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-0.3.2

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-0.3.2

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-0.3.2-

51 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-0.3.2.

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-0.3.2

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

Agradecimentos

Os agradecimentos principais são direcionados a meu orientador Prof. Marcello Peixoto Bax, que me concedeu a oportunidade de realizar este trabalho com muita liberdade e com extrema confiança em meu trabalho.

Agradecimentos especiais são direcionados aos meus pais por acreditarem em meu potencial e sempre aceitarem as escolhas feitas por mim, sempre me incentivando de um modo ou outro a procurar sempre novos conhecimentos e experiências.

Agradeço também de modo geral todas as pessoas da ECI, tanto pelas amizades feitas quanto também por me apoiarem, me ouvirem e me incentivarem a acreditar no meu trabalho e a fazer um estudo de qualidade.

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
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

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.

Com o crescimento da pesquisa em todo o mundo um grande número de publicações foram inseridas no meio, fazendo com que uma infinidade de material esteja disponível em poucos segundos. Deste modo, a necessidade de centralização automatizada dos dados e a contribuição dentre os pesquisadores é inerente ao desenvolvimento desta área.

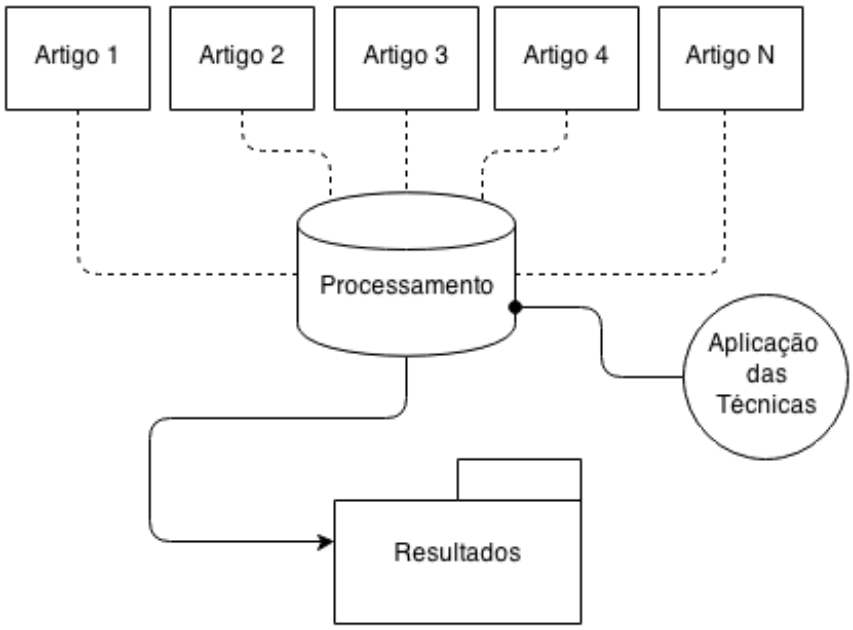
No âmbito acadêmico sempre contribuímos de uma forma ou outra com a formação de nossos colegas e parceiros de pesquisa. Esta contribuição pode ser feita com base em uma conversa informal ou até mesmo com uma ajuda em documentação ou sugestão de um texto para leitura. Esta sugestão de leitura geralmente 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

Figura 1 – Processo de Extração de Metadados



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 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

Visando obter uma revisão de literatura eficaz foi feita uma Revisão Sistemática de Literatura, com base no Relatório Técnico de (KITCHENHAM, 2004). Desde modo o projeto se torna mas abrangente do ponto de vista de pesquisa literária e ao mesmo tempo mais restritivo, realmente realizando o estudo dos objetos que são importantes para a pesquisa em si.

Basicamente, como o objetivo do trabalho tem foco na resultado prático, geralmente as bases relacionadas às áreas mais técnicas, como Ciência da Computação, devem ser bem focadas, como IEEE e ACM, por exemplo. Outras bases de natureza mais genérica também foram pesquisadas, geralmente através do site da CAPES, mas com caráter mais de apoio teórico realmente.

Embora a amplitude desta área, alguns detalhes necessitam ser observados para evitar assim redundância e perda de foco no trabalho. Basicamente a pesquisa necessitava ser focada em técnicas de extração de dados em artigos científicos, somente. Estas técnicas necessitam ser reais, de maneira a existir realmente uma forma prática de serem testadas, fazendo assim com que os resultados obtidos sejam comparados e confrontados para verificar então a eficácia do processo.

Como a pesquisa necessita de um resultado prático eficaz os critérios que serão adotados para demonstrar a relevância de um estudo serão seus resultados. Com base em uma técnica potencialmente eficaz, sua implantação deve ser realizada, independente da linguagem de programação, e então testada juntamente com um grupo de artigos previamente selecionados como teste base. Estes artigos já possuirão seus dados mapeados de maneira a poder comparar os resultados obtidos com cada uma das técnicas e os resultados então esperados. Portanto os critérios utilizados serão de natureza explicitamente prática, com foco em resultados concretos.

Seguindo as orientações propostas em (KITCHENHAM, 2004) foram mapeados alguns eventos para serem pesquisados, a fim de encontrar trabalhos relevantes para a área. Sendo assim, foram mapeados os seguintes eventos:

- KDIR (International Conference of Knowledge Discovery and Information Retrieval)
- ICDAR (International Conference on Document Analysis and Recognition)
- PDCAT (nternational Conference on Parallel and Distributed Computing, Applications and Technologies)
- IAPR (International Workshop on Document Analysis Systems)

- ACM Conference on Digital Libraries

As necessidades de tornar artigos mais conectados é cíclica, e permite que novas técnicas sejam descobertas ou criadas por meio de necessidades de grupos de pesquisadores desejando obter informações precisas cada vez mais rápido.

2.1 Técnicas e Algoritmos

Algumas técnicas e algoritmos de extração de dados são utilizadas em diversos projetos, de maneira a serem citadas em momentos onde exige-se uma precisão maior.

Estas técnicas se baseiam basicamente na classificação de dados com base nas suas representações escritas, tanto baseadas em padrões preestabelecidos ou até mesmo com base em um dicionário de palavras capaz de reconhecer ocorrências em diversas partes de um documento.

2.1.1 Support Vector Machines (SVM)

Vários algoritmos de extração possuem referências na técnica de SVM descrita por (HAN C. LEE GILES; FOX, 2003). Esta técnica é baseada na identificação de campos previamente selecionados no cabeçalho de um documento, do qual se deseja obter os metadados.

Esta técnica analisa diversos campos chamando-os de classes, e atribui a cada classe uma característica que a permite ser identificada. Deste modo cada linha do cabeçalho do documento é classificada em uma ou mais classes.

Estas classes seguem o padrão (WEIBEL, 1999) estabelecido pelo Dublin Core[1], que define por padrão 15 elementos para descrever um recurso de maneira digital.

Seymore et al. (SEYMORE; ROSENFIELD, 1999) definiu também 15 tags para esta definição do cabeçalho de um documento. Porém destas 15 tags definidas somente 4 correspondem ao padrão da Dublin Core e estão ilustradas na Tabela 1.

Para o caso de extração de metadados em artigos científicos utilizando *Support Vector Machines* (HAN C. LEE GILES; FOX, 2003) as tags de Seymore et al. (SEYMORE; ROSENFIELD, 1999) são utilizadas para representação destas classes.

Deste modo, com base nestas classes são definidas características de suas classes vizinhas, como por exemplo, elementos que ficam perto de outros elementos, que possuem uma sequência lógica geral de exibição. Com base nestas informações, que são feita classe

[1] Iniciativa existente a fim de padronizar metadados para descrever um objeto digital. <http://dublincore.org>

Tabela 1 – Relação de classes utilizadas e comparação com o padrão Dublin Core.

Classe (Tag)	**Referência Dublin Core**	**Descrição**
Title	Title	Título do artigo
Author	Creator	Nome do autor do documento
Affiliation		Afiliação do autor
Address		Endereço do autor
Note		Frases de reconhecimentos, *copyright*
Email		Endereço de e-mail do autor
Date		Data da publicação
Abstract Introdution	Description	A parte de introdução do artigo
Phone		Telefone do autor
Keyword	Subject	As palavras-chave do documento
Web		Endereço na Internet do autor
Degree		Associação com o grau acadêmico
Pubnum		Número da publicação do documento
Page		O final da página

por classe, estes padrões vão sendo encaixados a cada linha do cabeçalho analisado, permitindo que os metadados sejam extraídos com uma grande precisão.

Além da análise linha por linha são utilizadas análises de palavras dentro de um contexto previamente selecionado. Assim foi criado um *cluster* de palavras comuns que facilitam na identificação destas classes nos cabeçalhos analisados. Este *cluster* basicamente é composto de:

- Dicionário online padrão em sistemas Linux;
- 8441 nomes e 19613 sobrenomes;
- Sobrenomes chineses;
- Nome dos estados do Estados Unidos e das províncias canadenses;
- Nomes das cidades dos Estados Unidos;
- Nome dos países do mundo, de acordo com World Fact Book[2];
- Nome dos meses e suas respectivas abreviações.

Para cada uma das classes analisadas são feitas correlações com o tipo de dado esperado, de maneira a permitir que endereços de e-mail, por exemplo, sejam extraídos com base em expressões regulares utilizadas em linguagens de programação.

[2] Disponível em <https://www.cia.gov/library/publications/the-world-factbook/index.html>

2.1.2 Hidden Markov Models (HMM)

A teoria básica de Markov foi conhecida próximo dos anos 80 por engenheiros e matemáticos com grande aplicação inicialmente em processamento da fala mas com vasta amplitude em outras áreas onde descoberta de padrões pode ser aplicada (RABINER; JUANG, 1986).

O processo é baseado na identificação de modelos observáveis que representem e caracterizem a ocorrências de símbolos observáveis, ou seja, padrões. Se um sinal foi observado ele pode ser utilizado para futuras referências, de acordo com o padrão observado.

Um exemplo prático citado por Rabiner e Juang (RABINER; JUANG, 1986) é o caso de uso do jogo *Cara e Coroa*. Toma-se um observador em um quarto fechado com uma cortina totalmente fechada para outro cômodo. Este observador não consegue ver nada que acontece no outro cômodo, onde está uma outra pessoa jogando uma moeda pra cima, relatando sempre o resultado obtido (cara ou coroa). Neste caso o problema é construir um Hidden Markov Model (HMM) para explicar ao observador a sequência dos resultados obtidos.

Este exemplo é baseado tanto no estado de cada resultado (cara ou coroa) e em probabilidades matemáticas de ocorrência destes estados, neste caso 50%. Assim desenha-se modelos onde os estados são representados com base nas inúmeras possibilidades existentes, levando inclusive em consideração a sequência dos fatos, ou estados.

No âmbito da extração da informação, o HMM pode ser aplicado conforme é apresentado por Seymore et al. (SEYMORE; ROSENFIELD, 1999), onde um modelo construído manualmente contendo múltiplos estados por campos (título, autor, etc), pode ser mais eficiente do que um modelo com um estado por campo.

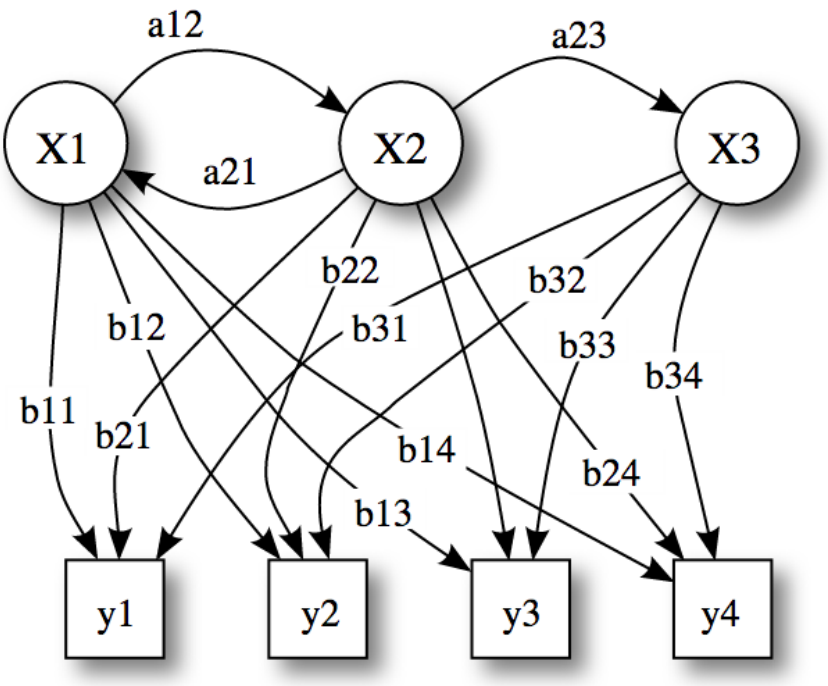
Um dos pontos positivos deste modelo é que por serem baseados em estatística eles são muito bem empregados em domínios de linguagem natural, aliando os resultados positivos à excelente performance computacional. Como desvantagem deste método podemos citar o fato de, por ser baseado em estatística, uma grande quantidade de dados deve ser utilizada a título de treino para obter os número significativos para então ser aplicados de maneira final.

Deste modo, para extração dos metadados o HMM pode ser utilizado aplicando um marcador (*label*) em cada palavras do cabeçalho de um documento (artigo científico), relacionando cada uma a uma classe, como título, autor, etc.

2.1.3 Word Clustering

Utilizando de trabalhos tradicionais, como (HAN C. LEE GILES; FOX, 2003), Han et al. (HAN EREN MANAVOGLU; ZHANG, 2005) apresentou uma ideia de um

Figura 2 – Exemplo de modelo HMM, onde X são os estados, Y as observações possíveis, A as probabilidades de mudança de estado e B as saídas das probabilidades.



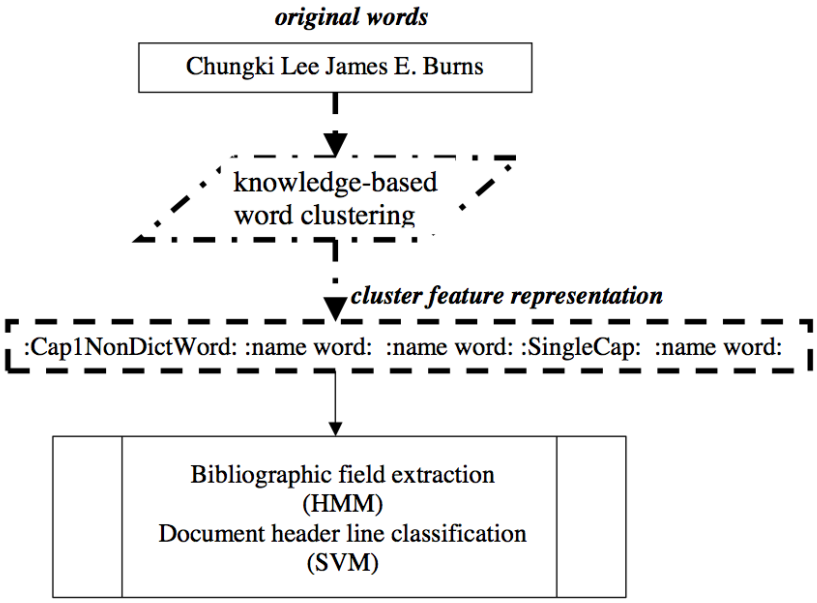
cluster de palavras para promover a extração de metadados de documentos, indo de maneira relativamente contrária às propostas tradicionais, baseando apenas na ocorrência e estatísticas de palavras isoladas.

Técnicas baseada em *cluster* de palavras geralmente apresentam performance maiores do que as técnicas tradicionais (HAN EREN MANAVOGLU; ZHANG, 2005). Este grupo de palavras demonstra relação entre palavras semelhantes dentro de um contexto, permitindo que a extração dos metadados possa ocorrer de maneira mais eficaz.

Han et al. agrupou bases de dados de domínios diversos incluindo também propriedades ortográficas de palavras, que possuem conhecimento prévio de classes específicas, como autor, título, etc. Deste modo palavras encontradas nos documentos vão sendo comparadas com palavras deste *cluster*, permitindo identificar por grupos características de metadados, como por exemplo, a palavra "Mary"faz parte do contexto de "nomes", portanto existe uma probabilidade maior de ela, juntamente com seu grupo de palavras ao redor fazerem parte da classe "autor"por exemplo. Esta lógica é apresentada para outras classes, como "e-mail"por exemplo, que pode ser identificado, em sua grande maioria, com a presença do caractere "@".

Han et al. ainda utiliza da técnica de SVM (Support Vector Machines) para classificação de linhas de um cabeçalho de um documento, tanto em função dos bons resultados obtidos quanto também pela boa performance apresentada. Deste modo, cada linha obtida se transforma em um vetor de palavras, que é comparado com o *cluster*,

Figura 3 – Workflow da extração de metadados usando *cluster* de palavras



identificando mais facilmente os metadados.

Além disso é utilizada também a técnica de HMM (RABINER, 1989) para a extração das referências, observando sempre a ocorrência de padrões como título e autor para identificação dos mesmos em referências existentes no documento.

Basicamente a *Rule-based Word Clustering* se resume em 3 (três) etapas. A primeira etapa se resume na construção das bases de dados, assim como (HAN C. LEE GILES; FOX, 2003), onde foram utilizadas também como bases externas os nomes apresentados em seu *cluster* de maneira a unir não somente estas bases externas mas também bases construídas dentro de um domínio específico, como palavras que fazem parte de um conjunto finito, não genérico.

A segunda etapa é chamada de *Cluster Design*. Esta etapa é onde os *clusters* são arquitetados, de maneira a contemplar também propriedades ortográficas das palavras, como se funcionasse de maneira geral como uma expressão regular de palavras, formando então um *cluster* com base nas características apresentadas.

A terceira etapa é chamada de *Rule Design*, que resume-se basicamente na representação de cada palavras dentro de seu contexto de apresentação. Por exemplo, nomes devem começar com a primeira letra maiúscula para então serem classificadas como do grupo de "nomes".

Como resultado este *cluster* permite um ganho considerável de performance, além de permitir uma precisão maior dos resultados, visto que eles são apresentados com base nestas bases focadas em um domínio específico, perfazendo um contexto mais definido e com resultados mais garantidos ao se comparar por exemplo com técnicas de *cluster*

distribuídos.

Por outro lado a utilização desta técnica possui uma falha em semântica dos dados, visto que quando um dígito ou conjunto deles é substituído por ":number"ele se torna apenas um número, sem um contexto específico, ou seja, pode ser tanto uma referência a alguma página ou até mesmo um mês de um ano.

2.1.4 Conditional Random Fields (CRFs)

Conditional Random Fields é um framework proposto por Lafferty et al. (LAFFERTY; PEREIRA, 2001) criado para construir modelos probabilísticos e dados marcados em sequência (*label sequence data*), geralmente utilizados no reconhecimento de padrões e aprendizado de máquinas (*machine learning*).

Sua representação é puramente matemática com modelos gráficos a fim de maximizar as probabilidades condicionais que se desejam aplicar.

Esta técnica é comparada e quase sempre utilizada juntamente com a HMM (Hidden Markov Models), de maneira a possuir algumas vantagens sobre esta última, como a habilidade de relacionar pressupostos independentes nos modelos, ou seja, relacionar observações e/ou interpretações.

Além disso, CRFs são utilizadas também em marcação e parseamento de dados sequenciais, com uma ordem determinada, como linguagem natural, sequências biológicas (como os genes) ou estados computacionais.

Sua aplicação na extração de metadados foi apresentada por Peng et al. (PENG; MCCALLUM, 2004), como uma maneira eficaz de extrair padrões em cabeçalhos e referências de artigos científicos. Deste modo, através da identificação destes padrões sequenciais pode-se determinar os tipos de dados existentes e então identificá-los, seguindo uma lógica/ordem pré-determinada.

A utilização de CRFs na extração de metadados mostra-se muito eficaz por reduzir bastante os erros em algumas métricas, aumentando o sucesso da aplicação desta técnica nesta área.

2.2 Projetos de Destaque

Alguns projetos baseiam sua extração em padrões pré-definidos de maneira a identificar dados relevantes dentro de uma região específica, facilitando a procura e consequentemente aumentando a velocidade no resultado.

Estes projetos geralmente permitem uma variedade muito grande de layouts, embora nem todos já estejam previamente definidos. Geralmente novos layouts são inseridos em

novas versões ou até mesmo por contribuições das mais diversas, como é o caso dos projetos de código livre, os chamados projetos *open source*.

Abaixo segue uma relação dos principais projetos relacionados à área de extração de metadados de artigos científicos, com informações sobre seu funcionamento e algoritmos que são utilizados.

2.2.1 Cermine

Um destes projetos é o recente CERMINE (TKACZYK PAWEŁ SZOSTEK; BOLIKOWSKI, 2014), uma biblioteca *open source* desenvolvida na linguagem de programação Java que permite que sejam extraídos os metadados de artigos científicos em formato digital PDF, oferecendo ainda a possibilidade de cruzamento de dados por meio de referências e títulos, permitindo assim identificar citações bem como a relevância de um determinado documento.

O CERMINE ainda possui um mecanismo de aprendizagem da própria máquina, permitindo que na medida que dados forem sendo alterados ele consiga absorver os detalhes e permitir assim uma mudança de sua maneira de extrair os dados. Deste modo ele permite que seja adaptado para novos padrões de layouts, o que permite de maneira geral que uma grande gama de modelos seja então abrangida.

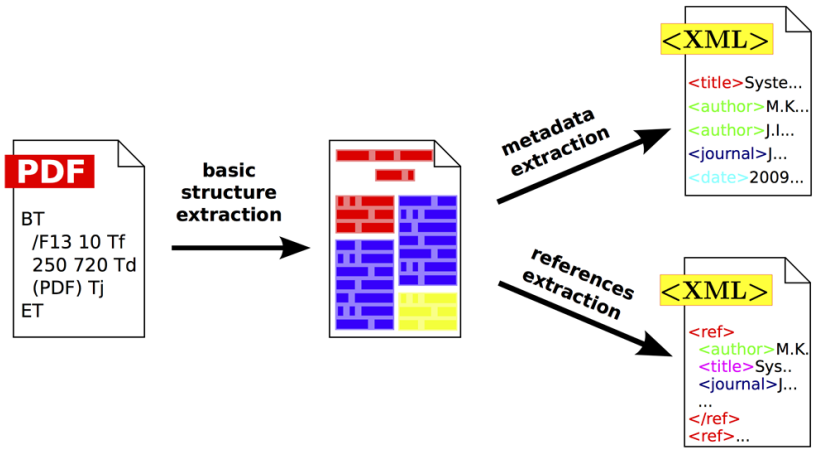
Seu grande diferencial em comparação com as demais técnicas é que ele não somente extrai os metadados de um artigo, mas sim analisa todo o seu conteúdo, incluindo citações a outros artigos, que podem ser facilmente cruzados por meio de informações como título e autor(es).

Seu mecanismo considera somente arquivos PDF com texto gerado de maneira pura, sem a utilização de imagens. Basicamente ele considera regiões, linhas e páginas como pontos estratégicos para a extração de informações. As bases destas regiões possuem padrões que são utilizados juntamente com técnicas de SVM (HAN C. LEE GILES; FOX, 2003). Com base nisso ele condensa um layout onde as informações geralmente estão dispostas, permitindo assim que em um determinado local do arquivo esteja, provavelmente, o título e o nome dos autores.

Com estas regiões definidas o CERMINE extrai as informações com base em padrões preestabelecidos, de maneira a gerar então sua saída para os metadados e sua saída para as referências encontradas. A saída trabalhada pelo projeto é no formato XML, permitindo assim que possa ser compartilhado com outros sistemas por possuir uma leitura semântica e ao mesmo tempo fácil de ser interpretada pelas linguagens de máquinas. A figura 4 demonstra como o processo de extração do CERMINE funciona.

Com o mapeamento definido ele identifica regiões de acordo com seu conteúdo, as quais ele chama de *zones*. Esta regiões são determinadas a fim de extrair as informações

Figura 4 – CERMINE Extraction Workflow



relevantes para cada uma, de maneira a separar, por exemplo, a área destina aos metadados do arquivo. O CERMINE divide estas *zones* da seguinte maneira:

- **Metadata:** É a região mais ao alto do documento, onde obtém os metadados, que seriam o resumo, *bib_info*, tipo, título, afiliação, autores, datas, editores e palavras-chaves.
- **References:** Região responsável por identificas detalhes de referências que foram utilizadas no artigo, como título e autores, por exemplo
- **Body:** O texto geral do artigo, incluindo equações, imagens e tabelas.
- **Other:** Outros detalhes menos significantes semanticamente, como número das páginas, dentro outros.

A extração das referências abrange também seus próprios metadados. Tanto no texto corrido (*Body*) quanto na lista de referências do artigo o *parser* do CERMINE analisa linha a linha, permitindo uma extração de dados mais eficaz. Das referências são extraídos os seguintes dados: autor, título, nome do *journal*, volume, *issue*, páginas, *publisher*, localização e o ano.

2.2.2 TeamBeam

Outros projeto de destaque é o TeamBeam (KERN KRIS JACK; GRANITZER, 2012), cuja base ideológica possui objetivos bem sociais, de contribuir com o compartilhamento de conhecimento. Basicamente o objetivo do projeto é extrair metadados de artigos científicos, porém focado apenas nestes, de maneira a extrair título, nome do *journal*, resumo e informações sobre os autores, como nome, endereço de e-mail e afiliações.

O projeto também é de código livre (*open source*) e é baseado na extração de pequenos blocos de texto. A manipulação dos arquivos PDF são feitas pela biblioteca PDFBox [3], que fornece meios eficazes de extrair textos com base nas regiões desejadas.

O algoritmo do TeamBeam utiliza o algoritmo de *Maximum Entropy* (BERGER; PIETRA, 1996), que utiliza basicamente de tarefas de classificação sequencial como ferramenta principal para obtenção de padrões. A base deste algoritmo está na utilização de CRFs (LAFFERTY; PEREIRA, 2001), principalmente no que diz respeito à extração dos metadados (PENG; MCCALLUM, 2004).

O processo de extração é feito basicamente em duas etapas. A primeira é a etapa de classificação de blocos de texto (*text block classification*), onde geralmente já é possível obter algum dado concreto de resultado. Neste etapa o objetivo é associar certos blocos de texto a um dos seguintes marcadores: *Title Block*; *Sub-Title Block*; *Journal Block*; *Abstract Block*; *Author Block*; *E-Mail Block*; *Affiliation Block*; *Author-Mixed Block*; e *Other Block*.

Dependendo do layout do artigo alguns metadados podem vir divididos em blocos de texto diferentes, necessitando de um processamento posterior, como é o caso, geralmente, dos blocos com informações sobre os autores. Neste caso também é realizada a etapa de classificação de token (*token classification*), que se resume na classificação de palavras individualmente de acordo com um dos seguintes marcadores: *Given Name*; *Middle Name*; *Surname*; *Index*; *Separator*; *E-Mail*; *Affiliation-Start*; *Affiliation*; e *Other*.

Kern et al. defendem a ideia dos excelentes resultados do TeamBeam ao ser comparado com outros projetos. Este fato é dado em virtude das características que são levadas em consideração no processamento do algoritmo, utilizando de dicionários, informações de layout e modelo de linguagem.

2.2.3 ParsCit

Escrever sobre.

2.2.4 Layout-CRFs

Escrever sobre.

2.2.5 Mendeley Desktop

Escrever sobre.

[3] Biblioteca de manipulação de arquivos PDF mantida pela Fundação Apache. Disponível em <https://pdfbox.apache.org/>

2.2.6 CiteSeer

Escrever sobre.

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.

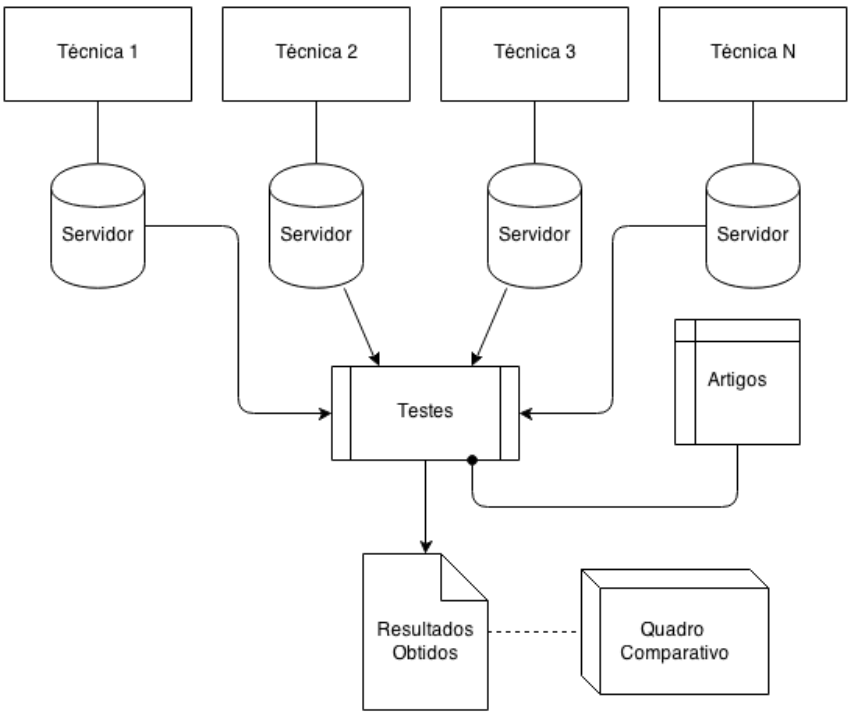
Desta maneira, a pesquisa de modo padrão não traz alteração nos ambientes pesquisados, apenas os analisa e os compara com base em padrões estabelecidos como sendo de resultado adequado. Assim, o projeto em si trata muito mais de pré-seleção de documentos e técnicas a serem testadas como também dos resultados 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 5.

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 5 – 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 2.

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 2 – 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 3, 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 3 – 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 4 – 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 4.

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

BERGER, S. A. D. P. A. L.; PIETRA, V. J. D. A maximum entropy approach to natural language processing. 1996. Citado na página 36.

HAN C. LEE GILES, E. M. H. Z. Z. Z. H.; FOX, E. A. Automatic document metadata extraction using support vector machines. 2003. Citado 4 vezes nas páginas 28, 30, 32 e 34.

HAN EREN MANAVOGLU, H. Z. K. T. C. L. G. H.; ZHANG, X. Rule-based word clustering for document metadata extraction. p. 1049–1053, 2005. Citado 2 vezes nas páginas 30 e 31.

KERN KRIS JACK, M. H. R.; GRANITZER, M. Teambeam — meta-data extraction from scientific literature. v. 18, n. 7/8, 2012. Disponível em: <http://www.dlib.org/dlib/july12/kern/07kern.html>. Citado na página 35.

KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. [S.l.], 2004. Citado 2 vezes nas páginas 24 e 27.

LAFFERTY, A. M. J.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ACM (Ed.). *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*. [S.l.]: ACM, 2001. p. 282–289. Citado 2 vezes nas páginas 33 e 36.

PENG, F.; MCCALLUM, A. Accurate information extraction from research papers using conditional random fields. In: *HLT-NAACL04*. [S.l.: s.n.], 2004. p. 329–336. Citado 2 vezes nas páginas 33 e 36.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. v. 77, n. 2, 1989. Citado na página 32.

RABINER, L. R.; JUANG, B. H. An introduction to hidden markov models. 1986. Citado na página 30.

SEYMORE, K.; ROSENFIELD, R. Learning hidden markov model structure for information extraction. p. 37–42, 1999. Citado 2 vezes nas páginas 28 e 30.

TKACZYK PAWEł SZOSTEK, P. J. D. M. F. D.; BOLIKOWSKI Łukasz. Cermine — automatic extraction of metadata and references from scientific literature. 2014. Citado na página 34.

WEIBEL, S. The dublin core: a simple content description format for electronic resources. p. 117–119, 1999. Citado na página 28.