Uma arquitetura hibrida para a indexação de documentos do Diário Oficial do Município de Cachoeiro de Itapemirim

A hybrid architecture for document indexing of the Municipal Official Gazette in Cachoeiro de Itapemirim

Bruno Missi XAVIER[1]
Alcione Dias da SILVA[1]
Geórgia Regina Rodrigues GOMES[2]

Resumo

Técnicas de Mineração de Textos vêm sendo amplamente utilizadas para processamento de grandes volumes de documentos. Contudo, ainda há uma grande defasagem na tentativa de definir uma arquitetura para sistemas transacionais com elementos de inteligência computacional. Este trabalho tem o objetivo de apresentar uma proposta de arquitetura para a construção de um sistema computacional que utiliza técnicas de Mineração de Textos para indexar conteúdos da base do Diário Oficial do município de Itapemirim, no estado do Espírito Santo, transformando a informação antes disponível em linguagem natural para um formato estruturado, passível de ser persistido. Para validar a arquitetura, foi desenvolvido um protótipo em linguagem Java acessível no ambiente *Web*. Para avaliação da ferramenta, o estudo de caso proposto contou com uma base composta por 22 documentos, contendo 198 atos normativos da base daquele Diário Oficial, para os quais foram identificados bons níveis de precisão e abrangência na recuperação da informação. Este trabalho contribui com a apresentação de uma arquitetura híbrida, composta por elementos do modelo de sistemas transacionais e elementos da Mineração de Textos, além da utilização de padrões de projetos de *software*.

Palavras-chave: Diário Oficial de Cachoeiro de Itapemirim. Indexação de documentos. Mineração de textos. Recuperação da informação.

Abstract

Text mining techniques have been widely used to process large volumes of documents. However, there is still a large gap when defining the architecture for systems with transactional elements of computational intelligence. The aim of the paper is to outline a proposed architecture to build a computational system that uses text mining techniques to index content from the database of the Official Gazette in the city of Cachoeiro de Itapemirim *in the state of* Espírito Santo*, transforming the information previously available in natural language into a structured format that can be persisted. To validate the architecture we developed a prototype in Java accessible in the Web environment to evaluate the tool. To evaluate the tool, a case study featured a database composed of 22 documents, containing 198 normative acts from the database of the Official Gazette, in which good levels of accuracy and coverage of information retrieval were identified. This study contributes to the presentation of a hybrid architecture consisting of components of the model of transactional systems and elements of text mining, in addition to the use of software design patterns.*

Keywords: *Official Gazette* Cachoeiro de Itapemirim. *Indexing documents. Text mining. Information retrieval.*

[1] Companhia de Tecnologia da Informação de Cachoeiro de Itapemirim. R. 25 de Março, 28, Centro, 29300-970, Cachoeiro de Itapemirim, ES, Brasil. Correspondência para/*Correspondence to*: B.M. XAVIER. *E-mail*: <bmissix@gmail.com>.

[2] Universidade Cândido Mendes, Faculdade de Engenharia de Produção, Departamento de Pesquisa Operacional e Inteligência Computacional. Campos dos Goytacazes, RJ, Brasil.

Recebido em 26/11/2013, reapresentado em 12/3/2014 e aceito para publicação em 20/3/2014.

Introdução

Técnicas de Mineração de Textos (MT) vêm sendo aplicadas em grandes volumes de documentos disponíveis em formato textual, com o objetivo de recuperar informação previamente oculta em arquivos não estruturados. Essas técnicas auxiliam na estruturação de documentos antes escritos em linguagem natural, através de etapas de pré-processamento e indexação de termos-chave encontrados no texto. Dessa forma, torna-se possível a persistência das informações através de bases de dados, bem como a recuperação das mesmas, utilizando as estruturas de índices previamente construídas.

Apesar de comumente utilizadas, as técnicas de Mineração de Textos vêm sendo pouco aplicadas a sistemas transacionais, a fim de conferir a eles maior inteligência computacional. Recentemente, alguns avanços têm sido feitos na tentativa se desenvolver uma arquitetura para MT em bases de documentos textuais proprietárias (Passini, 2012; Villalon & Calvo, 2013; Webber *et al*., 2013) e, ainda, com a utilização de bases abertas, como o *Microsoft Academic Search* e o *Google Scholar* (Kaur *et al*., 2012; Sun *et al*., 2012).

Os governos municipais são, em maioria, carentes de recursos tecnológicos de infraestrutura, e ainda mais carentes de recursos de inteligência computacional. É comum, na administração pública, a ausência de recursos de sistemas informacionais que possam organizar e gerir a informação gerada pela administração municipal. Em alguns outros municípios que podem ser vistos como privilegiados, apesar de existirem sistemas transacionais básicos, muito pouco se pode encontrar de inteligência computacional relacionada a esses sistemas. Isso é demonstrado claramente por *Wireless Mundi* (Ranking..., 2012). Em contraposição à defasagem tecnológica dos governos municipais, a *Internet*, em número de conteúdo disponibilizado através de *websites*, tem crescido em média 141% a cada ano, totalizando, em 2013, 634 milhões de *websites* e 2,4 bilhões de usuários da rede mundial de computadores (Internet..., 2012).

A indexação de grandes bases de documentos, ou indexação documentária, é descrita como o conjunto de atividades e processos responsáveis por identificar os traços descritivos de um documento através de seus termos representativos, resumidos aos conjuntos de unidades léxicas (Fujita, 1999; Pinto, 2001).

A indexação manual é uma tarefa árdua e sujeita à capacidade do analista indexador em reconhecer o contexto tratado em cada documento, e a ele associar termos-chave capazes de identificar adequadamente seu conteúdo. Nesse processo, o modo de leitura, apesar de pouco explorado e avaliado pela literatura científica, é a fase principal da operação, sendo considerado fator determinante para seu sucesso (Silva & Fujita, 2004). Ainda podem ocorrer divergências entre indexadores que atribuem diferentes termos-chave a um mesmo documento, ou ainda um mesmo indexador atribuindo diferentes termos-chave a um documento em momentos diferentes (Maron, 1977; Guedes, 1994; Fujita, 1999).

A indexação automática de termos é um tema bastante discutido nas áreas da Ciência da Informação e da Biblioteconomia, que apresentam uma convergência para o uso dessa técnica, em razão de algumas vantagens quando comparada ao processo manual. Dentre tais vantagens, destaca-se a redução da subjetividade do processo de indexação e a agilidade com que uma coleção de documentos é indexada (Swanson, 1960; Maron, 1961; Edmundson 1969; Salton, 1971; Vieira, 1988; Guedes, 1994; Araújo Júnior & Tarapanoff, 2006).

No cenário da administração pública municipal, o setor do Diário Oficial Municipal tem a atribuição de indexar e publicar todo o conteúdo gerado pelo Poder Executivo através dos Atos Normativos contidos em uma edição do Diário Oficial. Trata-se de uma tarefa nada trivial que consome tempo e recursos do setor, e quase sempre é auxiliada por ferramentas computacionais não específicas, o que compromete ainda mais a eficiência do processo.

As ferramentas de indexação automática de documentos representam uma esperança para solucionar um processo lento e, em alguns casos, deficiente, além de altamente dependente da presença do usuário indexador. Para Anderson e Perez-Carballo (2001), a indexação automática apresenta resultados bastante adequados em um ambiente de grandes coleções de documentos, além de uma significativa redução no tempo de indexação e na subjetividade do processo.

http://dx.doi.org/10.1590/0103-37862015000100008

Este trabalho propõe uma arquitetura para indexação de bases de documentos textuais, utilizando técnicas de MT para identificação de índices que representem adequadamente o contexto do documento. Além disso, a arquitetura descrita apresenta componentes dos sistemas transacionais para validação e armazenamento da informação, bem como elementos da MT para transformação da informação não estruturada em um conjunto de objetos estruturados, capazes de serem persistidos em bases de dados. A relevância deste estudo está associada à exposição de uma arquitetura computacional, para indexação de bases documentárias, relacionada com técnicas da MT, e à aplicação de inteligência computacional em sistemas transacionais com foco na gestão e indexação de documentos públicos.

Mineração de textos

Knowledge Discovery in Textutal Databases (KDT) ou Descoberta de Conhecimento Textual, ou ainda Mineração de Textos, surge no contexto do processamento de grandes volumes de documentos, com o objetivo de extrair informações significativas para a formação do conhecimento. Frawley *et al*. (1991) definem MT como o processo de extração não trivial de informação implícita, previamente desconhecida e potencialmente útil. Aranha e Passos (2006) citam que MT é a área de pesquisa dos sistemas de informação responsável por organizar, navegar e descobrir informações relevantes em documentos textuais. O processo de Mineração de Textos tem sua origem na Mineração de Dados, ou *Knowledge Discovery in Databases* (KDD), que busca extrair padrões relevantes através da aplicação de algoritmos em bases de dados estruturadas (Fayyad *et al*., 1996).

Apesar da complexidade do processo de Mineração de Textos, Monteiro *et al*. (2006) afirmam que ele pode ser dividido minimamente em três etapas: Pré-processamento; Análise e Extração do Conhecimento; e Pós-processamento. Aranha (2007) apresenta um modelo em que é acrescentada uma etapa de coleta de dados, com o objetivo de selecionar as informações que vão compor a base textual do trabalho.

Esse modelo genérico da Mineração de Textos não é novo. Lopes (2004) já havia apresentado um modelo contendo as mesmas etapas da adaptação dos modelos de Monteiro *et al*. (2006) e Aranha (2007). A primeira etapa do modelo proposto por Lopes (2004) é a preparação dos dados, na qual é selecionada a base textual de interesse do trabalho. Segue-se a etapa de análise dos dados, a fim de estruturar a base composta na primeira etapa. A seguir, a etapa de processamento dos dados executa as tarefas de acordo com os objetivos da mineração. Por fim, a etapa de pós-processamento avalia as descobertas efetuadas pela etapa anterior. Esta etapa também é responsável pela visualização das informações obtidas.

Para Feldman e Sanger (2007), a Mineração de Textos tem seu foco na coleção de documentos selecionada. Pode ser considerada uma coleção de dados válida, ou um agrupamento de textos de diversos documentos. Essas coleções, frequentemente chamadas de *corpus* ou *corpora*, são classificadas como estáticas ou como dinâmicas, conforme permaneçam inalteradas ou sofram alterações durante o processo.

Selecionar uma coleção de dados que represente adequadamente o domínio do problema abordado nem sempre é uma tarefa trivial. Para Ramos e Bräscher (2009), a seleção representa a tarefa mais trabalhosa em todo o processo, cuja principal dificuldade está em identificar a origem de armazenamento dos documentos a serem utilizados. Outra grande dificuldade está na dimensão que a coleção de documentos pode atingir. Utilizar um volume muito grande de documentos pode ter impacto no tempo de processamento, enquanto uma coleção demasiadamente pobre pode interferir negativamente na descoberta do conhecimento.

Pré-processamento: Aplicações de Mineração de Textos eficientes são baseadas em técnicas sofisticadas de pré-processamento. A MT é um processo tão dependente dos pré-processamentos utilizados em cada etapa, que se pode dizer que ela é a própria composição desses processos. Existe uma grande variedade de técnicas disponíveis para essa etapa, que são aplicadas e combinadas na tentativa de estruturar, total ou parcialmente, a coleção de documentos (Feldman & Sanger, 2007).

Tokenização: Tokenização ou Atomização é normalmente o primeiro processo da etapa de pré-processamento. Nesse processo, o fluxo contínuo de caracteres é dividido em unidades mínimas, conhecidas como *tokens* (Rehman *et al*., 2013). Um *token* é gerado descar-

tando-se os caracteres de "espaço" e pode corresponder a uma palavra, símbolo ou pontuação. Para Feldman e Sanger (2007), esse processo pode dividir os documentos em capítulos, sessões, parágrafos, frases, palavras ou até sílabas e fonemas, de acordo com a necessidade.

Stopwords: O número de palavras diferentes contidas em um documento é relativamente grande. Palavras que não possuem valor semântico para o texto são conhecidas como *stopwords* e podem ser de diferentes classes gramaticais, incluindo-se preposições, artigos, conjunções, algumas vezes verbos, nomes, adjetivos e advérbios.

Listas de *stopwords,* conhecidas como *stoplists,* normalmente são pré-definidas e utilizadas com a finalidade de reduzir o número de termos analisados nas etapas posteriores. Essas listas podem ser construídas a partir da experiência de um especialista no domínio do assunto, ou ainda criadas de forma automática, através da frequência de aparições dos termos no conjunto de documentos. Para isso, Carrilho Junior (2007) cita duas possibilidades: (i) termos com aparições acima de um percentual pré-definido podem ser eliminados e (ii) termos com aparições muito raras na coleção de documentos são supostamente irrelevantes para o *corpus* analisado.

Normalização: A normalização de palavras é uma etapa de grande importância para a Mineração de Textos. Nessa etapa, palavras de um mesmo significado morfológico são reduzidas a um único termo, chamado de forma canônica. A principal função desse processo é permitir o mapeamento de palavras que sejam semântica e morfologicamente relacionadas, assim permitindo agrupar, num termo único, uma maior quantidade de palavras que tenham o mesmo sentido.

Para a normalização de termos podem ser aplicadas as seguintes técnicas:

a) Radicalização: A radicalização do termo permite a utilização de vocábulos primitivos anteriores às variações, como plurais e inflexões verbais (Porter, 2006).

b) Lematização: A lematização é o processo de representação das palavras através de seu morfema. Termos extraídos a partir desse processo são comumente chamados de lema ou forma canônica. Dias e Malheiros (2005) ressal-tam que a redução à forma canônica, através do processo de lematização, não causa perdas morfológicas da categoria original da palavra, ao contrário do processo de radicalização, que pode incorrer em erros de *overstemming* e *understemming*.

c) Dicionário de vocábulos (tesauros): Baeza-Yates e Bertier (1999) definem tesauros como um conjunto de termos importantes para o domínio da aplicação, sendo associada a cada termo uma lista de palavras relacionadas. A aplicação de dicionários de vocábulos envolve conceitos de normalização, visto que busca a representação comum da palavra.

Análise e extração do conhecimento: A etapa de Análise e Extração do Conhecimento é responsável por extrair informação relevante da coleção de dados selecionada e estruturada (ou semiestruturada), sendo passível de ser armazenada em Bancos de Dados.

Os processos executados nessa etapa variam de acordo com o objetivo da mineração. Lopes (2004) aborda que os processos disponíveis nessa etapa são indexação, extração da informação, extração de características, sumarização, categorização, classificação e c*lustering* de documentos.

Alguns autores definem a tarefa de indexação como parte do Pré-processamento, pois seus resultados servem de apoio para outras tarefas (Aranha, 2007; Soares *et al*., 2008). Já Lopes (2004) e Marinho e Girardi (2003) definem essa tarefa como inerente à etapa de Análise e Extração do Conhecimento.

Pós-processamento: Feldman e Sanger (2007) descrevem o Pós-processamento como a etapa de visualização e análise das informações obtidas no processo da Mineração de Textos. Devido à natureza subjetiva da avaliação dos resultados, ferramentas gráficas sofisticadas, porém de simples representação dos dados, auxiliam o processo de análise e interpretação dos resultados gerados.

Tentando reduzir a subjetividade das análises e avaliações dessa etapa, são utilizadas algumas métricas de eficácia para avaliação do desempenho, provenientes da área de Recuperação da Informação (RI) e baseadas no conceito de relevância (Soares *et al*., 2008). As medidas de Precisão, Abrangência e Média-F são comumente utilizadas nesse contexto.

http://dx.doi.org/10.1590/0103-37862015000100008

Medidas de avaliação: *precisão* - A medida de Precisão (do inglês *Precision*) corresponde ao percentual de documentos relevantes recuperados em relação ao conjunto de todos os documentos recuperados (Corrêa, 2003; Silva, 2007; Barion & Lago, 2008; Moraes & Ambrósio, 2008; Soares *et al*., 2008). Essa medida tem o objetivo de avaliar o quanto de lixo (documentos irrelevantes) a consulta retorna. O valor da precisão pode variar entre 0 e 1, sendo 0, nenhum documento relevante recuperado, enquanto 1 equivale a todos os documentos relevantes recuperados. A Equação 1 representa o cálculo da medida de Precisão.

$$Precisão = \frac{Nº\ de\ documentos\ relevantes\ recuperados}{Total\ de\ documentos\ recuperados} \quad (1)$$

Abrangência: A medida de Abrangência (do inglês *Recall*) é chamada por Aranha e Passos (2006) de Eficiência. Essa medida representa o percentual de documentos relevantes recuperados em relação ao conjunto de documentos relevantes (Silva, 2007). O objetivo é avaliar a qualidade do resultado do processo de RI. O valor da abrangência varia entre 0 e 1 sendo 0, nenhum documento relevante recuperado, e 1, todos os documentos relevantes recuperados. A medida de Abrangência pode ser expressa através da Equação 2.

$$Abrangência = \frac{Nº\ de\ documentos\ relevantes\ recuperados}{Total\ de\ documentos\ relevantes} \quad (2)$$

Média-F: A Média-F (do inglês *F-Means*) fornece uma maneira de combinar as medidas de Precisão e Abrangência em uma única métrica. Podem-se estabelecer alguns pesos diferentes para cada medida (Precisão e Abrangência), dando flexibilidade para a definição de critérios de importância (Chinchor, 1992; Sasaki, 2007). Chinchor (1992) define a Média-F pela Equação 3.

$$F = \frac{(\beta^2 + 1.0)\ x\ Precisão\ x\ Abrangência}{\beta^2\ x\ Precisão + Abrangência} \quad (3)$$

Nessa equação, β é o parâmetro que define o balanceamento entre as medidas Precisão e Abrangência. Se esse parâmetro for igual a 1 ($\beta = 1$), então as medidas serão equivalentes. Caso o parâmetro seja maior que 1 ($\beta > 1$), a medida de Abrangência terá maior influência no sistema. Por fim, se o parâmetro for menor que 1 ($\beta < 1$), o sistema adotará a medida de Precisão como majoritária.

Sasaki (2007) define a Medida-F em sua fórmula harmônica, na qual as medidas de Precisão e Abrangência têm a mesma influência no sistema. A Equação 4 apre-senta a Medida-F de acordo com Sasaki (2007).

$$F = \frac{2\ x\ Precisão\ x\ Abrangência}{Precisão + Abrangência} \quad (4)$$

Arquitetura proposta

A arquitetura proposta é composta por quatro macrocamadas. A primeira camada é responsável pelo armazenamento dos dados em bases tabulares, com o objetivo de persistir a estrutura de índices processados pela aplicação de Mineração de Textos e os textos de forma não estruturada. A segunda camada é a camada transacional, responsável por gravar e recuperar as informações processadas na base de dados. A terceira camada é a camada de Mineração de Textos, responsável por processar as informações textuais não estruturadas, criando a estrutura de índices e a estrutura de recuperação da informação a partir dos índices gerados. Por fim, a quarta camada é responsável pela interface com o usuário. A Figura 1 demonstra as camadas da arquitetura proposta.

Juntas, as camadas de Mineração de Textos e Transacional, apresentadas na Figura 1, formam a macrocamada de inteligência da aplicação. Essa macrocamada é responsável pelas regras de negócios da aplicação transacional, regras de acesso a dados e regras de indexação e recuperação da informação através de índices pré-estabelecidos.

Camada de persistência de dados: Na camada de persistência de dados, responsável pelo armazenamento das informações a serem utilizadas no processo de Mineração de Textos, o modelo conceitual apresenta as entidades Grupo, Documento, Atos Normativos e Termo. Os Grupos definem o espaço de trabalho da MT, sendo que um Grupo representa uma coleção de Documentos delimitada através de um domínio ou um subdomínio

definido pelo usuário. Os Documentos representam agrupamentos de Atos Normativos, sendo que um Documento pode ter um ou muitos Atos Normativos associados. Os Atos Normativos são o alvo de toda busca, pois são eles que contêm a informação que o usuário procura. Os Termos representam cada palavra considerada relevante para o domínio da aplicação, sendo identificados como seus atributos: a posição da palavra no texto, o radical, o lema e o sinônimo da palavra. A Figura 2 representa o modelo conceitual da camada de persistência e suas principais entidades.

Ainda são armazenados alguns índices na entidade Termo:

- O TF (*Term Frequency*): representa o número de ocorrências do termo no documento;

- O IDF (*Inverse Document Frequency*): representa a frequência inversa de um termo em um documento;

- O TF-IDF (*Term Frequency - Inverse Document Frequency*): pondera a aplicação das duas me-didas para o grupo do documento.

Camada de inteligência: A macrocamada de inteligência é composta por duas camadas: a camada Transacional e a camada de Mineração de Textos. A disposição dessas camadas, do ponto de vista do fluxo da informação, é fundamental para conferir ao modelo a inteligência necessária para as aplicações dos conceitos da MT. A Figura 3 demonstra o fluxo da informação sob a ótica da macrocamada de inteligência.

O fluxo da informação é originário da interface, seja um formulário de cadastro ou alteração de dados, seja uma solicitação de busca de informação no banco de dados. A requisição (do inglês *Request*) é feita na camada de visão, passando por cada componente da macrocamada de inteligência até atingir a camada de dados. Nesse momento, retorna uma resposta (do inglês

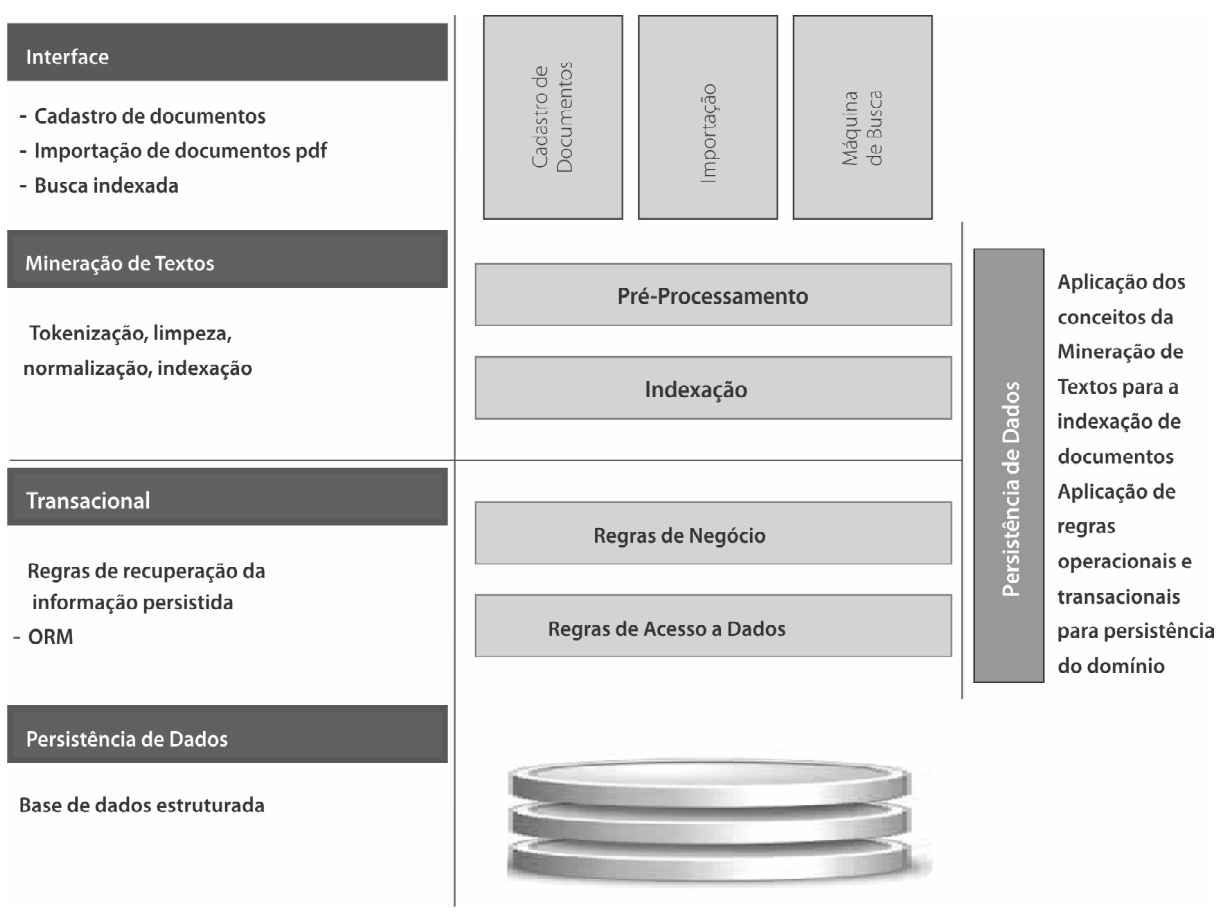

Figura 1. Arquitetura proposta.

Fonte: Elaborado pelos autores (2013).

Nota: ORM: *Object Relational Mapping*.

http://dx.doi.org/10.1590/0103-37862015000100008

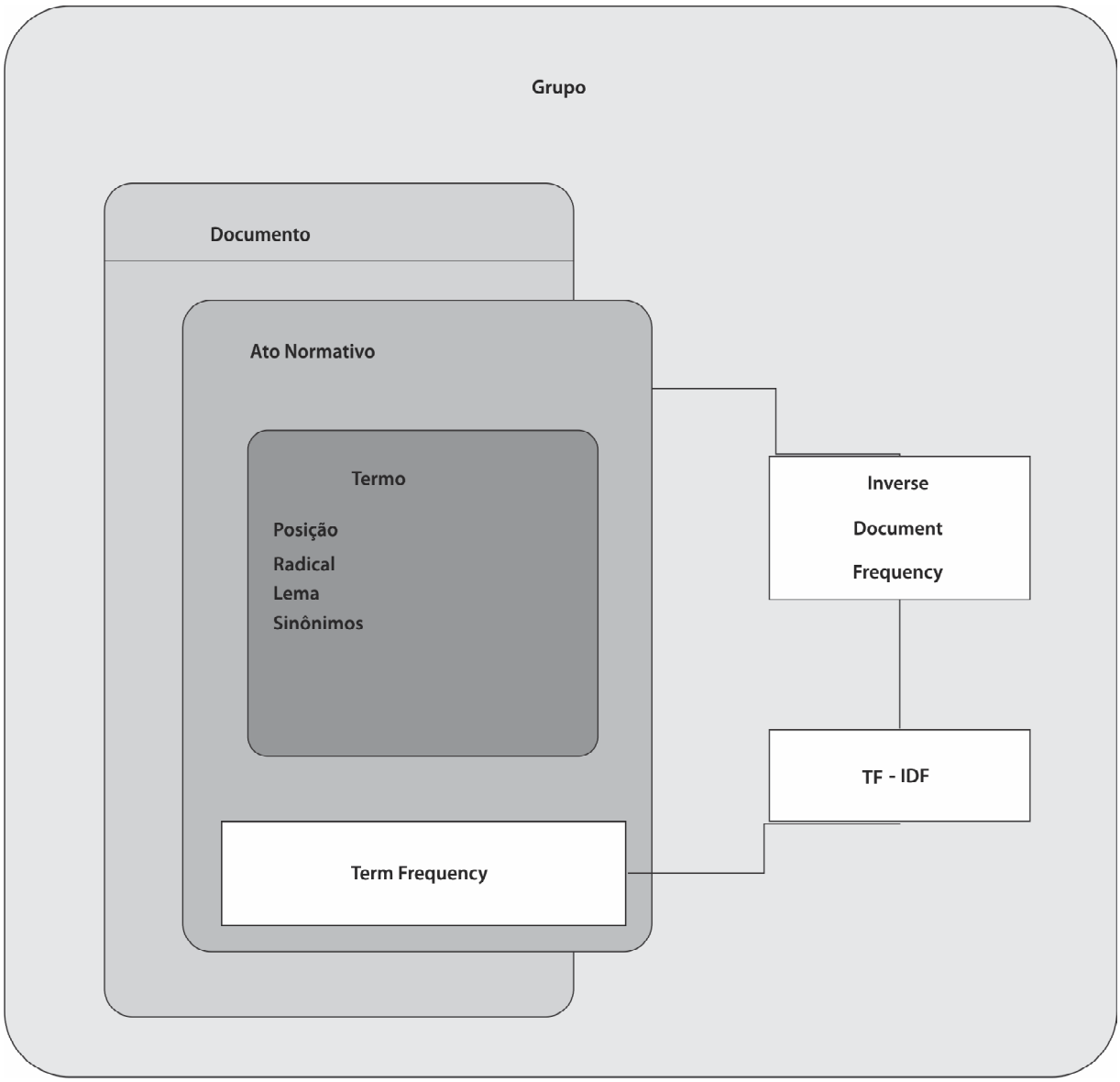

Figura 2. Modelo conceitual da camada de persistência.

Fonte: Elaborado pelos autores (2013).

Response) que faz o caminho inverso da requisição, até atingir seu alvo na camada de interface (Oracle, 2012).

Essa macrocamada é ainda considerada o núcleo computacional do modelo proposto, por se tratar de duas camadas que agregam as regras de acesso a dados, as regras de negócios do modelo transacional e os comportamentos referentes à Mineração de Textos.

Camada transacional: A camada transacional é responsável pelas regras de persistência da informação e pela aplicação das regras de negócios da aplicação. A divisão lógica e física dos componentes dessa camada traz vantagens associadas principalmente aos conceitos da Orientação a Objeto e dos padrões de projeto, sobretudo os padrões do Java2 *Enterprise Edition* (J2EE) definidos por Alur *et al*. (2003).

Camada de mineração de textos: os componentes da camada de Mineração de Textos são desenvolvidos com base em técnicas e algoritmos de extração da infor-

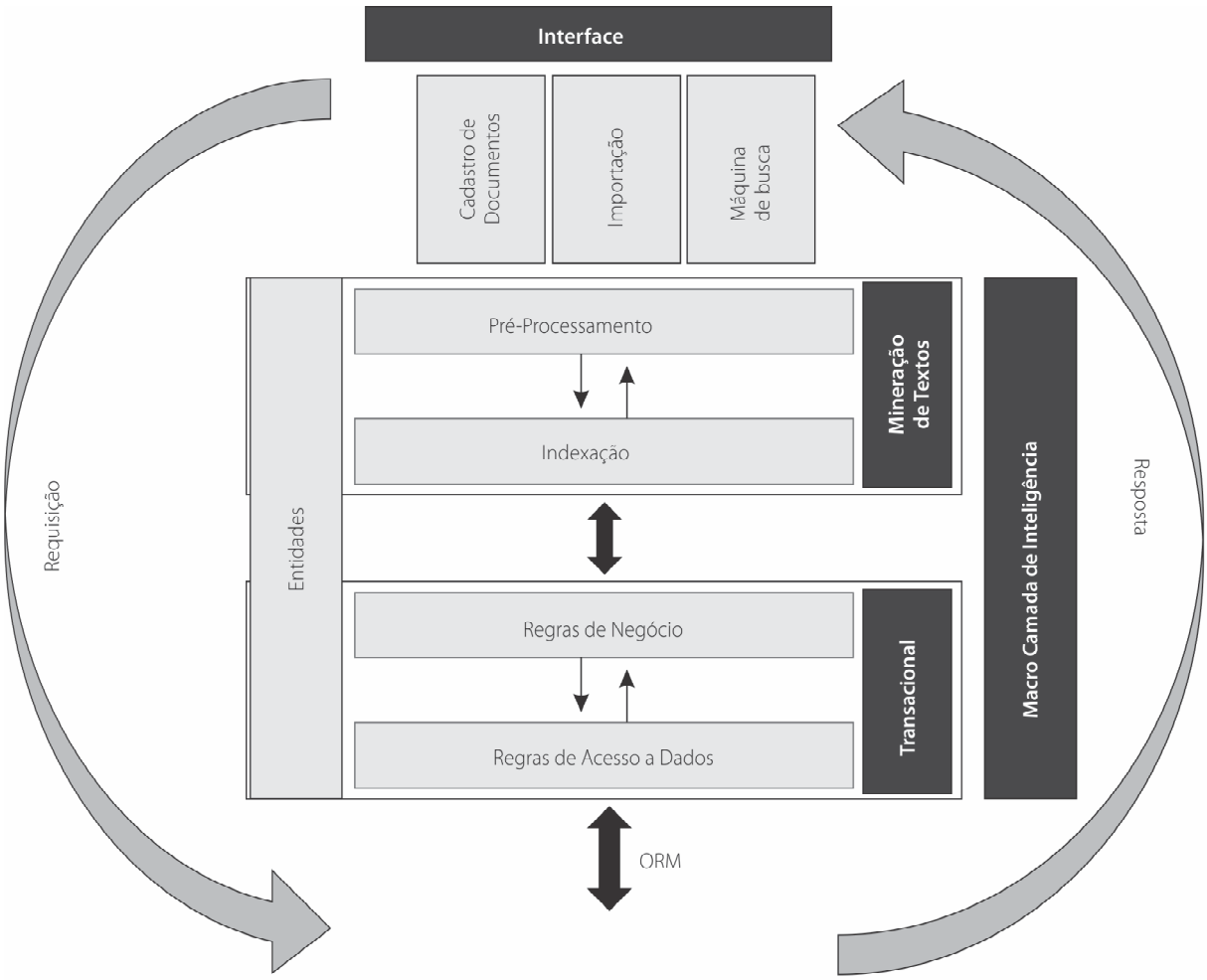

Figura 3. Fluxo da informação através do modelo.

Fonte: Elaborado pelos autores (2013).

Nota: ORM: *Object Relational Mapping*.

mação e indexação de documentos. O objetivo dessa camada é analisar o conjunto de textos fornecido pelo usuário, gerando uma estrutura de indexação de dados a ser armazenada para futuras consultas à base. A Figura 4 apresenta os componentes da camada de Mineração de Textos.

A Figura 4 define o fluxo do componente de Mineração de Textos, que recebe um documento em formato textual e, após a aplicação de cada tarefa das etapas de Pré-processamento e Processamento, é obtida, como resultado, uma coleção de objetos estruturados passiveis de serem persistidos em um repositório de dados. O componente de indexação de textos é composto por duas tarefas. A primeira refere-se aos pesos atribuídos a cada termo do documento. Esses pesos se utilizam das medidas TF-IDF. A segunda tarefa é referente aos termos-índice que serão utilizados como elementos de indexação. Além do termo, pode-se utilizar o radical, o lema, ou ainda o sinônimo associado ao termo em questão.

Camada de interface: A camada de interface é subdividida em três componentes básicos: cadastro de documentos, importação de arquivos pdf e máquina de busca. Cada componente fornece ao usuário um ambiente de interação em linguagem natural para que, na ma-

http://dx.doi.org/10.1590/0103-37862015000100008

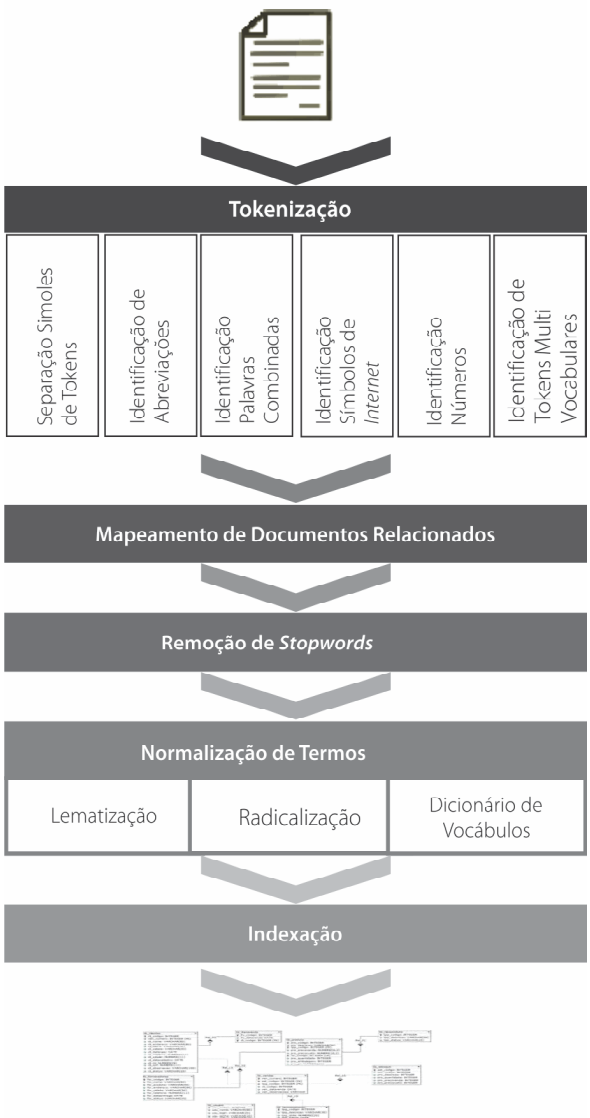

Figura 4. Componentes da camada de Mineração.

Fonte: Elaborado pelos autores (2013).

crocamada de inteligência, essas informações textuais sejam tratadas e transformadas em índices para armazenamento ou recuperação de dados.

Cadastro de documentos: O componente de cadastro de documentos da camada de interface permite que o usuário realize funções básicas de manutenção de dados. Nesse ponto o usuário autenticado terá permissão para cadastrar atos normativos em linguagem natural, associar os atos normativos a documentos e cadastrar sinônimos para os termos.

Depois de cadastrado um novo ato normativo, a macrocamada de inteligência inicia a execução dos componentes da camada de Mineração de Textos com as tarefas definidas.

Importação de documentos Portable Document Format (PDF): O componente de Importação de Documentos PDF permite que documentos já existentes possam ser importados e relacionados a grupos de forma fácil. Esse componente da camada de Interface é responsável por transformar os documentos existentes em formato PDF em entradas válidas para a macrocamada de inteligência.

Após a importação de um documento em formato PDF, a interface converte seu conteúdo para um formato textual ainda em linguagem natural, entregando a responsabilidade pela informação aos componentes da camada de Mineração de Textos que, após tratar a informação, transfere a responsabilidade para a camada de negócios.

O processo de importação de documentos termina quando a camada de negócios entrega a informação, antes textual e agora estruturada, para ser persistida em um repositório adequado.

Máquina de busca: O componente Máquina de Busca da camada de interface tem o objetivo de recuperar a informação de forma otimizada, através dos mecanismos de indexação propostos pelo modelo.

Inicialmente, deve-se selecionar o grupo de documentos alvo da pesquisa. Essa é a definição do espaço de busca da consulta. Após, devem-se informar as chaves de busca que podem concatenar expressões booleanas e/ou frases.

A pesquisa pode ser feita unindo quatro elementos básicos:

a) *Palavra*: a palavra deve ser consultada utilizando radical, lema e sinônimos associados à palavra;

b) *Frase*: consulta-se cada palavra utilizando o radical, lema e sinônimos associados, e levando em consideração a proximidade entre as palavras;

c) *Expressões booleanas*: podem ser adicionados dois tipos de expressões booleanas: "*and*", expressão pela qual a máquina de busca pesquisará por uma palavra;

"*or*", expressão que indicará pesquisas por uma palavra ou outra; "*not*", excluindo os documentos que contenham a palavra subsequente à expressão. Ex: "universidade *and* mestrado *not* doutorado";

d) *Frequência da palavra*: pode-se definir uma frequência mínima para cada palavra da consulta.

A combinação dos critérios de busca, definidos através da concatenação dos elementos básicos da pesquisa, gera uma lista dos Atos Normativos associados à consulta.

Métodos

O desenvolvimento da metodologia foi subdividido em 4 etapas que representam os procedimentos executados de acordo com a ocorrência dos processos:

Para a validação da arquitetura proposta e a avaliação da ferramenta por parte do setor do Diário Oficial Municipal, foi desenvolvido um protótipo em linguagem Java 1.7.x, disponível na *web* e acessado através de um navegador (*browser*).

Xavier *et al*. (2013) compararam três diferentes radicalizadores para a língua portuguesa, com o objetivo de definir qual melhor executa a tarefa de radicalização para uma instância de palavras, recuperadas a partir da análise dos Diários Oficiais. Esse trabalho demonstrou que, comparando os resultados gerados pelos algoritmos, Porter e Removedor Sufixal da Língua Portuguesa (RSLP) apresentaram alta qualidade para a tarefa, enquanto Savoy foi considerado o algoritmo mais leve e de menor qualidade. O trabalho ainda apontou Porter como o algoritmo mais apropriado para a instância de dados utilizada. Dessa forma, o algoritmo de Porter foi implementado e incorporado à camada de MT para a etapa de radicalização dos termos.

Da mesma forma, o lematizador descrito por Nunes (2007) foi adicionado à camada de MT. O autor descreve um lematizador capaz de normalizar palavras para a língua portuguesa. O algoritmo é eficiente em lematizar verbos desconhecidos da mesma forma que verbos conhecidos, utilizando regras da língua portuguesa e um dicionário de dados contendo grande parte das formas verbais flexionadas.

Confecção da instância de dados: A instância de testes para validação da ferramenta foi construída a partir dos apontamentos do usuário indexador. Trata-se de documentos oficiais do Diário Oficial do Município de Cachoeiro de Itapemirim (ES), que estão disponíveis para consulta pública <http://www.cachoeiro.es.gov.br/>. Este usuário selecionou para importação 22 documentos correspondentes ao mês de janeiro de 2012, compreendendo um total de 198 Atos Normativos do Poder Executivo.

Identificação dos termos de busca e Atos Normativos relacionados: Dentre o universo gerado a partir da instância de teste, o usuário indexador identificou dois conceitos de grande importância que constantemente requerem, por parte do setor, respostas imediatas à administração pública. São eles: "Convocações em geral" e "Exoneração de servidores". Para cada conceito o usuário indexador selecionou, a partir do método de indexação manual, um grupo de Atos Normativos relacionados. Em relação ao conceito "Convocações em geral" foram identificados nove Atos Normativos dispostos em oito documentos. Para o conceito "Exoneração de servidores", o método de indexação manual identificou dez Atos Normativos dispostos em sete documentos.

Resultados

Para a avaliação do método de indexação automática proposto, os resultados do processamento da ferramenta foram confrontados com os resultados obtidos pelo método manual e submetidos às medidas de Precisão, Abrangência e Média-F. Para o usuário indexador, é de extrema importância que a ferramenta proposta recupere todos os documentos envolvidos no conceito que se busca, trazendo o mínimo possível de documentos não pertinentes ao contexto.

Após o desenvolvimento da aplicação *Web*, o estudo de caso foi inteiramente orientado pelo usuário indexador a fim de validar a arquitetura proposta e verificar o atendimento da ferramenta construída. O experimento resultante deste estudo teve o objetivo de confrontar os resultados obtidos através do método de indexação manual e o método de indexação automática através de ferramentas de Mineração de Textos.

http://dx.doi.org/10.1590/0103-37862015000100008

Após a importação da instância de testes selecionada para o estudo, os documentos formam submetidos à indexação automática da ferramenta.

Para o primeiro conceito submetido à ferramenta, a chave de busca foi "Convocação de candidatos". Após o processamento da máquina de busca, a chave retornou dez Atos Normativos, nove deles relacionados ao conceito pesquisado, e um não relacionado.

Para o segundo conceito pesquisado - "Exoneração de servidores"-, foi adicionado à chave de busca o operador lógico "*and*" unindo os termos "exoneração" e "servidores". Assim, a máquina de busca passou a identificar apenas Atos Normativos que contivessem os dois termos ou suas variações. O processamento dessa chave de busca resultou em dez Atos Normativos, todos relevantes para o contexto da pesquisa.

A Tabela 1 apresenta, para cada conceito, o número de Atos Normativos relevantes, o número de Atos Normativos recuperados, e as medidas de Precisão, Abrangência e Média-F.

Para o assunto "Contratações em geral", nove Atos Normativos foram previamente identificados como relevantes para o assunto. Através da ferramenta de indexação automática, dez Atos Normativos foram recuperados, sendo nove relevantes ao contexto e um irrelevante. Através da Média-F, pôde-se estabelecer uma medida de qualidade da ferramenta analisada, calculando as medidas Precisão e Abrangência. Dessa forma, a medida de Precisão alcançou 1 ponto, enquanto a medida de Abrangência alcançou 0,9 ponto, e a Média-F, 0,95 ponto.

Tabela 1. Avaliação da recuperação da informação.

Medida	Exoneração de servidores (n)	Contratações em geral (n)
Documentos relevantes	10	9
Documentos recuperados	10	10
Precisão	1	0,9
Abrangência	1	1
Média-F	1	0,95

Fonte: Elaborado pelos autores (2013).

Relacionados ao assunto "Exoneração de servidores", foram previamente identificados dez Atos Normativos relevantes através do método de indexação manual. A máquina de busca encontrou exatamente os dez registros associados ao conceito pesquisado. Para esse contexto, a Média-F alcançou1 ponto, sendo recuperados todos os documentos relevantes (Precisão = 1) e sendo considerados relevantes todos os documentos recuperados (Abrangência = 1);

Conclusão

Este trabalho avaliou as técnicas de Mineração de Textos, a fim de indexar documentos municipais publicados no Diário Oficial do Município de Cachoeiro de Itapemirim-ES. Através das etapas de pré-processamento e indexação, propostas pelo modelo, pôde-se recuperar informação, antes apresentada em formato de linguagem natural e posteriormente transformada em uma coleção de objetos passível de ser persistida. A etapa de avaliação ou pós-processamento demonstrou a eficiência do processo anterior e a qualidade na recuperação de documentos, através da estrutura de índices previamente criados.

A proposta de definição de uma arquitetura híbrida, mesclando características do sistema transacional com o modelo de Mineração de Textos, aproxima o ambiente corporativo, voltado para a construção de sistemas de entrada e armazenamento de dados, à proposta de construção de aplicações inteligentes. A arquitetura apresentada foi composta por quatro camadas: a persistência, responsável pelo armazenamento da informação; a camada transacional, responsável pela validação das regras de negócios e regras de acesso a dados; a camada de Mineração de Textos, onde acontecem as transformações de textos em linguagem natural para a informação estruturada; e a camada de visão, interagindo com o usuário da aplicação, seja com o objetivo de gerar conteúdo para o domínio ou recuperar a informação através dos índices criados.

A indexação automática documentária contribuiu para o aprimoramento de dois processos do setor do Diário Oficial Municipal. Por um lado, a indexação de documentos, inicialmente realizada de forma manual através da leitura e análise dos textos produzidos pelo

município, torna-se um processamento informacional, no qual os termos-chave identificados são persistidos em estruturas de índices. Por outro lado, o processo de recuperação da informação, que anteriormente utilizava os termos-chave identificados no processo manual na localização de documentos, passa a utilizar a estrutura gerada para identificar os conceitos de busca através dos termos-chaves, sinônimos, lemas e radicais, aprimorando a busca no contexto pesquisado.

Os resultados obtidos excedem a expectativa do usuário especialista no assunto, indexando conteúdos de forma eficiente e sem a necessidade da supervisão humana. Os tempos computacionais do processo não foram avaliados devido ao alto grau de complexidade do método manual. Contudo, o método de indexação automática gera uma base de índices em poucos minutos, enquanto o processo manual consome horas de trabalho de um analista.

Referências

Alur, D.; Crupi, J.; Malks, D. *Core J2EE patterns, best practices and design strategies*. 2nd. Palo Alto: Sun Microsystems Press, 2003.

Anderson, J.D.; Perez-Carballo, J. The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing and Management*, v.37, n.2 p.231-254, 2001. doi: 10.1016/S0306-4573(00)00026-1

Aranha, C.N. *Uma abordagem de pré-processamento automático para mineração de textos em português*: sob o enfoque da inteligência computacional. 2007. Tese (Doutorado em Inteligência Computacional) - Faculdade de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Aranha, C.; Passos, E. A tecnologia de mineração de textos. *Revista Elerônica de Sistemas de Informação*, v.5, n.2, p.1-8, 2006. Disponível em: <http://revistas.facecla.com.br/index.php/reinfo/article/view/171>. Acesso em: 5 out. 2013. doi: 10.5329/RESI

Araújo Júnior, R.H.; Tarapanoff, K. *Precisão no processo de busca e recuperação da informação*: uso da mineração de textos. *Ciência da Informação*, v.35, n.3, p.263-247, 2006.

Baeza-Yates, R.; Bertier, R.N. *Modern information retrieval*. Harlow: Addison-Wesley, 1999.

Barion, E.C.N.; Lago, D. Mineração de textos. *Revista de Ciência Exatas e Tecnologias*, v.3, n.3, p.123-140, 2008.

Carrilho Junior, J.R. *Desenvolvimento de uma metodologia para mineração de textos*. 2007. Dissertação (Mestrado em Engenharia Elétrica) - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Chinchor, N. Muc-4 evaluation metrics. In: Conference on Message Understanding, 4th, 1992, San Francisco. *Proceedings*… San Francisco: Morgan Kaufmann Publishers, 1992. p.22-29.

Corrêa, H.L. *Teoria geral da administração*: abordagem histórica da gestão de produção e operações. São Paulo: Atlas, 2003.

Dias, M.A.L.; Malheiros, M.G. *Estudo de técnicas de radicalização para a língua portuguesa*. 2005. Disponível: <http://ensino.univates.br/~mald/artigowet.pdf>. Acesso em: 5 out. 2014.

Edmundson, H.P. New methods in automatic extracting. *Journal of the ACM*, v.16, n.2, p.264-285, 1969.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge discovery and data mining: Towards a unifying framework. In: International Conference on Knowledge Discovery and Data Mining, 2.,1996, Portland. *Proceedings*… Menlo Park: AAAI Press, 1996. p.82-88.

Feldman, R.; Sanger, J. *The text mining handbook*: Advanced approaches in analyzing unstructured data. New York: Cambridge University Press, 2007.

Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. Knowledge discovery in databases: An overvie1. Available from: <http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers//>. Cited: Oct. 2, 2013.

Fujita, M.S.L. A Leitura do indexador: estudo e observação. *Perspecivas em Ciência da Inforrmação*, v.4, n.1, p.101-116, 1999.

Guedes, V.L.S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. *Ciência da Informação*, v.23, n.3, p.318-326, 1994.

Internet 2012 in numbers. *Royal Pingdom*. 2012. Available from: <http:// royal.pingdom.com/2013/01/16/internet-2012-innumbers//>. Cited: Oct. 2, 2013.

Kaur, J. *et al*. Scholarometer: A social framework for analyzing impact across disciplines. *PLoS ONE*, v.7, n.9, p.1-13, 2012. doi:10.1371/journal.pone.0043235

Lopes, M.C.S. *Mineração de dados textuais utilizando técnicas de Clustering para o idioma português*. 2004. Tese (Doutorado em Engenharia Civil) - Faculdade de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

Marinho, L. B.; Girardi, R. *Mineração na web*. Porto: Universidade do Porto, 2003. Disponível em: <http://paginas.fe.up.pt/~mgi03006/ARI/MineracaoNaWeb.pdf>. Acesso em: 5 out. 2013.

Maron, M.E. Automatic indexing: An experimental inquiry. *Journal of the ACM*, v.8, n.3, p.404-417, 1961.

Maron, M.E. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, v.28, n.1, p.38-43, 1977.

Monteiro, L.O.; Gomes, I.R.; Oliveira, T. Etapas do processo de mineração de textos: uma abordagem aplicada a textos em português do Brasil. In: Congresso da Sociedade Brasileira de

Computação, 26., 2006, Campo Grande. *Anais*... Campo Grande: SBC, 2006. p.78-81.

Morais, E.A.M.; Ambrósio, A.P.L. Automatic domain classification of jurisprudence documents. In: Euro American Conference on Telematics and Information Systems, 8., 2008, New York. *Proceedings*… New York: ACM, 2008. Available from: <http://dl.acm.org/citation.cfm?id=1621103>. Cited: Oct. 11, 2013.

Nunes, F.V. *Verbal lemmatization and featurization of portuguese with ambiguity resolution in context*. 2007. Dissertation (Master Engenharia Informática) - Departamento de Informática, Universidade de Lisboa, Portugal, 2007.

Oracle. *The life cycle of a JavaServer faces page*, 2012. Available from: <http://docs.oracle.com/javaee/1.4/tutorial/doc/JSFIntro10.html>. Cited: Sept. 20, 2012.

Passini, M.L.C. *Mineração de textos para organização de documentos em centrais de atendimento*. 2012. Dissertação (Mestrado em Engenharia Civil) - Faculdade de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2012.

Pinto, V.B. Indexação documentária: uma forma de representação do conhecimento registrado. *Perspectiva em Ciência da Informação*, v.6, n.2, p.223-234, 2001.

Porter, M.F. An algorithm for suffix stripping. *Program*: *Electronic Library and Information Systems*, v.40 n.3, p.211-218, 2006.

Ramos, H.S.C.; Brascher, M. Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores informétricos para a área de C&T. *Ciência da Informação*, v.38, n.2, p.56-68, 2009.

Ranking das 100 cidades digitais. *Wireless Mundi*. 2012. Disponível em: <http://wirelessmundi.inf.br/component/content/article/51-edicoes/edicao-n-9/904-ranking-cidades-digitais>. Acesso em: 11 nov. 2012.

Rehman, Z. *et al*. Morpheme matching based text tokenization for a scarce resourced language. *PLoS ONE*, v.8, n.8, p.1-8, 2013. Available from: <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0068178#pone-0068178-g001>. Cited: Oct. 2, 2013.

Salton, G. *The SMART retrieval system-experiments in automatic document processing*. Upper Saddle River, NJ: Prentice-Hall, 1971.

Sasaki, Y. *The truth of F-measure*: Teaching, tutorial materials, version: 26th. 2007. Available from: <http://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf>. Cited: Oct. 2, 2013.

Silva, F.R.G. *Geodiscover*: mecanismo de busca especializado em dados geográficos. 2007. Tese (Doutorado em Computação Aplicada) - Departamento de Computação Aplicada, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2007.

Silva, M.R.; Fujita, M.S.L. A prática de indexação: análise da evolução de tendências teóricas e metodológicas. *Transinformação*, v.16, n.2, p.133-161, 2004.

Soares, M.V.B.; Prati, R.C.; Monard, M.C. *PreTexT II*: descrição da reestruturação da ferramenta de pré-processamento de textos. São Carlos: USP, 2008. Disponível em: <http://www.icmc.usp.br/~biblio/BIBLIOTECA/rel_tec/RT_333.pdf>. Acesso em: 5 out. 2013.

Sun, X. *et al*. Ambiguous author query detection using crowdsourced digital library annotations. *Information Processing and Management*: *An International Journal*, v.49, n.2, p.454-464, 2012. doi:10.1016/j.ipm.2012.09.001

Swanson, D.R. Searching natural language text by computer. *Science*, v.132, n.3434, p.1099-1104, 1960.

Vieira, S.B. Indexação automática e manual: revisão de literatura. *Ciência da Informação*, v.17, n.1, p.43-57, 1988.

Villalon, J.; Calvo, R.A. A decoupled architecture for scalability in text mining applications. *Journal of Universal Computer Science*, v.19, n.3, p.406-427, 2013.

Xavier, B.M.; Gomes, G.R.R.; Silva, A.D. Análise comparativa de algoritmos de redução de radicais e sua importância para a mineração de texto. *Pesquisa Operacional para o Desenvolvimento*, v.5, n.1, p.84-99, 2013.

Webber, C.G. *et al*. UnderMine text miner: uma ferramenta de mineração de texto para área educacional. *Renote*, v.11, n.1, 2013. Disponível em: <http://seer.ufrgs.br/renote/article/view/41635>. Acesso em: 10 out. 2013.