GT 8: Informação e Tecnologia

AVALIAÇÃO DO DESEMPENHO DE UMA FERRAMENTA AUTOMATIZADA DE BUSCA QUE UTILIZA COMO DESCRITORES AS EXPRESSÕES MULTIPALAVRAS.

Comunicação Oral

Edson Marchetti da Silva - Centro Federal de Educação Tecnológica de Minas Gerais
Renato Rocha Souza - Fundação Getulio Vargas

rsouza.fgv@gmail.com

Avaliação do desempenho de uma ferramenta automatizada de busca que utiliza como descritores as expressões multipalavras.

Resumo

Este trabalho visa testar um método alternativo para recuperação de documentos através do uso de Expressões Multipalavras (EM) extraídas de um documento base, para serem utilizadas como descritores de busca em um Sistema de Recuperação da Informação (SRI). Neste sentido, diferentemente dos métodos que consideram o texto como um conjunto de palavras, do inglês *bag of words*, utilizamos um método que leva em consideração as características da estrutura física do documento. Os bigramas são extraídos pelo uso de uma técnica algorítmica exaustiva para serem utilizados como descritores de busca em um *corpus* visando encontrar documentos similares. Tomando como base um *corpus* com documentos em formato digital foram realizados dois experimentos: o primeiro para avaliar a influência do uso de um ponto de corte no resultado da busca. Um segundo, para avaliar qual a influência que o número de bigramas utilizado, tem no resultado da busca. O resultado se mostrou promissor, ao apresentar respostas mais concisa do que as obtidas pela busca por palavras-chave, com a vantagem de utilizar um processo totalmente automatizado. Além do que mantem a parte cognoscente da escolha do documento, a cargo do usuário.

Palavras-chave: Extração de Expressões Multipalavras, Busca comparada, Sistema de Recuperação de Informação, Modelos Estruturados.

Performance evaluation of an automated search tool that use as descriptors the multiword expression.

Abstract

This paper aims to test an alternative method for retrieving documents through the use of Expressions Multiwords (MWE) extracted from a document base, to be used as descriptors in a search Information Retrieval System (IRS). In this sense, unlike methods that consider the text as a set of words, “bag of words”, we use a method that takes into account the characteristics of the physical structure of the document. The bigrams are extracted by using a exhaustive algorithmic technique to be used as descriptors in a corpus search aimed at finding similar documents. Based on a corpus of documents in digital format two experiments were conducted: the first to assess the influence of using a cutoff in the search result. A second, to assess what influence the number of bigram used, has the search result. The result is promising, presenting answers more concise than those obtained by searching for keywords, with the advantage of using a fully automated process. Furthermore, keeping the part of the choice of document by the user.

Keywords: Extraction of Expressions Multiwords, Compared Search, Information Retrieval System, Structured model.

1 Introdução

A partir da necessidade de organizar grandes volumes de conhecimento, registrados em meios impressos, de tal forma que pudessem ser armazenados e recuperados eficientemente surgiram os primeiros sistemas de recuperação de informação informatizados. Cabe a esses sistemas receber os dados, organizá-los e classificá-los de tal forma que possam ser recuperados e apresentados ao usuário requisitante a fim de suprir a demanda de informação desejada.

Embora as atuais ferramentas de busca, que trabalham com palavras chave, têm uma enorme importância, elas produzem um conjunto de respostas que normalmente é intratável para um ser humano, sem falar na quantidade de resultados totalmente irrelevantes. Portanto, a proposta deste trabalho é apresentar uma alternativa de busca que utiliza um documento de referência em vez de palavras-chave. Ou seja, a busca se dará a partir de um documento que contenha o assunto de interesse do requisitante. E, de forma automatizada a ferramenta proposta irá extrair um sentido desse documento. Esse processo se dará através da extração de palavras que co-ocorrem em uma frequência acima de um limite definido, as Expressões Multipalavras (EM).

Esse tipo de ferramenta, de maneira geral, pode ser acoplada aos sistemas convencionais de busca por palavras-chave, como uma funcionalidade complementar, ou mesmo substituindo parte de suas funcionalidades. O método proposto possui como vantagem o fato de ser independente de idioma e de manter a cargo o usuário a escolha do documento de referência para busca, e partir daí processar a recuperação de forma totalmente automatizada.

2 Trabalhos correlatos

Diversos trabalhos que visam identificar as EM foram publicados, dentre eles destacamos: Dias, Lopes e Guilloré (1999), que visa à extração de EM de forma independente de linguagem e baseado 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 Calzolari et al. (2002), Sag et al. (2002), Ramich (2009), Zhang et al. (2009) e Villavicencio et al. (2010) que aplicam o conceito de EM para tradução automática através do uso de alinhamento lexical das expressões, a fim de verificar de que forma comparar o mesmo texto em idiomas distintos, pode fornecer pistas relevantes para a identificação dessas expressões.

Tomando como base esses trabalhos verificamos a existência de uma lacuna, no que se refere à extração de EM, que leve em consideração as características físicas intrínsecas dos documentos e que seja independente de idioma. É a partir dessa ideia que buscamos obter as EM de um documento, e utilizá-las como descritores da busca comparada para recuperação automatizada de documentos similares.

Para melhor descrever os experimentos este trabalho está estruturado nas seguintes seções nas quais são apresentados os seguintes conteúdos: seção 3 - referencial teórico sobre as EM; seção 4 - metodologia; seção 5 – apresentação e análise dos resultados; seção 6 – conclusões; seção 7 - recomendações para trabalhos futuros.

3 Referencial teórico

Conforme citado por Zhang et al. (2009), a capacidade de expressar sentido de uma palavra depende das demais palavras que a acompanham. Portanto, quando uma palavra ocorre acompanhada por um mesmo conjunto de termos repetidas vezes, 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. Para Sag et al. (2002 p. 2), as EM são: “interpretações idiossincráticas que cruzam os limites (ou espaços) entre as palavras”. Uma outra descrição encontrada na literatura é apresentada a seguir.

> O termo expressão multipalavra vem sendo utilizado para descrever um grande número de construções distintas, mas fortemente relacionadas, tais como verbos de suporte (fazer uma demonstração, dar uma palestra), compostos nominais (quartel general), frases institucionalizadas (pão e manteiga), e muitos outros. [...] EM engloba um grande número de construções, tais como: expressões fixas, compostos nominais e construções verbo-partícula. (VILLAVICENCIO et. al 2010 p. 16)

Segundo Ranchhod (2003, p. 2), as expressões fixas são objetos linguísticos que apresentam divergências terminológicas e a ausência de critérios de análise que as levaram ser consideradas como objetos linguísticos excepcionais, não integráveis na gramática das línguas. Neste trabalho consideramos as EM como sendo formações compostas de duas ou mais palavras que co-ocorrem numa freqüência acima do acaso, que quando associadas possuem uma expressividade semântica mais forte do que quando cada um de seus termos são postos separadamente.

Tem ocorrido um crescente interesse, sobretudo na área de Processamento de Linguagem Natural (PLN), sobre EM afinal essas formas fixas são numerosas em qualquer tipo de texto. Portanto, não podem ser ignoradas. Essas características das EM, as tornam relevantes no tratamento dos recursos lexicais, as quais são importantes insumos informacionais para muitas aplicações relacionadas ao PLN, tais como: tradução automática, sumarização de texto, etc. Nesse sentido, Villavicencio et. al (2010) destaca que muitas pesquisas têm buscado formas de automatização na aquisição lexical. Esses trabalhos buscam entender a formação dos recursos lexicais, uma área ainda carente de pesquisas.

Para Sag (2002 p. 4), as EM podem ser classificadas em:

- Expressões Fixas – são aquelas que não apresentam flexões morfossintáticas e não permitem modificações internas. Elas desafiam as convenções da gramática e interpretação composicional, pois ao tratá-las na forma de palavra por palavra não teríamos a representação da expressão composta, aquela que tem um sentido próprio dado pela composição.
- Expressões Semi-Fixas – são aquelas que possuem restrições na ordem das palavras e composição, mas admitem eventuais variações léxicas na flexão, na forma reflexiva e na escolha de determinantes. Esse tipo de EM é subdividida em três categorias. A primeira categoria, **expressões não-decomponíveis** do inglês *non-decomposable idioms*, ocorre quando se juntam duas ou mais palavras para formar uma expressão que possui um novo significado, distinto daquele obtido pelas palavras de forma isolada. Exemplo: “chutar o balde”, que tem como significado composto a ideia de “desistir”. Nesse caso há variabilidade da expressão idiomática. A segunda categoria, **compostos nominais** do inglês *compound nominals*, são similares às expressões não-decomponíveis, entretanto se caracterizam como unidades sintaticamente inalteráveis que na maioria dos casos podem ser flexionadas em número. Vejamos como exemplo as expressões: “presidente da república” e “deputado federal”. Na primeira expressão, somente presidente pode ser flexionado. Enquanto que, na segunda, ambas as palavras são passíveis de flexão. A terceira categoria, **os nomes próprios** do inglês *proper names*, são sintaticamente altamente idiossincráticos. Vejamos por exemplo: o composto “Espírito Santo” pode estar relacionado ao estado federativo do Brasil ou pode ser um sobrenome.
- Expressões Sintaticamente Flexíveis – são expressões que admitem variações sintáticas na posição de seus componentes. Os tipos de variação possível são: **cons-**

truções verbo-partícula, que consistem de construções de um verbo e uma ou mais partículas, que podem ser semanticamente idiossincráticos ou composicional; **expressões idiomáticas decomponíveis**. Um exemplo é "tirar o cavalinho da chuva". O termo decomponível é utilizado, por que nesse caso, o significado "desistir da ideia" pode ser decomposto em "tirar" (desistir de), "o cavalinho da chuva" (a ideia); **construções verbo-leve**, do inglês *light-verbs*, é um verbo considerado semanticamente fraco estando sujeito à variabilidade sintática completa, incluindo a passivação. Eles são altamente idiossincráticos, pois existe uma notória dificuldade em predizer qual verbo-leve combina com qual substantivo.

- Expressões Institucionalizadas – São expressões composicionais, do inglês *collocation*, que podem variar morfológica ou sintaticamente e que normalmente possuem alta ocorrência estatística.

Segundo Moon (1998 citada por Villavicencio et al.), as EM são unidades léxicas formadas por um amplo contínuo entre os grupos composicionais e os não-composicionais ou idiomáticos. Nesse contexto, expressão composicional são aquelas que a partir das características de seus componentes determinam as características do todo. E, não-composicional ou expressões idiomáticas são aquelas cujo significado conjunto das palavras nada tem haver com o significado individual de cada uma delas. Dada essas características das EM, ao usarmos as palavras-chave de modo independente, como é feito em geral nas ferramentas de busca, certamente ira trazer anomalias para o processo de RI.

Dentre as diferentes abordagens do PLN que lidam com EM destacamos: aquelas que utilizam os métodos simbólicos Calzolari et al. (2002); e as que usam uma abordagem estatística Evert e Krenn (2005). Ambas buscam interpretar os conteúdos textuais escritos em linguagem natural, mas seguem caminhos diversos obtendo resultados de custo computacional[1] e de conteúdos diferentes. Dessa maneira, as vantagens e desvantagens de cada um desses métodos depende do contexto para o qual estão sendo utilizados. A abordagem simbólica visa interpretar o texto buscando o sentido sintático, morfológico e pragmático baseando-se em um dicionário controlado de palavras e em um conjunto de regras. Nesse caso, o processamento é fortemente dependente do idioma e do domínio do *corpus*. Enquanto que, a abordagem estatística procura dar um tratamento ao texto através do reconhecimento de padrões de comportamento baseados na frequência de co-ocorrência das

[1] Custo computacional neste contexto se relaciona ao consumo de recursos computacionais de processamento demandados numa relação direta com o tempo de resposta.

palavras. Ou seja, as EM são um conjunto de palavras que co-ocorrem numa frequência acima do acaso.

Calzolari et al. (2002 p. 1934) corrobora com a classificação apresentada por Sag (2002) e ainda inclui um "etc" no final. Ou seja, como os próprios autores definem EM é utilizada para descrever diferentes, mas relacionados fenômenos, que podem ser descritos como uma sequência de palavras que agem como uma unidade em algum nível de análise linguístico e que apresentam alguns ou todos dos seguintes comportamentos: reduzida transparência sintática e semântica; redução ou ausência de composicionalidade; mais ou menos estável; passível de violação de alguma regra geral sintática; elevado grau de lexicalização (dependendo de fatores pragmáticos); alto grau de convencionalidade. Ainda segundo esses mesmos autores, as EM estão situadas na interface entre a gramática e o léxico. Eles apresentam também algumas das causas das dificuldades ocorridas no âmbito teórico e computacional para o tratamento das EM, como sendo: a dificuldade de estabelecer limites claros para o domínio das EM; a falta de léxicos computacionais de tamanho razoável para auxiliar no PLN; a dificuldade, em muitas da vezes, perante a perspectiva multilingue não ser possível encontrar uma correspondência direta lexical equivalente; dificuldade generalização dos léxicos (geral e terminológico) para um contexto específico.

O trabalho de Calzolari et al. (2002) utiliza uma abordagem focada nas EM que são produtivas por um lado e que por outro demonstram regularidades que possam ser generalizadas para as classes de palavras com propriedades semelhantes. Particularmente, eles buscam encontrar dispositivos gramaticais que permitam a identificação de novas EM motivados pelo desejo do reconhecimento o mais automatizado possível desse processo. Nesse sentido, a pesquisa desses autores estudou em profundidade dois tipos de EM: os verbos de suporte e os substantivos compostos (ou complexos nominais). Pois segundo eles, esses dois tipos de EM estão no centro do espectro de variação em composicionalidade que pode ser observado pela coesão interna juntamente com um elevado grau de variabilidade em lexicalização e variação dependente do idioma.

Já a abordagem utilizada por Evert e Krenn (2005) é baseada no cálculo estatístico das medidas de associação das palavras contidas no texto.

A pesquisa conduzida por Villavicencio et al. (2010) busca extrair as EM combinando duas abordagens distintas: a abordagem associativa e a abordagem baseada em alinhamento lexical[2]. Na primeira, as medidas de associação são aplicadas para todos os bigramas e tri-

[2] Dois textos escritos em idiomas distintos são considerados como alinhados quando eles possuírem marcas que identifiquem os pontos de correspondência entre o texto original e a sua tradução.

gramas gerados a partir do *corpus* e o resultado dessas medidas são utilizados para avaliação. A segunda abordagem extrai de forma automatizada as EM tomando como base os alinhamentos lexicais das versões de um mesmo conteúdo escrito nos idiomas português e inglês. Para combinar os resultados obtidos pelas duas abordagens os autores utilizaram as redes bayesianas.

A abordagem estatística para extração de EM, através da co-ocorrência de palavras em textos, tem sido empregada em diversos trabalhos dentre os quais destacamos: Pearce (2002); Evert e Krenn (2005); Pecina (2006); Ramisch (2009) e Villavicencio et al. (2010).

A abordagem de alinhamento lexical verifica se a EM encontrada em um documento escrito em um idioma também ocorre na versão correspondente escrita em outro idioma. Para ser possível essa análise os documentos necessitam estar alinhados através da correspondência das palavras entre as diferentes versões expressas em idiomas distintos. Entretanto, para que o alinhamento seja possível é necessário que os documentos sejam analisados a partir de seus aspectos morfológicos tratados através de um pré-processamento de etiquetação[3]. Desse modo, as classes gramaticais são utilizadas como informação adicional no processo de identificação das EM.

A meta é obter a semântica do documento representado pelas EM e utilizá-las como descritores do processo de busca.

4 Metodologia

A proposta deste trabalho é de utilizar a ferramenta de busca comparada descrita e implementada no trabalho anterior apresentado por Silva e Souza (2012). Essa ferramenta utiliza uma técnica denominada Heudet, proposta pelos autores, que considera algumas características estruturais do documento para identificação de EM. Nesse mesmo trabalho citado foram comparados os resultados obtidos pela extração das EM através do uso da técnica Heudet com o da utilização de treze medidas de associação estatísticas produzidas pelo pacote de *software* Ngram Statistics Package (NSP) proposto por Pedersen et al. (2011). Como resultado, Heudet apresentou melhor precisão e desempenho. O ganho medido empiricamente foi de 1,5% na precisão das EM, através da redução de ruídos, além de uma redução aproximada de 35% no tempo de processamento consumido se comparado com as técnicas estatísticas.

Cabe ainda destacar, que a técnica Heudet leva em consideração às características da estrutura física dos documentos, tais como:

[3] Programa de computador conhecido genericamente como etiquetador de categorias gramaticais. Gera uma saída, normalmente em XML, associando cada palavra à sua classe gramatical: substantivo, verbo, artigo, etc.

- se as palavras que compõem a EM pertencem a uma mesma sentença;
- elimina conteúdos repetitivos do texto tais como cabeçalhos, etc;
- ponderá a relevância das palavras escritas com letras maiúsculas;
- identifica siglas e as transforma em texto.

A integração entre a ferramenta de busca comparada proposta, com um SRI que executa busca por palavras-chave em bibliotecas digitais ou quaisquer outros sistemas de busca em grandes bases de documentos, pode ser feito da seguinte forma. A primeira recebe o documento de referência, identifica as EM e as submete ao sistema pré-existente como uma sucessão de buscas por palavras chave. O resultado obtido deverá ser tratado de tal forma a se criar um ranqueamento e um filtro para apresentar apenas os documentos mais relevantes. É importante observar que o método de indexação dos documentos implementado pelo sistema legado deverá ser capaz de retornar como resposta a localização dos termos nos documentos nos quais eles forem encontrados. Ou seja, utilizar o método *positional index*, descrito por Manning, Raghavan & Schütze (2009, p. 41-43), no processo de indexação dos termos. Isso é necessário para que o processamento da resposta possa avaliar a distância entre os termos dos bigramas. Caso isso não seja possível, duas soluções poderão ser utilizadas: substituir totalmente o modelo de indexação ou implantar a nova indexação em um ambiente separado.

A figura 1 apresenta um diagrama da estrutura de integração do *software* proposto, onde se apresenta o módulo de busca comparada adicional, em destaque, e a sua interface com os sistemas convencionais de busca por palavras-chave.

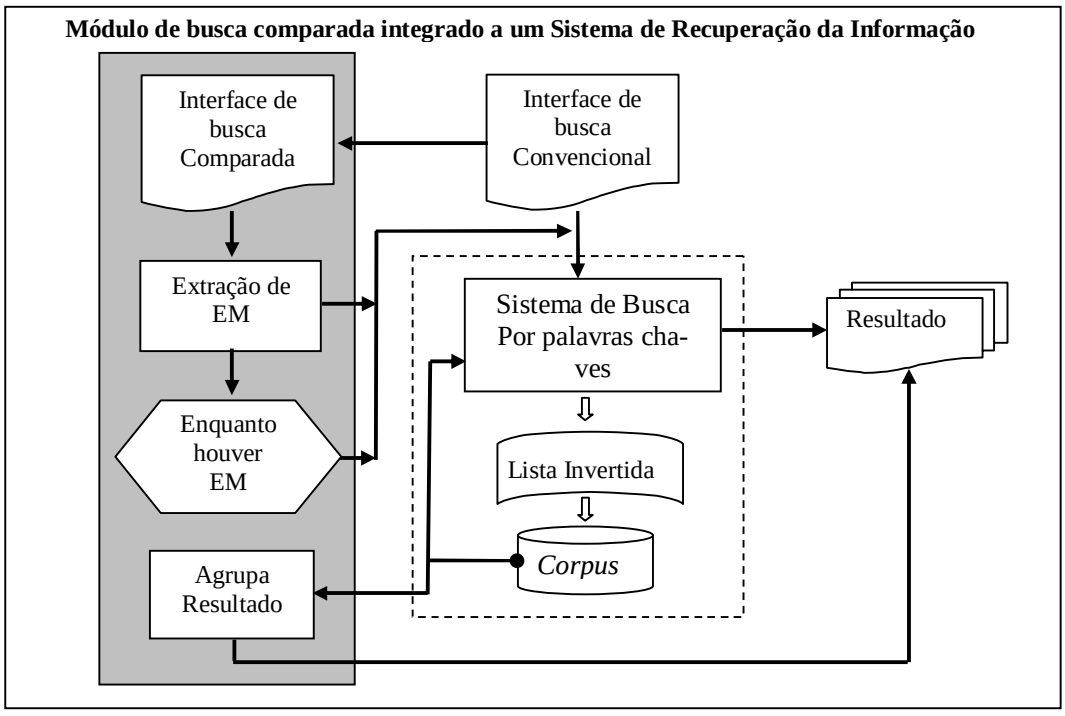

FIGURA 1 – Módulo de busca comparada agregado ao SRI.
Fonte: Elaborada pelos autores.

Para realizar este experimento foi utilizada a ferramenta proposta materializadas através de dois componentes de *software* principais elaborados em C++: 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. Esse componente representa um SRI que busca por palavras-chave. O Client é o responsável por receber o documento de referência para a busca, extrair dele as EM, enviar a requisição de consulta e retornar a resposta com os documentos similares.

Para a aplicação funcionar foi elaborada uma camada de apresentação, implementada através de uma página Web, que serve como interface do usuário final processar a busca comparada dos documentos. Essa página acessa um componente de *software* elaborado na linguagem PHP denominado "Busca". Essa interface se encarrega de receber o documento, fazer o *upload* do mesmo e executar uma requisição ao Client passando o documento como parâmetro do processamento. O documento PDF é convertido em bigramas normalizados e as requisições são passadas ao Server. A figura 2 apresenta um esboço da tela dessa interface.

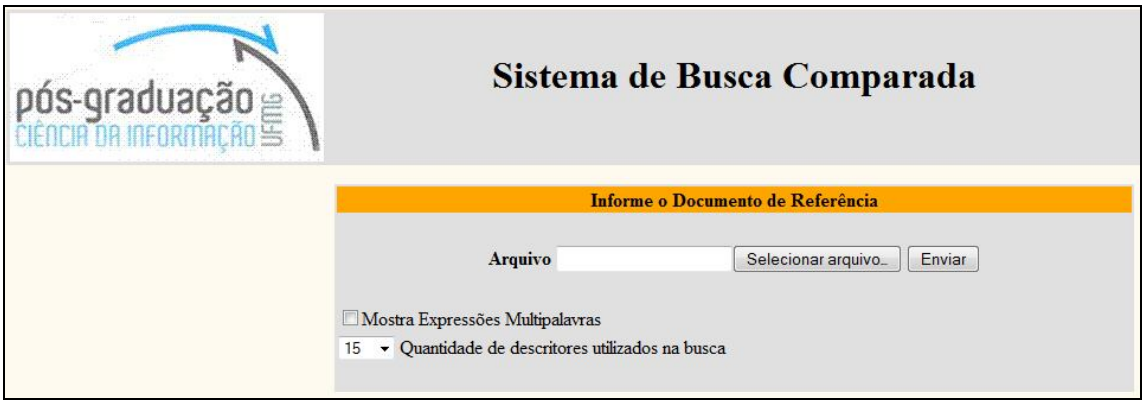

FIGURA 2 – Tela da ferramenta, onde é informado o documento utilizado na busca comparada.
Fonte: Elaborada pelos autores.

A proposta dessa metodologia é validar a ferramenta implementada, a fim de avaliar seu desempenho e refinar o parâmetro que define a quantidade de EM utilizadas no *software* identificando de que forma podem contribuir para a precisão das respostas da busca comparada.

A figura 3 mostra como que o resultado da busca é apresentado na tela do usuário da consulta.

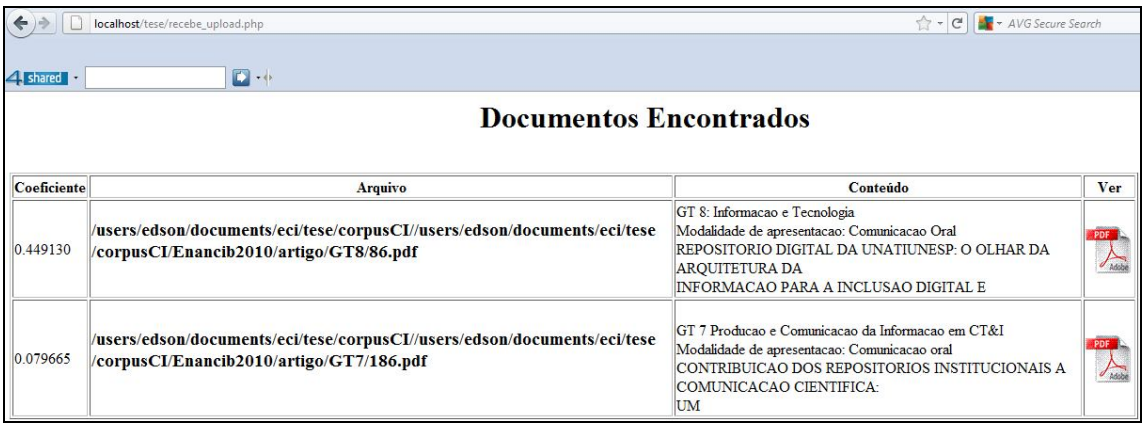

FIGURA 3 –Tela de respostas com os documentos encontrados
Fonte: elaborada pelos autores

5 Apresentação e análise dos Resultados

Para realizar os testes empíricos foi utilizado um *corpus* composto pelos artigos completos publicados no principal encontro científico da área da Ciência da Informação (ENANCIB) do ano 2010. Todos os documentos foram obtidos em formato *Portable Document Format* (PDF) e armazenados em um sistema de arquivos informatizado organizado em pastas e subpastas de forma hierarquizada pelos grupos de trabalho (GT). No total, o *corpus* utilizado possui 194 artigos, contendo tipicamente entre 20 a 25 páginas, totalizando 687.490 termos normalizados, sendo 7970 distintos.

Para atender aos objetivos deste trabalho o foco passou a ser a validação do uso da ferramenta, medido através de experimentos controlados. Esses experimentos são descritos a seguir.

5.1 Primeiro experimento exploratório

Primeiramente, para avaliar o uso desse aplicativo foram realizadas buscas no *corpus* utilizando como referência vinte documentos aleatórios, dois para cada GT. O cálculo de similaridade, o qual utiliza a técnica *Cosine Similarity Vector*, realizado pelo aplicativo é apurado pelo somatório dos coeficientes de correlação apurados para cada um dos bigramas extraídos do documento de referência e confrontados com *corpus*. De modo que, desde que haja pelo menos um bigrama coincidente, entre os extraídos do documento de referência com o *corpus*, ele já passa a ser considerado com parte da resposta. Portanto podem haver muitos documentos calculados com valores de coeficiente residual. Ou seja, valores bem pequenos. Em sendo assim, é conveniente que o 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 a cada busca. 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 o ponto em que apenas um único documento é retornado. Ou seja, apenas o mais similar.

O Gráfico 1 foi gerado utilizando 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, representado pela curva rosa. Dessa forma, cada coordenada mostrada no gráfico relaciona o valor obtido das duas 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 gráfico 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 nos 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*.

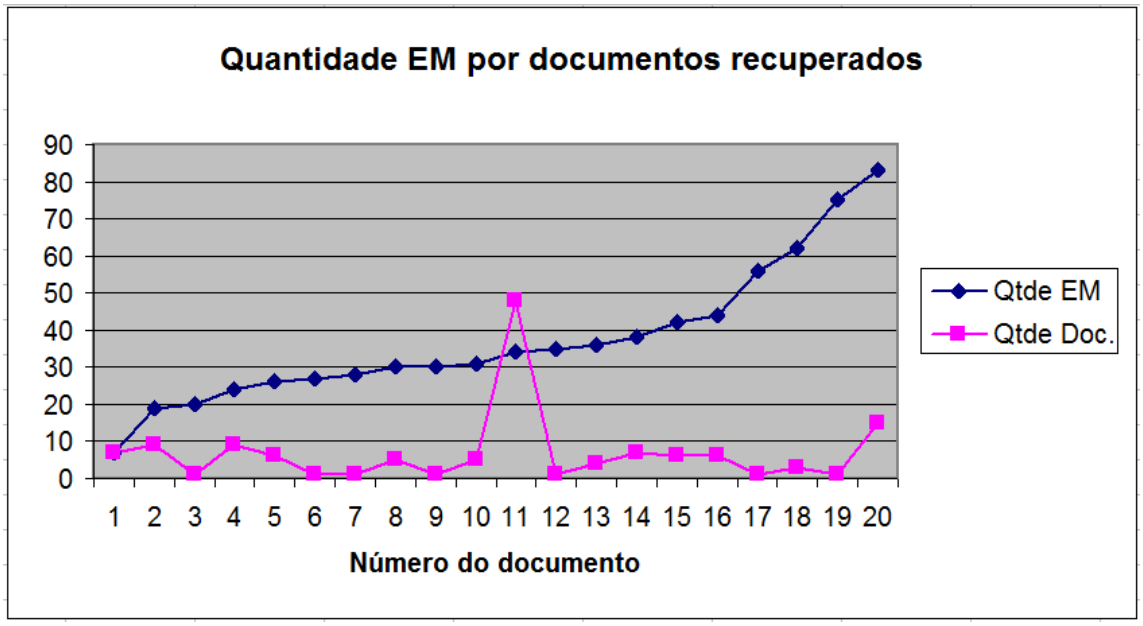

Gráfico 1 – Quantidade de EM extraídas *versus* quantidade de documentos recuperados.
Fonte: Elaborado pelos autores

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 por 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: Elaborado 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, nesse experimento o valor utilizado foi quatro;
- O número de EM extraídas, quanto maior esse número, maiores são as chances de se encontrar 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 encontrar similares;
- Os critérios adotados no cálculo do coeficiente de similaridade, que interferem no cálculo da relevância.

5.2 Segundo experimento exploratório

Visando avaliar outros aspectos do comportamento desse aplicativo resolveu-se 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 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.

Desse modo, 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 que serão considerados no processo de busca. Dessa maneira, a cada documento de referência processado, os seus bigramas são extraídos e inseridos em uma estrutura de dados em forma de 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 fo-

ram realizadas 194 buscas para cada valor de "*n*" arbitrado podendo assim calcular a quantidade de documentos recuperados e a partir desses apurar o valor médio de documentos recuperados para cada valor de *n.* A tabela 3 mostra os valores calculados para os vários valores de "*n"* 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. 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 referencia 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. Para cada valor de *n* é calculado o valor médio da quantidade de documentos retornados. Para realizar esse experimento foram processadas 3.298 consultas de documentos no *corpus.*

Tabela 3 – comparação da quantidade de descritores *versus* documentos retornados.

Sequência	Limite *n* de bigramas usados na busca	Média de documentos retornados pela busca
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: Elaborado pelos autores

Ao analisar os dados apresentados na tabela 3 verifica-se que ao usar apenas um bigrama, o que na prática funciona com um busca convencional por palavras-chave, são retornados em média 31,78 documentos. Na medida em que aumentamos 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 nos 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 do resultado. Essa estratégia melhora o desempenho da busca.

6 Conclusões

Os resultados dos experimentos demonstram que o uso da técnica Heudet 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. 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.

Portanto, na comparação da busca por palavras-chave com o uso do 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 da resposta.

7 Recomendações para trabalhos futuros

O resultado obtido pela técnica Heudet se mostrou promissor, pois apresentou melhores respostas que os obtidos por busca por palavras-chave e com a vantagem de utilizar um processo totalmente automatizado. A parte cognoscente da busca que fica a cargo do usuário da ferramenta se dá através de uma boa escolha do documento de referência. Como a ferramenta considera a estrutura física do documento durante o processo de extração de EM os resultados podem ainda ser melhorados a partir da criação de novas heurísticas que visam identificar características inerentes da estrutura física do documento, e que possam ajudar na identificação de EM. Tais, avaliações da estrutura textual poderá melhorar ainda mais os resultados obtidos. Recomenda-se também a realização de mais testes utilizando uma base de documentos maior.

Referências

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

DIAS, Gaël; LOPES, José Gabriel Pereira; GUILLORÉ, Sylvie. Mutual expectation: a measure for multiword lexical unit extraction. In proceedings of Vextal, 1999.

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

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich An introduction to information retrieval. Ed. Cambridge online, 2009.

PEARCE, Darren. A comparative evaluation of collocation extraction techniques. In proceedings of the Third (LREC 2002), Las Palmas, Canary Islands, Spain, May, 2002.

PECINA, Pavel; SCHLESINGER, Pavel. Combining Association Measures for Collocation Extraction. In ACL'06, page 652, 2006.

PEDERSEN, Ted et. Al. *The Ngram Statistics Package*. Disponível em: http://www.d.umn.edu/~tpederse/nsp.html. Acesso em: ago. 2011.

PORTELA, Ricardo José Rosa; MAMEDE Nuno; BATISTA, Jorge. Mutiword Identificação. In Terceiro Simpósio de Informática (INFORUM 2011), Oct. 2011, pp.

RANCHOLD, Elisabete M. O lugar das expressões 'fixas' na gramática do Português. In Castro, I. and I. Duarte (eds.), Razão e Emoção, vol. II, Lisbon: INCM, pp. 239-254, 2003.

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

SAG, Ivan. A. et al. Multiword expressions: a pain in the neck for nlp. In 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.

SILVA, Ferreira. J. LOPES Pereira 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.

SILVA, Edson Marchetti; SOUZA, Renato Rocha. Sistema de recuperação da informação por busca comparada, que utiliza como descritores expressões multipalavras obtidas através de uma técnica que avalia a estrutura do documento. In proceedings of the 9rd International Conference on Information Systems and Technology Management (Contecsi 2012), São Paulo.

VILLAVICENCIO, Aline et al. Identificação de expressões multipalavra em domínios

específicos. Linguamática, v. 2, n. 1, p. 15-33, abril, 2010.

ZHANG, Wen et al. Improving effectiveness of mutual information for substantival multiword expression extraction. Expert Systems with Applications, Elsivier, v. 36, 2009.