Comparing three different techniques to retrieve documents using multiwords expressions

Edson Marchetti da Silva (Centro Federal de Educação Tecnológica de Minas Gerais, CEFET-MG, Brasil) edson@div.cefetmg.br

Renato Rocha Souza (Fundação Getúlio Vargas, RJ, Brasil) rsouzaufmg@gmail.com

Abstract

This paper aims to compare the results obtained in the retrieval of documents in a corpus using as calculating the similarity function three different techniques adapted for use as descriptors Multi Words Expressions (MWE) extracted from a reference document. The techniques tested were Term Frequency-Inverse Document Frequency (TF-IDF), Cosine Similarity Vector (CVS) and Okapi BM25. The goal is to propose an alternative means of Information Retrieval Systems (IRS) find in a digital database compound of thesis, dissertations and articles, similar documents from a reference document informed by user. In this sense, are considered some structural features of the document. As results, are presented a comparison and analysis of empirical results obtained.

Keywords: Multiwords Expressions, Term Frequency – Inverse Document frequency, Cosine Similarity Vector, Okapi BM25.

Comparando três diferentes técnicas para recuperar documentos utilizando expressões multipalavras

Este trabalho visa comparar os resultados obtidos na recuperação de documentos em um *corpus* utilizando como cálculo da função de similaridade três diferentes técnicas adaptadas para usar como descritores as Expressões Multipalavras (EM) extraídas de um documento de referência. As técnicas testadas foram Term Frequency-Inverse Document Frequency (TF-IDF), Cosine Similaririty Vector (CVS) e BM25. O objetivo é propor um meio alternativo dos Sistemas de Recuperação da Informação (SRI) encontrar, em uma base digital de teses, dissertações e artigos, documentos similares a partir de um documento de referência informado pelo usuário. Nesse sentido, são consideradas algumas características estruturais do documento. Como resultados são apresentados uma comparação e análise dos resultados empíricos obtidos.

Palavras-chave: Expressões Multipalavras, Term Frequency – Inverse Document frequency (TF-IDF), Cosine Similarity Vector (CSV), Okapi BM25.

1 Introdução

Um dos grandes desafios atuais enfrentados por pesquisadores e estudantes é lidar com o excesso de informação. Com o aumento da capacidade de armazenamento dos computadores atuais tornou-se relativamente barato armazenar grandes bases de documentos em formato digital. Entretanto, o gargalo está em encontrar para uma dada consulta quais documentos são verdadeiramente relevantes. Na prática, normalmente, os sistemas de recuperação de informação sobrecarregam o usuário com respostas que retornam centenas, milhares ou muito mais documentos gerando um volume de informações que são intratáveis por seres humanos. Uma das razões que explica essa quantidade de documentos retornados como resposta é que, normalmente, as funções de similaridade dos SRI lidam com as palavras-chave fornecidas na consulta dos usuários que são tratadas como sendo compostas por unigramas independentes. Ou seja, a comparação, ou casamento dos termos é feito através de técnicas estatísticas que consideram o texto com um saco de palavras[1]. Desse modo as palavras são tratadas de forma independentes umas das outras. Entretanto, sabemos que o sentido de um texto é dado pela combinação das palavras em uma determinada ordem e agrupadas por sentenças. Além disso devem ainda ser considerados muitos outros conhecidos fatores que tornam o processo de lidar com o processamento de linguagem natural uma tarefa não trivial. Dentre esses podemos citar a sinonímia, a polissemia dentre muitos outros.

Portanto, a proposta deste trabalho é de apresentar uma forma de representar a consulta do usuário através de um documento de referência fornecido. A partir desse documento, identificar de forma automática um conjunto de EM que melhor o representam. Esse conjunto de bigramas é por fim usado como descritores da busca por documentos similares em um *corpus*. Desse modo, ao combinar vários bigramas para serem comparados espera-se aumentar a aderência dos documentos recuperados em relação ao documento utilizados como referência pelo usuário de tal forma a restringir a quantidade de documentos retornados como resposta.

Neste trabalho foram utilizadas três diferentes técnicas adaptadas para lidar com bigramas, ou termos dependentes, para calcular o coeficiente de similaridade. Como conclusão os resultados obtidos são comparados e analisados.

Para melhor descrever os experimentos este trabalho está estruturado nas seguintes seções nas quais são apresentados os seguintes conteúdos: seção 2 – são apresentados alguns dos trabalhos correlatos; na seção 3 - referencial teórico sobre as EM e as técnicas empregadas; na seção 4 a metodologia; na seção 5 – a conclusão; seção 6 - recomendações para trabalhos futuros.

[1] Saco de palavras é uma expressão traduzida do inglês *bag of words*, em que o texto é considerado com sendo um amontoados de palavras independente umas da outras.

2 Trabalhos correlatos

Diversos trabalhos que buscam identificar as EM foram publicados, dentre eles destacamos Dias, Lopes e Guilloré (1999) que visa a extração de EM de forma independente de linguagem, baseados exclusivamente em métodos estatísticos; Silva, Lopes (1999) que visa extrair *n*-gramas a partir da análise do texto em um contexto local denominado LocalMaxs; Portela, Mamede e Batista (2011) o qual leva em consideração as características morfo-sintáticas do texto, e que por isso demandam intensivo uso de recursos computacionais, dentre outros.

Também podemos citar pesquisas que aplicam o conceito de EM para tradução automática através do uso de alinhamento lexical das expressões, as quais visam verificar de que forma comparar o mesmo texto em idiomas distintos pode fornecer pistas relevantes para a identificação dessas expressões: Calzolari et al. (2002); Sag et al. (2002); Ramich (2009); Zhang et al. (2009); Villavicencio et al. (2010). Tomando como base esses trabalhos verificamos a existência de uma lacuna no que se refere à extração de EM que levam em consideração as características físicas intrínsecas dos documentos e que seja independente de idioma. É a partir dessas ideias que propusemos obter as EM de um documento base e utilizá-las como descritores da busca comparada para recuperação automatizada de documentos similares.

3 Referencial teórico

Segundo Sarmento (2006), o texto não é um simples amontoado aleatório de palavras. A ordem da colocação das palavras no texto é que produz o sentido. Portanto, o estudo da co-ocorrência das palavras traz consigo uma informação importante. Isso pode indicar que as palavras estão relacionadas diretamente por composicionalidade ou afinidade ou indiretamente por semelhança. Portanto, a base da linguística empírica consiste em encontrar a partir da frequência de co-ocorrências observada, as dependências significativas entre os termos. Evert (2005 citado por Sarmento) aponta como sendo quatro esses grupos de medidas:

– testes de significância estatística;

– coeficientes de associação;

– baseadas em conceitos da teoria da informação;

– baseadas em heurísticas diversas.

Conforme citado por Zhang et al. (2009), a capacidade de expressar sentido de uma palavra depende das demais palavras que a acompanham. Quando uma palavra aparece acompanhada por um conjunto de termos, maiores são as chances desse conjunto possuir um significado relevante. Isso indica que não apenas a palavra, mas também a informação contextual é útil para o processamento de informações. É a partir dessa ideia simples e direta que pesquisas sobre EM são motivadas. Desse modo espera-se capturar conceitos semânticos relevantes do texto expressos pelas EM.

Apesar de haverem diversos trabalhos publicados sobre o tema, não existe uma definição formal consensual na literatura sobre EM. Neste trabalho considera-se as EM como: formações compostas de duas ou mais palavras, baseado na frequência de co-ocorrência no texto, que quando associadas possuem uma expressividade semântica mais forte do que, quando cada um de seus termos são postos separadamente. A seguir apresenta-se alguns dos conceitos relacionados com as técnicas para apuração da relevância e as suas adaptações necessárias empregadas no cálculo do coeficiente de relevância.

Singhal et al. (1996), o cálculo do peso de relevância do documento é uma das partes mais importantes de um SRI. Esses métodos de cálculo do peso geralmente se baseiam em:

- Term Frequency (*tf*) – frequência do Termo correspondendo ao número de ocorrências do termo no documento;
- Document Frequency (*df*) – frequência do Documento correspondendo ao número de documentos da coleção em que o termo ocorre;
- Collection Frequency (*cf*) – frequência da Coleção correspondendo ao número total de ocorrências do termo na coleção.

Uma das formas mais intuitivas de buscar documentos similares é verificar a frequência de ocorrência em que os termos da consulta ocorrem nos documentos. Intuitivamente aqueles com maiores ocorrências do termo são os mais relevantes. Entretanto, essa é uma forma ingênua por considerar que todos os termos têm igual importância, além de desconsiderar outros fatores que também contribuem na identificação dos documentos relevantes. Um desses fatores é a identificação do quão frequente o termo no *corpus.* Afinal, termos muito frequentes são pouco importantes na geração de uma boa resposta. Foi a partir dessas ideias que Salton e McGill (1983, p. 201, 211) propusseram como forma de melhorar a precisão das respostas obtidas nas consultas, a adoção de técnicas que consideram as características em comum dos documentos (*intra-document*) e as características para fazer a distinção entre os documentos (*inter-document*). Dessa forma, os pesos são usados para computar as similaridades entre os documentos são relativizados pela dissimilaridade, ou capacidade discriminatória do termo. Em outras palavras, quanto maior a frequência do termo em documentos distintos, menos importante ele é para discriminar um documento dentro da coleção. Essa estratégia é conhecida como ponderação Term Frequency Inverse Document Frequency (TF-IDF). A forma adotada neste experimento é mostrada na expressão 3.1.

$$\boxed{\begin{array}{l} IDF = \log\left(\frac{N}{n}\right) \\ w = TF * IDF \end{array}} \qquad (3.1)$$

Onde: *TF* é a frequência em que o termo foi encontrado no documento; *IDF* o inverso da frequência do documento calculado pelo logaritmo da razão entre *N* (número total de do-

cumentos da coleção) por *n* (número de documento que contém o termo); finalmente o peso é calculado pelo produto de *TF* * *IDF*.

Nesse contexto, *TF* corresponde às características intradocumentos do termo e *IDF* dá uma medida de distinções interdocumento.

Uma outra abordagem utilizada para representar os termos como informação local (no nível do documento) e global (no nível da coleção de documentos) é o modelo de espaço vetorial. Manning e Schütze (1999, p. 539) definem o modelo de Espaço Vetorial como sendo um dos modelos mais amplamente utilizados devido à sua simplicidade conceitual e ao uso da metáfora que relaciona proximidade espacial com a proximidade semântica. Essa abordagem permite um casamento parcial delineado através dos pesos que expressam o grau de similaridade entre a consulta e os documentos. Cada documento é representado como um vetor de termos, sendo que cada um deles possui um peso associado que indica o seu grau de importância no documento. Os pesos podem ser calculados de diversas formas e servem para especificar a magnitude do vetor.

Para um melhor entendimento desse modelo faz-se necessário apresentar alguns conceitos básicos que o sustentam.

O plano cartesiano é formado por duas retas perpendiculares, que representam um eixo horizontal *x* ou eixo das abscissas e um eixo vertical *y* ou eixo das ordenadas. A representação de um ponto é feita por pares ordenados com valores de *x* e de *y*, dado por $P(x_1, y_1)$. O produto escalar dos pontos $P_1 \bullet P_2$ pode ser obtido pelo somatório da multiplicação das coordenadas dos pontos em cada eixo, dado pela expressão (3.2).

$$\boxed{P_1 \bullet P_2 = x_1 * x_2 + y_1 * y_2} \tag{3.2}$$

Se os pontos P_1 e P_2 estiverem definidos em *n* dimensões então o produto escalar é dado pela expressão (3.3).

$$\boxed{P_1 \bullet P_2 = x_1 * x_2 + y_1 * y_2 + \ldots + n_1 * n_2} \tag{3.3}$$

Outro conceito importante a ser apresentado é como calcular a distância entre dois pontos num plano e em seguida projetar esse conceito para a distância entre dois pontos num espaço *n*-dimensional. Nesse sentido, considerando P_0 como sendo a posição de coordenada (0, 0), tem-se d_1 como sendo a distância entre os pontos P_0 e P_1 e d_2 como sendo a distância entre P_0 e P_2, conforme mostrado na figura 1.

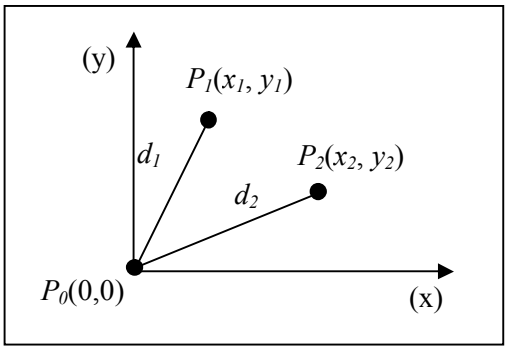


FIGURA 1 – Distâncias d_1 e d_2
Fonte: Elaborada pelos autores.

Para obter-se a magnitude da distância d_1 entre os pontos P_0 e P_1 e a distância d_2 entre os pontos P_0 e P_2, também chamada de distância Euclidiana, pode-se utilizar a expressão (3.4).

$$d_1 = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2} \text{ e } d_2 = \sqrt{(x_2 - x_0)^2 + (y_2 - y_0)^2} \quad (3.4)$$

Entretanto, como o objetivo é utilizar o modelo de Espaço Vetorial para expressar os dados da coleção deve-se adaptar representação das retas e substituí-las por vetores, conforme mostra a figura 2.

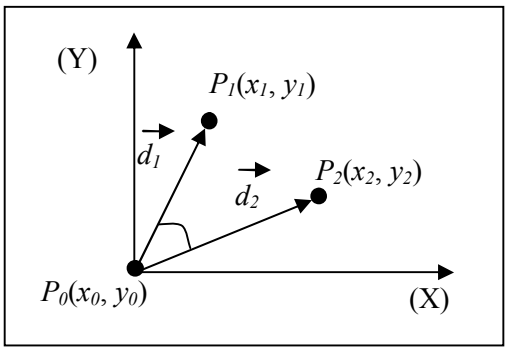


FIGURA 2 – Representação dos vetores no plano
Fonte: Elaborada pelos autores.

Como um vetor representa duas grandezas: o valor/modulo e a sua direção/sentido, o modelo considera que o peso *w* está representado pelo seu módulo e a direção representada no eixo. Sendo que cada eixo representa um termo do espaço *n*-dimensional. Desse modo, um documento é representado por um vetor resultante da projeção de todos os seus termos no espaço *n*-dimensional. Considerando que uma consulta de usuário *q* também pode ser representada da mesma forma, pela projeção de seus termos produzindo um vetor $\vec{q}$, quanto maior for o grau de similaridade entre os vetores resultantes da consulta com o documento, maior será a

proximidade dos vetores projetados no espaço *n*-dimensional que representam essas duas grandezas.

Conforme apresentado por Manning e Schütze (1999, p. 146), esse modelo quantifica a similaridade entre dois vetores resultantes pela projeção de seus termos no espaço *n*-dimensional considerando os pesos. Sendo que o peso representa a importância que cada termo tem para o documento. Entretanto, essa medida apresenta uma desvantagem por representar dois documentos muito similares através de vetores resultantes que podem ser significativamente bem diferentes, simplesmente por que um documento é muito maior do que o outro. Isso ocorre, pois, mesmo com uma idêntica distribuição em ambos os documentos, os valores absolutos do peso podem ser distintos. Portanto, para compensar esses efeitos do tamanho do documento é necessário fazer uma ponderação pelo tamanho do documento. Para padronizar o valor do coeficiente calculado utiliza-se a técnica Cosine Vector Similarity (CVS). Ou seja, para normalizar o valor calculado da similaridade entre os dois vetores emprega-se o resultado obtido pelo coseno do ângulo formado pela interseção dos vetores resultantes das projeções, da consulta e do documento, no espaço *n*-dimensional. Esses valores calculados variam entre 0 e 1. Desse modo, quando o ângulo entre os vetores for se aproximando de zero isso indica que o resultado do coseno aproxima-se do valor um, representando máxima similaridade. Por outro lado, se o ângulo entre os dois vetores aproxima-se de 90 graus, o coseno tende ao valor zero, desse modo não há similaridade. A figura 3 apresenta um esboço da representação do modelo de espaço vetorial considerando apenas $n = 3$. Ou seja, é uma simplificação, em três dimensões, para que possa ser representado através de um desenho. Os termos que compõem o dicionário estão representados nos eixos T_1, T_2 e T_3. Os pontos w_1, w_2 e w_3 mostrados nos eixos T_1, T_2 e T_3 correspondem repectivamente à magnitude dos pesos de cada um desses termos para o documento d_1. Os pontos w_4, w_5 e w_6 mostrados nos eixos T_1, T_2 e T_3 correspondem respectivamente à magnitude dos pesos de cada um desses termos para o documento d_2. Os vetores d_1 e d_2 representam à projeção resultante dos termos dos documentos. O q representa o vetor resultante da projeção dos termos da consulta. Desse modo, os cosenos dos ângulos $_1$ e $_2$ representam a similaridade entre o documento d_1 e a consulta q e entre o documento d_2 e a consulta q respectivamente.

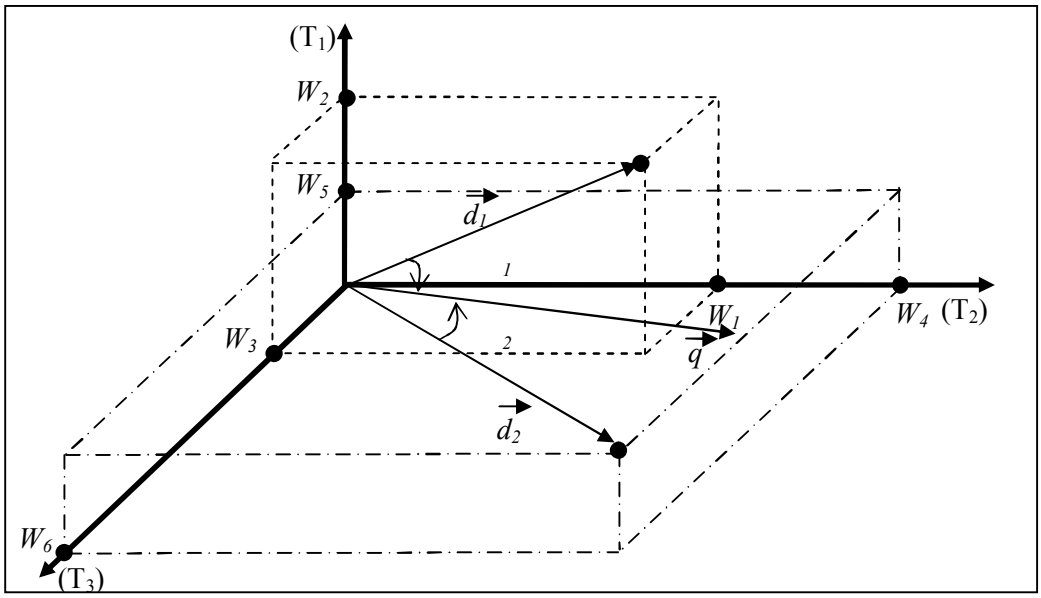


FIGURA 3 – Representação do *cosine vector similarity*
Fonte: Elaborada pelos autores.

Se considerarmos que o ângulo $_1$ é menor que $_2$ isso significa dizer que o documento d_1 é mais relevante para a consulta q do que o documento d_2. A expressão para calcular essa medida de similaridade é mostrada em (3.5).

$$Sim(\vec{d}_1, \vec{d}_2) = \cos eno\Phi = \frac{P_1 \bullet P_2}{|P_1||P_2|} = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{(x_1)^2 + (y_1)^2} * \sqrt{(x_2)^2 + (y_2)^2}} \quad (3.5)$$

Ou seja, a medida de similaridade mostra o quão próximo os vetores que representam os documentos se aproximam do vetor que representa a consulta. Conforme pode ser visto na expressão (3.5), o valor da similaridade pode ser obtido, pela razão entre o produto escalar dos vetores pelo produto da raiz quadrada das coordenada em cada dimensão ao quadrado.

Para apurar a similaridade deve-se calcular o coseno do ângulo formado entre a consulta com cada um dos vetores resultantes, que representam cada um dos documentos da coleção e classificar o resultado em ordem decrescente. Isso, pode ser representado em um espaço *n* dimensional de bigramas. Onde cada eixo representa uma EM composta por termos do dicionário e o valor do peso é atribuído para cada documento que possuir o bigrama. Desse modo, o espaco *n*-dimensional terá o valor de *n* correspondente à quantidade de bigramas extraidos do documento de referência.

O cálculo da similaridade também pode ser realizado através da expressão (3.6).

$$Similaridade(Q,D) = \frac{\sum_{k=1}^{n} w_{qd} * w_{dk}}{\sqrt{\sum_{k=1}^{n} (w_{qk})^2 * \sum_{k=1}^{n} (w_{dk})^2}} \quad (3.6)$$

Onde:

- Q representa o vetor de bigramas da consulta;
- D *representa o* vetor de bigramas de cada documento;
- W_{qk} peso do bigrama k para a consulta;
- W_{dk} peso do bigrama k para o documento.

Um outro aspecto que deve ser considerado, em especial para o foco deste trabalho, é que em uma base de documentos que contém teses, dissertações e artigos, se caracteriza por possuir documentos de tamanhos diversos. Tipicamente com documentos que poderão variar desde cinco até pouco mais de três centenas de páginas. Essa variação de tamanho faz com que documentos maiores possam ser sobre estimados no cálculo do peso. Singhal et al. (1996) afirma que durante anos pesquisadores têm trabalhado com a suposição de que a relevância do documento é independente do seu tamanho. Entretanto, seus estudos mostraram que os documentos longos têm maiores chances de serem julgados como relevantes em uma consulta de usuário do que os documentos menores. Desse modo, visando normalizar documentos de tamanhos diferentes no processo de busca, Robertson e Spark (1976) propusseram um modelo que considera o peso local de um termo. Foi a partir dessa ideia que foi desenvolvido um importante modelo de RI, conhecido como Okapi BM25, após ter sido implementado pelo SRI Okapi nos anos 1980 e 1990 pela London's City University. A fórmula utilizada para o cálculo é apresentado em (3.7).

$$score(d,q) = \sum_{i-1}^{n} \frac{TF_{di} * IDF_{di}}{TF_{di} + length_i} \quad (3.7)$$

Sendo que *length* corresponde ao cálculo da razão do tamanho do documento, em palavras, pelo tamanho médio dos documentos da coleção. Ou seja, documentos grandes acima do valor médio do *corpus* terão o valor de tamanho (*length*) maior do que um. Esse valor, somado ao denominador da expressão produz uma redução no valor da relevância. Desse modo, quanto maior for a diferença de tamanho do documento em relação ao valor da média de tamanho, maior será essa redução.

Conforme apresentado por Manning, Raghavan & Schütze (2009, p. 232-243), novos parâmetros foram introduzidos no esquema de ponderação BM25 como forma de construir um modelo probabilístico sensível às variações de tamanho do documento. Algumas dessas variações são apresentadas a seguir em (3.8) e (3.9).

Onde:

- df_t representa o número de documentos que contem o termo t;
- df_{td} representa a frequência do termo t no documento d;
- L_d representa o tamanho do documento, em palavras;
- L_{avg} representa o valor do tamanho médio dos documentos da coleção;
- k_1 é um parâmetro positivo de ajuste para calibrar a escala da frequência do termo no documento;
- b é um parâmetro de ajuste que determina a escala pelo tamanho do documento.

$$RSV_d = \sum_{t \in q} \log\left[\frac{N}{df_t}\right] * \frac{(k_1+1)tf_{td}}{k_1\left[(1-b)+b*\left(L_d / L_{avg}\right)\right]+tf_{td}} \tag{3.8}$$

Se o valor de k_1 for igual a zero, fará com que o valor do termo que multiplica a expressão no numerador seja igual a um e no denominador irá zerar a expressão entre colchetes. Desse modo, a expressão da frequência do termo no documento será desconsiderada. Por outro lado, aumentar o valor de k_1 trará uma maior participação da frequência do termo no resultado do cálculo.

Se b for igual a um, indica o uso total do valor de ponderação de tamanho calculado para cada documento, correspondente ao peso do termo pelo tamanho do documento. Enquanto que, se b for igual a zero a não haverá normalização de tamanho.

$$RSV_d = \sum_{t \in q} \log\left[\frac{N}{df_t}\right] * \frac{(k_1+1)tf_{td}}{k_1\left(\left((1-b)+b*\left(L_d / L_{avg}\right)\right)+tf_{td}\right)} * \frac{(k_3+1)}{k_3+tf_{tq}} \tag{3.9}$$

Onde:

- df_{tq} representa o número de documentos que contêm o termo t na consulta q;
- k_3 é um outro parâmetro positivo de ajuste que calibra a escala da frequência do termo na consulta.

Nessa expressão, não é necessário incluir a normalização de tamanho da consulta, pois no processo de RI sempre a consulta tem tamanho fixo.

O ajuste desses parâmetros em ambas as expressões possibilita otimizar o desempenho das respostas para cada *corpus* específico. Segundo esses mesmos autores, testes empíricos mostram que valores dos parâmetros que dão um bom desempenho são para k_1 e k_3 na faixa

entre 1,2 e 2, e do parâmetro b, um valor igual a 0,75. Eles afirmam ainda que BM25 tem sido usada com muito sucesso em uma diversidade de tarefas de busca e em diferentes coleções de documentos.

No contexto deste trabalho três diferentes formas de calcular o coeficiente de relevância foram implementadas: TF-IDF, CVS e BM25. Entretanto todos esses modelos foram adaptados para ponderar o peso da frequência dos termos de um bigrama a um coeficiente estrutural (Ce). Dessa forma, os termos que compõem o bigrama têm seu peso relativizado de acordo com a sua estrutura encontrada no documento original. Ou seja, a frequência dos termos será majorada para capturar essas características estruturais do termo no documento, tais como: termo todo em maiúsculas, com somente a primeira letra em maiúsculas ou totalmente em minúsculas. Adicionalmente, uma outra adaptação das técnicas do cálculo de relevância aplicadas neste trabalho se fez necessária. Afinal, em vez dos descritores serem representados por termos isolados, eles representados por bigramas constituídos por termos dependentes. Portanto, para ser considerado como adjacentes os termos têm de estar em uma mesma sentença. A seguir, na seção da metodologia, são apresentados os métodos empíricos utilizados.

4 Metodologia

Para realizar o experimento foi utilizado um *corpus* composto por 194 artigos completos publicados no ano de 2010 no principal encontro científico da área da Ciência de Informação (ENANCIB). Todos os arquivos foram obtidos em formato Portable Document Format (PDF) e armazenados em um sistema de arquivos informatizado. Adicionalmente implementou-se dois componentes de *software:* Um denominado Server e um outro denominado Client. O Server é o responsável por indexar o *corpus* e disponibilizar um serviço de consulta para recuperação da informação. O Client é o responsável por receber o documento de referência para a busca, extrair as EM, enviar a requisição de consulta e retornar a resposta com os documentos similares.

Este trabalho dá sequência ao trabalho de [Silva e Souza, 2012], o qual apresentou uma forma determinística de extração de EM. Sendo que, a meta agora é comparar o uso de três diferentes técnicas de cálculo de relevância utilizando como descritores EM extraídas de um documento de referência fornecido pelo usuário. Portanto, a meta é verificar empiricamente qual é a melhor técnica para cálculo da relevância baseada no casamento de bigramas.

Nesse sentido, serão extraídas as EM de um documento de referência a serem utilizadas como descritores de busca em um SRI. Essa proposta possibilita ao usuário uma alternativa de busca. Em que, em vez de informar palavras-chave como elemento de busca, caberá ao usuário informar um documento. Em outras palavras, as buscas serão feitas a partir dos bigramas extraídos de um documento de referência. Essa estratégia alternativa simplifica o trabalho do usuário, que passa a utilizar documentos conhecidos sobre o tema de seu interesse para servir como base da busca comparada na recuperação de documentos similares.

Inicialmente foram realizadas buscas no *corpus* utilizando-se como referência vinte documentos aleatórios. O cálculo de similaridade utilizado foi o CSV. Essa técnica calcula o somatório dos coeficientes de correlação apurados para cada um dos bigramas extraídos do documento de referência e identificados no *corpus*. De modo que, desde que haja pelo menos um bigrama coincidente, entre os extraídos do documento de referência com o *corpus*, ele passa a ser considerado com parte da resposta. Portanto, pode haver muitos documentos calculados com valores de coeficiente residual. Ou seja, com valores bem pequenos. Sendo assim, é conveniente que todo processamento de seleção deva trabalhar com um ponto de corte. A definição desse limiar permite selecionar como resposta apenas os documentos em que o cálculo do seu coeficiente de similaridade seja maior que um percentual parametrizado em relação ao valor do máximo coeficiente apurado entre todos os documentos obtidos como resposta. A tabela 1 apresenta um estudo exploratório desse comportamento, em que foi observada a quantidade de documentos retornados, considerando vinte buscas realizadas para diversos limites utilizados como ponto de corte.

TABELA 1 – Documentos retornados considerando o ponto de corte.

Ponto de Corte (%)	Quantidade média de documentos retornados
1	6,9
10	3,5
20	2,1
30	1,8
40	1,3
50	1,2
60	1,15
70	1,1
80	1
90	1
100	1

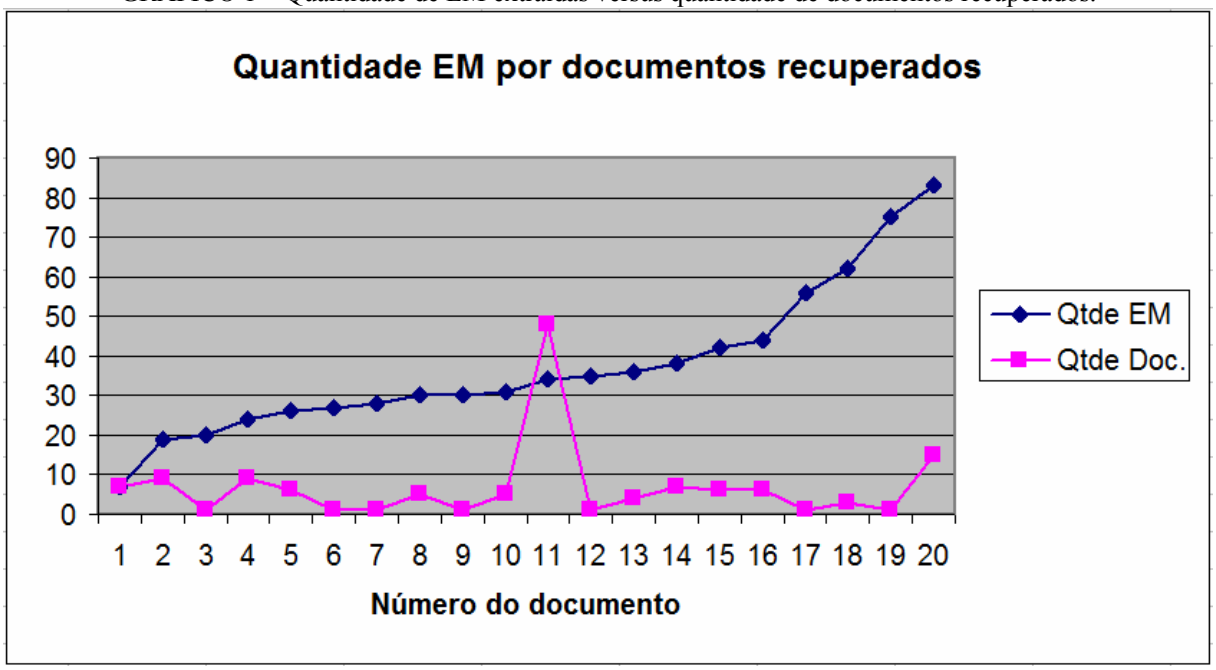
Fonte: Elaborada pelos autores.

Ao analisar a Tabela 1, percebe-se que o número de documentos retornados reduz à medida que o limiar de corte aumenta, até que apenas um único documento é retornado. Ou seja, apenas o mais similar.

O Gráfico 1 foi gerado utilizando-se como ponto de corte o valor igual a 1%. Ele expressa no eixo das abscissas cada um dos vinte documentos pesquisados e no eixo das ordenadas duas grandezas em valores absolutos. Sendo a primeira, a quantidade de EM identificadas no documento, representada pela curva em azul e, a segunda, a quantidade de documentos retornados pela busca, representada pela curva rosa. Dessa forma, cada coordenada mostrada no gráfico relaciona o valor obtido dessas grandezas através da busca realizada para cada documento. Para facilitar a análise do comportamento dessas grandezas, os documentos foram ordenados de forma crescente pela quantidade de EM extraídas. Entretanto, ao analisar o grá-

fico constata-se que não há uma relação de dependência direta entre a quantidade de EM identificadas no documento com a quantidade de documentos recuperados. Esse resultado leva a supor que existem outros fatores que contribuem para influenciar esses comportamentos, como por exemplo: a frequência de ocorrência dos bigramas pesquisados nos demais documentos do *corpus*.

GRÁFICO 1 – Quantidade de EM extraídas versus quantidade de documentos recuperados.



Fonte: Elaborado pelos autores.

A Tabela 2 apresenta a quantidade de documentos em que foram extraídas as EM dentro de determinadas faixas de valores. Ou seja, de 1 a 10; de 11 a 20 etc. Ao analisar esses dados percebe-se que, da maior parte dos documentos foi extraído entre 20 a 60 EM, sendo que o valor médio encontrado foi de 43,5 EM por documento.

TABELA 2 – EM identificadas nos documentos.

EM identificadas por intervalo de quantidade	Total de documentos
10	4
20	19
30	51
40	38
50	26
60	27
70	16
80	41
90	1
100	1
110	0
120	1
130	0

140	1
Média por documento	43,5 EM

Fonte: Elaborada pelos autores.

Existem vários fatores que impactam no processo de RI utilizando EM, dentre os quais destacam-se:

- O fator de co-ocorrência utilizado, que é o limite inferior para se considerar um bigrama como sendo relevante, neste trabalho o valor utilizado foi quatro;
- O número de EM extraídas, quanto maior esse número, maiores são as chances de se encontrarem documentos similares. Portanto, maior a necessidade de se utilizar um ponto de corte para excluir os coeficientes de similaridade calculados com valores distantes do valor máximo;
- O tamanho do documento. Documentos menores normalmente possuem menor frequência de co-ocorrência dos bigramas;
- O tamanho do *corpus*. Quanto maior a quantidade de documentos existentes na base maiores são as chances de se encontrarem similares;
- Os critérios adotados no cálculo do coeficiente de similaridade, que interferem no cálculo da relevância.

Visando avaliar outros aspectos, elaborou-se um novo experimento que visa submeter uma busca para cada um dos 194 documentos existentes no *corpus*. Para realizar esse experimento, de busca exaustiva, um novo componente de *software* foi elaborado, denominado "ConsEM.exe". Esse programa funciona como um robô de consulta, evitando a necessidade de processar as buscas uma a uma através da interface do usuário. O objetivo é automatizar o processo de consulta dos documentos e do registro das respostas produzidas que servirão para avaliar os resultados obtidos.

Assim, esse novo programa foi utilizado para processar as 194 consultas e contabilizar quantas foram as EM extraídas em cada documento. A partir desse processamento, verificou-se como sendo 38,6 a quantidade média de bigramas extraídos dos documentos, sendo que os valores máximos e mínimos foram respectivamente 134 e 7 bigramas.

Para entender qual é a influência que a quantidade de bigramas utilizados como descritores impacta na quantidade de documentos recuperados, foi implementado no programa uma requisição solicitando a quantidade de bigramas a serem considerados para o processo de busca. Dessa maneira, a cada documento de referência processado, os seus bigramas são extraídos e inseridos em uma árvore binária, a qual os ordena de forma decrescente pela frequência de co-ocorrência. Consequentemente, no momento de processar a busca dos documentos similares, somente são considerados os "*n*" primeiros bigramas recuperados da estrutura de árvore. Ou seja, os mais frequentes. Portanto, nesse experimento foram realizadas 194 buscas para

cada valor de "*n*" arbitrado. Sendo que, como foram definidos 17 valores distintos para *"n*", cada documento de referência foi consultado 17 vezes, totalizando 3.298 consultas no *corpus.*

Desse modo, foi possível calcular a quantidade de documentos recuperados e a partir deles apurar o valor da quantidade média de documentos recuperados para cada valor de *n.* A tabela 2 mostra os valores calculados, considerando 1% como sendo o valor de ponto de corte do fator de relevância. Ou seja, são considerados apenas os documentos cujo coeficiente de similaridade calculado corresponda a um valor maior ou igual a 1% do maior coeficiente apurado por documento. Cabe ressaltar que as consultas são realizadas de tal forma que a cada instante é comparado um documento do *corpus* com os demais até que todos tenham sido consultados. Nesse contexto, a cada consulta sempre o documento retornado como sendo o mais relevante é o próprio documento utilizado como referência da busca. Afinal, nenhum documento pode ser mais similar do que o próprio documento. Isso faz com que, nesse caso, o valor do coeficiente similaridade seja máximo.

TABELA 3 – Comparação da quantidade de descritores *versus* documentos retornados.

Sequência	Limite *n* de bigramas usados na busca	Quantidade média de documentos
1	1	31,78
2	5	20,37
3	10	18,91
4	15	15,97
5	20	15,61
6	25	16,17
7	30	14,66
8	35	14,76
9	40	16,28
10	45	15,54
11	50	14,73
12	55	14,78
13	60	15,17
14	65	14,89
15	70	14,68
16	75	14,63
17	999	14,81

Fonte: Elaborada pelos autores.

Ao analisar os dados apresentados na tabela 3 verifica-se que ao usar apenas um bigrama - o que na prática funciona com uma busca convencional por palavras-chave - são retornados em média 31,78 documentos. Na medida em que aumenta-se o número de descritores, por exemplo 15, o número de documentos retornados cai pela metade. Ou seja, ocorre melhora na precisão da busca. A partir desse ponto, mesmo aumentando os descritores até atingir o valor total de EM extraídas, a variação da quantidade média de documentos retornados apresenta uma variação insignificante. Isso leva a concluir que não é necessário estender a busca para todos os bigramas extraídos. É possível limitar a busca para apenas uma parte das EM extraídas mantendo a precisão. Essa estratégia melhora o desempenho da busca.

Com o objetivo de verificar o uso da ferramenta de busca e avaliar os resultados retornados e o tempo de resposta demandado pela busca um outro experimento foi realizado. Nesse sentido, foram realizadas duas buscas aleatórias. Os resultados obtidos são descritos a seguir.

Em todas as pesquisas realizadas, foi informado como ponto de corte o percentual igual a 50%. Ou seja, são apresentandos como respostas da consulta apenas os documentos cujo coeficiente de relevância calculado atingir pelo menos 50% do valor da máxima relevância alcançada na busca. Desse modo, somente serão considerados como respostas aqueles documentos cujo ângulo formado pelos vetores do documento e da consulta estiverem na faixa de zero até cinquenta graus.

Ao pesquisar o documento intitulado “Uma abordagem baseada em métricas de redes complexas para estabelecimento do grau de influência de termos em documentos” foram encontrados 39 bigramas com as respectivas frequências de ocorrência conforme mostrados na tabela 4. Como o tema desse artigo é bem específico dentro da coleção utilizada, o que pode ser perceber pelos próprios bigramas mostrados na tabela que nessas condições parametrizadas nenhum documento similar foi encontrado.

TABELA 4 – Bigramas e frequência de ocorrência extraídas do documento 172.pdf.

Bigrama	F	**Bigrama**	F	**Bigrama**	F
redes complexas	24	information retrieval	5	correspondente numero	4
recuperacao informacao	16	peso termo	5	degree centrality	4
et al	13	maior grau	5	centralidade rede	4
termos documentos	12	rede medida	5	tal modelo	4
atribuicao pesos	11	pesos termos	5	numero ligacoes	4
termo documento	9	funcao utiliza	5	medida normalizada	4
coeficiente agrupamento	9	calculo similaridade	5	precisao interpolada	4
metricas redes	8	documentos consultas	5	relacoes sintaticas	4
grau proximidade	7	grau influencia	5	analise redes	4
modelo vetorial	6	metricas rede	5	consultas documentos	4

documentos coleção	6	complex networks	4	similaridade documentos	4
grau intermediação	6	distancas geodesicas	4	metodos tradicionais	4
grau centralidade	6	distancia geodesica	4	complexas palavras	4

Fonte: Elaborada pelos autores.

Por outro lado, ao pesquisar documentos que lidam com termos de uso mais comum do jargão da área relacionada à coleção, o número de respostas tende a crescer. Por exemplo, ao pesquisar o documento intitulado "Repositório digital da Unati-UNESP: o olhar da arquitetura da informação para a inclusão digital e social de idosos" foram encontrados 12 bigramas conforme mostrados na tabela 5.

TABELA 5 – Bigramas e frequência de ocorrência extraídas do documento 86.pdf.

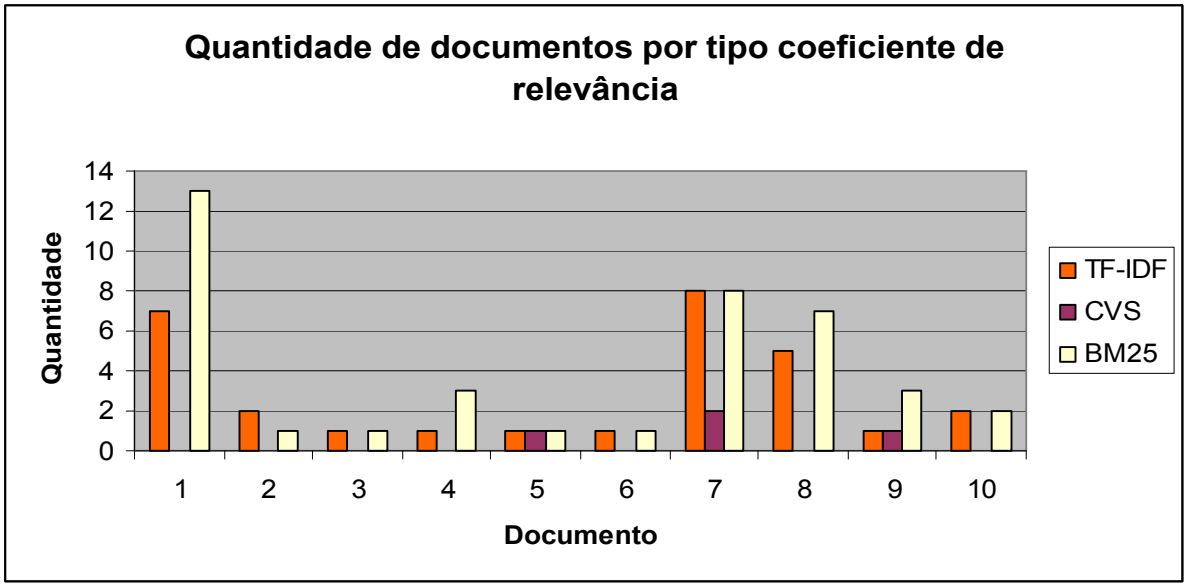
Bigrama	F	**Bigrama**	F	**Bigrama**	F
deste trabalho	4	construcao participativa	4	digitais idosos	4
producao intelectual	4	informacional digital	4	vidotti 2008	4
tecnologias informacao	4	acessibilidade comportamento	4	digital repositorio	4
grupos focais	4	meio grupos	4	digital idosos	4

Fonte: Elaborada pelos autores.

Nesse caso foram selecionados os seguintes artigos intitulados: "Representação interativa e folkosonomia assistida para repositórios digitais" (384.pdf) e "Contribuição dos repositórios institucionais à comunicação científica: um estudo da Universidade Federal do Rio Grande do Sul" (186.pdf).

Finalmente, foi realizando um experimento com objetivo é comparar as respostas produzidas pelas três diferentes formas de apurar o coeficiente de relevância implementado na ferramenta de busca: TF-IDF, CVS e BM25. Nesse sentido, foram retirados aleatoriamente do *corpus* dez documentos. O objetivo é fazer com que os documentos do *corpus* sejam diferentes dos documentos utilizados como referência da busca. Isso fará com que o coeficiente máximo calculado não seja o valor do casamento do documento de busca com ele próprio. Portanto, a coleção de documentos indexada passou a ser de 184 documentos. Desse modo, os dez documentos retirados foram utilizados, um por um, como sendo o documento de referência da busca comparada. Para cada documento foram realizadas três buscas, uma para cada técnica. No total foram realizadas trinta consultas através da interface padrão de busca do usuário. Para realizar as consultas os parâmetros de configuração definidos foram: ponto de corte igual a 60%; o limite de bigramas extraídos no documento de referência utilizado foi de até vinte bigramas; parâmetros de ajuste da técnica BM25 foram definidos como sendo K1 e K3 = 2 e B = 0.75. Os resultados obtidos são apresentados no gráfico 2, o qual mostra a quantidade de documentos retornados pelas três técnicas para cada um dos dez documentos de referência utilizados.

GRÁFICO 2 – Quantidade de documentos retornados por tipo de coeficiente de relevância



Fonte: Elaborado pelos autores.

Conforme demonstra o gráfico 2, as técnicas TF-IDF e BM25 tiveram resultados próximos em termos de quantidade de documentos retornados. Sendo que a BM25 retornou mais documentos do que a TF-IDF em quatro casos e menos, em apenas um caso. Nos demais casos, a quantidade retornada foi a mesma. Por outro lado, CVS não retornou nenhum documento em sete casos e em apenas um a quantidade foi a mesma das demais técnicas empregadas. Esses resultados mostram que a CVS foi a técnica mais seletiva no processo de seleção e que BM25 a que retornou mais documentos.

Esses resultados tão restritivos obtidos pela CVS podem ser melhor entendidos se considerarmos as características envolvidas em seu processo de cálculo. Em uma consulta composta por até 20 bigramas utilizando CVS, para que um documento seja retornado com o ponto de corte definido em 60%, é necessário que no documento exista cerca de doze bigramas iguais, dos vinte existentes no vetor de consulta. Ou seja, é necessário haver um casamento de muitos termos simultaneamente. Nesse caso, cujo coeficiente varia entre zero e um, para que a similaridade seja máxima, é necessário que todos os bigramas da consulta existam no documento. Por outro lado, as outras duas técnicas trabalham no sentido de encontrar o documento mais relevante. Esse valor apurado é que servirá de base para filtrar os demais documentos que tenham valor de coeficiente maior ou igual a 60% em relação ao valor máximo apurado. Portanto, sempre se terá pelo menos um documento de resposta, desde que ocorra o casamento com os documentos do *corpus* de pelo menos um dos bigramas que compõem a consulta.

Portanto, BM25 pode ser considerada, nesse caso da busca em base de teses, dissertações e artigos, como sendo a técnica mais apropriada para ser utilizada. Afinal, ela leva em consideração as diferenças de tamanho existentes entre esses tipos de documentos; considera

também o peso que a EM tem no documento de referência ao ponderar a frequência do bigrama da consulta no processo de cálculo do coeficiente.

5 Conclusão

Ao realizar uma reflexão sobre todo processo de pesquisa empreendido constatou-se que o desempenho do *software*, ou seja, o tempo de resposta da consulta do usuário está dentro de valores razoáveis. Em todos os testes realizados o tempo de busca não ultrapassou de 4 segundos. É importante enfatizar que o custo do algoritmo de consulta não cresce de forma linear com o tamanho do *corpus.* Afinal o acesso aos termos da busca é realizado numa estrutura em memória com acesso direto ou direto mais um curto caminhamento através de uma lista de colisões de uma estrutura de *hash.* Desse modo, a tendência é que o tempo de resposta *online* se mantenha mais ou menos constante em diferentes tamanhos de bases de dados indexadas.

Obviamente, que o mesmo não ocorre com o algoritmo de indexação da base que cresce linearmente com o tamanho do conjunto de documentos a ser indexado. Mas, esse processamento de carga é necessário apenas na inicialização do serviço de busca. Após o serviço estar disponibilizado ele pode permanecer disponível por tempo indeterminado e incluir novos documentos de forma incremental mantendo ao mesmo tempo o serviço de busca disponível.

Vistos esses números conclui-se que o algoritmo determinístico proposto apresenta vantagens em relação ao uso das técnicas estatísticas. Isso se dá pelo fato de ele levar em consideração a estrutura do documento. Afinal, ele considera o documento como um conjunto de sentenças, em vez de, como um conjunto de palavras.

Portanto, os resultados dos experimentos demonstram que o uso das EM melhora a precisão da busca, tendo em vista que ao combinar o uso de vários descritores no processo de busca faz com que o resultado apresentado como resposta seja a união entre as combinações de vários bigramas concomitantes. No teste empírico realizado verificou-se que com a partir de quinze EM utilizadas como descritores a quantidade de documentos retornados reduziu pela metade. Desse modo, quanto mais bigramas, dentre os extraídos do documento de referência, forem coincidentes com os encontrados nos documentos do *corpus,* maior será o valor apurado no cálculo da relevância. Afinal, a relevância é apurada pelo somatório das relevâncias de cada par de bigramas possibilita produzir um resultado mais restrito, que tenha não apenas um termo similar, mas sim um conjunto de bigramas relacionados.

A técnica BM25, considerada o estado da arte dos SRI, foi a técnica que apresentou o melhor resultado nos teste empíricos realizados, se mostrando como a mais apropriada para ser utilizadas em processos de recuperação de documentos de base de dados de teses, dissertações e artigos, por ponderar a relevância em função do tamanho do documento e por ponderar

o peso do bigrama da consulta em função de sua frequência no documento de referência. Desse modo, os bigramas da consulta têm sua importância relativizada correspondendo à sua relevância no documento da busca, o que não é possível de ser obtida pelas demais técnicas.

Os resultados demonstraram ao comparar a busca tradicional por palavras-chave com o método proposto de busca comparada, o número de documentos retornados tende a ser menor e baseado na construção de significado a partir de vários bigramas coincidentes trazendo uma melhora na precisão das respostas. Desse modo, a técnica elaborada de extração das EM para utilizá-las como descritores em uma ferramenta de busca se mostrou bastante viável podendo de ser implementada de forma integrada com ferramentas de RI em base de dados de documentos digitais.

6 Trabalhos futuros

Ao avaliar os resultados obtidos, verifica-se que eles não apenas atendem aos objetivos propostos, como ao mesmo tempo abrem um caminho para se retormar a motivação deste trabalho, as bibliotecas digitais de teses, dissertações e artigos. O protótipo desenvolvido materializa-se na concepção de uma metodologia automatizada de recuperação da informação que combina técnicas existentes e também trás novas ideias produzindo um conjunto teórico prático, possibilitando a implementação de uma ferramenta de *software* para uso real. Essa ferramenta poderá ser integrada a um processo de busca que considera metadados dos documentos o que aumenta a capacidade seletiva das respostas produzidas. Um outro aspecto a ser observado é que o projeto do *software* foi desenvolvido em camadas e numa arquitetura cliente servidor, o que possibilita uma boa escalabilidade. Desse modo, a ferramenta pode ser facilmente adaptada para realizar pesquisas em múltiplas bases de dados, sendo que cada base de dados considerada como um serviço de busca disponibilizado em um computador provedor de um serviço diferente.

Portanto, este trabalho poderá derivar no futuro em um trabalho aplicado a fim de se operacionalizar o uso dessa ferramenta em um ambiente real de uma biblioteca digital. Isso poderá ser feito através da integração dessa ferramenta proposta com uma aplicação existente, ou mesmo implementar outras funcionalidades a ela, de forma a agregar metadados dos documentos em um banco de dados relacional, a fim de utilizar para a busca a soma desses vários critérios e de forma simultânea. Tudo isso conduz ao encontro a motivação em tornar as ferramentas de busca cada vez mais precisas em seus resultados retornados.

Contudo, o trabalho de pesquisa não para aqui. Novas heurísticas que visam agregar a extração de sentido do texto, ou mesmo melhorar o processo de compressão de dados poderão ser implementadas e validadas possibilitando uma retro-alimentação entre o trabalho teórico e o prático.

Referências

1 CALZOLARI, Nicoletta et al. 2002. Towards best practice for multiword expressions in computational lexicons. Em Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), pp. 1934–1940, Las Palmas, Canary Islands.

2 DIAS, G.; LOPES, J.; GUILLORÉ, S. Mutual expectation: a measure for multiword lexical unit extraction. In Proceedings of Vextal, 1999.

3 EVERT, S. e B. KREEN. 2005. Using small random samples for the manual evaluation of statistical association measures. Computer Speech and Language, 19(4):450–466.

4 MANNING, C.D.; RAGHAVAN, P; SCHÜTZE, H. An introduction to information retrieval. Ed. Cambridge online, 2009.

5 PORTELA, R.; MAMEDE N.; BATISTA, J. Mutiword Identificação. In Terceiro Simpósio de Informáctica (INFORUM 2011), Oct. 2011, pp.

6 RAMISCH, Carlos. *Multiword terminology extraction for domain specific documents.* Dissertação – Mathématiques Appliqueées, École Nationale Supérieure d'Informatiques, Grenoble, 2009.

7 SAG, I. A. et al. Multiword expressions: a pain in the neck for nlp. Em Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing CICLing-2002), volume 2276 of (Lecture Notes in Computer Science), pp. 1–15, London, UK. Springer-Verlag.

8 SARMENTO, Luís. Simpósio Doutoral Linguateca 2006. Disponível em: http://www.linguateca.pt/documentos/SimposioDoutoral2005.html: out. 2011

9 SILVA, F. J. LOPES G. A local maxima method and fair dispersion normalization for extracting multi-word units from corpora. (1999). *Sixth meeting on Mathematics of Language*, pp. 369-381.

10 SILVA, Edson Marchetti; SOUZA, Renato Rocha. *Information retrieval system using multiwords expressions (MWE) as descriptors*. JISTEM - Journal of Information Systems and Technology Management Vol. 9, No. 2, Mai/Aug. 2012, *pp.213-234.*

11 VILLAVICENCIO, Aline et al. Identificação de expressões multipalavra em domínios específicos. Linguamática, v. 2, n. 1, p. 15-33, abril, 2010.

12 ZHANG, W.; et al. Improving effectiveness of mutual information for substantival multiword expression extraction. Expert Systems with Applications, Elsevier, v. 36, 2009.