Opala: uma biblioteca de indexação e busca de textos e imagens por conteúdo

Lidijanne de Miranda Santos (1); Aécio Solano Rodrigues Santos (2);
Ricardo Martins Ramos (3); Valéria Oliveira Costa (4)
Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Laboratório de Pesquisa em Sistemas de Informação.
Praça da Liberdade, 1597, 64000-040, Centro, Teresina (PI)
(1) lidijanne@ifpi.edu.br
(2) aecio@ifpi.edu.br
(3) ricardo@ifpi.edu.br
(4) valeria@ifpi.edu.br

RESUMO

Este trabalho apresenta uma ferramenta que tem por objetivo facilitar o desenvolvimento de sistemas de busca por conteúdo de texto e imagem. Ela foi desenvolvida no Laboratório de Pesquisa em Sistemas de Informação (LAPESI) do IFPI (Instituto Federal de Educação, Ciência e Tecnologia do Piauí) para a Biblioteca Digital da Rede Federal de Educação Profissional, Científica e Tecnológica (EPCT). O artigo mostra as tecnologias utilizadas no desenvolvimento, bem como sua arquitetura, principais funcionalidades e a aplicação da ferramenta.

Palavras-chave: biblioteca digital, search engine, lire, lucene, recuperação da informação

1 INTRODUÇÃO

Segundo Cunha (1997), as bibliotecas digitais têm como característica uma coleção de documentos que compartilham a informação por meio do uso de redes de computadores. O processo de recuperação de informação (*information retrieval*) consiste em identificar nestas coleções de objetos digitais, quais deles atendem a necessidade de informação do usuário. Neste contexto, surgiu a Biblioteca Digital da Rede Federal de Educação Profissional, Científica e Tecnológica (EPCT)[1] que visa a disseminação do material científico e tecnológico produzido nos institutos da rede. Ela é um projeto de pesquisa desenvolvido colaborativamente entre o Instituto Federal do Piauí (IFPI) e o Instituto Federal Fluminense (IFF).

A Biblioteca Digital da EPCT é desenvolvida utilizando tecnologias de código aberto e Arquitetura Orientada a Serviços (*Service Oriented Architecture*). O IFPI é responsável por desenvolver o serviço de indexação e busca de documentos, que podem ser textos ou imagens. Para o desenvolvimento deste serviço, foi implementada a biblioteca *Opala*. Nela, a recuperação de texto é feita baseada no conteúdo (*full-text search*), utilizando técnicas de índices invertidos e o modelo de recuperação de informação vetorial. Já a recuperação de imagens, é feita utilizando recuperação de imagens por conteúdo (*Content-Based Image Retrieval - CBIR*), que utiliza as características inerentes a própria imagem, como cor, forma e textura, para recuperar imagens em uma grande base de dados (ZHANG *et al*, 2002). Além disso, a busca pode ser feita através de metadados inseridos durante a indexação de um documento.

As próximas seções deste trabalho estão estruturadas da seguinte forma: a seção 2 apresenta a fundamentação teórica, na seção 3 é apresentada a descrição da proposta deste trabalho, a Opala e sua arquitetura, descrevendo sua arquitetura e funcionalidades; e a seção 4 apresenta as discussões e considerações finais do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Diante do grande fluxo de informações que se tem acesso atualmente, torna-se de fundamental importância ter sistemas de buscas que retornem as informações de interesse para o usuário. No entanto, os sistemas de banco de dados relacionais e objeto-relacional não suprem esta necessidade, pois a busca e os termos

[1] http://www.renapi.org/biblioteca-digital

recuperados são claramente definidos e estruturados, enquanto que em um sistema de recuperação de informação, os textos se encontram em linguagem natural, às vezes, ambígua e não bem estruturados. Costa (2001), diz que um sistema de recuperação da informação deve ser capaz de interpretar o conteúdo dos documentos e classificá-los por ordem de relevância de acordo com a consulta apresentada pelo usuário, enquanto um sistema de recuperação de dados não tem este compromisso.

De acordo com a literatura (HALAWAVI *et al,* 2006), os primeiros trabalhos em recuperação de imagem eram baseados em anotações de palavras chaves que descreviam a imagem, que nem sempre estão disponíveis ou não descrevem de maneira adequada o conteúdo visual da imagem, além de ser mais difícil encontrar resultados eficazes com consultas por texto. Enquanto que um sistema de recuperação de imagens baseado em conteúdo (CBIR) está preocupado em encontrar imagens com informações visuais similares a imagem submetida pelo usuário, em vez de apenas procurar pelos termos que descrevem a imagem.

A seguir descrevemos como ocorre o procedimento de recuperação de documentos textuais e a recuperação de imagens por conteúdo, além de apresentar as principais tecnologias utilizadas no desenvolvimento da biblioteca *Opala.*

2.1 Recuperação de documentos textuais por conteúdo

Para que se possa fazer buscas no conteúdo de um documento, é necessário ter um processo para construir uma representação das informações contidas no texto de forma organizada. Isto é feito através da extração ou associação de termos ou descritores utilizados para avaliar a relevância do documento, com o propósito de retornar um resultado satisfatório de forma rápida e precisa a uma consulta. Este processo é a indexação.

Segundo Costa (2001), um índice é uma estrutura de dados capaz de identificar todos os documentos que possuam as combinações de termos especificados numa pergunta ou consulta de busca. O software indexador é o responsável pela manutenção e alimentação do índice do mecanismo. É ele quem prepara os documentos para a indexação (através da aplicação de operações de normalização, como a remoção de *stopwords*[2], remoção de sinais de acentuação e identificação dos radicais das palavras), lê os documentos armazenados e extrai destes os termos que farão parte do índice – termos de índice.

A técnica de indexação utilizada para construir e manter os índices na biblioteca *Opala* é de arquivos invertidos. Arquivos invertidos são estruturas compostas por duas partes: o vocabulário e as ocorrências. O vocabulário incorpora o conjunto de todas as palavras distintas existentes no documento. Para cada palavra do vocabulário são construídas listas que contêm as exatas posições nas quais aparecem dentro do texto. O conjunto de todas as listas é denominado de ocorrências (BAEZA-YATES *apud* COSTA, 2001). Em uma coleção de documentos, para cada palavra existente é armazenado também o número do documento no qual ocorre. A busca num arquivo invertido se dá através da verificação das entradas do arquivo e a recuperação de todos os documentos que citam o termo usado.

Um modelo de recuperação de informação parte do princípio de que cada termo de um documento possui valor particular para representação do conteúdo semântico do documento no qual está contido. Esse valor é quantificado em forma de peso. Sendo assim, a cada termo de um documento é associado um peso que mede a sua importância para representar o documento que o contém. Estes pesos são aplicados à formula do modelo de recuperação da informação utilizado, para que seja recuperado o conjunto de documentos que melhor atende a necessidade informacional do usuário. (COSTA, 2001)

O modelo de recuperação de informação da *Opala* tem por base o modelo vetorial, que segundo Costa (2001) fundamenta-se na premissa de que tanto a consulta quanto os documentos indexados podem ser representados como vetores em um espaço *n*-dimensional, onde *n* é o número de termos distintos presentes no índice. O modelo vetorial utiliza pesos para calcular o grau de similaridade entre consulta e documentos armazenados, permitindo que documentos que não satisfazem integralmente a consulta sejam recuperados. A partir dos vetores de cada documento e da consulta, calcula-se um conjunto de resultados ordenado por relevância, e que apresenta os documentos de maior similaridade antes dos julgados de menor similaridade.

Portanto, o processo de busca é iniciado quando os usuários fazem uma consulta na coleção dos documentos indexados. Os termos da consulta são submetidos aos mesmos processos de normalização dos documentos

[2]Palavras com alta taxa de freqüência e que ocorrem na maioria dos documentos da coleção.

indexados na base de dados, pois estes termos serão procurados no índice. Em seguida, é executado um algoritmo de busca, o qual compara os documentos indexados na base de dados com a finalidade de encontrar os que satisfazem os termos fornecidos pelo usuário. Os documentos são selecionados e apresentados em ordem decrescente de relevância calculado pelo algoritmo de relevância, este é responsável por avaliar que documentos do índice possuem maior peso ou maior importância para uma consulta.

2.2 Recuperação de imagens por conteúdo (CBIR)

A recuperação de imagem baseada em conteúdo é a técnica na qual se utiliza do conteúdo visual para buscar imagens em uma grande base de dados conforme o interesse do usuário. Assim, o objetivo de um sistema CBIR é transformar o conteúdo visual do usuário (ou seja, as características intrínsecas à própria imagem como cores, formas, textura entre outros) em dados numéricos e/ou textuais gerando um conjunto de dados que representam o mais próximo possível a semântica expressada na imagem (LONG; ZHANG; FENG, 2003)

Um sistema de CBIR pode usar vários descritores para extrair o conteúdo das imagens. O descritor da imagem é um procedimento que caracteriza a extração de cada atributo inerente da imagem (cor, forma ou textura). A seleção dos descritores a serem utilizados constitui um fator importante para obtenção de um desempenho satisfatório nesses sistemas (SOUSA, 2009).

O processo de recuperação da imagem envolve duas fases: *indexação* e *busca*. A indexação inicia com a extração das propriedades visuais da imagem gerando um vetor multidimensional chamado vetor de características. Este armazena os valores numéricos dos descritores em um índice. Cada vetor de característica fica armazenado no índice de imagens indexadas que juntos formam o banco de imagens da aplicação. Durante o processo de busca, é gerado o vetor de características da imagem de consulta que é comparado aos demais vetores do banco para cálculo de similaridade. O resultado desta comparação é a recuperação do conjunto de imagens ordenado por relevância.

3 DESCRIÇÃO DA PROPOSTA

A *Opala* é uma biblioteca Java desenvolvida com a finalidade de provê recuperação de documentos de texto e imagem, possuindo funcionalidades de indexação e busca por conteúdo e metadados. Apesar de ter sido desenvolvida para uma biblioteca digital, ela pode ser utilizada no desenvolvimento de qualquer aplicação que precise de soluções de recuperação da informação por conteúdo e metadados. As tecnologias utilizadas são as bibliotecas Java Lucene e LIRe. A seguir, estas são descritas, bem como a arquitetura da *Opala*, suas principais funcionalidades e o Servidor XML-RPC que torna possível a utilização da ferramenta a partir de qualquer linguagem de programação.

3.1 Lucene

Segundo Hatcher (2004), o Lucene[3] é uma ferramenta *open-source* de recuperação de informação de alta performance, sendo responsável por indexar e pesquisar qualquer dado, desde que este seja convertido para o formato textual, podendo criar e armazenar informações no índice de texto. Ele utiliza a técnica de índices invertidos na indexação do conteúdo e o modelo de recuperação vetorial para ordenar os objetos de uma busca por relevância.

3.2 LIRe (*Lucene Image Retrieval*)

De acordo com Lux (2008), o LIRe[4] é uma biblioteca *open-source* que possibilita a recuperação de imagem baseada no conteúdo. O LIRe utiliza, dentre outros, os descritores (*Edge Histogram Descriptor - EHD; Color Layout Descriptor - CLD; Scalable Color Descriptor - SCD*) de imagem do padrão MPEG-7. A API armazena os vetores de características dos descritores em um índice do Lucene, a qual oferece rápidos serviços de indexação e busca. Os índices são armazenados em sistemas de arquivos. Isso traz vantagens em

[3] http://lucene.apache.org
[4] http://www.semanticmetadata.net/lire

relação ao armazenamento em banco de dados já que não há a necessidade de um servidor de banco de dados, estruturas de índice e gerenciamento de usuários, transações e acesso, por exemplo. O descritor utilizado na *Opala* é uma combinação dos descritores CLD/EHD. De acordo com pesquisas realizadas por Sousa (2009), a combinação destes dois descritores apresentaram um melhor desempenho, medido através do nível de revocação.

3.3 Arquitetura

A Figura 1 apresenta a arquitetura da biblioteca Opala. A Opala age como um facilitador para as *engines* de busca LIRe e Lucene, de maneira que seu usuário não necessite de um prévio conhecimento destas ferramentas para utilizá-las.

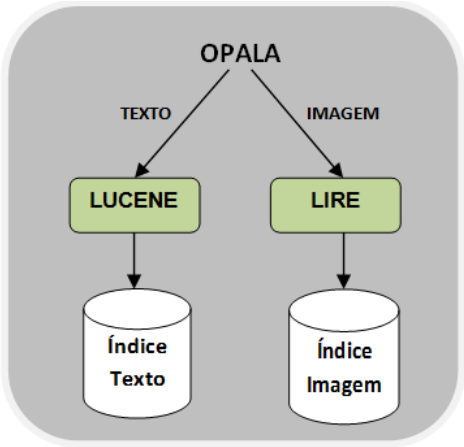

Figura 1 – Arquitetura da biblioteca Opala

Para usar a Opala basta saber a chamada dos métodos disponibilizados pelas suas interfaces de indexação e busca. O componente Opala provê as interfaces para a chamada dos métodos responsáveis pelas funcionalidades de indexação, exclusão, alteração e busca de documentos. Os principais métodos do estão descritos na Tabela 1.

Tabela 1 - Métodos principais da biblioteca Opala

MÉTODO	DESCRIÇÃO
TextIndexer.addText	Indexa um documento
TextIndexer.delText	Remove um documento do índice
TextIndexer.updateText	Atualiza um documento do índice
TextSearcher.searchText	Realiza busca por conteúdo no índice
ImageIndexer.addImage	Adiciona uma imagem no índice
ImageIndexer.delImage	Remove uma imagem do índice
ImageIndexer.updateImage	Atualiza uma imagem do índice
ImageSearcher.searchImage	Realiza busca de imagens por conteúdo no índice

Durante a indexação dos documentos é gerado um índice invertido para textos e outro para imagens, que são salvos no sistema de arquivos. Este índice é utilizado para realização das buscas.

3.4 Metadados

Para a Opala existem duas categorias de documentos: texto e imagem. Ambas podem ser buscadas por conteúdo e por metadados. Na *Opala*, os metadados são representados através dos atributos da classe

MetaDocument. Os atributos padrão estão descritos na Tabela 2. Além destes atributos, é possível adicionar metadados personalizados para o domínio de cada aplicação fornecendo uma chave identificadora do metadado e um valor para o mesmo. Os metadados padrão formam um subconjunto dos metadados do *DublinCore*[5].

Tabela 2 – Metadados padrão da Opala

PROPRIEDADES	DESCRIÇÃO
Id	Identificador de um documento
Author	Autor(es) do documento
Title	Título do documento
PublicationDate	Data da publicação
Format	Formato do documento
Type	Tipo de documento
Keywords	Palavras chaves de um documento

3.5 Principais funcionalidades

- **Indexação de Texto**: durante a indexação o texto passa por processo de análise, onde são removidas palavras irrelevantes para busca (*stop-words*). Além disso, cada palavras é reduzida ao seu radical. Após a análise, é gerado um índice invertido utilizando o texto processado e os metadados. Identificação dos autores.
- **Busca de Texto**: responsável por recuperar os registros a partir de uma consulta de palavras. Esta consulta é aplicada sobre a base de índices onde estão as informações que permitem a recuperação das informações que mais se aproximam das necessidades do usuário.
- **Indexação de Imagem**: nos sistemas CBIR as propriedades visuais de cada imagem são extraídas gerando um vetor multidimensional chamado vetor de características. Este vetor é guardado em um índice de imagens. Além disso, utiliza um conjunto de metadados da imagem para serem utilizados na busca textual.
- **Busca de imagem**: é gerado um vetor de características da imagem de consulta, o qual é comparado com os demais vetores presentes no índice. Deste processo, obtém-se uma medida que determina a similaridade entre as características da imagem de consulta e as contidas no índice. Por fim é retornado um conjunto de imagens similares.
- **Backup do índice**: responsável por realizar o backup dos índices de texto e imagem periodicamente. Em caso de corrupção do índice, é o backup mais atual é restaurado automaticamente. Algumas configurações como quantidade de backups podem ser alteradas em um arquivo de configuração.
- **Estatísticas**: Informa algumas estatísticas sobre o índice de texto e imagem, como a quantidade de documentos, o espaço ocupado pelo índice em disco e a quantidade de buscas realizadas.

3.6 Servidor XML-RPC

A Opala foi desenvolvida para prover um serviço para a Biblioteca Digital EPTC. No projeto são utilizadas outras tecnologias e plataformas além de Java. Portanto, foi desenvolvido um servidor de chamada de procedimento remoto baseado no protocolo XML-RPC que disponibiliza as funcionalidades do Opala para qualquer linguagem que possua uma implementação do protocolo XML-RPC. O protocolo XML-RPC foi escolhido por integrar diferentes tecnologias em um produto de software funcionalmente viável, permitindo que softwares executáveis em sistemas operacionais diferentes, rodando em ambientes diferentes, façam chamadas de processos na internet.

[5] *DublinCore* é um padrão de esquema de metadados. Mais informações em http://dublincore.org/

4 CONSIDERAÇÕES FINAIS

Diante da quantidade de informações que se tem acesso atualmente, torna-se uma necessidade fundamental para os sistemas, ter um mecanismo de recuperação de informação que retorne buscas satisfatórias ao usuário. Observando este fato, foi desenvolvida a Opala, uma ferramenta que tem por finalidade a recuperação de textos e imagens por conteúdo. A facilidade de uso é uma das principais características da Opala. Ela possibilita a implementação de sistemas de recuperação de informação baseada em conteúdo e metadados por qualquer aplicação sem a necessidade de conhecimento teórico aprofundado e entendimento de outras as ferramentas de buscas mais complexas, tornando o desenvolvimento mais simples e rápido.

5 AGRADECIMENTOS

Agradecemos ao MEC/SETEC pelo apoio financeiro para o desenvolvimento deste trabalho e a Antônio de Pádua F. F. Sousa, José Tavares de A. Filho, Dannylvan C. Guimarães e Mônica Regina da Silva, ex-bolsistas do Laboratório de Pesquisa em Sistemas de Informação do Instituto Federal de Educação, Ciência e Tecnologia do Piauí pela colaboração no desenvolvimento da ferramenta.

REFERÊNCIAS

COSTA, Valéria Oliveria. **Estudo Analítico Descritivo das Máquinas de Busca da Web Brasileira**. Dissertação. Universidade Federal de Minas Gerais - UFMG. Belo Horizonte, p. 209. 2001.

CUNHA, Murilo Bastos da. **Biblioteca digital: bibliografia internacional anotada**. *Ci. Inf.* [online]. 1997, vol.26, n.2 ISSN 0100-1965. doi: 10.1590/S0100-19651997000200013.

HALAWANI, A. et al. **Fundamentals and applications of image retrieval: An overview**. Datenbank-Spektrum, v. 18, p. 14-23, 2006.

HATCHER, E. **Lucene in Action**. [S.I.]: Maning Publications Co., 2004.

LONG, Fuhui; ZHANG, Hong-Jiang; FENG, David Dagan. Fundamentals of content-based image retrieval. In: FENG, David Dagan; SIU, Wan-Chi; ZHANG, Hong-Jiang. **Multimedia Information Retrieval and Management - Technological Fundamentals and Applications.** [S.I.]: Springer-Verlag, 2003. p. 1-26.

LUX, M.; CHATZICHRISTOfiS, S. A. Lire: Lucene image retrieval: an extensible java CBIR library. **MM'08: Proceeding of the 16th ACM internacional conference on Multimedia**, New York, 2008. 1085-1088.

SOUSA, A. D. P. F. F. **Análise da influência dos descritores visuais MPEG-7 no processamento de consultas por similaridade em CBIR aplicado à Biblioteca Digital**. Monografia. Instituto Federal de Educação,Ciência e Tecnologia do Piauí – IFPI. Teresina. 2009.

ZHANG, Q. et al. Content-based image retrieval using multiple-instance learning. **ICLM '02: Proceedings of the Nineteenth**, 2002.