Utilização de Técnicas de Recuperação da Informação na Construção de uma Ferramenta de Busca em Documentos do Arquivo Público de Biguaçu

Alessandro Mueller
alessandro@univali.br
UNIVALI

Luiz Gonzaga Ferreira Junior
luixjunior@hotmail.com
UNIVALI

Resumo:O Arquivo Municipal de Biguaçu tem aumentado gradativamente sua quantidade de documentos armazenados e esse aumento faz surgir a necessidade de se encontrar uma maneira eficiente para o processo de busca e localização destes documentos. Com a utilização da recuperação da informação, área da Ciência da Informação que tem como principal técnica a recuperação de referências aos documentos pesquisa de palavra chave, tornou-se possível o desenvolvimento de uma ferramenta de busca em documentos no Arquivo Público de Biguaçu. A ferramenta desenvolvida possui como diferencial o tratamento da língua portuguesa, permitindo que o conteúdo dos documentos seja analisado.

Palavras Chave: Recuperação - Informação - Mineração - Arquivos - Documentos

1. INTRODUÇÃO

A informação ocupa uma posição central no mundo contemporâneo, pois é utilizando a informação que cada parte da sociedade se organiza e define seus planos de ação. É a informação que serve de base para tomada de decisão de empresas e instituições de todo o mundo. Atualmente, a posse da informação é sinônimo de poder e as tecnologias da informação podem ser consideradas o retrato do novo mundo. A preservação de informações e o acesso às mesmas estão cada vez mais presentes no dia-a-dia das pessoas e por isso esse assunto tem merecido a atenção de cientistas em todo o mundo. Com isso, desenvolveu-se a Ciência da Informação, uma área de conhecimento voltada para a questão informacional. Seu objetivo é estudar o elemento fundamental do mundo moderno: a informação (RAMOS, 2008).

O Arquivo Público de Biguaçu vem aumentando gradativamente o número de documentos arquivados, tornando necessária sua informatização. Com esse grande volume de documentos armazenados, procurou-se desenvolver um meio eficiente para que aconteça a busca e localização desses documentos. Assim, a forma encontrada para que esse problema seja resolvido foi o desenvolvimento de uma ferramenta de busca em documentos do Arquivo Público de Biguaçu.

O desenvolvimento de uma ferramenta que possa indexar e buscar os documentos digitalizados do Arquivo Municipal de Biguaçu é muito importante, pois facilita e auxilia no processo de busca, fazendo com que os documentos possam ser encontrados com mais agilidade. Essa ferramenta é capaz de buscar os documentos indexados em um tempo compatível com a necessidade do Arquivo Público, levando-se se em conta que o Arquivo deve aumentar gradativamente o número de documentos informatizados.

Para que a ferramenta fosse desenvolvida, foi necessário um estudo sobre o caso especifico do Arquivo de Biguaçu, conhecendo-se assim o funcionamento geral do Arquivo, a forma como eram efetuadas as buscas antes do desenvolvimento da ferramenta, a maneira como os documentos são informatizados e sua respectiva indexação, para que se possa obter informações e aplicar os métodos adequados no processo de recuperação da informação.

A recuperação da informação pode ser entendida como um processo de recuperação de referências de documentos em resposta a alguma solicitação, onde os sistemas de recuperação da informação são sistemas de operações interligadas para identificar, entre um grande conjunto de informações, aquelas que são realmente úteis, isto é, que estão de acordo com a demanda expressa pelo usuário (ARAÚJO JUNIOR, 2007).

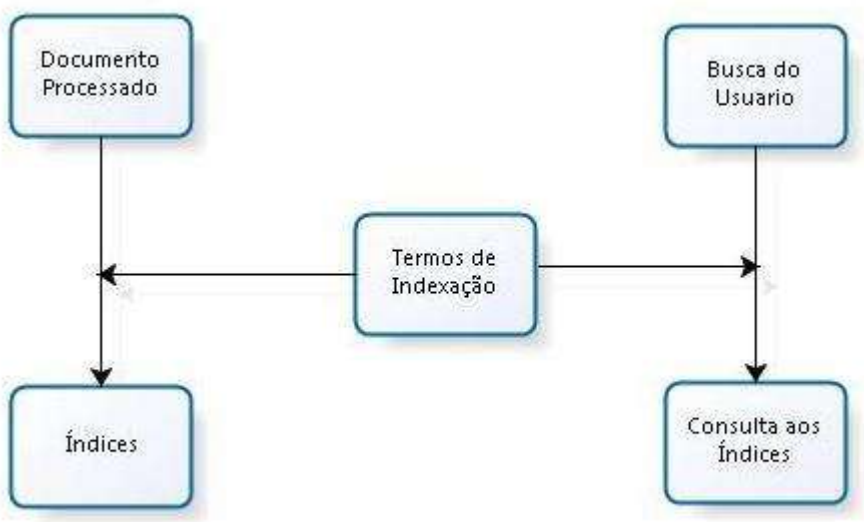

Figura 1 Representação do processo de recuperação de informação a partir de um índice.

Fonte: (Robredo 2005)

Para que as informações sobre determinado documento possam ser encontradas e recuperadas, é necessário que haja a indexação dos mesmos. A indexação é um processo que visa obter o acesso à informação dos documentos, por intermédio de termos ou códigos, atuando como ponto de partida para selecionar os próprios documentos. O índice é o mais importante instrumento para recuperar a informação, tendo em vista que o mesmo é como uma "chave" que dá acesso à informação contida nos documentos, ou como uma "ponte" entre o conteúdo de um acervo de documentos e os usuários (ROBREDO, 2005).

Durante a indexação são obtidos os conceitos do documento através da análise do seu conteúdo e traduzidos para uma linguagem de indexação, tesauros, cabeçalhos de assunto, etc. Esta representação identifica o documento, definindo seus pontos de acesso para a busca, podendo também substituir o documento (FERNEDA, 2003).

Após tomar conhecimento da maneira como é feita a indexação dos documentos no Arquivo de Biguaçu, além das necessidades que o mesmo possui para que sejam efetuadas as buscas, pode-se então tornar concreto o desenvolvimento da ferramenta que deve auxiliar e agilizar o trabalho de busca de documentos.

2. O PROJETO

Os sistemas de busca e recuperação da informação são tão variados quanto os métodos de organização da informação. Assim, se os documentos do acervo foram organizados por assuntos, um meio de busca e recuperação será a busca e escolha direta nas estantes. Se os documentos foram organizados em função de outro critério (por tipo de suporte, por exemplo, microfichas, filmes, etc., ou por tamanho ou cores), não é possível a

pesquisa direta na estante e torna-se necessário o uso de algum tipo de índice ou catálogo sistemático (ROBREDO, 2005).

Sendo assim, foi preciso efetuar uma análise da maneira em que os documentos do Arquivo de Biguaçu estão organizados, para que se possa compreender a melhor forma de se recuperar essa informação. Um fato importante de ser destacado, é que os documentos que foram informatizados de forma não automática – pela digitação – já estão indexados de acordo com a necessidade de busca do Arquivo de Biguaçu, isto é, todos os dados contidos nos documentos digitais devem ser mantidos em sua forma integral, para que palavras importantes para a pesquisa não sejam removidas.

Porém, notou-se que a análise do texto dos documentos teria grande importância já que melhoraria a qualidade dos resultados. Por isso procurou-se fazer o tratamento e análise do texto através das técnicas de remoção de stowords e stemming. Com a remoção de stopwords foi possível fazer a remoção de palavras com baixo valor semântico como artigos e preposições. Com a técnica de stemming é possível reduzir as variantes de uma palavra a um mesmo radical comum.

Todas as funcionalidades da ferramenta foram desenvolvidas para atender ao caso específico do Arquivo de Biguaçu, o que tornou necessária a implementação de alguns recursos como a utilização de tesauros, a tela de busca avançada e a criação e manipulação de índices. Além disso, foi necessária a utilização de ferramentas que auxiliassem nas fases de coleta de documento e indexação. Para essas atividades foram utilizadas respectivamente as ferramentas Apache POI e Apache Lucene.

2.1 FUNCIONAMENTO DA FERRAMENTA

A ferramenta foi desenvolvida de modo a proporcionar agilidade e facilidade no processo de busca e recuperação de documentos. Dessa forma, procurou-se fazer com que a tela inicial do programa, que está disponível para todos os usuários (administrador e usuário comum), seja exatamente a tela de pesquisa de documentos.

Nessa tela estão disponíveis todos os recursos que o usuário possui para pesquisa, permitindo a entrada dos termos que serão pesquisados, a escolha do conjunto de documentos sobre o qual será feita a pesquisa, a utilização ou não de tesauros, a quantidade máxima de resultados a serem apresentados e, ainda, a opção de pesquisa avançada.

No campo referente às buscas, as mesmas poderão ser realizadas através de termos e operadores. Dessa forma, através de operadores lógicos (booleanos), pode-se formar expressões de busca mais complexas e detalhadas. Com uma combinação entre termos e operadores pode-se obter uma busca mais específica ou mais abrangente dependendo da necessidade do usuário. O sistema retornará os resultados em ordem de similaridade entre os termos e os resultados encontrados. Um exemplo de busca com operadores booleanos pode ser visto na Figura 2.

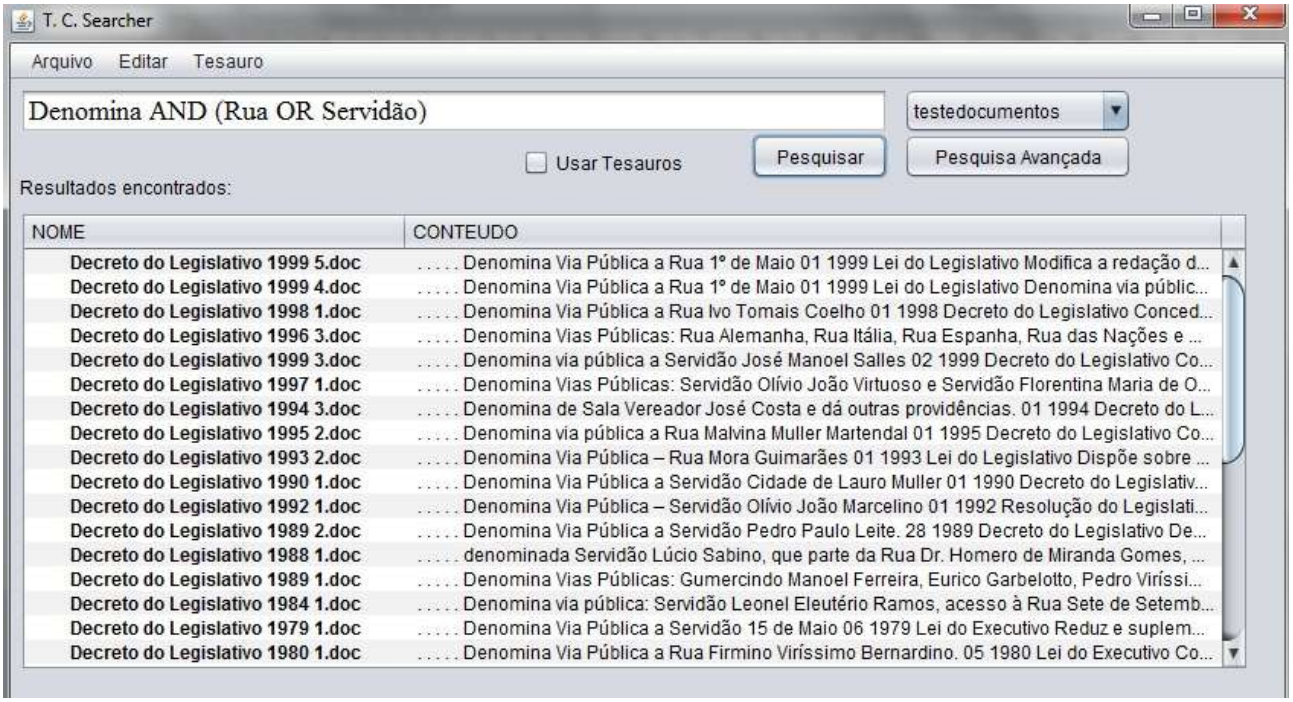

Figura 2 - Exemplo da utilização de operadores

Dessa forma pode-se utilizar uma grande quantidade de combinações de operadores permitindo que se tenha uma maior probabilidade de refinamento de resultados. A Tabela 1 mostra os operadores e como podem ser utilizados no contexto da ferramenta.

Tabela 1 - Lista de Operadores

Exemplo (Pesquisa)	**Resultado**
Hospital	Documentos com o termo **hospital**
Hospital AND Construção	Documentos que contenham os termos **Hospital** e **Construção**
Hospital OR Farmácia	Documentos com no mínimo um dos termos
+Lei –Executivo	Documentos que contenham o termo **Lei,** mas não contenham o termo **Executivo**
"Denominação de Rua"	Documentos que contenham a frase buscada
Mari?	Documentos que comecem com o termo **Mari** e possuam mais um caractere, como **Maria** e **Mario**
Constr*	Documentos que iniciem com o termo **Constr**

Parana~	Documentos com sonoridade semelhante, baseado no algoritmo de Edit Distance
"Luiz Gonzaga Junior"~1	Documentos que contenham os termos pesquisado com uma distância máxima de um termo entre cada.
João^3 Joaquim	Documentos que possuem os termos pesquisados, porém com um peso maior para o termo especificado.
(Alvará OR Habite-se) +Arlindo	Documentos que contenham os termos **Alvará** ou **Habite-se**, e contenham o termo **Arlindo**

Ao executar qualquer tipo de pesquisa o usuário deverá selecionar uma base de dados disponível, isto é, o conjunto de documentos a ser pesquisado e recuperado. Isso permite que o usuário possa optar pelo conjunto de documentos que lhe convém, tornando os resultados da recuperação da informação mais precisos.

A criação dessas bases de dados só estará disponível para administradores do sistema, que são responsáveis pela escolha dos conjuntos de documentos necessários. Assim, os administradores poderão criar, atualizar e excluir conjuntos de documentos indexados, conforme a necessidade das buscas indicar. A Figura 3 mostra como é a tela de criação de índices.

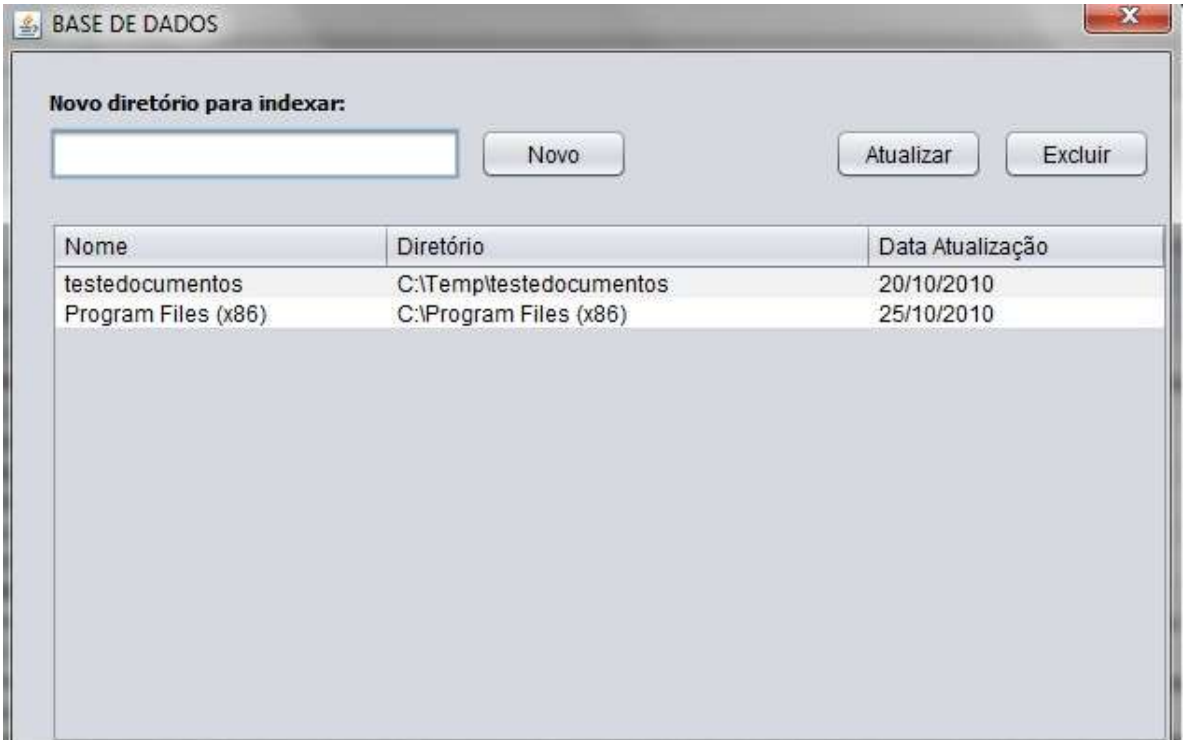

Figura 3 - Tela de criação de índices

2.2 DESENVOLVIMENTO

A primeira etapa no desenvolvimento do trabalho é a fase da coleta de documentos, onde são extraídos os textos dos documentos que serão utilizados pela ferramenta. Esse processo é feito com o auxílio da biblioteca Apache POI que é uma biblioteca de código aberto para Java, utilizada para leitura e escrita de texto nos formatos da Microsoft, como arquivos de Word, Excel e Power Point. Os seguintes campos são coletados de cada documento: título, conteúdo e diretório.

Após os documentos serem coletados tem-se a etapa de indexação onde foi utilizada a biblioteca Apache Lucene para auxiliar nesse processo. Essa biblioteca utiliza o sistema de lista invertida para armazenar seus índices. A Figura 4 simplifica o funcionamento desse sistema, onde cada termo é uma chave que aponta para os documentos correspondentes.

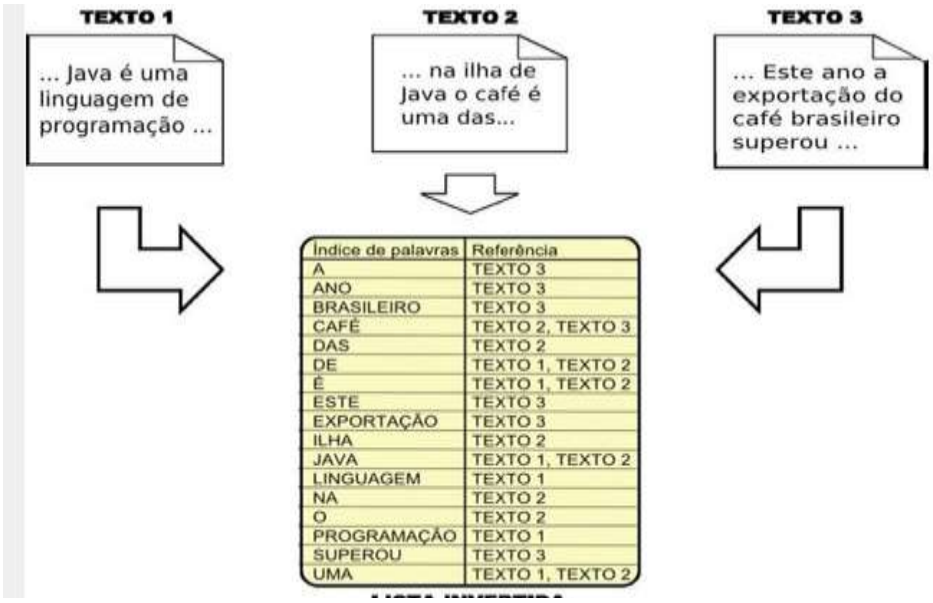

Figura 4 - Esquema simplificado de uma lista invertida

Fonte: (Veloso 2010)

Para fazer a análise do texto o Lucene permite a utilização de um analisador. O analisador é responsável por gerar ou retirar *tokens*, padronizar o texto, retirar *stopwords*, etc. Em geral permite realizar qualquer funcionalidade programável (KRAMER, 2008). Um trecho de código do processo de indexação pode ser observado na Figura 5

```
public void indexarDiretorio(File diretorio) throws Exception {
    File[] arquivos = diretorio.listFiles();
    if (arquivos==null)
        return;
    for (int i = 0; i<arquivos.length; i++) {
        File f = arquivos[i];
        if (f.isFile()) {
            if (f.getName().endsWith(".doc") && (f.isHidden()==false)) {
                WordReader w = new WordReader(f);
                insereArquivo(f.getName(), w.getConteudo(), f.getPath());
            }
        } else {
            indexarDiretorio(f);
        }
    }
}
private void insereArquivo(String titulo, String conteudo, String caminho) throws IOException, CorruptIndexException {
    Document doc = new Document();
    Field fieldTitulo = new Field(TITULO.getName(), titulo, Field.Store.YES, Field.Index.ANALYZED);
    Field fieldConteudo = new Field(CONTEUDO.getName(), conteudo, Field.Store.YES, Field.Index.ANALYZED);
    Field fieldCaminho = new Field(CAMINHO.getName(), caminho, Field.Store.YES, Field.Index.NO);
    doc.add(fieldTitulo);
    doc.add(fieldConteudo);
    doc.add(fieldCaminho);
    indexWriter.addDocument(doc);
```

Figura 5 - Processo de Indexação

Nesse trecho de código pode-se notar que o diretório de documentos é percorrido em busca de arquivos, sendo que, quando um diretório (pasta) é encontrado, o método usa recursividade, fazendo com que subpastas também sejam percorridas.

Então pode-se ver que o diretório dos documentos é percorrido em busca de arquivos, sendo que quando um diretório (pasta) é encontrado o método se chama novamente, fazendo com que as subpastas também sejam percorridas.

Quando um arquivo do tipo ".doc" é encontrado, então é chamado o método insereArquivo. Esse método utiliza a classe **Document** que é composta por *Fields* – campos onde estão as informações retiradas dos documentos – e faz a adição de um novo documento.

Para resolver o problema de palavras com significado semelhante ou palavras escritas de maneiras diferentes foi adotado o sistema de tesauros, onde os usuários podem efetuar o cadastro de palavras consideradas semelhantes. Assim, no processo de busca, caso o usuário opte por utilizar tesauros, o sistema também buscará pelas palavras semelhantes às palavras de busca. A tela de busca com tesauro pode ser vista na Figura 6.

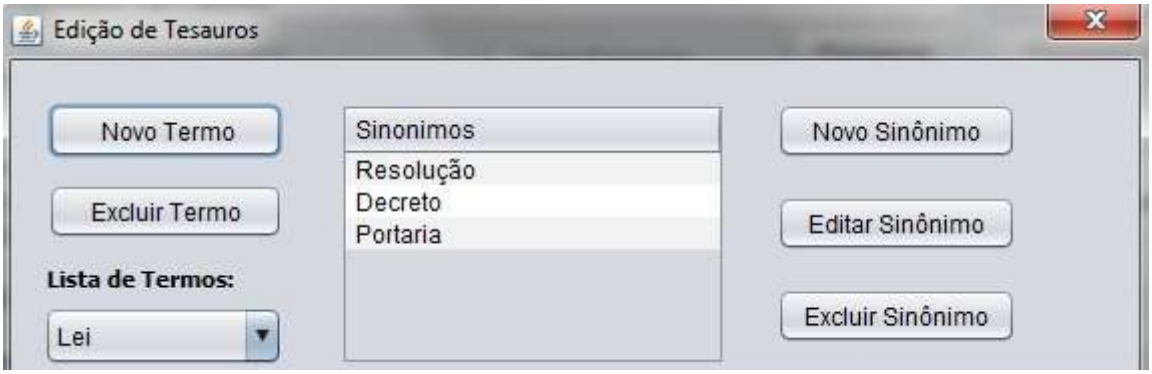

Figura 6 - Tela de utilização de tesauros

3. CONCLUSÕES

Apesar de ferramentas como o Lucene serem utilizadas mundialmente no processo de recuperação da informação, é necessário conhecer suas características e funcionalidades, o que demanda tempo de estudo do assunto para uma melhor utilização. O Lucene, que é uma biblioteca muito completa, contém apenas o núcleo do "motor" de busca. Dessa forma, foi preciso a colaboração de outras ferramentas e a implementação de outros procedimentos que auxiliassem a tornar o sistema mais completo.

Outro problema enfrentado – e esse é um dos problemas mais encontrados nesse tipo de ferramenta – foi o tratamento dos textos dos documentos. Muitos documentos possuem conteúdo que não é de interesse no processo de indexação e busca e precisa ser tratado. Figuras, tabelas, quebra de linha e outros códigos são desnecessários e acabam atrapalhando no processo da coleta e indexação de documentos.

Procurou-se tratar todos estes problemas para que se obtivesse um resultado satisfatório quando a ferramenta fosse concluída, sendo que para o atual contexto do Arquivo Público de Biguaçu, a ferramenta mostrou ser capaz de auxiliar e otimizar o processo de busca e recuperação da informação. A análise dos resultados obtidos e a utilização da ferramenta no Arquivo Público de Biguaçu provaram a eficiência da ferramenta ao tratar os documentos do mesmo, com o grande diferencial sendo o tratamento de documentos em língua portuguesa.

Todavia, isso não significa que a ferramenta não deva evoluir. Assim como a informação e o número de documentos crescem de forma vertiginosa, a tecnologia vem se aperfeiçoando para tratar com problemas cada vez mais complexos. Neste caso não deve ser diferente, pois trabalhos futuros podem e devem ser desenvolvidos.

Como a tendência do Arquivo Público de Biguaçu é a transformação automática de documentos em arquivos textuais com o uso da tecnologia OCR, o conteúdo dos documentos informatizados tende a mudar de característica, passando a ser um conteúdo formado por textos maiores, o que implicará em um maior tratamento.

Além disso, quando os documentos informatizados tiverem também sua imagem digitalizada, deve ser possível que a ferramenta localize além do conteúdo documento, a sua respectiva imagem. Isso poderá fazer com que a ferramenta forneça ao usuário, além

do conteúdo do documento, a sua reprodução fiel ao original. A ferramenta também poderá ser integrada com o scanner e o OCR, fazendo com que a mesma ferramenta possa digitalizar, transformar em texto, indexar, buscar e reproduzir os documentos.

Também é interessante ressaltar que a qualidade dos resultados da ferramenta desenvolvida ainda pode ser melhorada, pois várias técnicas de Inteligência Artificial apresentadas neste trabalho podem ser adaptadas para auxiliar a ferramenta neste processo.

Por fim, os objetivos do projeto foram alcançados e a ferramenta desenvolvida atendeu às necessidades atuais do Arquivo Público de Biguaçu, servindo de fundamento para que novas técnicas e funcionalidades possam ser estudadas e desenvolvidas.

Referências

Araújo Júnior, R. H. "Precisão no processo de busca e recuperação da informação". Brasília: Thesaurus, 2007

Ferneda, Edberto. "Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação". São Paulo, 2003

KRAMER, João. " Lucene". Disponível em: < http://projeto.lexml.gov.br/Members/joaolima/02_lucene>. Acesso em 02/08/2010, 2010.

Ramos, L. B. "Centros de cultura, espaços de informação: um estudo sobre a ação do Galpão Cine Horto". Belo Horizonte: Argvmentvm, 2008.

Robredo, Jaime. "Documentação de hoje e de amanhã: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas". Brasília: Edição de autor, 2005

VELOSO, S. "Conhecendo o Apache Lucene"**.** Disponível em: <http://www.devmedia.com.br/articles/post-8308-Artigo-Java-Magazine-49-Conhecendo-o-Apache-Lucene.html>. Acesso em: 28/10/2010.