo geral, foi pior. A ferramenta que obteve maior êxito na extração deste metadado foi a Cermine, que conseguiu obter apenas 46.62% de sucesso. Os resultados para este metadado obtidos pela ferramenta CiteSeer foram bem inferiores às expectativas, pois somente 4.17% dos endereços foram extraídos com sucesso, resultado inferior ainda à ferramenta ParsCit, que extraiu 12.16%. Como informado no capítulo "Resultados" (Capítulo 3) a ferramenta CrossRef não conseguiu realizar a extração de nomes de autores, endereços de e-mails e do resumo, sendo estes resultados desconsiderados nesta sessão.

Em virtude da variação de leiaute do Corpus e da ausência de padronização do metadado "resumo" (*abstract*), os resultados obtidos para este metadado superaram as expectativas. Exceto pela ferramenta CrossRef, todas as demais obtiveram resultados acima de 60%, chegando a 86.83% da ferramenta Cermine, a maior precisão.

Esses resultados podem ser considerados positivos, principalmente em virtude de alguns artigos apresentarem o metadado de maneira bem diferente, com posicionamento bem divergente do habitual, inclusive, sem indícios de que ali se apresentava o resumo do artigo.

Outro ponto onde as expectativas não foram atingidas foi na extração das "referên-

cias". A ferramenta Cermine, mais uma vez, demonstrou-se mais precisa, alcançando 73.81% de sucesso. A ferramenta CiteSeer, que utiliza a ParsCit para extração das referências, ao ser comparada com a própria ParsCit, produziu resultados pouco superiores, 54.75% e 53.58%, respectivamente.

A diferença nos resultados se deve ao fato da ParsCit necessitar de arquivos `.txt` como entrada de dados. No caso das extrações realizadas pela própria ferramenta, os arquivos `.txt` foram gerados pelo programa `pdftotext`, conforme detalhado no capítulo de "Resultados" (Capítulo 3), diferentemente da ferramenta CiteSeer, que transforma o arquivo `.pdf` em `.txt` de sua própria maneira, causando então uma pequena divergência nos resultados gerais (1.17%).

Já a ferramenta CrossRef obteve apenas 20.06% de precisão na extração das referências, o que era esperado em função de sua extração com poucos detalhes, com apenas um único campo com todas as informações de cada referência.

Embora os resultados da extração não tenham sido positivos, um detalhe interessante que merece atenção é a forma como a ferramenta trata as referências. A ferramenta permite que elas sejam comparadas com o banco de dados existente no <http://api.crossref.org>, possibilitando identificar exatamente quais artigos já foram catalogados pelo site, gerenciando o conteúdo, inclusive relacionando-o.

Para os artigos encontrados na base de dados do CrossRef é possível obter, inclusive, a descrição de cada um em formato BibTeX.

Para este trabalho, em virtude dos poucos resultados obtidos, esta funcionalidade não foi utilizada na extração e na comparação.

Em se tratando da separação dos resultados por área do conhecimento as ferramentas Cermine e CiteSeer obtiveram destaque, conseguindo 100% de acertos em 3 subáreas do conhecimento, porém para metadados diferentes.

A ferramenta Cermine acertou todos os títulos das áreas de Arquitetura e Urbanismo e Fonoaudiologia, além de 100% dos nomes dos autores da área de Geologia. A ferramenta CiteSeer conseguiu precisão total na extração dos títulos de Arquitetura e Urbanismo, Ciência da Computação e Fonoaudiologia.

Já a ferramenta CrossRef obteve melhor resultado na extração dos títulos dos artigos da área de Geologia, obtendo 97.62% de precisão, superando a ferramenta CiteSeer e ParsCit, que obtiveram 73.77% e 14.29% respectivamente.

Para a extração dos títulos dos artigos, os piores resultados foram encontrados nas áreas de Música (CiteSeer, com 49.02%), Zootecnia (Cermine, com 49.95% e CrossRef, com 32.05%) e as áreas Ciências Biológicas (Zoologia) e Arquitetura e Urbanismo (ParsCit, com nenhum acerto).

A ferramenta Cermine se destacou na extração de títulos de 8 (oito) subáreas do conhecimento - Arquitetura e Urbanismo, Ciências Biológicas (Zoologia), Enfermagem, Engenharia Mecânica, Fonoaudiologia, Geologia, História e Música -, obtendo resultados superiores a 99%, o que foi considerado excelente.

Para os autores, seus maiores destaques foram nas áreas de Geologia, Letras, Medicina Veterinária e Música, com resultados acima de 90%.

Na extração dos e-mails dos autores a ferramenta Cermine obteve resultados superiores a 85% somente na área de Medicina Veterinária, seu melhor resultado para este metadado. Além disso, a ferramenta destacou-se na extração dos resumos em 5 (cinco) áreas, com resultados acima dos 98%, e na extração das referências de Ciências Biológicas (Genética), onde obteve resultados acima de 96% de precisão.

Já a ferramenta CiteSeer foi bem eficiente na extração dos títulos de 5 (cinco) subáreas: Arquitetura e Urbanismo, Ciência da Computação, Fonoaudiologia, História e Letras, com resultados superiores a 99%.

Para a extração dos nomes dos autores o resultado foi relevante em apenas 2 (duas) subáreas: Arquitetura e Urbanismo e Ciência da Informação, com precisão acima de 90%.

Já para os e-mails dos autores os resultados deixaram a desejar para 13 (treze) das 16 (dezesseis) subáreas, com 0% de acerto, tendo resultados positivos apenas para as subáreas de Ciência da Computação, Ciências Biológicas (Genética) e Psicologia, porém com resultados abaixo de 29% de acerto.

Para a extração dos resumos a ferramenta CiteSeer também se mostrou bem eficiente, com resultados acima de 90% para 4 (quatro) subáreas: Arquitetura e Urbanismo, Ciência da Computação, Fonoaudiologia e Medicina Veterinária.

Para as referências (utilizando o ParsCit) os resultados deixaram a desejar, com resultados abaixo de 72%.

A ferramenta CrossRef mostrou resultados positivos apenas para a extração de títulos de artigos da subárea de Geologia, como já dito anteriormente, com 97.62% de acerto, não tendo resultados considerados satisfatórios para as demais áreas.

Para a extração das referências os resultados deixaram a desejar, com apenas 2 (duas) subáreas com precisão próxima de 30%: Ciências Biológicas (Zoologia), com 32.72% e Letras, com 32.58%.

Por fim, a ferramenta ParsCit obteve resultados abaixo dos 38% para os títulos, em todas as subáreas analisadas.

O acerto dos nomes dos autores também foi baixo, onde os melhores resultados ficaram entre 43% e 59%, para as subáreas de Ciência da Computação, Letras e Música.

Tabela 12 – Melhores ferramentas para o metadado “Título”

Subáreas do Conhecimento	**Ferramentas**	**Precisão**
Arquitetura e Urbanismo	Cermine/CiteSeer	100%
Ciência da Computação	CiteSeer	100%
Ciência da Informação	CiteSeer	84.44%
Ciências Biológicas (Genética)	Cermine	91.58%
Ciências Biológicas (Zoologia)	Cermine	99.78%
Enfermagem	Cermine	99.77%
Engenharia Civil	CiteSeer	97.81%
Engenharia Mecânica	Cermine	99.45%
Fonoaudiologia	Cermine/CiteSeer	100%
Geologia	Cermine	99.54%
História	CiteSeer	99.53%
Letras	CiteSeer	99.57%
Medicina Veterinária	Cermine/CiteSeer	85.71%
Música	Cermine	99.03%
Psicologia	CiteSeer	94.93%
Zootecnia	CiteSeer	71.43%

Para os e-mails dos autores o melhor resultado foi para os artigos da subárea de Ciência da Computação, com 47.62% de precisão.

Já para o resumo dos artigos, os resultados foram um pouco melhores, acima de 70% para 5 (cinco) subáreas do conhecimento.

Os resultados para as referências foram semelhantes aos obtidos pela ferramenta CiteSeer (utiliza a mesma ferramenta).

Os 2 (dois) melhores resultados foram para as subáreas de Ciência da Computação e Psicologia, com precisão de 69.72% e 68.84%, respectivamente.

Para melhor visualização dos resultados, as Tabelas 12, 13, 14, 15 e 16 apresentam as ferramentas que obtiveram os melhores resultados para cada subárea do conhecimento, separados por metadado.

Os resultados mostram que, no Corpus escolhido, para o metadado “Título”, a ferramenta Cermine foi superior, com 89.8% de precisão, seguida da CiteSeer, com 83.55%.

O mesmo acontece para o metadado “Autores”, onde a Cermine obteve os melhores resultados (76.34%), seguida da CiteSeer, com 71.93%.

Já para o metadado “E-mails” a Cermine foi sem dúvida a melhor, com 46.62% de acertos, deixando uma grande diferença da segunda colocada ParsCit, com apenas 12.66%.

Para o metadado “Resumo” a Cermine também se saiu melhor, com 86.83% de precisão, e em segunda posição a CiteSeer com 75.81%.

Tabela 13 – Melhores ferramentas para o metadado “Autores”

Subáreas do Conhecimento	**Ferramentas**	**Precisão**
Arquitetura e Urbanismo	CiteSeer	96.89%
Ciência da Computação	CiteSeer	83.75%
Ciência da Informação	CiteSeer	99.50%
Ciências Biológicas (Genética)	CiteSeer	83.15%
Ciências Biológicas (Zoologia)	Cermine	73.16%
Enfermagem	CiteSeer	52.82%
Engenharia Civil	Cermine	76.34%
Engenharia Mecânica	Cermine	75.97%
Fonoaudiologia	Cermine	77.75%
Geologia	Cermine	100%
História	Cermine	89.29%
Letras	Cermine	99.50%
Medicina Veterinária	Cermine	91.11%
Música	Cermine	90.61%
Psicologia	CiteSeer	83.85%
Zootecnia	CiteSeer	82.41%

Tabela 14 – Melhores ferramentas para o metadado “E-mails”

Subáreas do Conhecimento	**Ferramentas**	**Precisão**
Arquitetura e Urbanismo	Cermine	16.67%
Ciência da Computação	ParsCit	47.62%
Ciência da Informação	Cermine/ParsCit	28.57%
Ciências Biológicas (Genética)	Cermine	50.00%
Ciências Biológicas (Zoologia)	Cermine	42.86%
Enfermagem	Cermine	16.67%
Engenharia Civil	Cermine/ParsCit	37.50%
Engenharia Mecânica	Cermine	58.33%
Fonoaudiologia	Cermine	71.43%
Geologia	Cermine	66.67%
História	Cermine	50.00%
Letras	Cermine	42.86%
Medicina Veterinária	Cermine	85.71%
Música	Cermine	66.67%
Psicologia	Cermine	47.62%
Zootecnia	Cermine	42.86%

Tabela 15 – Melhores ferramentas para o metadado “Resumo”

Subáreas do Conhecimento	**Ferramentas**	**Precisão**
Arquitetura e Urbanismo	Cermine	99.01%
Ciência da Computação	CiteSeer	99.81%
Ciência da Informação	Cermine	78.02%
Ciências Biológicas (Genética)	Cermine	84.72%
Ciências Biológicas (Zoologia)	Cermine	84.74%
Enfermagem	Cermine	98.09%
Engenharia Civil	Cermine	94.18%
Engenharia Mecânica	Cermine	77.97%
Fonoaudiologia	Cermine	98.13%
Geologia	Cermine	53.66%
História	Cermine	65.59%
Letras	Cermine	82.10%
Medicina Veterinária	CiteSeer	98.88%
Música	Cermine	95.47%
Psicologia	Cermine	92.53%
Zootecnia	Cermine	87.40%

Tabela 16 – Melhores ferramentas para o metadado “Referências”

Subáreas do Conhecimento	**Ferramentas**	**Precisão**
Arquitetura e Urbanismo	Cermine	82.67%
Ciência da Computação	Cermine	77.25%
Ciência da Informação	CiteSeer	55.56%
Ciências Biológicas (Genética)	Cermine	96.11%
Ciências Biológicas (Zoologia)	Cermine	72.28%
Enfermagem	Cermine	81.69%
Engenharia Civil	Cermine	56.23%
Engenharia Mecânica	Cermine	82.87%
Fonoaudiologia	Cermine	80.05%
Geologia	Cermine	64.03%
História	CiteSeer	63.81%
Letras	Cermine	86.74%
Medicina Veterinária	Cermine	80.05%
Música	Cermine	68.50%
Psicologia	ParsCit	68.84%
Zootecnia	Cermine	81.99%

Por fim, para a extração do metadado "Referências" novamente a Cermine obteve o melhor resultado, com precisão de 73.81% dos resultados, seguida da CiteSeer com 54.75%.

4.1 Contribuições

Como dito, os resultados coletados após as comparações ficaram abaixo das expectativas, exceto pelo metadado "Título", onde os números foram expressivos.

Em virtude da grande diferença no posicionamento visual dos elementos dos artigos do Corpus, os resultados foram muito variáveis, não sendo possível definir, com precisão, que ferramenta se comporta melhor para uma determinada área ou subárea do conhecimento, mesmo que os resultados demonstrem, numericamente, o comportamento diferenciado de cada uma.

Estes resultados permitem aferir que as ferramentas de extração de metadados ainda tem espaço para evoluir, sendo necessários ajustes e adaptações para que uma maior quantidade de metadados seja extraída com sucesso.

Além disso, algumas ferramentas apresentam melhores resultados em algumas subáreas do conhecimento, mas sem generalização possível, o que demandaria uma análise mais aprofundada.

Todo o código utilizado na comparação está disponível em <http://github.com/jgrossi/met>, podendo ser utilizado para futuras pesquisas. É possível incluir novas ferramentas de maneira simplificada.

Ademais, todo o processo de comparação elaborado neste trabalho pode ser reutilizado, permitindo inclusive o cálculo do índice de confiabilidade segundo os critérios adotados pelo autor. O índice permite a classificação de uma ferramenta segundo pesos definidos para cada metadado (subseção 2.2.2).

Por fim, pôde-se observar que o comportamento das técnicas de extração utilizadas pelas ferramentas é muito variável. Uma parcela dos resultados parece ser influenciado pelo modo de uso da técnica em cada ferramenta.

Com efeito, deve-se levar em consideração a maneira como os algoritmos (das técnicas) são implementados bem como a maneira como os dados são tratados, tanto antes quanto depois da extração. Assim, com base nos resultados numéricos apresentados, não é possível determinar qual técnica é mais aplicada para a extração de cada metadado.

4.2 Trabalhos Futuros

Embora este trabalho tenha abrangido 16 (dezesseis) subáreas do conhecimento, com um total de 112 (cento e doze) artigos científicos, a variedade real de formatos e

leiautes vai muito além.

Poderia-se pensar em trabalhos mais detalhados para cada subárea do conhecimento, permitindo testar uma maior quantidade de artigos e padrões, de maneira a obter resultados mais próximos do real.

Um ponto interessante de pesquisa seria um trabalho de comparação para artigos de uma área do conhecimento específica, como Engenharia, por exemplo, onde um maior número de artigos desta área seria testado, objetivando identificar o comportamento destas ferramentas para esta área em específico, com um Corpus bem maior e variado, porém mais direcionado.

Um outro estudo possível seria a comparação por revistas ou bases de dados. Embora a diferenciação de leiaute para uma área do conhecimento seja muito ampla, geralmente existe uma padronização visual para uma determinada base, como é o caso da Elsevier <http://www.elsevier.com>, onde, independente da área do conhecimento, os artigos passam por uma padronização visual.

Apesar de selecionadas as quatro ferramentas aqui comparadas, existem muitas outras ferramentas que merecem atenção, possibilitando um estudo de caso focado para uma determinada ferramenta, aprofundando muito mais suas características e funcionalidades, permitindo conclusões mais direcionadas e inclusive críticas mais precisas quanto aos resultados por ela apresentados.

Seria ainda interessante estender a pesquisa considerando possíveis variações na extração manual dos metadados. Embora os dados tenham sido extraídos de maneira bem cautelosa é possível que as mesmas extrações, realizadas por pessoas diferentes, produzam resultados variados.

4.3 Considerações Finais

Em virtude dos resultados apresentados e com base nas comparações realizadas, sugere-se que as ferramentas ainda tem espaço para evoluir para abranger um maior número de artigos e áreas do conhecimento.

Algumas ferramentas se comportaram melhor para alguns padrões visuais, porém não sendo possível estabelecer uma regra ou afirmação com base nos resultados encontrados.

As fragilidades das ferramentas testadas sugerem que o desenvolvimento de uma nova solução poderia ser de interesse.

Referências

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of Machine Learning*. [S.l.: s.n.], 2012. ISBN 978-0-262-01825-8. Citado na página 11.

OLIVER, I. *Programming classics : implementing the world's best algorithms*. New York: Prentice Hall, 1993. ISBN 0-13-100413-1. Disponível em: <http://opac.inria.fr/record=b1084473>. Citado na página 18.

Anexos

ANEXO A – Elementos do padrão Dublin Core, versão 1.1.

Name	**Label**	**Definition**	**Comment**
title	Title	A name given to the resource.	
creator	Creator	An entity primarily responsible for making the resource.	Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity.
subject	Subject	The topic of the resource.	Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element.
description	Description	An account of the resource.	Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource.
publisher	Publisher	An entity responsible for making the resource available.	Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity.
contributor	Contributor	An entity responsible for making contributions to the resource.	Examples of a Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity.

date	Date	A point or period of time associated with an event in the lifecycle of the resource.	Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601 [W3CDTF].
type	Type	The nature or genre of the resource.	Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCTYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.
format	Format	The file format, physical medium, or dimensions of the resource.	Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
identifier	Identifier	An unambiguous reference to the resource within a given context.	Recommended best practice is to identify the resource by means of a string conforming to a formal identification system.
source	Source	A related resource from which the described resource is derived.	The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.
language	Language	A language of the resource.	Recommended best practice is to use a controlled vocabulary such as RFC 4646 [RFC4646].
relation	Relation	A related resource.	Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system.

coverage	Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.	Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges.
rights	Rights	Information about rights held in and over the resource.	Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights.