Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Expansão Semântica de Consultas com o Auxílio de Georreferenciamento de Termos

Vinícius de Araújo Porto

Orientadores:

Cláudio de Souza Baptista

Leandro Balby Marinho

Campina Grande – PB

2012

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Expansão Semântica de Consultas com o Auxílio de Georreferenciamento de Termos

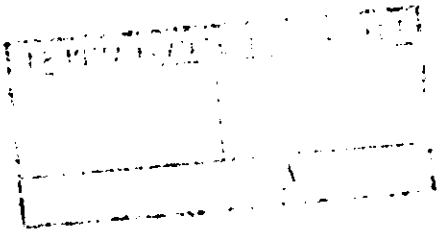
Vinícius de Araújo Porto

Dissertação submetida à Coordenação do Curso de pós-graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciência da Computação.

Orientadores: Cláudio de Souza Baptista e Leandro Balby Marinho

ÁREA DE CONCENTRAÇÃO: CIÊNCIA DA COMPUTAÇÃO
LINHA DE PESQUISA: METODOLOGIA DE TÉCNICAS DE COMPUTAÇÃO

Campina Grande – PB
2012



P853e Porto, Vinícius de Araújo.
Expansão semântica de consultas com o auxílio de georreferenciamento de termos / Vinícius de Araújo Porto. - Campina Grande, 2012.
85 f.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2012.
"Orientação : Prof. Ph.D Cláudio de Souza Baptista, Prof. Dr. Leandro Balby Marinho".
Referências.

1. Recuperação da Informação. 2. Sistemas de Recuperação da Informação. 3. Expansão de Consultas. 4. Geoprocessamento. 5. Contexto Geográfico. 6. Tesauro. 7. Dissertação - Ciência da Computação. I. Baptista, Cláudio de Souza. II. Marinho, Leandro Balby. III. Universidade Federal de Campina Grande - Campina Grande (PB). IV. Título

CDU 004.414.28(043)

"EXPANSÃO SEMÂNTICA DE CONSULTAS COM O AUXÍLIO DE GEORREFERENCIAMENTO DE TERMOS"

VINÍCIUS DE ARAÚJO PORTO

DISSERTAÇÃO APROVADA EM 10/09/2012

CLÁUDIO DE SOUZA BAPTISTA, Ph.D, UFCG
Orientador(a)

LEANDRO BALBY MARINHO, Dr., UFCG
Orientador(a)

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Examinador(a)

DAMIRES YLUSKA DE SOUZA FERNANDES, D.Sc, IFPB
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Diante do acelerado crescimento do volume de informações disponíveis na Web, conseguir recuperar, de maneira simples e eficiente, documentos que atendam às necessidades cada vez mais específicas dos usuários é o grande desafio no campo da Recuperação de Informação (RI). Além da grande quantidade de informação, outro aspecto que interfere nesse processo de aquisição, por parte do usuário, é o fato de essa necessidade de informação ser expressa através de consultas com poucas palavras-chaves. Nesse cenário, técnicas de expansão de consulta são aplicadas a fim de acrescentar novos termos à consulta original, com o objetivo de enriquecê-la semanticamente. Algumas dessas técnicas de expansão têm sido aplicadas em termos relativos a localidades geográficas, uma vez que o contexto geográfico tem papel importante na identificação da real necessidade do usuário. Todavia, essas técnicas se restringem a termos que representam uma localização geográfica e deixam de considerar o fato de que o vocabulário utilizado pelos usuários varia de acordo com a sua região geográfica. Portanto, se o usuário utilizar um termo de seu conhecimento, ou seja, da sua região, para realizar uma busca por documentos de outra região, e esse termo não for muito utilizado naquela região, o resultado da sua consulta será comprometido. Logo, este trabalho apresenta um sistema de recuperação da informação (SRI), que utiliza o contexto geográfico dos termos para expandir semanticamente as consultas dos usuários.

Palavras-chave:

Recuperação da informação, sistema de recuperação da informação, expansão de consultas, geoprocessamento, contexto geográfico, tesauro.

Abstract

Toward the quickly growing of information available on the web, to recover, in a simple and efficient way, documents that meet the increasingly specific needs of the users is the major challenge in Information Retrieval (IR). Besides the large amount of information, another issue that interferes on process of acquisition by the user is the fact, that information need be expressed with a few keywords. In this scenario, query expansion techniques are applied in order to add new terms to the original query in order to enrich it semantically. Most of these expansion techniques have been applied to geographic locations terms, because the geographical context plays an important role in identifying the real user needs. However, these techniques are restricted to terms that represent a geographic location and fail to consider that the vocabulary used by is different according to geographical region. So, if the user uses a term of his knowledge, in other words, from its region, to do a search for documents from other region, and this term isn't widely used in this region, the result of your query will be compromised. Therefore, this paper presents an information retrieval system, which uses the geographical context of the terms to semantically expand users' queries.

Keywords:

Information retrieval, information retrieval system, query expansion, geo-processing, geographical context, thesauri.

Agradecimentos

Primeiramente, agradeço a Deus, que me deu saúde e discernimento para chegar até aqui. A toda a minha família, em especial, a minha mãe, Rosa, pelo amor e apoio durante essa caminhada; ao meu pai, Carlos Ronaldo, por seu exemplo de dedicação e honestidade na sua profissão, qualidades que eu tento seguir todos os dias; a minha irmã, Vanessa, pelo apoio e pelo incentivo nessa conquista;

A minha avó, Júlia, por todas as orações intercedidas ao meu favor; aos meus avós paternos, Jorge e Carmelita, pelo carinho e por compreenderem minhas ausências aos domingos, nos almoços da família;

A minha noiva, Michelly, por todo amor, carinho, motivação e compreensão, em todos os momentos, inclusive nos que não pude dar a atenção que ela merecia durante a execução deste trabalho;

Aos Professores orientadores, Cláudio de Souza Baptista e Leandro Balby Marinho, pela paciência e dedicação durante a orientação deste trabalho;

Aos companheiros de laboratório: Hugo, Daniel, Ana Gabrielle, e aos parceiros do Icuriã: Fábio, Dimas, Jaindson, Damião, Tiago Leite, Tiago Brasileiro, Amilton, Tiago Eduardo e Luan;

Aos colegas de Mestrado, Odilon e Maxwell, pelas experiências compartilhadas durante o curso, e aos amigos, Paulo de Tarso e Eduardo que, desde a graduação, estiveram ao meu lado, me ajudando no que era preciso;

A todos os que compõem a COPIN, em especial, a Aninha, Vera e Rebeka, por sempre atenderem aos meus pedidos;

À CAPES, pelo apoio financeiro.

Lista de Símbolos

API - *Application Programming Interface*

DF – *Document Frequency*

HTML - *HyperText Markup Language*

IDF – *Inverse Document Frequency*

IP – *Internet Protocol*

JSON - *JavaScript Object Notation*

REST - *REpresentational State Transfer*

RI – *Recuperação da Informação*

RIG – *Recuperação da Informação Geográfica*

SRI – *Sistema de Recuperação da Informação*

TGN - *Thesaurus of Geographic Names*

XLDB - *EXtremely Large DataBases*

XML - *eXtensible Markup Language*

XSLT - *eXtensible Stylesheet Language for Transformation*

Lista de Figuras

Lista de Tabelas

Lista de Códigos Fonte

Conteúdo

Capítulo 1

Introdução

Com o aumento da disponibilidade de documentos *online*, conseguir encontrar informações que atendam a necessidades específicas é uma tarefa cada vez mais necessária entre os usuários. Diante disso, a área de Recuperação da Informação (RI) tem recebido cada vez mais atenção.

A maioria dos sistemas de recuperação de informação (SRI) tem uma *interface* padrão, que consiste de uma caixa de texto, na qual o usuário submete palavras-chave com o objetivo de encontrar documentos que as contenham. A construção dessa consulta é fundamental para o sucesso do processo de recuperação da informação. Nesse cenário, diversos problemas semânticos podem existir, como a sinonímia, em que palavras com grafia diferente têm o mesmo significado. Por exemplo, as palavras "*tv*" e "*televisão*" têm o mesmo significado, porém, caso o usuário escolha a palavra "*tv*" e realize uma busca em um SRI tradicional, documentos que contenham apenas o termo "*televisão*" serão desconsiderados pelo sistema. Outro problema existente diz respeito à polissemia, em que uma mesma palavra tem significados diferentes, dependendo do contexto em que está empregada. O vocábulo "*Java*", por exemplo, pode se referir à Ilha de Java, localizada na Indonésia, ou à linguagem de programação. Logo, quando essa palavra for submetida a um SRI, tanto resultados relacionados com a linguagem de programação quanto com a ilha asiática serão recuperados, mesmo que o usuário esteja interessado em uma semântica específica. Além desses problemas semânticos, também são frequentes os sintáticos, como erros de grafia, flexões de gênero e de número, entre

outros, que são frequentes em sistemas de recuperação da informação.

Uma solução muito utilizada para esse problema é a expansão de consultas que, basicamente, consiste na agregação de termos relacionados à consulta original do usuário, com o objetivo de melhorar os resultados obtidos (MANNING, RAGHAVAN, SCHÜTZE, 2008). Essa expansão pode minimizar tanto os problemas de sintaxe (erros de grafia, emprego do plural etc.) quanto os aspectos semânticos. Uma forma de expansão que contempla os aspectos semânticos é através de tesauros (IMRAN; SHARAN, 2009) ou de ontologias de domínio (BHOGAL *ET AL.*, 2007). Nesses casos, as palavras informadas pelo usuário são avaliadas de modo que termos relacionados (sinônimos, por exemplo) são descobertos e adicionados à consulta original, para que resultados mais precisos sejam obtidos.

Além da expansão de consultas, uma alternativa que vem melhorando os resultados dos sistemas de busca é o uso de informações de contexto, porquanto podem indicar a real intenção do usuário ao fazer uso de termos específicos em sua busca (XIANG *ET AL.*, 2010). O contexto é qualquer informação que pode ser utilizada para caracterizar algo que seja relevante para a interação entre o usuário e a aplicação (DEY, 2001). Por exemplo, o contexto geográfico pode ser fundamental em muitas buscas efetuadas pelos usuários em um SRI, pois os resultados considerados mais interessantes são aqueles relacionados à região dos próprios usuários. Nesse cenário, um usuário que deseja comprar um medicamento, ao realizar uma busca por farmácias, em geral, quanto mais próxima a farmácia estiver do usuário, mais relevante será esse resultado.

Outro aspecto importante do contexto geográfico é a influência dele no vocabulário utilizado pelos usuários ao realizarem uma busca nos SRI. Por exemplo, pessoas que residem na Região Nordeste empregam palavras que não são usadas por quem mora no Sul ou no Sudeste do país, onde a "macaxeira" também é conhecida como "mandioca" e "aipim". Porém cada um desses termos é mais utilizado em algumas regiões do país. Dessa forma, um usuário que reside no Nordeste brasileiro e que pretende encontrar receitas da Região Sul que tenham macaxeira, ao realizar uma busca de receitas em um SRI tradicional, pode ter como resultado um número restrito de documentos ou não encontrar nenhum resultado relevante. Isso ocorre porque, geralmente, o usuário não sabe que aquele termo é conhecido apenas em sua região.

Portanto, foram construídos alguns casos de teste com esse cenário para avaliar a abordagem proposta em comparação com um SRI tradicional e um SRI com uma técnica de expansão de consulta baseada em um tesauro comum.

1.1 Objetivos

Nesta seção é apresentado o objetivo geral do trabalho e as etapas necessárias (objetivos específicos) para a sua concretização.

1.1.1. Objetivo Geral

Construir um sistema de recuperação da informação (SRI) que tenha uma técnica de expansão semântica de consultas a partir do contexto geográfico.

1.1.2. Objetivos Específicos

- **Utilizar um tesauro como fonte de conhecimento (*background knowledge*) para a técnica de expansão de consulta**. Para auxiliar na inferência semântica dos termos de uma consulta, será utilizado um tesauro para expandi-los, a fim de que documentos com termos relacionados possam ser sugeridos ao usuário;

- **Georreferenciar os termos do tesauro**. Como parte da técnica de expansão de consulta, os termos do tesauro devem ser georreferenciados. Logo, uma técnica automática de georreferenciamento dos termos do tesauro é proposta para que o contexto geográfico dos termos seja utilizado como critério para filtrar os termos acrescidos à consulta;

- **Usar o contexto geográfico para expandir os termos da consulta**. A partir do tesauro georreferenciado, uma consulta que apresente termos relativos a uma localização geográfica terá os demais termos expandidos semanticamente por meio desse contexto geográfico;

- **Elaborar o protótipo de um sistema de recuperação da informação.**

Esse sistema é proposto para a utilização da técnica de expansão semântica de consultas com o auxílio de geoprocessamento;

1.2 Relevância

Atualmente, o contexto geográfico tem sido muito utilizado como estratégia para tornar mais eficientes os sistemas de recuperação da informação (CARDOSO; SILVA, 2007), (CAMPELO, 2008), (OVERELL, 2009), (FERNANDES, 2010). Muitos sistemas têm utilizado técnicas de expansão de consultas e o contexto geográfico a fim de melhorar a recuperação de documentos, porém, essas técnicas de expansão se restringem a termos que representam uma localização geográfica. Por exemplo, em uma consulta por "recôncavo baiano", as técnicas expandem o termo, visto que ele se refere a uma localidade geográfica, ou seja, a consulta é expandida com outros termos como: "*Salvador*", "*Madre Deus*", "*Santo Amaro*", "*Maragojipe*", cidades que fazem parte do recôncavo baiano.

Existem diversos cenários onde o vocabulário utilizado varia de acordo com a região geográfica do usuário. Determinado termo é mais utilizado e conhecido em uma região do que em outra. Logo, se o usuário utilizar um termo de seu conhecimento para realizar uma busca por documentos de outra região, e esse termo não for empregado naquela região, o resultado da sua consulta será comprometido. No cenário automobilístico, determinados carros têm nomes distintos em diferentes países. O fusca, por exemplo, é conhecido como bolha, na Finlândia, cucaracha, em alguns países da América Latina, sapo, na Romênia, entre outros. Portanto, se um usuário da Finlândia tivesse interesse de saber sobre esse carro na Romênia e submetesse a consulta pelo carro "bolha", em um SRI tradicional, na busca por documentos romenos relacionados com automobilismo, provavelmente nenhum resultado relevante seria encontrado, uma vez que esse carro é conhecido como sapo naquela região.

Diante disso, a relevância deste trabalho está na criação de um sistema de recuperação da informação onde haja uma expansão semântica de termos cuja grafia é diferente de acordo com o contexto geográfico. Para esse tipo de problema, a abordagem proposta obteve uma precisão média de 60,25%, nos casos de teste

avaliados, enquanto as abordagens comparadas obtiveram uma precisão média de 43,77% e 26,71%.

1.3 Estrutura da Dissertação

Esta dissertação é composta por cinco capítulos. O Capítulo 2 traz uma abordagem sobre os pressupostos teóricos que embasam esta pesquisa, incluindo os conceitos fundamentais da recuperação da informação; no Capítulo 3, são feitas considerações acerca das principais contribuições deste trabalho e dos principais trabalhos relacionados com o tema; no capítulo 4, apresentam-se a abordagem proposta neste trabalho e o protótipo construído para validá-la; no capítulo 5, é feito um estudo de caso desenvolvido para validar o protótipo proposto. Por fim, são apresentadas as conclusões e sugestões de possíveis trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Com a popularização da Internet, o número de documentos disponíveis cresceu rapidamente. Diante disso, conseguir recuperar e organizar esse grande volume de informação tornou-se uma atividade complexa, que levou o usuário a ter dificuldades de encontrar o que precisa utilizando sistemas de recuperação da informação. O principal problema é a falta de uma organização semântica de grande parte dessa informação. Nesse contexto, cresce o número de pesquisas na área de recuperação da informação (RI), com o objetivo de facilitar essa tarefa de busca de informação relevante, por parte do usuário.

O uso do contexto geográfico e de técnicas de expansão de consultas, nos sistemas de recuperação da informação, tem proporcionado resultados positivos aos usuários, uma vez que essas abordagens se apresentam como uma forma de organização semântica do grande volume de informações existentes nos sistemas.

Neste capítulo, são feitas considerações acerca dos seguintes aspectos: principais conceitos da área de Recuperação da Informação; principais métodos de expansão de consultas; uso de tesauros como forma de representação do conhecimento; apresentação dos conceitos fundamentais da ciência de contexto e um estudo sobre sistemas de recuperação de informação geográfica.

2.1 Recuperação da Informação

Baeza-Yates e Ribeiro-Neto (1999) concebem que a recuperação da informação (IR) trabalha com a representação, o armazenamento, a organização e o acesso à informação. A representação e a organização devem fornecer ao usuário fácil acesso à informação em que ele está interessado, embora a caracterização desse interesse, geralmente, não seja um problema simples. De acordo com Manning, Raghavan e Schtze (2008), recuperar a informação é uma atividade que consiste em encontrar, em grandes coleções de dados, documentos sem uma estrutura definida (geralmente, em forma de texto), com o objetivo de satisfazer a uma necessidade.

Atualmente, com o aumento do volume de informação existente na Web e a crescente complexidade dos objetos armazenados, exigem-se processos de recuperação cada vez mais elaborados. Diante disso, a RI apresenta, a cada dia, novos desafios e se configura como uma área de grande significância, no campo da Ciência da Computação.

Em um sistema de RI, espera-se que os resultados para uma determinada consulta do usuário apresentem o maior número possível de itens relevantes. Diante disso, várias técnicas foram desenvolvidas com o propósito de aumentar o grau de relevância dos resultados de uma consulta. Em geral, procura-se construir um índice de similaridade entre o conjunto de identificadores dos documentos e o conjunto de termos da consulta. Com base nessa similaridade, um *ranking* de documentos pode ser recuperado e apresentado de acordo com uma consulta solicitada.

Além do grau de similaridade, outras técnicas para se construir o ranking de relevância são adotadas, como por exemplo, a estrutura de ligações (links) entre as páginas da Web (BRIN;PAGE, 1998), muito utilizada nos motores de busca na Internet.

2.1.1. Avaliação de RI

Segundo Manning, Raghavan e Schtze (2008), três componentes básicos são necessários para avaliar os sistemas de recuperação da informação: (i) uma coleção de

documentos; (ii) um conjunto de testes, em forma de consultas; e (iii) um conjunto de julgamentos relativos à relevância dos documentos para cada consulta. Na literatura, existem várias bases de dados para avaliar os SRIs, entre elas, podemos destacar a primeira base de dados com esse propósito, conhecida como *Cranfield* (CLEVERDON, 1991), e duas bases de dados desenvolvidas pelo Instituto Nacional de Padrões e Tecnologia dos Estados Unidos (NIST): a TREC – *Text Retrieval Conference* (VOORHEES; HARMAN, 2005) e a GOV2 (CRASWELL, 2004).

A avaliação de um sistema de recuperação da informação consiste em medir a eficiência do processo de análise quanto à relevância de um documento, uma vez que um documento retornado por um sistema desse tipo pode ser relevante ou não. Além disso, nesse processo de busca, um documento pode ter sido recuperado ou descartado (não recuperado). Na Figura 1, são apresentados os possíveis resultados de uma consulta sobre o contingente total de documentos. Geralmente, duas métricas são associadas a um sistema de recuperação de informação: precisão (*precision*) e *recall*.

Figura 1 - Resultado de uma consulta sobre o total de documentos - *Adaptado de Kowalski, (1997).*

A *precisão (P)* é definida como a fração dos documentos recuperados que são relevantes, e pode ser definida de acordo com a Equação 1.

$$P = \frac{\text{Número relevantes recuperados}}{\text{Total recuperados}} \qquad (1)$$

Por sua vez, o *recall (R)* é definido como a fração de documentos relevantes que

são recuperados, e pode ser definida de acordo com a Equação 2.

$$R = \frac{\text{Número relevantes recuperados}}{\text{Total relevantes existentes}} \quad (2)$$

Ressalte-se, no entanto, que, nos Sistemas de Recuperação da Informação, é preciso haver um equilíbrio entre essas duas medidas. Por exemplo, é possível obter sempre o máximo *recall* total ($R = 1$), recuperando todos os documentos para todas as consultas, porém, nesse caso, a precisão será muito baixa. Dessa forma, é preciso utilizar uma medida que avalie a precisão e o *recall* conjuntamente. Van Rijsbergen (1979) propôs o uso da Medida-F que, por definição, é a média harmônica entre a precisão e o *recall* e pode ser explicada pela Equação 3.

$$F = \frac{2PR}{P + R} \quad (3)$$

2.1.2. Operações Textuais

Na maioria dos sistemas de recuperação da informação, os documentos são representados pelo seu conjunto de palavras. Essa forma de representar é a mais completa. No entanto, para grandes coleções de documentos, pode ser necessário diminuir esse conjunto de termos representativos, visando reduzir os custos computacionais. Essas operações são chamadas de operações textuais.

Uma estratégia utilizada com frequência é a eliminação de palavras muito usadas (Ex: artigos, preposições, pronomes, entre outros), que não têm relevância para se identificar um documento. Esses termos são conhecidos como *stopwords* (BAEZA-YATES; RIBEIRO-NETO, 1999) (Vide Figura 2).

Outra estratégia também utilizada, conhecida como *stemming*, é a redução de palavras para seus radicais (Ex: os vocábulos democracia, democrata e democrático podem ser reduzidos para o radical "*democ*"). Isso permite encontrar mais documentos sobre um mesmo assunto sem precisar usar variações linguísticas (plural, aumentativo, masculino/feminino). Em seguida, processos de compressão podem ser empregados. Com esses procedimentos, reduz-se a representação de um documento a um conjunto

de termos indexáveis.

```
"a","ainda","alem","ambas","ambos","antes",
"ao","aonde","aos","apos","aquele","aqueles",
"as","assim","com","como","contra","contudo",
"cuja","cujas","cujo","cujos","da","das","de",
"dela","dele","deles","demais","depois","desde",
"desta","deste","dispoe","dispoem","diversa",
"diversas","diversos","do","dos","durante","e",
"ela","elas","ele","eles","em","entao","entre",
"essa","essas","esse","esses","esta","estas",
"este","estes","ha","isso","isto","logo","mais",
"mas","mediante","menos","mesma","mesmas","mesmo",
"mesmos","na","nas","nao","nas","nem","nesse","neste",
"nos","o","os","ou","outra","outras","outro","outros",
"pelas","pelas","pelo","pelos","perante","pois","por",
"porque","portanto","proprio","propios","quais","qual",
"qualquer","quando","quanto","que","quem","quer","se",
"seja","sem","sendo","seu","seus","sob","sobre","sua",
"suas","tal","tambem","teu","teus","toda","todas","todo",
"todos","tua","tuas","tudo","um","uma","umas","uns"};
```

Figura 2 – Exemplos de Stopwords na língua portuguesa.

2.1.3. Indexação

Os índices são amplamente utilizados em sistemas de recuperação da informação. Eles são fundamentais para promover agilidade no processo de busca quando se manipulam grandes coleções de documentos. Atualmente, a técnica de indexação baseada em arquivos invertidos (*inverted files*) é a mais difundida e utilizada nas aplicações de RI, uma vez que é mais eficiente em relação às demais (ZOBEL; ALISTAIR, 2006).

A técnica de arquivos invertidos é um mecanismo orientado a palavras para indexar coleções textuais e promover mais desempenho na atividade de busca. Essa abordagem tem dois elementos básicos: vocabulário e ocorrências (CHOWDHURY, 2010). O vocabulário é o conjunto de diferentes palavras no texto. Para cada palavra, é armazenada uma lista contendo todas as posições em que a palavra aparece no texto. O conjunto de todas essas listas é chamado de ocorrências. Na Figura 3, demonstra-se um exemplo de um índice em arquivo invertido para um pequeno texto.

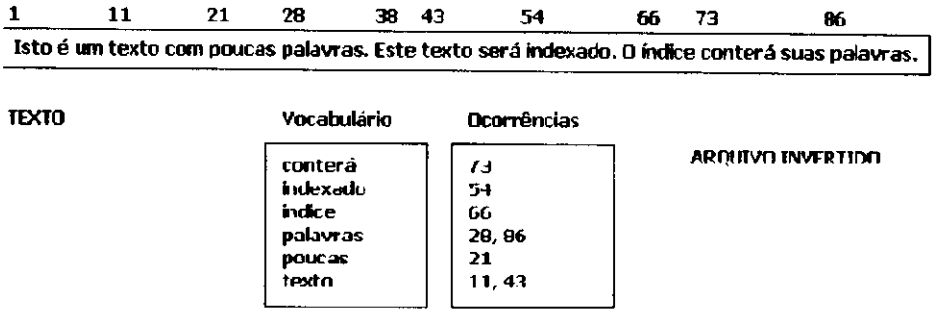


Figura 3 - Arquivo Invertido gerado a partir de um texto (CAMPELO, 2008).

2.1.4. Modelos Clássicos

Os modelos clássicos de RI consideram que cada documento é descrito por um conjunto de palavras-chaves, chamadas termos de indexação. Esses modelos assumem que um determinado termo do documento pode ser mais representativo do que outro, logo, para cada termo, é associado um valor numérico, conhecido como peso *(w)*, e em cada modelo, o cálculo desse peso é feito de forma diferente. Entre os modelos clássicos, podemos destacar o booleano (WALLER; KRAFT, 1979), o probabilístico (FUHR, 1992) e o vetorial (SALTON, 1971); (SALTON; BUCKLEY, 1988).

O modelo booleano é baseado na teoria dos conjuntos e na álgebra booleana. Nele, uma consulta (*query*) é uma expressão booleana composta de termos ligados por conectivos - OR, AND e NOT - os pesos relativos aos termos são binários, ou seja, $w \in \{0,1\}$, e os documentos recuperados são aqueles que contêm os termos que satisfazem à expressão lógica da consulta. As principais vantagens desse modelo são a facilidade de serem formalizados e a simplicidade. Porém, o fato de um documento ser considerado apenas relevante ou não relevante (não é permitida uma relevância parcial), muitas vezes, gera um número muito grande ou muito pequeno de documentos como resultado para uma determinada consulta. Além disso, nesse modelo, é muito difícil transformar a consulta do usuário em uma expressão booleana.

O modelo probabilístico parte da premissa de que, para uma dada consulta, existe um subconjunto de documentos da coleção que são considerados relevantes. Dessa forma, o modelo tenta estimar a probabilidade de um usuário considerar, a partir

de sua consulta, aquele documento relevante. A principal ferramenta utilizada nesses modelos é o *Teorema de Bayes* (JENSEN, 2001), cuja vantagem é a possibilidade de formar um ranking de acordo com a probabilidade de relevância dos documentos. Entretanto, a formação desse ranking depende da precisão das estimativas de probabilidade. Além disso, o método não explora a frequência do termo no documento, como fator para mensurar sua.

No modelo vetorial, cada termo de indexação em um documento tem um peso associado que quantifica a correlação entre os termos e esse documento. Nesse modelo, um documento é representado como um vetor dos seus termos, e cada posição do vetor representa um termo diferente. Ou seja, essa representação não considera a ordem em que as palavras aparecem no documento (Vide Figura 4). Por isso, esse modelo também é conhecido como "*bag of words*" (MANNING; RAGHAVAN; SCHÜTZE, 2008).

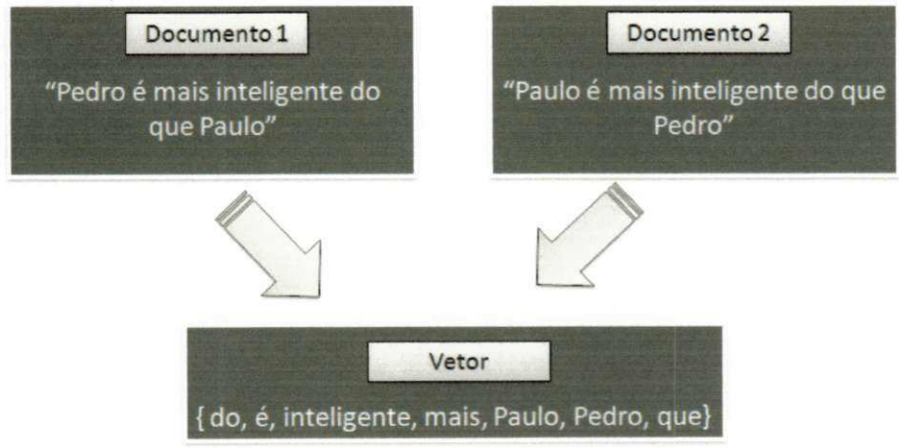


Figura 4 - Modelo de representação de documentos em forma de vetor.

Uma estratégia para atribuir peso aos termos de um documento é a frequência (*tf*) ou o número de vezes em que cada termo *(t)* aparece no documento (d). Porém, utilizar apenas a frequência dos termos não é uma boa estratégia, uma vez que, se um termo aparece cinco vezes em um documento e apenas uma em outro, isso não implica que aquele documento é cinco vezes mais relevante do que o outro. Portanto, não existe uma relação de proporção entre a relevância e a frequência do termo no documento.

Para corrigir esse problema, geralmente, é utilizado o logaritmo da frequência do termo para representar o peso (*w*). Logo, o peso do logaritmo da frequência de um termo (*t*) em um documento (*d*) é dado pela Equação 4.

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d}, & se \quad tf_{t,d} > 0 \\ 0, & caso \quad contrário \end{cases} \qquad (4)$$

Geralmente, termos que não costumam aparecer com frequência nos documentos são mais informativos do que os frequentes. Por exemplo, uma consulta pelo termo "*papiloscopista*", em um conjunto de documentos, onde esse termo não esteja tão presente, é provável que o documento que contenha o termo seja muito relevante para essa consulta. Assim, é interessante que os termos que não estejam tão presentes na coleção tenham um peso maior em relação aos frequentes.

Logo, foi proposta uma maneira de calcular o peso dos termos levando em consideração o número de documentos que contém um determinado termo. Essa medida é conhecida como *idf* (*inverse document frequency*) ou frequência inversa de documentos, e é representada pela Equação 5, onde N é o número total de documentos do sistema, e df_t é a frequência de documentos que contém o termo (t).

$$idf_t = \log\left(\frac{N}{df_t}\right) \qquad (5)$$

Combinando as definições da frequência de termos (*tf*), com a frequência inversa de documentos (*idf*), tem-se uma das formas mais utilizadas para associar pesos aos termos indexados de um documento, conhecida como *tf-idf*, conforme a Equação 6.

$$tf.idf = tf_{t,d} \times idf_t \qquad (6)$$

No modelo vetorial, termos e documentos são representados como vetores em um espaço vetorial, onde os termos são os eixos desse espaço, e os documentos são vetores. Assim, quando uma consulta (q) é realizada no sistema, representa-se como um vetor no espaço ($\vec{q}$) e ordenam-se os documentos ($\vec{d}_i$) de acordo com a proximidade em relação a essa consulta. Portanto, a medida de similaridade (*sim*) entre os vetores é o cosseno do ângulo formado pelos vetores, conforme visto na Equação 7.

$$sim(q,d) = \cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}\,\|\,\vec{d}\,|} \quad (7)$$

Além dos modelos clássicos, outros modelos mais avançados de recuperação da informação foram propostos, tais como os baseados em: bases de conhecimento (MYLONAS ET AL, 2008), lógica *fuzzy* (HORNG ET AL, 2005), redes neurais (RESHADAT; FEIZI-DERAKHSHI, 2012) e outras técnicas de aprendizagem de máquina, como SVM (*Support Vector Machines*) (YUE ET AL, 2007).

2.2 Recuperação da Informação Geográfica

Com a capacidade de processamento geográfico, novos desafios foram propostos aos processos de recuperação de informação, que exige a adaptação dos principais componentes de um sistema de RI. A recuperação de informação geográfica (RIG) pode ser vista como um ramo da área de recuperação de informação tradicional, que inclui todas as suas linhas de pesquisa, porém destaca a recuperação e a indexação de informações espaciais e geográficas (LARSON, 1996). O Objetivo da RIG é tratar dos problemas da recuperação de informações que contenham referências geográficas fundamentais para o significado de uma consulta.

De acordo com Sanderson e Han (2007), entre os grupos de palavras mais utilizadas em consultas está o uso de termos geográficos. Isso acontece devido ao fato de que as localidades geográficas têm uma importância semântica alta. Por exemplo, sites que contêm informações sobre restaurantes e bares são mais interessantes para usuários próximos daquela localidade.

O fato de a informação na Internet ter um caráter semiestruturado na Web dificulta o acesso a informações geográficas. Entre as dificuldades, podemos destacar: (i) o contexto geográfico é incluído nas descrições via linguagem natural; (ii) os nomes de lugares são ambíguos e confundidos com nomes de pessoas, de animais e de ruas (Ex: Patos-PB, Juarez Távora-PB); (iii) dependência da existência e relação com os termos do texto; (iv) interpretação das relações espaciais ("próximo", "a oeste", etc.);

(v) construção de ranking específico para definir a relevância geográfica (HILL, 2007). Contudo um grande volume de dados presentes na Internet pode ter associado um contexto geográfico. Existem várias formas de deduzir o contexto geográfico com base, por exemplo, no conteúdo das páginas e na estrutura de links da Web. De acordo com Markowetz *et al.* (2005), esse processo pode ser dividido em três etapas: extração, mapeamento e propagação. Na primeira, são identificados os elementos utilizados para referenciar localidades geográficas, como por exemplo, nomes de lugares, números de telefone, códigos postais etc. Na segunda, associa-se cada referência detectada em uma localidade geográfica válida. Na terceira etapa, por meio da estrutura de links, são realizadas propagações da localidade geográfica para páginas que não tenham referência geográfica a partir da sua hierarquia de links.

Para Buyukkokten *et al.* (1999), o escopo geográfico é atribuído a uma página a partir dos dados coletados por meio do endereço IP dos servidores hospedeiros e representado por uma coordenada geográfica referente ao centroide da região resultante do processo de análise da página. Já Ding *et al.* (2000) atribuem os escopos através dos conteúdos dos documentos e da estrutura de links da Internet, utilizando-se uma hierarquia que contempla cidades, estados e países, a partir das regiões administrativas dos Estados Unidos. Por sua vez, Markovetz, Brinkhoff e Bernhard (2005) relatam que existe grande relevância nos dados da seção *admin-c* (seção que contém dados de contato do administrador do domínio) do *whois*. No entanto, outras partes não são tão importantes, visto que estão relacionadas com a localização dos servidores, e muitas companhias pequenas ou mesmo indivíduos terceirizam o serviço de hospedagem, fazendo com que, muitas vezes, o local onde está armazenado o site não tenha nenhuma relação com o local de sua criação ou com o seu conteúdo.

No geral, os sistemas de informação geográfica costumam atribuir um único escopo geográfico a cada documento presente na coleção. Porém, essa abordagem é bem restritiva. Por isso, Batista et al (2010) propõem a associação de uma "*assinatura geográfica*", que pode ser definida como uma lista de referências geográficas encontradas nos documentos.

2.2.1. Sistemas de Recuperação de Informação Geográfica

De acordo com Overell (2009), o sistema de RIG é composto por três componentes principais: analisador geográfico, indexador textual e o motor de busca (Figura 5). O primeiro é responsável pelo processamento da coleção de documentos por meio da identificação de referências geográficas, desambiguação (geralmente realizadas através de tesauros e *Gazzeters*) e construção do índice geográfico. Assim como os sistemas de recuperação da informação tradicionais, o indexador é responsável pela criação do índice textual. A partir da consulta, o motor de busca é responsável por construir um ranking dos documentos existentes na coleção com a combinação dos valores de relevância geográfica e textual dos documentos.

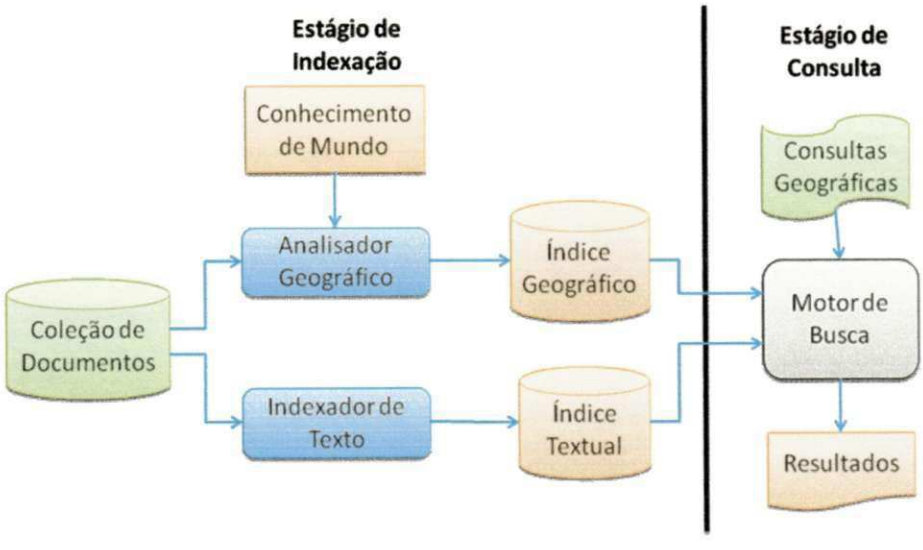


Figura 5 - Componentes de um Sistema de RIG (Adptado de (OVERELL, 2009)).

Nos últimos anos, diversos sistemas de recuperação de informação geográfica e métodos de busca têm sido desenvolvidos. Entre eles, algumas bibliotecas digitais como a *Alexandria Digital Library* (HILL; GOODCHILD; JANEE, 2004) e a *Perseus Digital Library* (SMITH; CRANE, 2001), foram criadas com a possibilidade de associar um contexto geográfico aos documentos adicionados e fornecer uma interface com mapas para buscá-los.

Entre 2002 e 2005, através do projeto SPIRIT, desenvolvido na Universidade de Sheffield (JONES *ET AL.*, 2004), foi criado um motor de busca geográfico com uma interface que provê tanto a entrada de dados textuais quanto a interação com um mapa que relaciona o contexto geográfico dos documentos recuperados. Além disso, no sistema são utilizadas ontologias para modelar o espaço geográfico, ranking de relevância baseados nas ontologias geográficas, técnicas de expansão de consultas, índices espaciais para as coleções de documentos e um mecanismo de aprendizagem para extração de contexto geográfico a partir dos documentos para a geração de metadados espaciais.

Na mesma época, o Grupo XLDB da Universidade de Lisboa, desenvolveu o *GeoTumba!*, um sistema de busca em páginas da web portuguesa (SILVA *ET AL.*, 2006), no qual foram incorporadas diversas heurísticas para detectar as referências geográficas em páginas da Web. Essas heurísticas são relativas ao texto das páginas Web, aos hiperlinks, ao ambiente da Internet, e ao uso das referências geográficas.

Campelo (2008) propõe a criação de um sistema chamado GeoSEn (GEOgraphic Search ENgine), que permite: (i) detectar referências a localizações geográficas; (ii) modelar o escopo geográfico dos elementos da Web; (iii) realizar indexação espaço-textual; (iv) recuperar os documentos utilizando operações textuais, como continência, adjacência e distância, bem como suas respectivas negações; (v) elaborar um ranking de relevância espaço-textual; (vi) uma interface multimodo. Fernandes (2010) estende esse sistema, chamando-o de GeoSEn_Tags, que permite a realização de buscas semânticas por meio de ontologias e de técnicas de expansão de consulta utilizando-se de tags em uma *folksonomia.*

Além dessas iniciativas acadêmicas, desde 2005 que a Microsoft e o Google investem no desenvolvimento dos seus próprios sistemas de recuperação de informação geográfica, Bing Maps[1] e GoogleMaps[2] respectivamente. Esses motores de busca permitem que os usuários realizem consultas por produtos e serviços em uma área geográfica específica e são financiados pelos anúncios exibidos juntamente com os respectivos resultados da consulta (Figura 6).

---

[1] http:// www.bing.com/maps/
[2] http://maps.google.com.br/

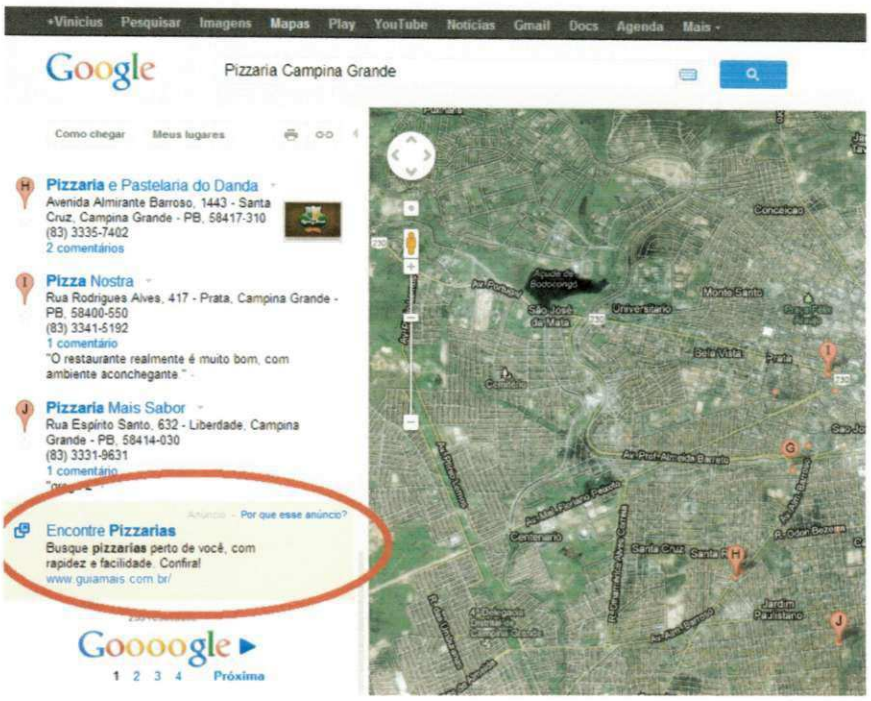


Figura 6 - Propagandas exibidas no GoogleMaps.

2.3 Expansão de Consultas

A relativa ineficiência dos sistemas de recuperação de informação é causada, em grande parte, pela dificuldade do usuário na escolha dos termos da consulta criada para modelar a sua real necessidade de informação. Além disso, nem sempre os autores dos documentos usam as mesmas palavras que os usuários, quando se referem a um mesmo conceito. Na área de RI, esse problema é conhecido como "*Word Mismatch*" (XU; CROFT, 1996).

Assim, sem o conhecimento prévio do conteúdo dos documentos da coleção, grande parte dos usuários encontram muitas dificuldades na formulação de consultas que tornem o processo de recuperação da informação mais eficiente. Geralmente, nos motores de busca, os usuários tendem a encontrar as informações depois de vários refinamentos da consulta original. Uma forma de amenizar esse problema é aumentando o número de termos na consulta. Todavia, segundo Carpineto e Romano (2012), o tamanho médio das consultas é de 2,3 palavras.

Um método bem conhecido para superar esta limitação é a expansão de consulta,

que consiste do aumento dos termos da consulta original, por novos termos, com um significado semelhante (CARPINETO; ROMANO, 2012). Ou seja, é a expansão da consulta original, com outras palavras que identifiquem melhor a intenção real do usuário ou que simplesmente, produzam uma consulta em que é mais provável se recuperarem documentos relevantes.

Além de ser uma das estratégias mais utilizadas para melhorar a eficiência do processo de ranking de documentos nos SRIs, a expansão de consultas vem sendo utilizada em diversas outras áreas de RI, tais como: sistemas de informação multimídia (NATSEV *ET AL.*, 2007), filtragem da informação (ARGUELLO *ET AL.*, 2008), sistemas de recuperação de informação multilíngue (CAO *ET AL.*, 2007), entre outras.

Manning, Raghavan e Schtze (2008) asseveram que os métodos de expansão de consulta podem ser divididos em dois grupos: os globais e os locais. Nos métodos locais, o próprio resultado da busca é utilizado para ajustar a consulta inicial. Já nos métodos globais, a reformulação da consulta independe da consulta ou dos seus resultados.

Entre os métodos locais, podemos destacar a técnica de *feedback de relevância*. Nessa abordagem, o usuário realiza uma consulta, e o motor de busca retorna um conjunto de documentos. Dentre esses documentos, o usuário assinala os documentos que ele considerou relevantes e a partir disso, o motor de busca gera uma nova consulta levando em consideração esses documentos. Por fim, o motor de busca processa essa nova consulta e exibe os resultados. Esse processo pode se repetir por várias vezes.

Como vimos no início desse capítulo, os documentos podem ser representados como vetores, isto é, como pontos em um espaço de várias dimensões, onde o número de dimensões vai ser o número de termos distintos existentes na coleção de documentos. O conceito principal do *feedback de relevância* é o centro de massa de um conjunto de pontos (documentos): o centroide, que pode ser definido como a média dos pesos dos vários termos presentes nos documentos de uma coleção. Logo, seja D, o conjunto de documentos, e $\vec{v}(d) = \vec{d}$, o vetor que representa um documento d, então o centroide desse conjunto de documentos é definido de acordo com a Equação 8.

$$\vec{\mu}(D) = \frac{1}{|D|} \sum_{d \in D} \vec{v}(d) \qquad (8)$$

A principal implementação da técnica de *feedback de relevância* é o algoritmo de Rocchio (ROCCHIO, 1971), cujo objetivo é o de encontrar a melhor consulta que represente o interesse de informação do usuário. Dessa forma, seja D_r o conjunto dos documentos que o usuário marcou como relevantes, e D_{nr}, o conjunto dos demais documentos apresentados como resultado da consulta inicial que não foram marcadas (não relevantes), logo, tem-se que o vetor de consulta ótimo é:

$$\vec{q}_{ótimo} = \mu(D_r) + [\mu(D_r) - \mu(D_{nr})] \qquad (9)$$

Já entre os métodos globais de expansão de consulta, existem as técnicas baseadas em estruturas ou fontes de conhecimento (*background knowledge*), como os tesauros e as ontologias. Essas técnicas consistem de uma base de dados que armazenam sinônimos e termos relacionados. Isso significa que, nesses métodos, para cada termo *t* na consulta, há uma expansão da consulta com os termos presentes nessa estrutura de conhecimento que estão semanticamente relacionados ao termo *t*. Embora essa técnica possa aumentar significativamente o *recall* do SRI, conjuntamente pode piorar a precisão do sistema caso a expansão tenha termos ambíguos.

2.4 Tesauros

O principal problema da expansão de consultas está na escolha de quais termos devem ser usados para expandir a consulta original do usuário e consequentemente melhorar o seu resultado. Os tesauros são frequentemente utilizados nos sistemas de recuperação da informação para identificar termos e expressões que tenham uma relação semântica.

Os *tesauros* consistem de uma base de dados que armazenam sinônimos e termos relacionados. Isso significa que, nesses métodos, para cada termo *t* na consulta, há uma

expansão da consulta com os termos presentes no *tesauro* que estão semanticamente relacionados ao termo *t*. Embora essa técnica possa aumentar significativamente o *recall* do SRI, conjuntamente pode piorar a precisão do sistema caso a expansão tenha termos ambíguos.

Basicamente, consistem de uma lista pré-definida de palavras importantes para um determinado domínio de conhecimento, e para cada palavra presente nessa lista um conjunto de termos sinônimos (IMRAN; SHARAN, 2009).

Os tesauros podem ser classificados de duas formas: (*i*) de acordo com a forma de expansão das consultas (global ou local); (*ii*) de acordo com a forma de construção dele (manual ou automático).

Quanto à forma como a expansão de consulta é realizada, na abordagem global (RUGE, 1992), a expansão é feita a partir da coocorrência dos termos e seus relacionamentos na coleção de documentos. Ou seja, num tesauro global a expansão dos termos é realizada antes do processo de indexação. Já nas abordagens locais (XU; CROFT, 2000), a expansão é realizada por meio de determinado número (*k*) de documentos retornados pela própria consulta do usuário. Nessa abordagem, a reestruturação da consulta é feita durante o processo de consulta, e seu resultado é empregado para modificar apenas ela mesma.

Quanto ao tipo de construção do tesauro, as abordagens com construção manual, requerem um trabalho intenso e a relação das palavras é feita de forma manual, geralmente com a presença de especialistas no domínio específico do tesauro. Já nas abordagens automáticas, tem-se uma subdivisão de acordo com a técnica utilizada, são as baseadas em: coocorrência de termos ou medidas de similaridade. Na técnica de coocorrência, os termos são relacionados baseados numa hipótese de associação. Por exemplo, dois termos estão relacionados semanticamente, se eles se relacionam com uma mesma palavra ("pirarucu" ≈ "tambaqui", pois ambos os termos se relacionam com "rio" e "peixe". Logo, devem ser similares). Um tesauro gerado automaticamente por meio de uma técnica de similaridade considera a relação dos termos na consulta para calcular a medida de similaridade que determinará a expansão dos termos, ou seja, os termos selecionados para expansão levam em consideração toda a consulta, ao contrário do modelo de coocorrência que avalia termo a termo.

Segundo Ding, Ghowdhury e Foo (2000), a combinação das técnicas de

UFCG/BIBLIOTECA/BC

construção manual e automática apresenta bons resultados para a expansão de consulta, uma vez que técnicas manuais refletem a organização dos termos baseado na inteligência humana, enquanto as técnicas automáticas podem capturar mudanças dinâmicas que possam vir a existir no relacionamento entre os termos, como por exemplo, o surgimento de neologismos.

2.5 Sensibilidade ao Contexto

A computação ubíqua foi uma das áreas pioneiras no estudo e na utilização do conceito de contexto e, com isso, demonstrou o potencial da aplicação desse aspecto nos sistemas computacionais. Atualmente, esse conceito tem sido estudado em diversas áreas da Computação, como: Recuperação da Informação, Sistemas Colaborativos, Inteligência Artificial, Interação Homem-máquina, entre outras, que buscam estudar como o contexto pode ser aplicado nos sistemas computacionais.

A palavra sensibilidade refere-se à capacidade de um sistema em detectar a ocorrência de eventos ou objetos existentes no seu entorno (PERNAS, ET AL., 2010). Já o contexto é qualquer informação que pode ser utilizada para caracterizar o estado de uma entidade. Essa entidade pode ser um lugar, um objeto ou uma pessoa que é considerada importante para a relação entre um usuário e uma aplicação, inclusive os próprios usuários e as aplicações (DEY, 2001). Por exemplo, um telefone celular não irá emitir um sinal sonoro em sala de aula, caso o sistema reconheça sua localização e o horário da aula. Bazire e Brézillon (2005) elencaram cerca de 150 definições de contexto, originadas de diferentes domínios, e chegaram à conclusão de que o contexto atua como um conjunto de restrições que influenciam o comportamento de um sistema em uma dada tarefa, e que essa definição depende da área de conhecimento à qual pertence.

De um modo geral, as informações contextuais podem ser classificadas em seis dimensões básicas, a partir das quais é possível contextualizar uma determinada atividade: *who* (identificação), *where* (localização), *when* (tempo), *what* (ação, atividade), *why* (motivação das ações) e *how* (como os elementos de contexto são coletados) (TRUONG; ABOWD; BROTHERTON, 2001).

Em um sistema ciente de contexto, geralmente existe uma série de sensores de software e de hardware que monitoram o ambiente e os dispositivos de interesse e recolhem informações que serão enviadas a um conjunto de serviços de contexto, onde são processadas e modificadas para que possam ser entregues às aplicações. Por ser uma tarefa complexa, o desenvolvimento desse tipo de aplicação requer o uso de técnicas de modelagem de sistemas. Por isso, diversas técnicas vêm sendo utilizadas na literatura por diferentes sistemas para representar o contexto, como: par chave-valor (SCHILIT; ADAMS; WANT, 1994), baseados em lógica (GHIDINI; GIUNCHIGLIA, 2001), linguagem de marcação (BUCHHOLZ; HAMANN; HUBSCH, 2004), modelos gráficos (HENRICKSEN; INDULSKA; RAKOTONIRAINY, 2002), orientação a objetos (SCHMIDT; BEIGL; GELLERSEN, 1999) e ontologias (WANG *ET AL.*, 2004). Segundo Vieira *et al.* (2006), não há uma técnica ideal que se aplique a todos os sistemas cientes de contexto, visto que cada tipo de sistema impõe restrições distintas.

2.6 Considerações Finais

Neste capítulo, foram apresentados os principais conceitos relacionados à recuperação da informação clássica e geográfica. Esses conceitos são essenciais para a compreensão do presente trabalho. No próximo capítulo, serão apresentados alguns trabalhos recentes do estado da arte, que utilizam o contexto geográfico para melhorar a eficiência dos Sistemas de Recuperação da Informação.

Capítulo 3

Trabalhos Relacionados

Encontrar informação relevante, em meio à grande quantidade de conteúdo, é uma tarefa difícil, porque a representação dessa necessidade de informação por parte dos usuários consiste de uma consulta formada por algumas palavras submetidas aos motores de busca. Por essa razão, essa consulta (geralmente composta de poucas palavras) deve caracterizar sua necessidade para que somente os recursos relevantes sejam apresentados como resultado. Porém, esse é um problema complexo e, por isso, algumas abordagens vêm sendo desenvolvidas na área de RI, quais sejam: a expansão semântica dos termos da consulta e o uso de informações contextuais como uma forma de expansão.

Atualmente, as informações geográficas estão se tornando cada vez mais importantes em pesquisas na web, porquanto os engenhos de busca, muitas vezes, podem retornar resultados melhores para os usuários, por meio da análise de características, como a sua localização ou a presença de termos geográficos nas consultas e em páginas da internet. Além disso, essas informações têm grande valor comercial, pois permitem que empresas produzam publicidade para determinados usuários de uma localidade ou em páginas relativas a um contexto geográfico específico. Como resultado disso, as empresas responsáveis por motores de busca têm investido com recursos significativos em tecnologias de busca geográfica (GAN ET AL, 2008).

Neste capítulo, são apresentados alguns trabalhos recentes na área de RI que tratam da melhoria dos mecanismos de busca, através de técnicas de expansão de

consulta que utilizam o contexto geográfico. Na seção 3.1, apresentamos trabalhos que utilizam informações geográficas no processo de expansão de consultas; a seção 3.2 traz as considerações finais sobre o levantamento bibliográfico do estado da arte, para o desenvolvimento deste trabalho, bem como suas principais contribuições.

3.1 A Expansão de Consultas e o Contexto Geográfico

Expandir a consulta é a tarefa de adicionar palavras consideradas sinônimas ou relacionadas com os termos de consulta do usuário, com o objetivo de recuperar os documentos mais relevantes (BAEZA-YATES; RIBEIRO-NETO, 1999). Como já referimos, as técnicas de expansão de consulta são classificadas em métodos globais e locais. Os métodos globais expandem os termos, independentemente da consulta e dos seus resultados. Por sua vez, os métodos locais expandem os termos de consulta em relação aos documentos do topo do ranking retornados como resposta para a consulta.

A motivação para expandir termos geográficos nas consultas é a recuperação de documentos que são considerados relevantes para um determinado local, mesmo sem mencioná-lo na consulta original. Atualmente, existem duas classes de estratégias de expansão de consultas que contemplam aspectos geográficos em dois grupos: as baseadas em fontes de conhecimento e as baseadas em *feedback* de relevância.

3.1.1. Expansão Baseada em Fontes de Conhecimento

Esse tipo de expansão envolve a adição de termos à consulta original, oriundos de uma fonte de conhecimento geográfico, como *Gazetteers*, com o objetivo de recuperar mais documentos relevantes dentro da localidade de interesse do usuário. Ou seja, ao se fazer uma consulta utilizando algum termo relativo à localidade geográfica, esse termo é considerado como o contexto geográfico da consulta e consequentemente,

o contexto de interesse do usuário. No trabalho de Buscaldi, Rosso, e Arnal (2006), os termos geográficos das consultas são expandidos por meio de sinônimos presentes na *WordNet*, uma ontologia da língua inglesa, composta por 155 mil palavras, mapeadas em 118 mil redes de sinônimos (*synSet*), em que cada conjunto representa um conceito diferente. Essas redes de sinônimos são divididas em 45 categorias, de acordo com a sua classe gramatical (verbo, adjetivo, advérbio, substantivos). Além disso, existe um mapeamento das relações entre as redes de sinônimos, ou seja, é possível saber se um conjunto é antônimo, hipônimo ou instância de outra rede.

Neste trabalho, para cada termo da consulta, inicialmente, é avaliado se esse termo tem como hipernímia no *WordNet* os termos: "Cidade", "Estado". Ou seja, é identificado se o termo é relativo a uma localização geográfica e, a partir disso, são adicionados à consulta original os termos sinônimos existentes na ontologia. Além disso, é proposta a construção de um índice dos termos relativos às localidades geográficas dos documentos da coleção, que identifica os termos geográficos existentes nos documentos, com o objetivo de indexá-los por esses termos. Portanto, o sistema de recuperação é composto por dois índices: o textual-geográfico e o puramente textual, que indexa os demais termos que não são contemplados pelo primeiro índice.

Embora seja proposto um índice dos termos relativos às localidades geográficas, não foi avaliado o seu impacto no processo de recuperação da informação, porquanto apenas a técnica de expansão foi avaliada. Na avaliação, constatou-se que houve uma melhoria do *recall* do sistema em relação às abordagens tradicionais, porém, devido ao fato de a expansão ser feita sem nenhum critério de filtragem dos termos, houve uma degradação considerável na precisão do sistema, em alguns casos, inferior às abordagens tradicionais.

Buscaldi e Rosso (2009) georreferenciaram as redes de sinônimos do *WordNet* relativas às localidades geográficas, denominaram-nas de *GeoWordNet* e propuseram um sistema de recuperação de informação geográfica, chamado de *GeoWorSE - (Geographical Wordnet Search Engine)*, composto por dois índices: um geográfico e um textual. Durante a fase de indexação, os documentos são examinados com o objetivo de encontrar termos geográficos, que serão georreferenciados a partir da *GeoWordNet* e acrescentados ao índice geográfico, enquanto os demais termos são armazenados no índice textual. Portanto, quando uma consulta é submetida ao sistema proposto, o

coeficiente de relevância do documento em relação à consulta é calculado com base nos dois índices. Ou seja, caso exista na consulta algum termo geográfico, ele é georreferenciado e submetido ao índice geográfico, a fim de recuperar os documentos geograficamente mais próximos à coordenada relativa ao termo, enquanto os demais são submetidos ao índice textual. Logo, o coeficiente de relevância é a soma entre o coeficiente de proximidade geográfica e o coeficiente de proximidade textual.

O trabalho de Buscaldi e de Rosso (2009), assim como o de Buscaldi, Rosso, e Arnal (2006), concentram as técnicas de expansão de consultas apenas nos termos geográficos, e quando comparados com o sistema de recuperação da informação tradicional, não houve uma melhoria significativa da precisão e do *recall* desses sistemas.

Larson, Gey e Petras (2006) propõem uma técnica que identifica regiões geográficas e expande as consultas com localidades que fazem parte daquela região. Assim, caso um termo relativo a um estado seja identificado, esse é expandido com cidades presentes naquele estado. Para essa tarefa, é utilizado um *Gazetteer* derivado do *World Gazetteer*[3]. Com o objetivo de auxiliar a recuperação de documentos relativos a uma determinada área geográfica, Delboni *et al.* (2007) propuseram a expansão de termos que representam relações espaciais. O termo relacional *perto* é expandido por outros com a mesma semântica, como *próximo*, *em frente*, *nos arredores* etc.

Cardoso e Silva (2007) apresentaram um sistema de informação geográfico capaz de expandir consultas com aspectos geográficos através de uma ontologia geográfica. Esse sistema captura termos geográficos e as relações espaciais existentes na consulta e mapeia-os em conceitos da própria ontologia. Ou seja, um termo que represente uma relação espacial é mapeado em uma relação da própria ontologia. Porém, nessa abordagem, é necessário que o usuário expresse a sua consulta como uma tripla <o que, relação, onde>, para que essa relação seja mapeada numa relação da ontologia, e o termo relativo à coordenada geográfica seja expandido. Logo, em consultas que não haja essa relação, não haverá expansão. Além disso, nenhuma avaliação, em termos de precisão e do *recall* da técnica proposta, foi realizada.

---

[3] http://world-gazetteer.com/

Da mesma forma, Fu *et al.* (2005) propuseram uma técnica de expansão dos termos geográficos baseado na tripla *<o que, relação, onde>* e acrescentaram outra ontologia de domínio para contemplar termos que não se referem a características geográficas. Nessa abordagem, tanto se expandem os termos geográficos quanto os relativos à ontologia de domínio. Porém, a expansão dos termos relativos à ontologia de domínio não considera o contexto geográfico da consulta, portanto, pode apresentar problemas quando os termos expandidos não forem característicos desse contexto.

Além dos termos relativos às localidades geográficas, outros domínios de conhecimento podem ter termos diretamente relacionados a uma geografia específica. Por exemplo, condições climáticas específicas podem estar associadas a determinada região. Diante disso, Leite e Ricarte (2008) descreveram uma técnica que, por meio de relações *fuzzy*, relacionaram várias ontologias de domínio geográfico, com o objetivo de expandir a consulta inicial do usuário. A relação entre as ontologias é medida de acordo com a distribuição do clima no território brasileiro e essa relação pode ser entre uma região geográfica e uma zona climática ou entre os estados brasileiros e a *classificação climática de Köppen* (MCKNIGHT; HESS, 2000). Por exemplo, para se medir a relação entre o clima tropical e a região norte, inicialmente, coleta-se o tamanho que ocupa a zona climática no Brasil (59.811 pixels do mapa) e o tamanho que ocupa esse clima na região norte (30.616 pixels), logo a relação entre o clima Tropical e a região norte é dado pelo quociente 30616 / 59.811 = 0.51. Ou seja, isso quer dizer que o conceito "Região Norte" implica no Clima tropical com um peso de 0.51. Já o peso da associação inversa é dado pelo quociente entre o tamnho que o clima ocupa na região norte (30.616 pixels) e a área ocupada pela região norte no mapa (43.737 pixels) resultando em 0.7. (Vide Figura 7)

Assim, em consultas que utilizem termos relativos a um clima específico, pode ser expandida para os locais que têm esse clima. A Região Nordeste do Brasil, por exemplo, apresenta o clima semiárido e, em muitas situações, ao invés de se empregar a expressão "*Região Nordeste*", utiliza-se "*região do semiárido*". Embora o contexto geográfico possa ser expandido por termos de outro domínio, em alguns casos, os demais termos da consulta podem não ser característicos dessa região, o que levará a uma busca por termos que não costumam ser encontrados em documentos de certa região.

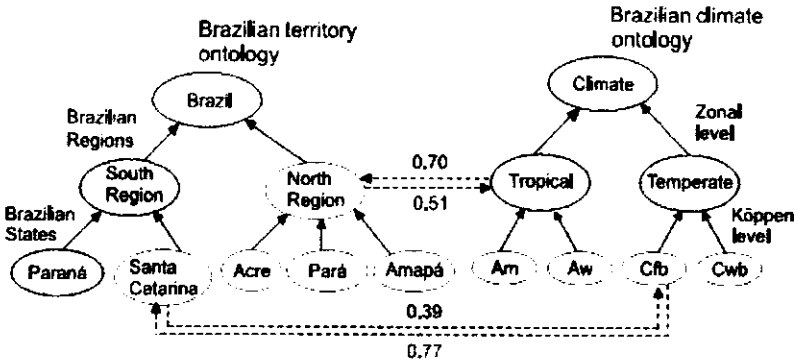


Figura 7 - Relacionamento de Ontologias de Domínio diferentes. (LEITE; RICARTE, 2008).

3.1.2. Expansão Baseada em *Feedback* de Relevância

O *feedback* de relevância é uma técnica bastante popular para a expansão de consultas. Por meio dela, a expansão dos termos é realizada a partir dos documentos retornados pela consulta original do usuário. Uma dos tipos dessa técnica, conhecido como *pseudo-feedback* de relevância, assume que os *n* documentos retornados como resposta à consulta são relevantes e, a partir deles, são colctados os *m* termos mais frequentes para expandir a consulta inicial.

Larson, Gey e Petras (2006) utilizaram essa técnica para melhorar a recuperação da informação de um sistema de recuperação de informação geográfica alemão, adicionando os 30 termos que mais aparecem nos 20 primeiros documentos retornados pela consulta original. Por meio de um *gazzeteer* derivado do World Gazzeteer, os termos relativos a localidades geográficas existentes nas consultas são expandidos por todos os termos de hierarquia inferior no *gazzeteer*. Por exemplo, uma consulta com o termo "Europa", será expandida para os países que fazem parte daquela região (Itália, Inglaterra, França, Espanha, etc.).

Andogah (2010) propôs uma técnica de expansão restrita ao escopo geográfico dos documentos. Em uma fase anterior, os documentos são marcados com escopos geográficos, que, em linhas gerais, representam o quão relevantes são esses documentos para cada localidade. Inicialmente, a consulta é submetida ao sistema e são selecionados

os escopos mais frequentes nos *n* documentos mais relevantes retornados como resultado. Depois, selecionam-se os termos geográficos existentes nos documentos que pertencem aos escopos mais frequentes para se expandir a consulta inicial, gerando uma nova consulta. Por fim, a nova consulta é submetida ao índice de documentos e o ranking dos documentos mais relevantes são apresentados ao usuário (Vide Figura 8). Em termos de precisão, a técnica de expansão restrita ao escopo geográfico proporcionou um melhoria de 33% em relação à técnica baseada apenas em *feedback* de relevância, e 9%, em relação a um sistema de busca tradicional.

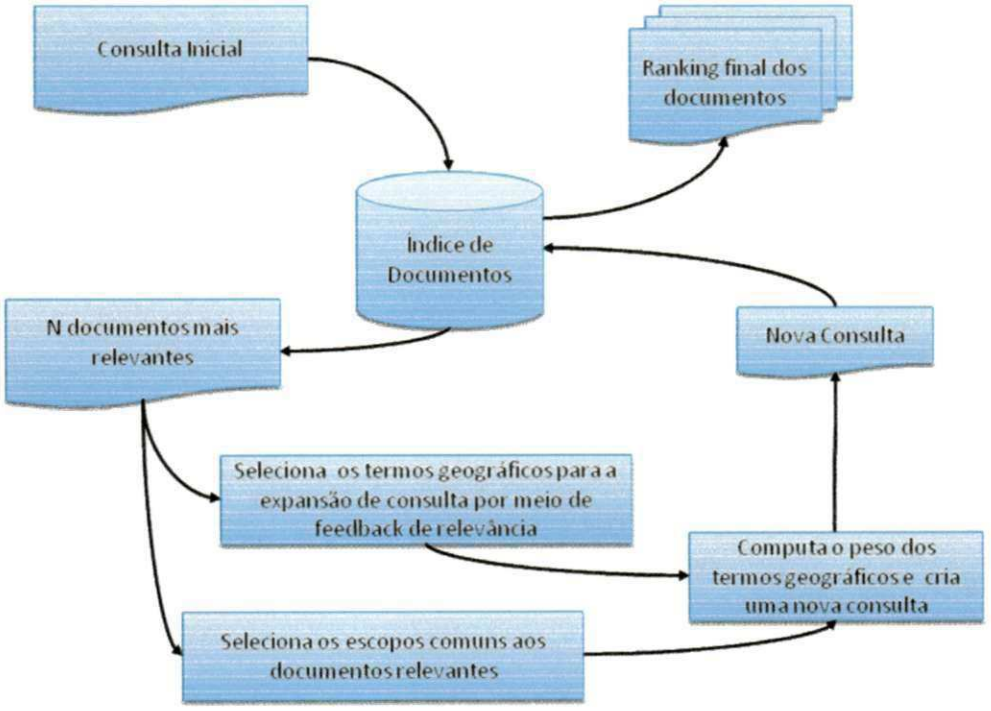


Figura 8 - Processo de expansão de consulta restrita ao escopo geográfico dos documentos - Adaptado de Andogah (2010).

Tanto a técnica de Andogah (2010) quanto a de Larson, Gey e Petras (2006) assume que os *n* documentos retornados pela consulta inicial são realmente relevantes para o usuário, porém pode haver alguns documentos que não sejam relevantes à consulta original, como, por exemplo, o uso de algum termo ambíguo na consulta, o que ocasionaria um possível erro na etapa seguinte, em que se escolhem os escopos a serem expandidos para a consulta inicial.

3.2 Considerações Finais

Neste capítulo, apresentamos alguns trabalhos recentes da área de Recuperação da Informação, que tratam da melhoria dos métodos de busca através do uso do contexto geográfico e de métodos de expansão de consultas. Nesses trabalhos, um documento pode ser relevante em termos geográficos a uma consulta de duas formas: (*i*) envolvendo um termo geográfico, que é considerado um nome alternativo ao que aparece na consulta; e (*ii*) envolvendo localidades geográficas que satisfaçam a alguma relação espacial (ex, *próximo a*, *perto de*, *etc*) com o termo presente na consulta (FU; JONES; ABDELMONTY, 2005).

Na Tabela 1 é apresentado um comparativo das principais características encontradas nos sistemas de recuperação da informação descritos neste capítulo com o presente trabalho. Logo, é possível observar, que nos trabalhos revisados, o termo geográfico presente em uma consulta é expandido por outros relacionados, a fim de encontrar documentos que estejam nesse mesmo escopo local. Os demais termos, embora possam ter alguma semântica geográfica associada, são ignorados.

Então, em muitos domínios do conhecimento, em que o vocabulário utilizado pelos usuários para descrever um mesmo conceito varia de acordo com a sua localidade geográfica, a expansão de consulta desses termos pode não trazer nenhuma melhoria para processo de recuperação da informação. Em uma consulta como "*Juçara + Parahyba*", por exemplo, o termo "*Parahyba*", por meio de ontologias geográficas, poderia ser expandido para "*Paraíba*" ou "*PB*". Ainda que a consulta expandida pudesse encontrar mais documentos relativos à Paraíba, o número de documentos referentes ao termo "Juçara" continuaria não sendo apresentado, pois é sinônimo de "Açaí" e característico dos Estados do Maranhão e do Piauí. Portanto, em um sistema que identificasse esse tipo de problema, além da expansão do termo geográfico, haveria a termo "Juçara" para "Açaí", uma vez que ele é mais característico da região de interesse da busca.

No próximo capítulo, apresentaremos, em detalhes, a técnica que propusemos neste trabalho, que visa resolver esse problema.

UFCG/BIBLIOTECA/BC

Tabela 1 - Comparativo das características dos SRI verificadas nos trabalhos relacionados.

Trabalhos	Técnica de Expansão de consulta	Fonte de Conhecimento	Índice	Expansão de Termos Geográficos	Expansão com base no contexto geográfico
Fu et al. (2005)	Fonte de Conhecimento	Ontologia Geográfica + Ontologia de Domínio	2 Índices: Textual e geográfico	SIM	NÃO
Buscaldi, Rosso e Arnal (2006)	Fonte de Conhecimento	Dicionário + Tesauro (WordNet)	2 Índices: Textual e textual c/ termos geográficos	SIM	NÃO
Larson, Gey e Petras (2006)	Feedback de relevância	World Gazetteer	2 Índices: Textual e textual c/ termos geográficos	SIM	NÃO
Cardoso e Silva (2007)	Fonte de Conhecimento	Ontologia Geográfica	2 Índices: Textual e geográfico	SIM	NÃO
Leite e Ricarte (2008)	Fonte de Conhecimento	Ontologia Geográfica + Ontologia Climática	Não mencionado no trabalho.	SIM	NÃO
Buscaldi e Rosso (2009)	Fonte de Conhecimento	WorldNet Georeferenciado	2 Índices: Textual e geográfico	SIM	NÃO
Andogah (2010)	Feedback de relevância	Geonames.org[4]	2 Índices: Textual e geográfico	SIM	NÃO
Abordagem Proposta	Fonte de Conhecimento	Tesauro Georrefernciado	Índice Textual	SIM	SIM

[4] http://www.geonames.org

Capítulo 4

Expansão Semântica com o Auxílio do Contexto Geográfico

Este capítulo apresenta um novo método de expansão de consultas baseado no contexto geográfico. A principal característica da técnica sugerida é a possibilidade de associar o contexto geográfico a termos que não se referem a uma localização geográfica. Ou seja, existem palavras que são características de algumas regiões geográficas, e a existência delas em uma consulta pode sugerir o contexto geográfico do usuário.

O capítulo traz, ainda, uma visão geral a respeito da abordagem proposta e uma descrição detalhada do algoritmo de expansão de consultas e do método para referenciar geograficamente os termos da base de conhecimento.

4.1 Visão Geral

A expansão semântica de consultas, com o auxílio de geoprocessamento, foi desenvolvida com o objetivo de resolver um grande desafio da recuperação da informação em grandes coleções de dados, conhecido como "*word mismatch*" (MANNING; RAGHAVAN; SCHÜTZE, 2008). Em linhas gerais, esse problema se refere ao fato de que um mesmo conceito é descrito pelo usuário, em sua consulta, por

termos diferentes dos presentes nos documentos. Uma das razões que pode interferir no termo escolhido pelo usuário é o contexto geográfico onde ele está inserido.

A abordagem proposta consiste, basicamente, de dois módulos auxiliares e um módulo principal, conforme apresentado na Figura 9. O primeiro módulo auxiliar é responsável pela extração de características de todos os documentos da coleção, ou seja, os documentos são representados como vetores, e cada posição dessa estrutura representa um termo distinto existente no documento. No segundo módulo, os termos de um tesauro tradicional são georreferenciados e essa fonte de conhecimento serve como parte da técnica de expansão de consulta, ou seja, palavras semanticamente correlacionadas com as da consulta inicial são expandidas e utilizadas para refinar a consulta originalmente submetida. Por último, no módulo principal, ou seja, na etapa de expansão da consulta, os termos sugeridos pelo tesauro são filtrados de acordo com o contexto geográfico definido na consulta.

A partir dessa etapa, a nova consulta será submetida ao sistema, onde é calculado um índice de relevância dos documentos para essa consulta e uma lista de documentos ordenada de acordo com esse índice é apresentada ao usuário.

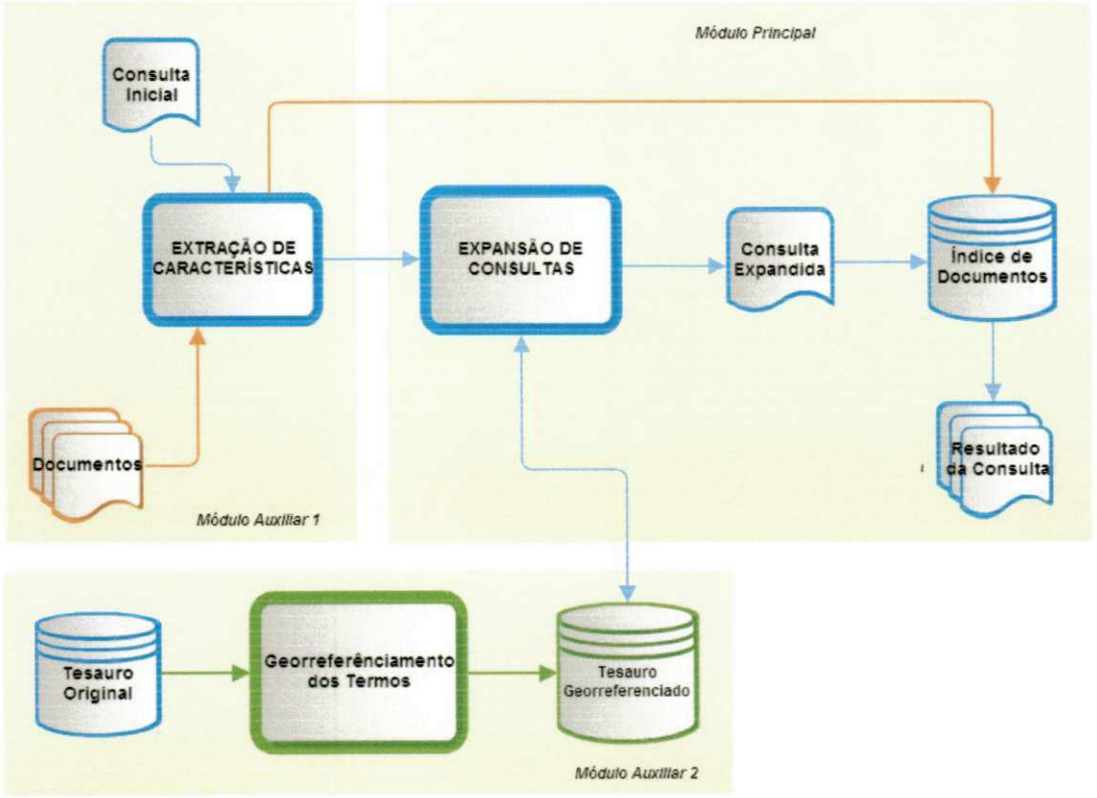


Figura 9- Visão geral da abordagem proposta.

Para a construção dessa abordagem, foi utilizada a biblioteca *Apache Lucene*[5] que é um dos *frameworks* mais utilizados para indexação e busca em arquivos de textos com código aberto. Essa biblioteca é desenvolvida e JAVA e pode ser utilizada por qualquer aplicação, inclusive aplicações desenvolvidas em outras linguagens, como Python, Delphi e PHP. Na abordagem proposta, essa biblioteca foi utilizada diretamente em parte da fase de extração de características dos documentos e da consulta (módulo auxiliar 1, na Figura 9) e no módulo principal, na etapa indexação dos documentos e ranking dos documentos retornados como resultado da consulta. Os demais módulos foram construídos e, nas seções subsequentes, todos os módulos do sistema serão mais bem detalhados.

4.2 Extração de Características

O processo de extração de características (Vide módulo auxiliar 1, na Figura 9) tem como objetivo preparar os documentos para construir o índice, por meio da redução do número de termos utilizados para indexar os documentos. O presente trabalho utiliza a estratégia de arquivo invertido, utilizada pelo *Apache Lucene*, para indexar o conteúdo dos documentos. Nesse método, todos os documentos são transformados em uma estrutura ordenada de palavras-chave, onde cada palavra tem uma lista de ponteiros para os documentos que contêm aquele termo. Esse processo consiste de duas etapas: (*i*) Análise textual e (*ii*) *Stemming*, como demonstrado na Figura 10.

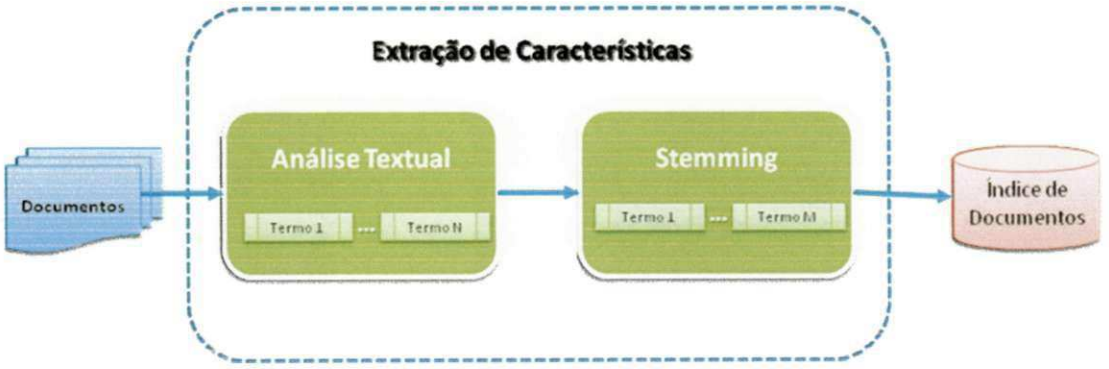


Figura 10- Etapas do Processo de Extração de Características.

[5] http://lucene.apache.org/

Na primeira etapa, os documentos são submetidos ao analisador textual, que identifica cada sequência de caractere, incluindo os espaços em branco e a pontuação. Assim, o documento é transformado numa sequência de palavras (*tokens*). Na Figura 11, é apresentado um exemplo de saída do texto após a execução do analisador léxico, que o separa em sequência de caracteres. Nessa etapa, foi utilizado o módulo léxico do próprio Apache Lucene.

Figura 11 - Conteúdo do Documento convertido em tokens.

Após a conversão do texto lido para uma sequência de *tokens*, eles são encaminhados para um módulo de filtragem textual. Nessa etapa, é realizada a verificação de termos irrelevantes que devem ser descartados, comumente conhecidos como *stop-words*. Na abordagem proposta, são consideradas *stop-words* as classes de palavras: pronomes, artigos, conjunções, variações de alguns verbos (ex. ter, ser, estar, dizer e fazer), entre outras (Vide Tabela 2). Nessa fase, são retiradas as pontuações e os acentos, e letras maiúsculas são substituídas por minúsculas. Logo, para essa etapa foi necessária a customização do módulo de detecção de *stop-words* do Lucene, para que as palavras presentes na Tabela 2 fossem contempladas pelo sistema, uma vez que a quantidade de palavras do lucene consideradas como *stop-words* era bastante reduzida.

Na Figura 12, apresenta-se a saída do documento depois que os termos apresentados na Figura 11 foram analisados. Isso significa que, no processo, os termos considerados irrelevantes foram descartados, os acentos presentes em alguns termos foram removidos, e os caracteres maiúsculos foram convertidos em minúsculos.

Figura 12 - Termos do Documento após a fase de Filtragem.

Tabela 2 - Stop Words utilizadas pela abordagem proposta.

STOP WORDS								
A	Contra	Deviam	Este	Mesmas	Nunca	Porque	Si	Últimos
agora	Contudo	Disse	Estes	Mesmo	O	Posso	sido	Um
ainda	Da	Disso	Estou	Mesmos	Os	Pouca	só	Uma
alguém	Daquele	Disto	Eu	Meu	Ou	Poucas	sob	Umas
algum	Daqueles	Dito	fazendo	Meus	Outra	Pouco	sobre	Uns
alguma	Das	Diz	Fazer	Minha	Outras	Poucos	sua	Vendo
algumas	De	Dizem	Feita	Minhas	Outro	Primeiro	suas	Ver
alguns	Dela	Do	Feitas	Muita	Outros	Primeiros	talvez	Vez
ampla	Delas	Dos	Feito	Muitas	Para	Própria	também	Vindo
amplas	Dele	E	Feitos	Muito	Pela	Próprias	te	Vir
amplo	Deles	Ela	Foi	Muitos	Pelas	Próprio	tem	Vos
amplos	Depois	Elas	For	Na	Pelo	Próprios	tendo	Vós
Ante	Dessa	Ele	Foram	Não	Pelos	Quais	tenha	
antes	Dessas	Eles	Fosse	Nas	Pequena	Qual	ter	
Ao	Desse	Em	Fossem	Nem	Pequenas	Quando	teu	
Aos	Desses	Enquanto	Grande	Nenhum	Pequeno	Quanto	teus	
Após	Desta	Entre	grandes	Nessa	Pequenos	Quantos	ti	
aquela	Destas	Era	Há	Nessas	Per	Que	tido	
aquelas	Deste	Essa	Isso	Nesta	Perante	Quem	tinha	
aquele	Deste	Essas	Isto	Nestas	Pode	São	tinham	
aqueles	Destes	Esse	Já	Ninguém	Pôde	Se	toda	
aquilo	Deve	Esses	No	Podendo	Seja	Todas	todo	
As	Devem	Esta	Nos	Poder	Sejam	Todavia	todos	
Até	Devendo	Está	Lá	Nós	Poderia	Sem	tu	
através	Dever	Estamos	Lhe	Nossa	Poderiam	Sempre	tua	
Cada	Deverá	Estão	Lhes	Nossas	Podia	Sendo	tuas	
coisa	Deverão	Estas	Ló	Nosso	Podiam	Será	tudo	
coisas	Deveria	Estava	Mas	Nossos	Pois	Serão	última	
Com	Deveriam	Estavam	Me	Num	Por	Seu	últimas	
como	Devia	estávamos	Mesma	Numa	Porém	Seus	Último	

Nesse trabalho, como processo de *stemming*, é utilizada uma adaptação do algoritmo de Porter para a língua portuguesa, chamada de "*removedor de sufixos da língua portuguesa*" ou RSLP (ORENGO; HUYCK, 2001). Esse algoritmo consiste de oito etapas; cada uma delas tem um conjunto de regras a serem avaliadas em sequência, mas apenas uma delas é aplicada. Em cada regra, são estabelecidos os sufixos a serem removidos, o tamanho mínimo do radical, um sufixo a ser adicionado ao radical, caso seja necessário, e a lista de palavras que são uma exceção à respectiva regra.

Na Figura 13, apresentam-se as oito etapas desse algoritmo, que são:

- **Redução do plural:** Nessa etapa, remove-se a letra "*s*" presente no final das palavras que não estão na lista de exceções. Entre as palavras presentes nessa lista, estão aquelas que não se referem à flexão numérica (Ex. gás, mês, lápis);
- **Redução do gênero feminino:** Aqui, as palavras do gênero feminino são transformadas no correspondente masculino. Nessa fase, apenas as palavras que terminem com a letra "*a*" são convertidas (Ex. Francesa →Francês);
- **Redução do advérbio:** Nessa etapa, reduzem-se as palavras que terminam com o sufixo "*mente*". Todavia, assim como na primeira etapa, existe uma lista de exceções, uma vez que nem todas as palavras que terminam com esse sufixo são advérbios (Ex. Semente);
- **Redução do aumentativo/diminutivo:** São considerados os sufixos mais comuns que indicam diminutivo ou aumentativo. Por exemplo, "*-inha*", "*-inho*", "*-ão*", etc. Assim como na primeira etapa, há uma lista de exceções, como as palavras: coração, sensação, entre outras;
- **Redução do substantivo:** São consideradas 61 terminações de palavras que indicam substantivos e adjetivos. Caso haja a remoção do sufixo nessa etapa, os próximos passos não serão executados. Por exemplo, sufixos como "agem" → "coragem", "carruagem", "chantagem";
- **Redução do verbo:** Nessa etapa, os verbos são reduzidos para o seu radical, uma vez que são formados de acordo com a estrutura: radical + vogal temática + desinência (Ex. Venderam → *Vend* + *e* + *ram*).; e
- **Remoção da vogal:** Nessa etapa, as últimas vogais das pmalavras que não sofreram processo de redução nas fases de redução do substantivo e de redução do verbo serão removidas. Por exemplo, a palavra "garoto", que não foi reduzida em nenhuma das etapas anteriores teria a última letra "o" removida nessa etapa. Desta forma, o *stem* resultante seria "garot", assim como outras variações da palavra, como "garota", "garotão", "garotinho".

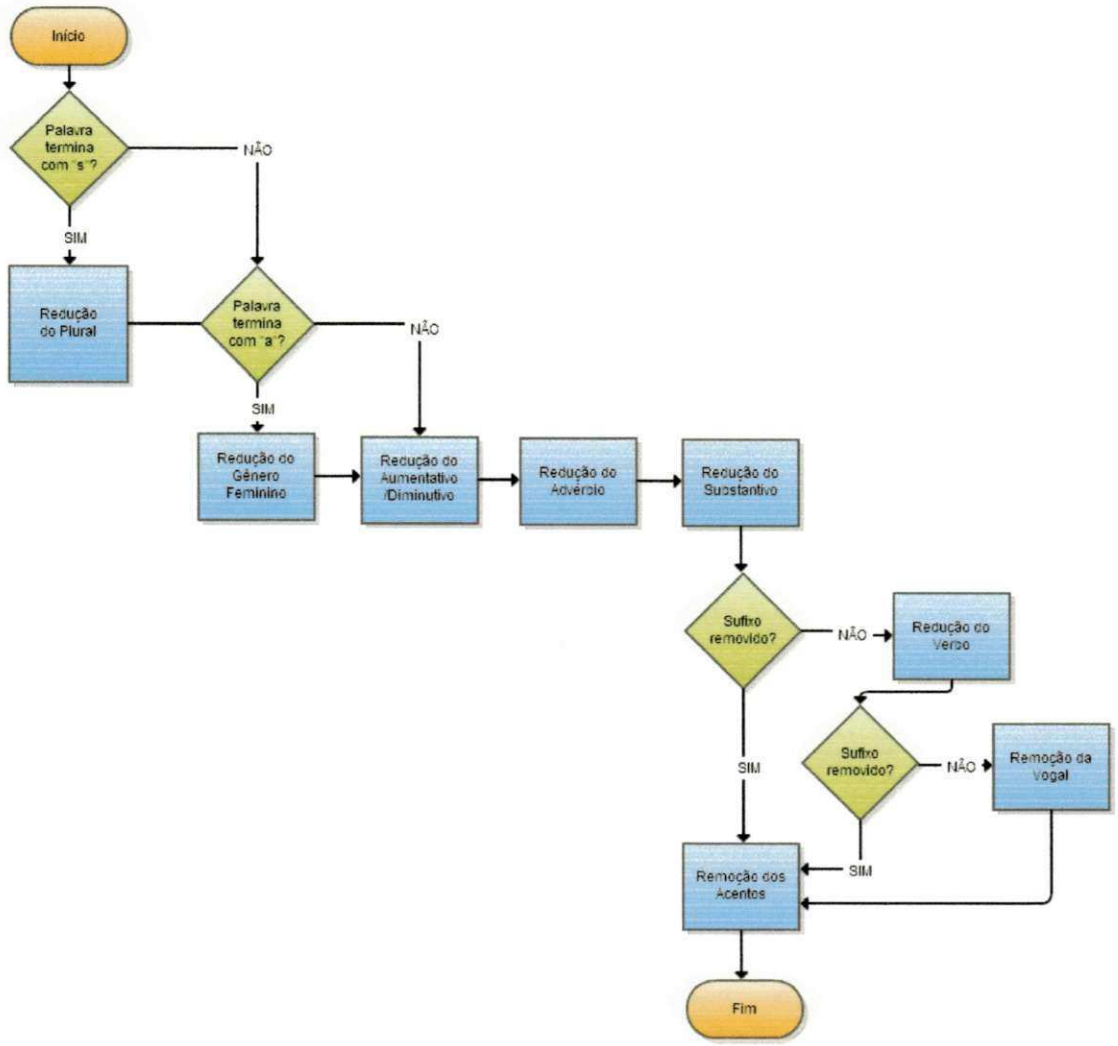


Figura 13 - Etapas do algoritmo de Stemming (Adaptado de (ORENGO; HUYCK, 2001))

Como a biblioteca *Apache Lucene* utiliza a técnica de Porter original como algoritmo padrão para a fase de *Stemming* e esse possui um desempenho inferior ao RSLP, foi necessária a customização do processo de *stemming* da biblioteca com esse algoritmo.

Na abordagem proposta, o conjunto de documentos do sistema é representado como vetores em um mesmo espaço vetorial com N dimensões, em que N representa a quantidade de termos distintos existentes na coleção de documentos. Depois das fases de análise textual e de *stemming*, todos os documentos são transformados em vetores, e nas posições que correspondem aos termos dos próprios documentos, existe um peso associado, que quantifica a relevância desse termo no documento. Nas posições relativas aos termos ausentes no documento, é atribuído o valor zero. Por exemplo:

numa coleção de documentos composta por três documentos - A, B e C - onde o documento A tem o texto "casa grande azul", o documento B é formado pelo texto "carro pequeno azul", e no documento C, tem o texto "carro pequeno verde". Para cada documento será construído um vetor com seis posições, que representa cada termo distinto existente na coleção, e nas posições relativas a termos existentes nos respectivos documentos, haverá um peso associado (Vide Figura 14).

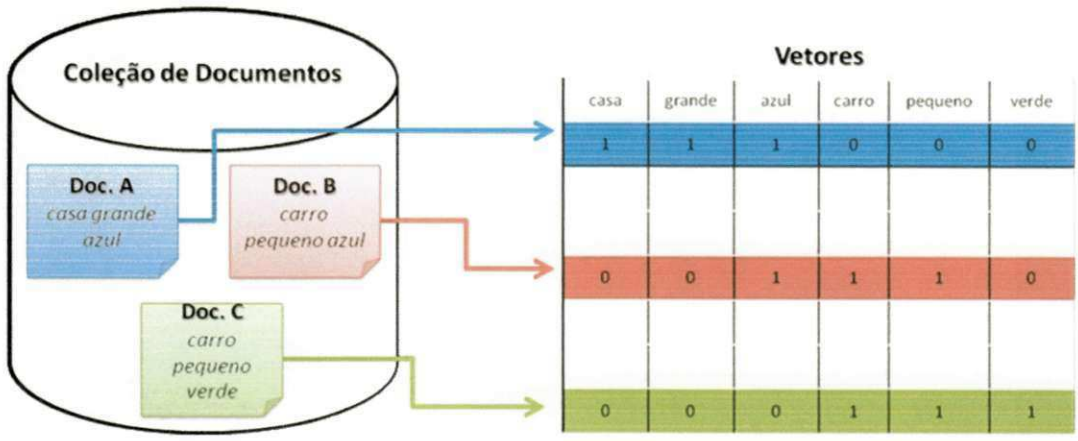


Figura 14 - Representação dos documentos em Vetores.

Neste trabalho, como peso para cada termo do documento, será utilizado o produto entre a frequência do termo no documento e o número de vezes em que esse termo aparece em toda a coleção, também conhecido como *tf-idf*. Considere, por exemplo, um documento com 100 termos, onde a palavra "farinha" aparece seis vezes. Nesse caso, a frequência do termo (*tf*) farinha é de $tf = 6/100 = 0{,}06$. Considerando que tenhamos 10 milhões de documentos e que, em 1000 deles, apareça a palavra "farinha", a frequência inversa do documento (*idf*) é calculada como sendo $\text{idf} = \log(10000000/1000) = 4$. Logo, o peso (*tf-idf*) do termo "farinha" será $tf - idf = 0{,}06 \times 4 = 0{,}24$.

Quando uma consulta q é submetida ao sistema, é calculada uma medida de relevância dos documentos da coleção em relação à consulta e, a partir dela, é construído um ranking de documentos relevantes que compõem o resultado dessa consulta. Portanto, assim como os documentos, a consulta também é representada como um vetor e, consequentemente, é calculada uma medida de distância entre o vetor da consulta $\vec{v}(q)$ e os vetores dos documentos $\vec{v}(d)$. Neste trabalho, a medida de

distância utilizada foi o cosseno entre os vetores, que pode variar entre 0 e 1. Isso quer dizer que, quanto mais próximo de 1 o valor dessa medida, mais relevante no documento é a consulta submetida. Essa medida é definida de acordo com a Equação 10.

$$relevancia(q,d) = \frac{\vec{v}(q) \cdot \vec{v}(d)}{|\vec{v}(q)||\vec{v}(d)|} \quad (10)$$

No algoritmo da construção do ranking de documentos relevantes a uma consulta submetida ao sistema, inicialmente, é construída uma lista que contém os documentos e o valor do cálculo da relevância entre o vetor da consulta e o vetor de cada documento da coleção. Depois dessa etapa, a lista é ordenada de forma decrescente, através da medida de relevância, e os documentos são exibidos ao usuário de acordo com essa ordem (Ver Código 1).

```
1  RelevanciaCosseno(q)
2      medidasRelevancia[D] = 0
3      for each documento d in D do
4          medidaRelevancia[d] = medidaCosseno(q,d)
5      ordenaMaisRelevantes(medidasRelevancia[D])
6      return medidasRelevancia[D]
```

Código 1 - Algoritmo elaboração do ranking de relevância dos documentos.

4.3 Georreferenciamento do Tesauro

Os tesauros vêm sendo utilizados frequentemente, nos sistemas de recuperação da informação, para identificar termos ou expressões que tenham uma relação semântica. Geralmente, eles são utilizados no processo de expansão da consulta para selecionar os termos que serão acrescentados na consulta original do usuário, com o objetivo de retornar mais documentos relevantes. Porém, o problema dessa técnica é selecionar quais palavras devem ser expandidas e se as palavras acrescentadas realmente contribuem para a melhoria, uma vez que esses termos podem não fazer parte

do contexto de intenção do usuário, o que ocasiona a recuperação de documentos irrelevantes.

É importante ressaltar que, além das relações semânticas já encontradas nos tesauros, existem outras informações relativas ao contexto que podem auxiliar na escolha dos termos que devem ser acrescentados à consulta original do usuário. Portanto, este trabalho propõe o uso do contexto geográfico como forma de filtrar os termos a serem acrescentados na consulta original, visto que determinados termos são mais utilizados e, consequentemente, mais característicos de localidades geográficas específicas.

Atualmente, existem diversos tesauros geográficos, ou seja, os que têm informações geográficas associadas aos termos. Porém quase todos são voltados para termos que identifiquem locais geográficos, como, por exemplo, o *Getty TGN*[6]. Nesse sentido, este trabalho propõe um método de georreferenciamento automático dos termos do tesauro. Para isso, é necessário definir qual o menor e o maior contexto geográfico possível associado aos termos. Aqui, serão utilizados contextos geográficos relativos ao Brasil, e os menores contextos geográficos são os vinte e seis estados da Federação, juntamente com o Distrito Federal. O maior será o país como um todo. Além desses contextos, serão considerados contextos intermediários relativos às cinco regiões geográficas existentes no país (Ex. Norte, Sul, Nordeste, Centro-oeste e Sudeste). Na Figura 15, pode ser vista a hierarquia desses contextos. Juntamente com cada contexto escolhido, é armazenada a coordenada geográfica (latitude e longitude) associada. Esse mapeamento entre os nome do contexto geográfico e sua coordenada foi realizado através do Gazetteer GeoNames[7] e a sua utilidade será descrita na Seção 4.4.

Para o georreferenciamento automático dos termos do tesauro, utilizou-se, neste trabalho, a *Google Custom Search API*[8], um framework de desenvolvimento de sites e aplicações para realizar busca e exibir esses resultados por meio do *Google*. Com essa API, é possível realizar requisições RESTful (FIELDING, 2000) e obter os resultados da busca no Google em formato JSON[9]. O método consiste, basicamente, em tentar

---

[6] *http://www.getty.edu/research/tools/vocabularies/tgn/index.html*

[7] *http://www.geonames.org*

[8] *https://developers.google.com/custom-search/v1/overview*

[9] *http://www.json.org/*

UFCG/BIBLIOTECA/BC

estimar o contexto geográfico do termo, com base no número de ocorrências encontradas numa consulta composta pelo termo e um escopo geográfico no Google.

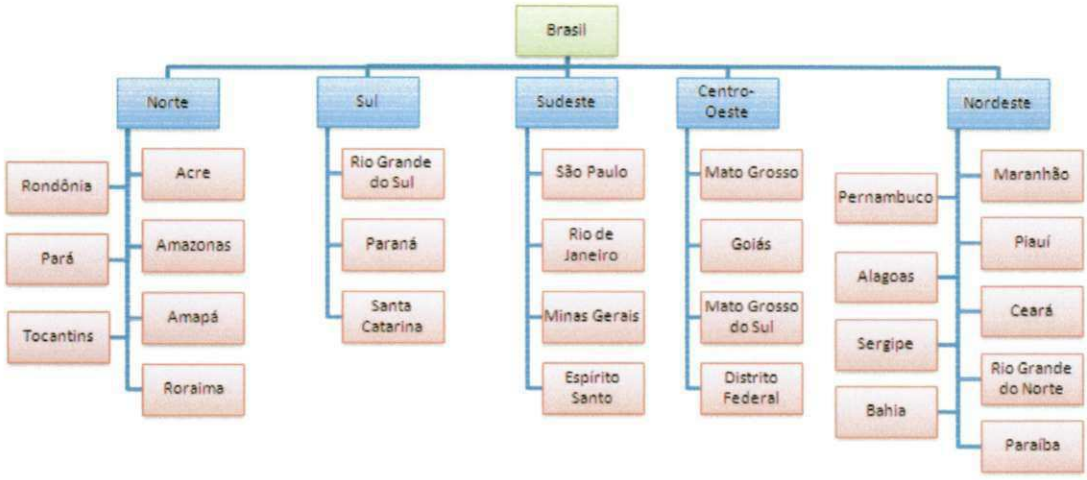


Figura 15 - Hierarquia dos Contextos Geográficos.

Dessa forma, foram construídas consultas compostas por um termo do tesauro e uma localidade geográfica, as quais foram submetidas ao motor de busca do Google por meio de uma requisição RESTful. Nessa requisição, alguns parâmetros são necessários, como: uma chave de uso da API, gerada a partir de uma conta Google, e os termos da consulta que se deseja submeter ao motor de busca. Na Figura 16 é apresentado a requisição utilizando a consulta "Castanha do Pará + Acre" e o resultado da busca no formato JSON são apresentados na Figura 16.

A partir dos resultados retornados, foi construído um módulo (Vide módulo auxiliar 2, na Figura 9) para a leitura desses arquivos e, para cada termo do tesauro, foi feita uma lista de localidades geográficas e o número de resultados (páginas web) relativo ao termo. Essas listas serão utilizadas para determinar os escopos geográficos do termo.

Assim, seja T um tesauro composto pelo conjunto n de rede de sinônimos S, onde cada rede de sinônimos (s_i) é composta por m termos semanticamente relacionados. Seja E o conjunto de escopos geográficos de menor nível na hierarquia dos escopos. Na abordagem proposta, em cada rede de sinônimo (s_i), são criadas consultas compostas

pelos seus termos (t) e um escopo geográfico ($e \in E$) que são submetidas a um motor de

busca, e o número de resultados retornados pela consulta é armazenado. Nesse trabalho, considera-se rede de sinônimos o termo e as palavras semanticamente relacionadas, ou seja, as palavras que possuem relação de sinonímia definida pelo tesauro.

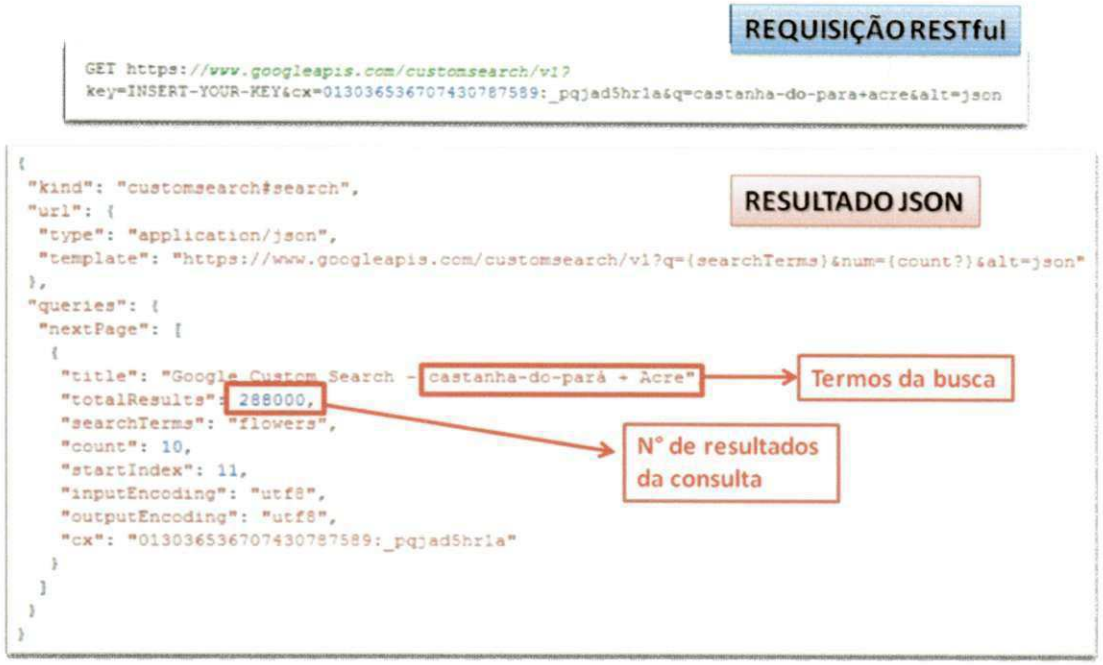


Figura 16 - Requisição e o resultado de uma busca utilizando Google Custom Search API.

. Por exemplo, os termos "*munguzá*", "*canjica*" e "*canjicão*" se referem a um prato culinário à base de grãos de milho cozidos em leite de coco, polvilho e canela em pó e que, no tesauro compõem uma rede de sinônimos. Considere um conjunto de escopos geográficos com apenas três localidades geográficas: Paraíba, São Paulo e Minas Gerais. Dessa forma, serão criadas para cada termo, três consultas, são elas: "*munguzá+Paraíba*", "*munguzá+Minas Gerais*", "*munguzá+São Paulo*", "*canjica+Paraíba*", "*canjica+Minas Gerais*", "*canjica+São Paulo*", "*canjicão+Paraíba*", "canjicão+Minas Gerais" e "canjicão+São Paulo". Para cada consulta, é coletado o número de resultados retornados pelo motor de busca.

Em seguida, dentro de cada rede de sinônimos e de cada escopo geográfico, é feita uma normalização através da consulta "termo + escopo geográfico", que retornou mais resultados. Após essa etapa, para cada termo, a lista de escopos geográficos é ordenada em ordem decrescente, e os k primeiros escopos da lista são considerados os escopos geográficos do respectivo termo. No exemplo anterior, caso a consulta "*munguzá+Paraíba*" retorne 500 resultados, e as consultas "*canjica+Paraíba*" e

"*canjicão+Paraíba*" retornem, respectivamente, 250 e 100, o número de resultados da primeira consulta será utilizado como fator de normalização. Assim, para o escopo "*Paraíba*", o índice relativo do termo "*munguzá*" terá valor igual a 500/500 = 1; para o índice do termo "*canjica*", o valor será igual a 250/500=0,5; e para o termo "*canjicão*", o valor será de 100/500=0,2. Esse mesmo processo se repete nos demais escopos e, no final, para cada termo, tem-se uma lista ordenada, em que os k primeiros termos são o escopo geográfico do respectivo termo. Para uma lista em que o k=1, o termo "*munguzá*" teria como escopo geográfico a Paraíba, como demonstrado na Figura 17.

No Código 2, apresenta-se o algoritmo responsável por calcular o índice de relevância dos termos do tesauro para os escopos geográficos. A relevância de um escopo *e* para um termo *t* é medida em relação aos demais termos da rede de sinônimos da qual ele faz parte.

```
1  CalculaEscopoGeografico(Si)
2      For each escopo e in E do
3          maxResultado = 0
4          For each termo t in Si do
5              If maxResultado < t.getNumeroResultados(e) then
6                  maxResultado = t.getNumeroResultados(e)
7          For each termo t in Si do
8              t.normalizaResultado (maxResultado, e)
9              t.ordenaResultados()
```

Código 2 - Algoritmo para o Cálculo do Contexto Geográfico dos termos do tesauro.

Caso os *k* primeiros escopos sejam todos pertencentes a uma mesma região geográfica, são substituídos pelo contexto relativo à região. Por exemplo, se um termo tem, nas três primeiras posições, os contextos Rio Grande do Sul, Paraná e Santa Catarina, eles serão substituídos pelo escopo geográfico relativo à Região Sul.

Por fim, é importante destacar que o processo de georreferenciamento do tesauro é uma etapa externa ao processo de recuperação da informação, uma vez que a abordagem proposta precisa do tesauro georreferenciado para que a expansão de consultas com base no contexto geográfico seja realizada.

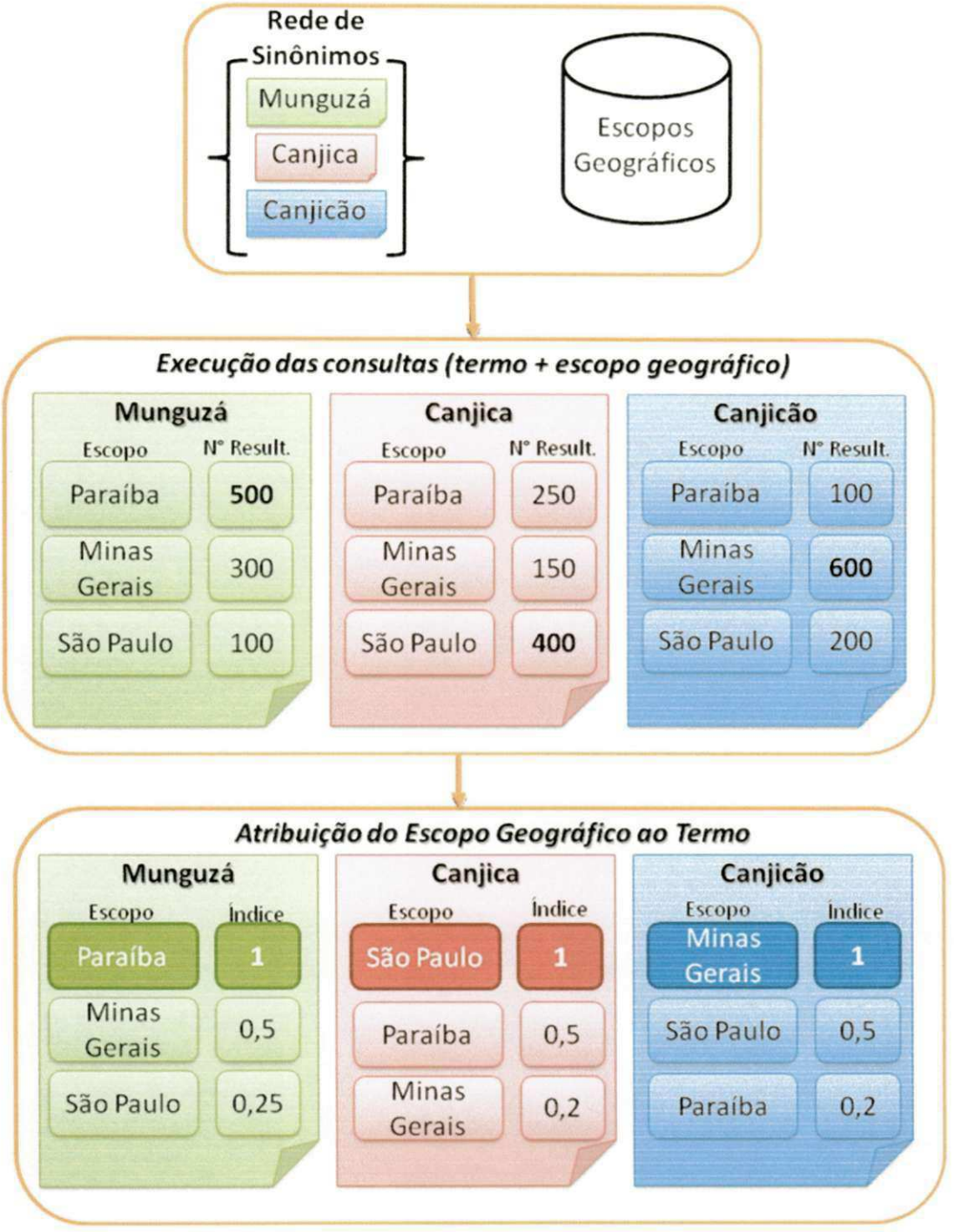


Figura 17 - Processo de Georreferenciamento Automático dos Termos.

4.4 Expansão de Consultas

De maneira geral, se, em uma consulta, as palavras informadas pelo usuário não estiverem presentes nos documentos da coleção, eles não serão recuperados pelo

sistema de busca. Logo, o principal objetivo da expansão de consulta é reduzir a incongruência dos termos do documento com os termos informados pelo usuário, por meio de palavras com significado similar.

O algoritmo de expansão proposto (Vide módulo principal, na Figura 9) neste trabalho é composto de duas etapas (Figura 18). Na primeira, é feita uma avaliação semântica dos termos da consulta, de modo a se obterem termos relacionados. Por fim, é feita uma filtragem da lista dessas palavras sugeridas no passo anterior e, como resultado, gera-se uma nova consulta composta pelos termos informados originalmente pelo usuário e os termos filtrados.

Figura 18 - Etapas do Processo de Expansão de Consultas.

Assim como os documentos da coleção, a consulta também é representada como um vetor, onde cada termo tem um peso associado e é uma posição desse vetor. Na primeira etapa do processo de expansão de consulta, o vetor dos termos da consulta original é avaliado à procura de termos que tenham sinônimos presentes no tesauro. Assim, caso sejam encontrados, esses termos são adicionados ao vetor da consulta. Por exemplo, uma consulta por "*bolo de tangerina*" é submetida ao sistema; essa consulta é transformada em um vetor, em que as posições relativas aos termos "*bolo*" e "*tangerina*" terão um peso associado. Depois disso, é identificado que a palavra "*tangerina*" tem termos relacionados (bergamota, laranja cravo, etc.) no tesauro. Portanto, o vetor original passa a ter, além dos termos "*bolo*" e "*tangerina*", os vocábulos "*bergamota*" e "*laranja cravo*".

Nesse cenário, não há uma distinção entre os novos termos que foram acrescidos à consulta original. Por exemplo, resultados que contenham o termo "laranja cravo" ou "bergamota" serão apresentados, o que pode levar a um número muito grande de resultados encontrados, porém sem um critério específico de relevância. Por essa razão,

neste trabalho, o processo de expansão de consulta apresenta uma segunda etapa, responsável por reduzir esse número de resultados apresentados que, muitas vezes, não é relevante para a consulta do usuário.

Na segunda etapa, caso se encontre, na consulta original, algum termo relativo a um contexto geográfico, os termos adicionados na primeira fase são filtrados de acordo com a proximidade, ou seja, a distância entre o contexto geográfico da consulta e o contexto geográfico dos termos do tesauro. No exemplo anterior, se a consulta original fosse "*Bolo de Tangerina do Rio Grande do Sul*", apenas o termo "*bergamota*" seria adicionado à consulta original, uma vez que o termo "laranja cravo" não é característico dessa localidade. Porém, em alguns cenários, pode não haver nenhum termo relacionado com o contexto geográfico informado pela consulta, como, por exemplo, se a consulta fosse "*Bolo de Laranja do Uruguai*", nesse cenário, nenhum dos termos sinônimos de laranja teria o Uruguai como contexto geográfico. Consequentemente, nenhum termo seria expandido.

Assim, a abordagem proposta usa, como critério para a filtragem dos termos, a distância entre a localização geográfica informada na consulta e os contextos geográficos dos termos do tesauro. Portanto, quando uma localização geográfica é detectada nos termos da consulta, ela é mapeada para sua respectiva coordenada de latitude e longitude, da mesma forma, como os contextos geográficos dos termos do tesauro foram mapeados no processo de georreferenciamento. Para o cálculo da distância, caso haja mais de um contexto associado ao termo do tesauro, é considerado o mais próximo à localidade geográfica existente na consulta.

No exemplo anterior, para a consulta "*Bolo de Tangerina do Uruguai*", o termo "*tangerina*" será expandido para "*bergamota*", uma vez que esse termo tem como contexto o Rio Grande do Sul, que é mais próximo do Uruguai do que os contextos dos demais termos da rede de sinônimos.

No final do processo de expansão da consulta, ela é submetida ao índice de documentos, e os mais relevantes à consulta são ranqueados e apresentados de acordo com a medida de relevância descrita na Equação 10.

4.5 Considerações Finais

Apresentamos, neste capítulo, a abordagem proposta - um sistema de recuperação que utiliza uma técnica de expansão de consulta e o contexto geográfico de palavras que não se referem diretamente a uma localidade geográfica para melhorar a eficiência no processo de recuperação da informação. Inicialmente, apresentamos uma visão geral do sistema; na sequência, descrevemos os principais módulos do sistema e as tecnologias utilizadas. No capítulo seguinte, tecemos algumas considerações sobre a avaliação experimental apresentada, em relação a um sistema de recuperação da informação tradicional baseado no modelo vetorial e o mesmo sistema com um processo de expansão de consulta baseado em redes de sinônimos.

Capítulo 5

Avaliação Experimental

Neste capítulo, apresentamos os resultados de alguns experimentos realizados com a abordagem proposta no Capítulo 4. Os testes foram realizados com o objetivo de mensurar a eficiência da abordagem proposta em relação a sistemas tradicionais de recuperação da informação.

Para avaliar a abordagem proposta, foi escolhido o domínio das receitas culinárias, uma vez que o vocabulário utilizado nas receitas varia de acordo com a localização geográfica de origem da receita. Nesse cenário, diversos ingredientes e receitas têm nomes diferentes de acordo com a região do país. Por exemplo, a "*abóbora*", em algumas regiões, é conhecida como "*jerimum*"; já em outras, é identificada pelo termo "*cucúrbita*".

Inicialmente, são apresentadas as construções da coleção de documentos do sistema de recuperação e do tesauro necessário para o processo de expansão semântica das consultas. Em seguida, são analisadas a precisão e o *recall* do processo de recuperação da informação, com o auxílio da técnica de expansão de consultas proposta em relação às abordagens tradicionais. Também é feita uma análise do impacto do processo de georreferenciamento dos termos do tesauro na abordagem proposta. Por fim, são feitas considerações sobre os experimentos realizados.

5.1 Base de Dados

A base de receitas foi construída a partir de dois sites: (*i*) Comida e Receitas[10]; e (*ii*) Receitas Típicas[11]. A escolha por esses dois sites se justifica devido à existência de uma divisão das receitas por região geográfica, mais precisamente, entre os 26 estados brasileiros. Essa característica é importante, pois garante que o sistema terá receitas distribuídas por toda a extensão geográfica considerada pelo sistema e a possibilidade de validar a hipótese de que os termos utilizados pelo usuário são influenciados pela sua região de origem.

A coleta dessas receitas foi feita através da ferramenta de extração de dados da Web, *WebHarvest*[12]. Essa ferramenta possibilita coletar páginas de internet e extrair delas informações desejadas a partir de técnicas e tecnologias para manipulação de XML e HTML, como XSLT, *XQuery* e expressões regulares. O processo de extração foi descrito através de um arquivo XML de configuração para cada site, porque eles apresentam estruturas lógicas distintas. Na Figura 19, é apresentado um exemplo desse arquivo.

Cada receita extraída foi transformada em uma estrutura XML, composta por três atributos: título, ingredientes e o modo de preparo. Após a extração, foi realizada uma etapa de pré-processamento, com o objetivo de remover as quantidades dos ingredientes, posto que essas quantidades não são relevantes para identificar as receitas. Nessa fase, também é acrescentado um atributo de identificação única na receita, para facilitar a recuperação e a identificação da receita no sistema. Na Figura 20, é possível observar destacadas, em vermelho, as quantidades dos ingredientes da receita e, em verde, outras modificações necessárias realizadas na fase de pré-processamento, como a separação de termos que são concatenados no processo de coleta das receitas nas páginas de internet.

---

[10] http://www.comidaereceitas.com.br/

[11] http://www.receitastipicas.com/

[12] http://web-harvest.sourceforge.net/

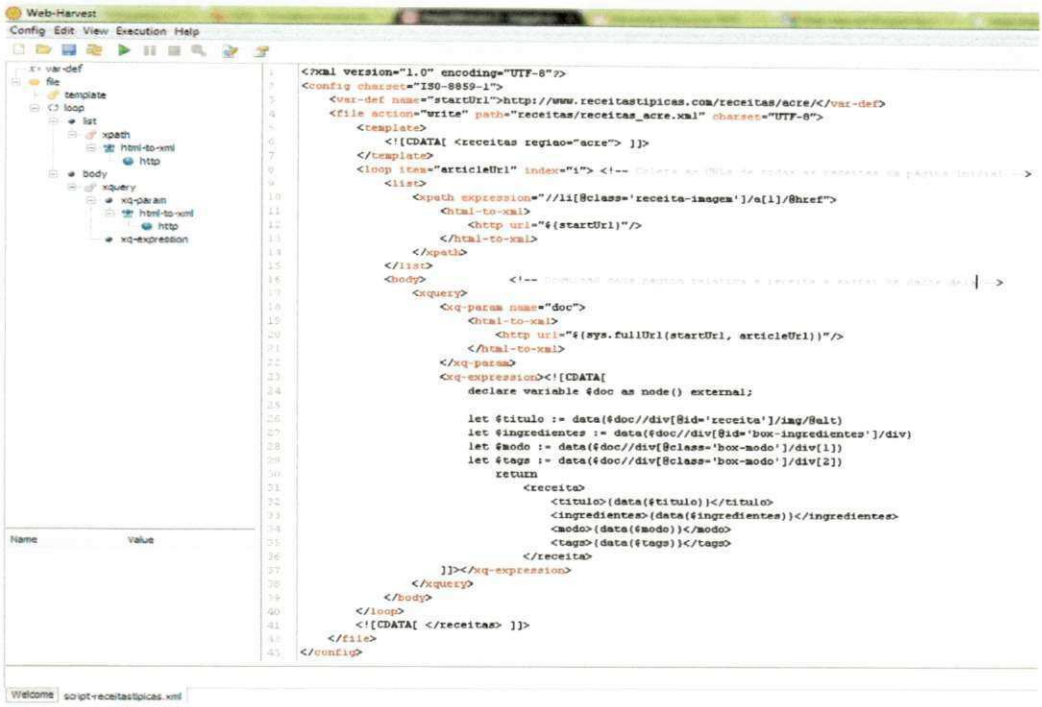


Figura 19 - XML de Configuração do *WebHarvest* para Coleta das Receitas do Site Receitas Típicas.

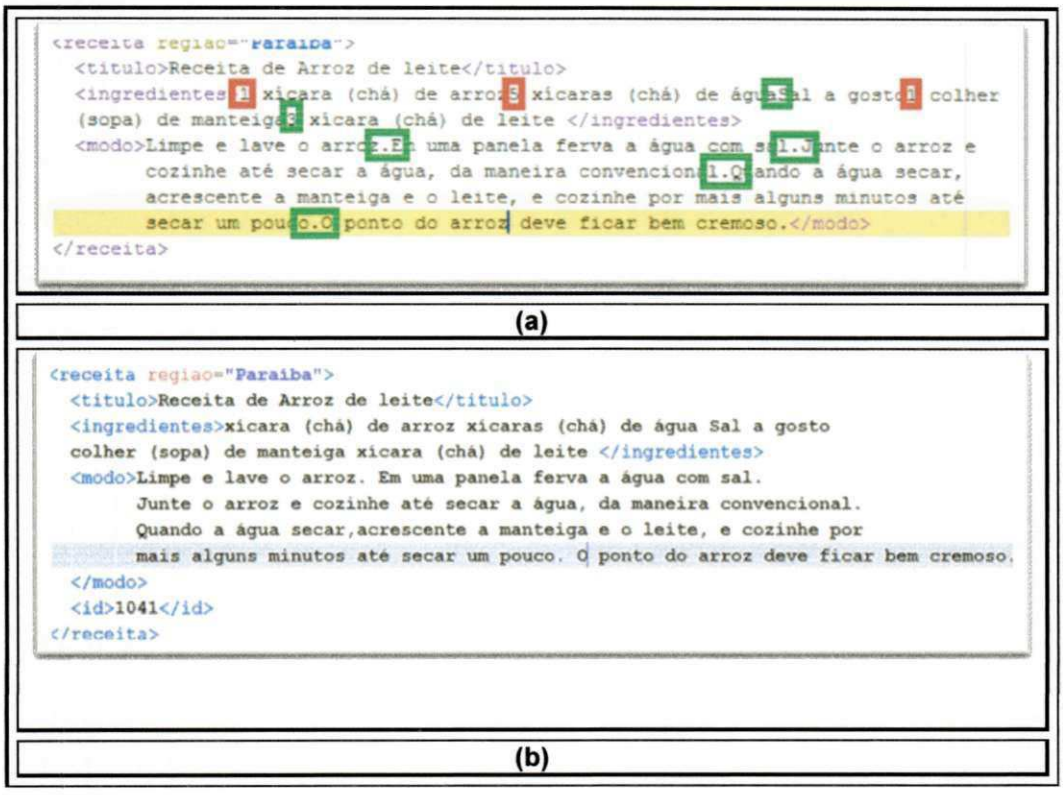


Figura 20- Receitas em XML: (a) XML extraído da página web; (b) XML resultante após a fase de reprocessamento.

No total, foram coletadas 2.292 receitas, sendo 1.783 retiradas do site Comida e

Receitas e 509 receitas coletadas do site Receitas Típicas. Na Tabela 3, apresenta-se um detalhamento da coleta das receitas por estado.

Tabela 3 - Detalhamento da Base de Receitas por Estado.

Estado	Sigla	N° de Receitas "Comidas e Receitas"	N° de Receitas "Receitas Típicas"	Total
Acre	AC	37	20	**57**
Alagoas	AL	75	7	**82**
Amapá	AP	0	10	**10**
Amazonas	AM	87	20	**107**
Bahia	BA	83	46	**129**
Ceará	CE	93	17	**110**
Distrito Federal	DF	0	0	**0**
Goiás	GO	78	34	**112**
Espírito Santo	ES	99	20	**119**
Maranhão	MA	69	10	**79**
Mato Grosso	MT	108	12	**120**
Mato Grosso do Sul	MS	0	44	**44**
Minas Gerais	MG	225	46	**271**
Pará	PA	87	40	**127**
Paraíba	PB	90	19	**109**
Paraná	PR	84	14	**98**
Pernambuco	PE	75	12	**87**
Piauí	PI	45	12	**57**
Rio de Janeiro	RJ	78	23	**101**
Rio Grande do Norte	RN	75	6	**81**
Rio Grande do Sul	RS	90	27	**117**
Rondônia	RO	0	10	**10**
Rorâima	RR	0	6	**6**
São Paulo	SP	75	27	**102**
Santa Catarina	SC	85	13	**98**
Sergipe	SE	45	6	**51**
TOTAL		**1783**	**509**	**2292**

5.2 Tesauro

A abordagem proposta neste trabalho usa um tesauro para expandir as consultas submetidas ao Sistema. Como o domínio escolhido para validarmos a abordagem

UFCG/BIBLIOTECA/BC

proposta foi o das receitas culinárias, é necessário um tesauro do mesmo domínio, que tenha termos relativos a possíveis ingredientes das receitas, nomes de receitas, modos de preparo etc. Devido à especificidade do domínio, nenhum tesauro foi encontrado no estado da arte para esse domínio.

Diante disso, para validar o presente trabalho, foi escolhido um tesauro relativo à cadeia alimentícia, desenvolvido pela empresa *Infothes*[13] e disponível na internet. Esse tesauro foi construído com base nos conceitos da cadeia produtiva, incluindo matérias-primas, máquinas, produtos intermediários e finais, a distribuição e a comercialização desses produtos. Esses conceitos são divididos em dez categorias (Tabela 4), porém, para este trabalho, foram utilizadas apenas duas, porque podem ser empregadas no domínio das receitas culinárias: Alimentos e Bebidas.

Tabela 4 - Categorias do tesauro Cadeia Alimentícia.

CATEGORIAS
ALIMENTOS
BEBIDAS
COMPONENTES DOS ALIMENTOS
DOENÇAS E SAÚDE
HIGIENE DO ALIMENTO
POLÍTICAS PÚBLICAS PARA ALIMENTAÇÃO
PRODUÇÃO E COMERCIALIZAÇÃO DE ALIMENTOS
PROPRIEDADES DOS ALIMENTOS
QUÍMICA DOS ALIMENTOS
SEGURANÇA ALIMENTAR E NUTRICIONAL

Assim como na construção da base dados, foi preciso utilizar o *WebHarvest* para coletar os termos do tesauro, uma vez que ele está disponível apenas na Internet e não pode ser utilizado diretamente no sistema. Assim, foram coletados todos os termos do tesauro e armazenados em XML, como mostrado na Figura 21. Juntamente com cada termo, são coletadas as relações de hiperonímia ou hiponímia indicadas pelas tags TG (Termo Geral) e TE (Termo Específico), pelas relações de equivalência representadas pelas tags UP (Usado Para) e USE (Use) e pelas relações associativas, ou seja, as

[13] http://www.thesaurus.eti.br/

relações não hierárquicas, como por exemplo, antonímia, indicadas pela tag TR (Termo Relacionado) e as notas utilizadas para conceituar ou fornecer outras informações que ilustram o termo, incluindo: NE (Nota de Escopo), responsável por restringir ou expandir a aplicação desse termo, NC (Nome Científico), quando o termo tem um nome científico, como, por exemplo, as frutas, e NI (Informações Complementares), responsável por fornecer informações que complementam o significado desse termo e o individualizam em relação aos demais.

```
<palavra letra="A">
  <termo>ABACAXI</termo>
  <FAC>FR FRUTAS </FAC>
  <NC>Ananas Comosus  da família das Bromeliácea</NC>
  <NE>Fruta cítrica rica em potássio, magnésio, cálcio e vitaminas A, B1 e C.
  Fruto do abacaxizeiro,  tem forma cilíndrica ou cônica (frutos maiores na base),
  com rebentos na base e coroa de folhas no ápice, sua polpa é sucosa,
  saborosa e ligeiramente ácida, de consumo in natura. </NE>
  <NI>A norma de classificação do abacaxi, estabelecida pela Instrução Normativa Nº 1,
  de 1 de fevereiro de 2.002, do MAPA, caracteriza o abacaxi em 2 grupos, de acordo
  com a cor da polpa e 4 subgrupos, de acordo com a cor da casca. A coloração da
  casca do abacaxi depende das características climáticas da região produtora, da época
  de produção e do período de colheita.</NI>
  <TE>ABACAXI HAVAÍ | ABACAXI PÉROLA | ABACAXI PEROLERA | ABACAXI PRIMAVERA</TE>
  <TG>FRUTA CÍTRICA</TG>
  <TR>CÁLCIO | COMPOTA | FÓSFORO |  FRUTA TROPICAL |
  GELÉIA DE FRUTA | VITAMINA A | VITAMINA C</TR>
  <UP>ANANÁS</UP>
  <USE></USE>
  <id>2</id>
</palavra>
```

Figura 21 - Termo do tesauro em XML

Para a abordagem proposta, foram utilizadas apenas as tags de equivalência, a partir das quais foram criados redes de sinônimos, que serão utilizados no algoritmo de expansão descrito no capítulo anterior. Logo após esse processo, o tesauro ficou com 506 termos agrupados em 188 redes de sinônimos. Dentre todos os termos presentes no tesauro, os que não possuíram sinônimos, de acordo com o tesauro, foram descartados nesse processo.

5.3 Abordagens Comparadas

Em nível de comparação com o trabalho proposto, foram escolhidas duas abordagens a fim de se comparar com a abordagem proposta. A primeira é um sistema

de recuperação da informação tradicional, baseado no modelo vetorial, ou seja, os documentos e as consultas a ele submetidas são representados como vetores, e cada posição dessa estrutura corresponde a um termo existente no documento/consulta. Além disso, cada termo tem um peso associado, que representa a sua relevância em relação ao documento. Para computar esse peso, foi utilizado o TF-IDF, que representa o produto entre a frequência do termo no documento e a frequência inversa desse termo em todo o conjunto de documentos. Assim como a abordagem proposta, através da medida cosseno é calculada a relevância dos documentos para a consulta inicial e a partir dessa medida um ranking de documentos é retornado como resultado da consulta. A Figura 22 dá uma visão geral dessa abordagem, que foi escolhida com o objetivo de simular um motor de busca tradicional - o Google.

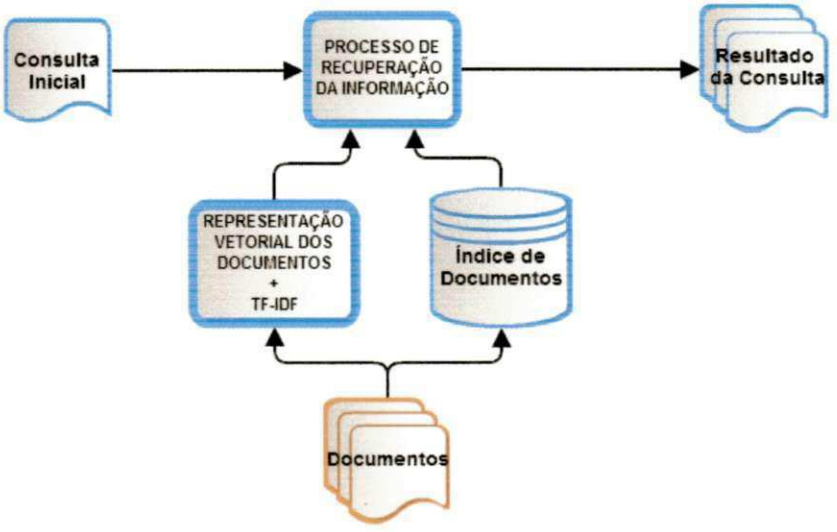


Figura 22 - Visão geral da Abordagem tradicional.

Assim como a primeira, a segunda abordagem também é baseada no modelo vetorial com TF-IDF para representar o peso dos termos da coleção de documentos. Além disso, utiliza-se uma técnica de expansão de consultas tradicional com um tesauro, que é o mesmo empregado pela abordagem proposta neste trabalho. Porém os termos não são georreferenciados. Dessa forma, o processo de expansão de consulta consiste, basicamente, no acréscimo de termos relacionados aos presentes nas consultas que também estão nesse tesauro. Assim como a abordagem tradicional, a partir da medida cosseno é elaborado um ranking de documentos relevantes à consulta expandida como resultado do processo de recuperação da informação. A Figura 23 apresenta uma

visão geral dessa abordagem.

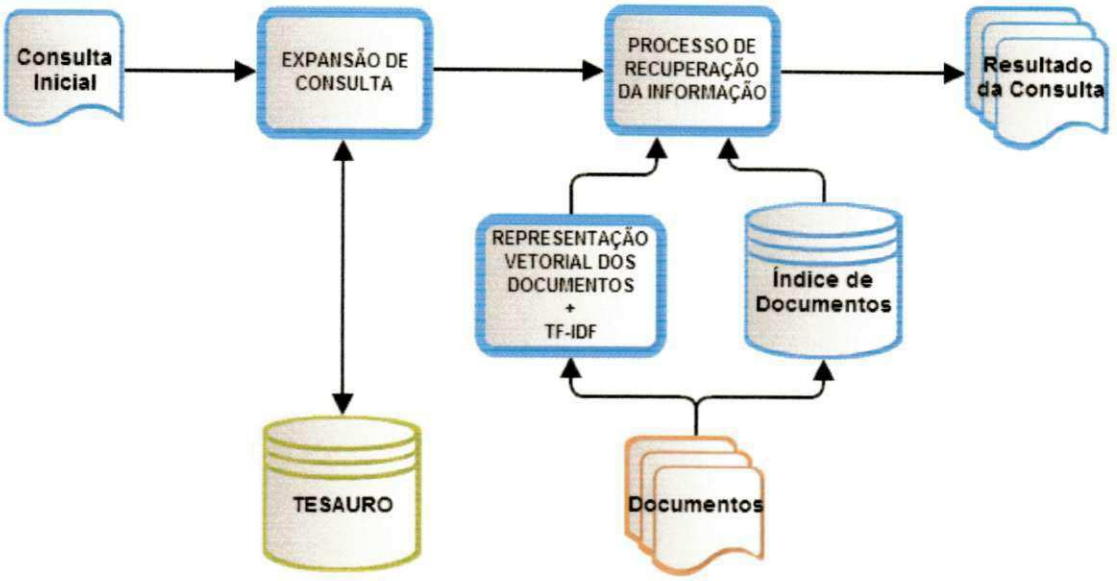


Figura 23 - Abordagem Tradicional com Expansão de Consultas.

5.4 Expansão Semântica de Consultas com o Auxílio de Geoprocessamento

Para avaliar a abordagem proposta, foram escolhidos casos de testes que envolvem o contexto geográfico e que, geralmente, estão presentes nos motores de busca. O cenário escolhido foi o seguinte: "*Um usuário vai visitar um novo lugar e está interessado em receitas características dessa região. Porém, ao realizar a busca no sistema, ele utiliza um vocabulário específico da sua região que, por sua vez, não é conhecido no destino a ser visitado*". Em outras palavras, em sua busca, o usuário usa um ingrediente ou um nome de uma receita, determina o seu objeto de interesse e, juntamente com um termo relativo à localidade geográfica, informa em que local ele deseja buscar as informações.

Foram construídos 131 casos de testes compostos por uma consulta contendo: um ou mais termos relativos ao domínio avaliado e um termo relativo à localização geográfica; e uma lista de receitas esperadas como resultado daquela consulta, a qual foi construída manualmente, com base nas conversas com pessoas de diferentes perfis de

alguns locais espalhados pelo Brasil, entre eles, podemos destacar a Paraíba, o Acre, Tocantins, o Maranhão, o Rio Grande do Sul, o Ceará, São Paulo e o Rio de Janeiro. Além disso, nos próprios sites consultados para a construção da base de dados, havia lista de receitas relacionadas para cada receita coletada, que foram utilizadas como base para a elaboração de alguns casos de teste. Na Tabela 5, é apresentada a lista de receitas esperadas para o caso de teste cuja consulta é "Aipim Paraíba".

Tabela 5 – Lista de receitas esperadas para o caso de teste cuja consulta é "Aipim Paraíba".

ID	Lista de Receitas Esperadas	
32	Título	Delícia de macaxeira
	Ingredientes	- 1 1/2 kg de macaxeira ou aipim - 1 pote de requeijão - 350 g de charque - 1 cebola média cortada em rodelas - 1 colher (sopa) de margarina - sal a gosto - queijo parmesão ralado para polvilhar
	Modo de Preparo	Cozinhe a macaxeira com um pouco de sal até que fique bem macia, quando ela estiver cozida e transparente; amasse e reserve. Corte em cubos e escalde o charque e despreze a água. Em seguida, corte a cebola em rodelas e frite o charque até que a cebola comece a murchar. Separe a cebola e o óleo do charque. Em uma tigela misture a macaxeira já amassada com o charque e em um refratário coloque uma camada da macaxeira com charque cubra com o requeijão e coloque outra camada da macaxeira com charque cobrindo novamente com o requeijão e polvilhe com o queijo parmesão ralado. Leve ao forno para dourar. O charque pode ser substituído por carne do sol, bacon, linguiça defumada ou o recheio ao seu gosto.
36	Título	Macaxeira Frita
	Ingredientes	1 kg de macaxeira 2 1/2 xícaras de óleo Sal a gosto
	Modo de Preparo	Descasque o aipim, corte ao meio, retire a fibra central, corte em pedaços, coloque em uma panela, cubra com água, leve ao fogo alto e cozinhe até os pedaços ficarem macios, mas sem se desfazerem. Tire do fogo e escorra bem. Coloque óleo em uma frigideira, leve ao fogo alto, aqueça bem, junte o aipim aos poucos e frite até dourar. Retire com uma escumadeira, deixe escorrer sobre papel absorvente, polvilhe queijo ralado.
44	Título	Bolo de Macaxeira
	Ingredientes	1 kg de macaxeira 2 xícaras (chá) de leite de gado 1 coco ralado 4 ovos inteiros 2 colheres (sopa) de manteiga derretida 1 xicara (chá) de farinha de trigo 1 colher (sopa) de fermento em pó Sal a gosto
	Modo de Preparo	Rale a macaxeira, esprema e coloque numa vasilha.Passe no liquidificador metade do coco com o leite de gado, para tirar o leite de coco.A outra metade do coco ralado, junte com a macaxeira, assim como os 4 ovos, a manteiga, o leite de coco, a farinha de trigo e o fermento.Leve ao forno em forma untada.Corte em pedaços e sirva.

Como exemplo desse cenário, suponha que um usuário do Rio de Janeiro

pretende visitar João Pessoa na Paraíba. Como ele gosta bastante de aipim, pretende encontrar receitas que tenham o mesmo ingrediente na Paraíba. Para isso, criou uma consulta "*receitas aipim Paraíba*" e a submeteu aos três sistemas de recuperação. Na abordagem tradicional com TF-IDF, ao realizar a consulta, muitas receitas que utilizam o aipim foram retornadas, até as receitas da Paraíba. Porém, entre as receitas paraibanas, não foi encontrada nenhuma que contivesse aipim. Além disso, as receitas paraibanas ficaram no final da lista das consideradas relevantes (Vide Figura 24). Nenhuma receita paraibana com aipim foi considerada relevante na consulta do usuário, porquanto esse termo não é muito comum na região. Esse comportamento é o mesmo nos motores de busca mais famosos existentes na Internet, como o *Google* e o *Bing*.

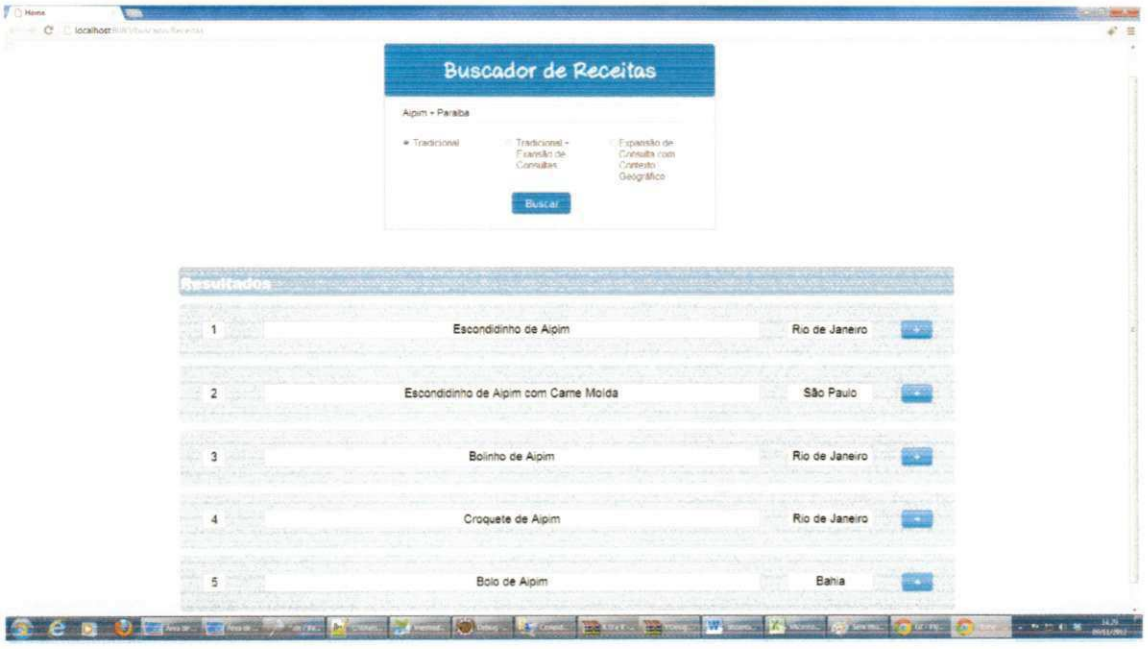


Figura 24 – Resultados da Consulta no sistema de busca tradicional.

Na abordagem com TF-IDF e a técnica simples de expansão de consulta, a consulta original é expandida, mais especificamente, o termo "aipim", porque existe no tesauro e tem como sinônimos os termos: mandioca e macaxeira. Dessa forma, a consulta submetida ao sistema passa a ser: "receita aipim mandioca macaxeira Paraíba". Como resultado dessa consulta, o sistema retorna receitas que contêm mandioca, macaxeira e aipim. Entre os resultados, algumas receitas da Paraíba são retornadas, uma vez que macaxeira é um termo muito utilizado nesse escopo geográfico, porém, além delas, receitas de outros estados que utilizam os termos mandioca e aipim também são apresentadas e, às vezes, aparecem em uma posição melhor do ranking de documentos relevantes, como se vê na Figura 25.

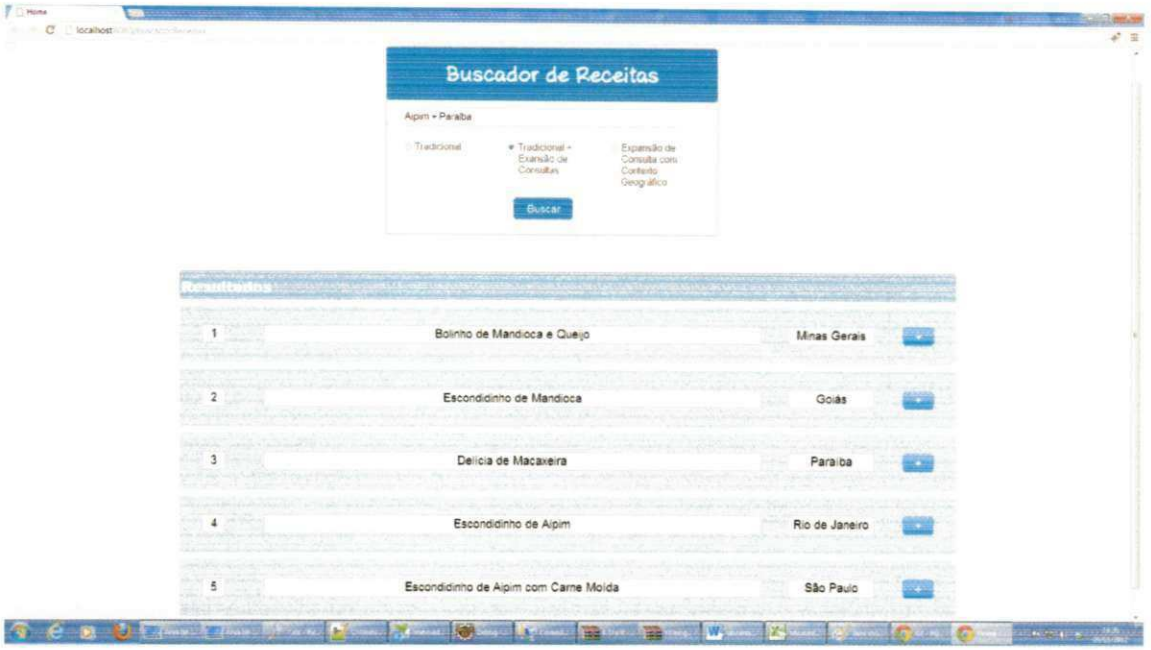


Figura 25 Resultados da consulta pelo sistema com TF-IDF e expansão de consulta.

Por fim, na abordagem proposta, assim como na anterior, a consulta original é expandida. Porém, antes da expansão, o termo Paraíba é identificado, e o termo aipim só é expandido para os sinônimos cujo escopo geográfico seja da Paraíba. Portanto, a consulta resultante da expansão é "receita aipim macaxeira mandioca Paraíba". Como resultado, são retornadas receitas que contêm o termo macaxeira, e como esse termo é característico do estado da Paraíba, os primeiros resultados do sistema são receitas de macaxeira (aipim) da Paraíba, embora existam alguns resultados de estados vizinhos que também o empregam (Vide Figura 26). Nesse caso de uso, a abordagem proposta consegue retornar para o usuário as receitas do seu interesse.

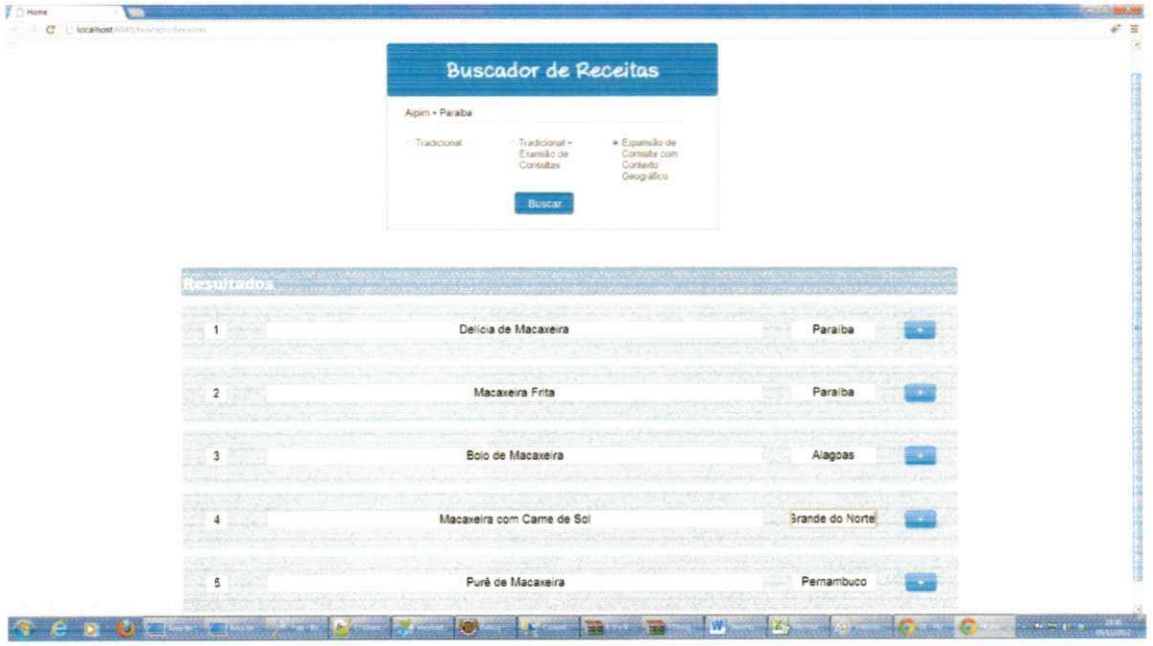


Figura 26 - Resultado da abordagem proposta para a consulta: *"Aipim Paraíba"*.

5.4.1. Métricas de Avaliação

As consultas foram submetidas à abordagem apresentada neste trabalho, e o conjunto de resultados retornados pelo sistema foi comparado com a lista de receitas esperadas. A partir dela, foram avaliadas a precisão, o *recall* e a média harmônica das duas medidas por meio da medida F, para a abordagem proposta.

A precisão é a capacidade do sistema de manter os documentos irrelevantes fora do resultado de uma consulta, ou seja, mede-se a quantidade de documentos relevantes dentre os itens retornados para a consulta (Vide Equação 1).

Por sua vez, o recall é utilizado para medir a capacidade do sistema de recuperar os documentos mais relevantes para o usuário, quer dizer, mede-se a fração de documentos relevantes em relação à quantidade de documentos total retornados pelo sistema (Vide Equação 2).

Por fim, a medida F é uma métrica que leva em consideração tanto o *recall* quanto a precisão, porque, é muito importante que um sistema de recuperação da informação alinhe essas duas medidas. Logo a medida F é definida como a média harmônica dessas duas medidas (Vide Equação 3).

5.4.2. Avaliação da Precisão

Na Tabela 6, é apresentada a precisão das três abordagens, considerando-se os resultados mais relevantes retornados por cada uma delas. Na média, a abordagem proposta apresenta uma melhoria de 33,54% em relação à baseada apenas em TF-IDF e 16,48%, em relação à abordagem que utiliza a expansão de consulta tradicional. Na Figura 27, é possível observar o gráfico que exibe esses resultados, onde o eixo x representa o tamanho da lista considerada relevante a busca e o eixo a precisão média para os cenários avaliados.

Essa melhoria se deve, principalmente, ao fato de que a abordagem proposta é mais restritiva ao expandir a consulta, posto que considera a localização geográfica

presente na consulta como uma informação semântica relevante. Além disso, é possível observar que com o aumento do tamanho da lista de documentos considerados relevantes para o sistema, a precisão tende a diminuir para as três abordagens, uma vez que documentos com coeficiente de relevância inferior à consulta tendem a aparecer.

Tabela 6 - Medida de precisão considerando-se os resultados mais relevantes.

Nº de Resultados Relevantes	Técnicas		
	TF-IDF	TF-IDF + EXPANSÃO DE CONSULTA	ABORDAGEM PROPOSTA
1	29,34 %	45,34 %	62,11 %
2	29,82 %	43,82 %	60,77 %
3	28,33 %	44,33 %	61,33 %
4	27,77 %	44,98 %	60,88 %
5	27,27 %	44,24 %	60,27 %
6	26,07 %	44,97 %	60,15 %
7	25,95 %	43,82 %	59,82 %
8	24,63 %	42,63 %	59,63 %
9	24,51 %	42,08 %	59,09 %
10	23,44 %	41,44 %	58,44 %
MÉDIA	**26,71%**	**43,77%**	**60,25%**

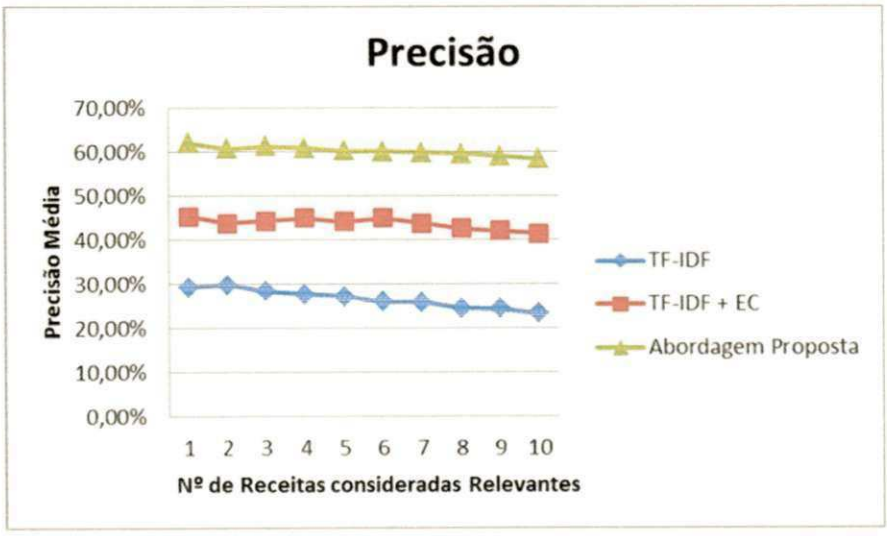


Figura 27 - Gráfico de precisão das três abordagens avaliadas.

5.4.3. Avaliação do Recall

Em relação à medida de *recall*, na média, a técnica proposta neste trabalho obteve uma melhoria de 28,03%, em relação à abordagem baseada apenas no TF-IDF. Porém, houve uma diminuição de 1,31% em relação ao método de expansão de consulta tradicional (Vide Tabela 7). Essa diminuição ocorre, porque a abordagem com expansão simples considera todos os sinônimos existentes no tesauro. Assim, a probabilidade de recuperar é maior, porém, como foi visto, a precisão é degradada. A Figura 28 apresenta o *recall* das três abordagens e o comportamento acima descrito. Nesse gráfico, o eixo x representa o tamanho da lista considerada como documentos relevantes a consulta e o eixo y é o *recall* médio para os cenários avaliados.

Tabela 7 - Medida de recall considerando os resultados mais relevantes.

N° de Resultados Relevantes	Técnicas		
	TF-IDF	TF-IDF + EXPANSÃO DE CONSULTA	ABORDAGEM PROPOSTA
1	29,20 %	59,11 %	57,42 %
2	31,11 %	60,62 %	59,81 %
3	34,44 %	63,65 %	63,03 %
4	37,94 %	66,94 %	64,02 %
5	39,72 %	68,63 %	66,35 %
6	42,26 %	72,17 %	70,14 %
7	43,98 %	73,59 %	72,66 %
8	44,87 %	74,02 %	73,91 %
9	46,78 %	75,78 %	75,16 %
10	49,33 %	78,46 %	77,39 %
MÉDIA	**39,96%**	**69,30%**	**67,99%**

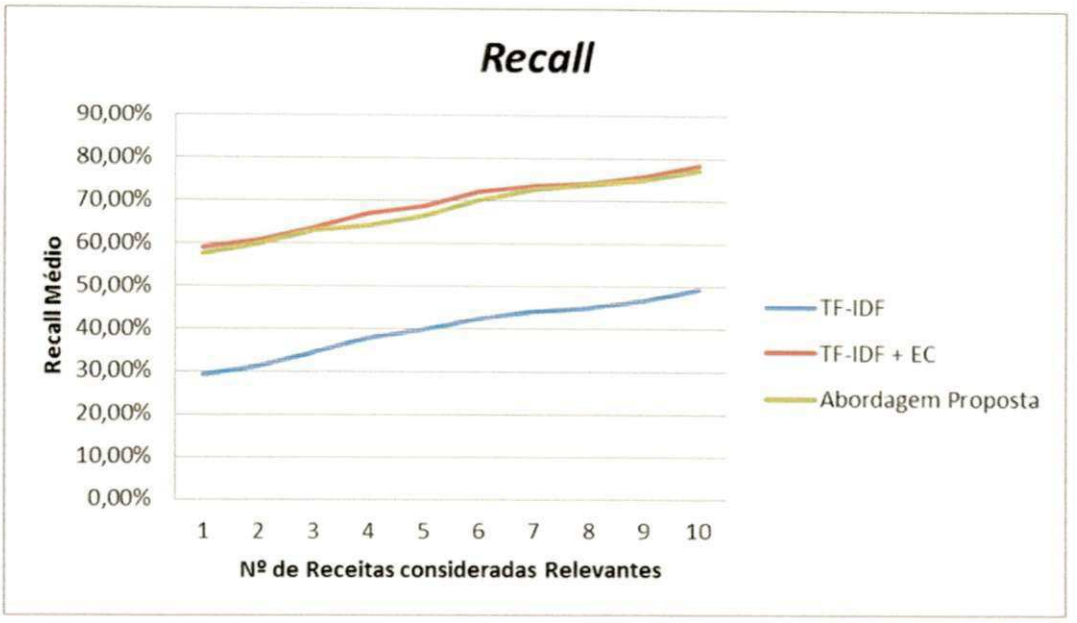


Figura 28 - Gráfico do *Recall* das três abordagens avaliadas.

5.4.4. Avaliação da Medida F

Embora o *recall* da abordagem proposta não tenha sido superior a todas as técnicas comparadas, quando avaliada em conjunto com a precisão, por meio da medida F, obteve-se uma melhora (Vide Figura 29), na média, de 32,09% em relação à abordagem baseada apenas em TF-IDF e 10,20% em relação à técnica de expansão de consulta tradicional, como apresentado na Tabela 8.

Para concluir que a abordagem proposta proporciona uma melhora na recuperação de documentos relevantes pelo usuário, foi realizado um teste-t. A hipótese nula é de que as médias amostrais para as duas abordagens são iguais, e a hipótese alternativa é de que a média dos resultados da medida F da abordagem comparada é menor do que a média dos resultados deste trabalho. Logo, foram realizados dois testes-t: o primeiro, entre a abordagem tradicional com TF-IDF e a abordagem proposta; o segundo, entre a abordagem com TF-IDF e a expansão de consultas e a abordagem proposta com TF-IDF. Na Tabela 9, são apresentados os resultados do teste-t realizado.

Em um nível de confiança de 95% e 5% de significância (α), ao realizar o teste-t para a abordagem tradicional, identificada pelo (*grupo 1*), com a abordagem proposta (*grupo 2*), obtiveram-se como *p-valor* $8{,}15\times10^{-19}$, portanto, menor do que o

UFCG/BIBLIOTECA/BC

nível de significância de 5%. Devido a isso, rejeita-se a hipótese nula de que as médias são iguais e se aceita a hipótese alternativa de que a média da abordagem tradicional com TF-IDF é menor do que a média da medida F deste trabalho. Diante disso, em um nível de confiança de 95%, pode-se afirmar que este trabalho apresenta uma melhora na recuperação de documentos relevantes por parte do usuário em relação à abordagem

Tabela 8 - Medida F considerando-se os resultados mais relevantes.

Nº de Resultados Relevantes	Técnicas		
	TF-IDF	TF-IDF + EXPANSÃO DE CONSULTA	ABORDAGEM PROPOSTA
1	29,27 %	51,32 %	59,67 %
2	30,45 %	50,87 %	60,29 %
3	31,09 %	52,26 %	62,17 %
4	32,07 %	53,81 %	62,41 %
5	32,34 %	53,80 %	63,16 %
6	32,25 %	55,41 %	64,76 %
7	32,64 %	54,93 %	65,62 %
8	31,80 %	54,10 %	66,01 %
9	32,17 %	54,11 %	66,16 %
10	31,78 %	54,23 %	66,59 %
MÉDIA	**31,59%**	**53,48%**	**63,68%**

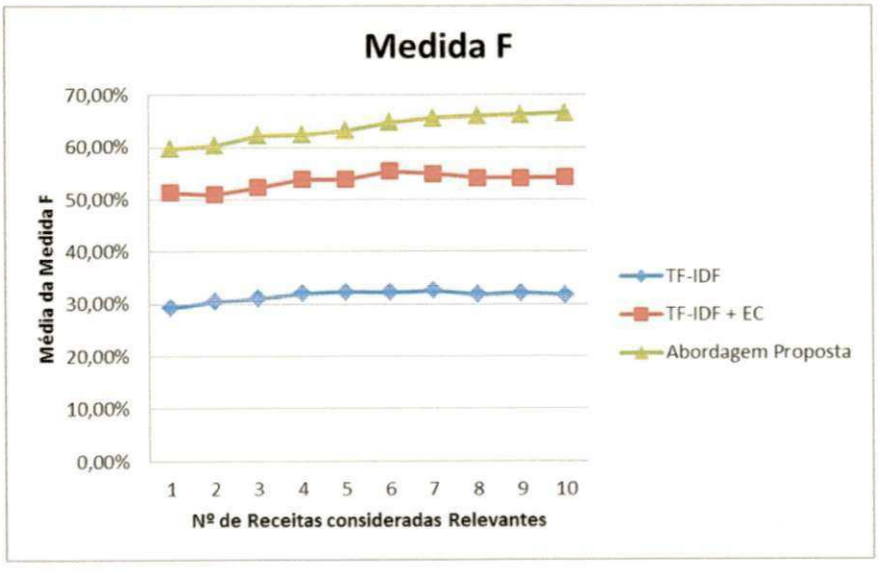


Figura 29 - Gráfico da medida F das três abordagens avaliadas.

Tabela 9 - Resultados do teste-t entre a abordagem tradicional e o trabalho proposto.

Informação	Valor
T	-37,37149192
Graus de Liberdade	18
P-valor	8,15E-19
Média no grupo 1:	0,31586
Média no grupo 2:	0,63684
Desvio padrão amostral do grupo 1:	0,010370599
Desvio padrão amostral do grupo 2:	0,025102643
Desvio padrão agrupado:	0,019205364
Intervalo de Confiança	95%
Limite Superior	-0,306086301

Da mesma forma, em um nível de confiança de 95% e 5% de significância (α), foi realizado o teste-t para a abordagem com TF-IDF e expansão de consulta, identificada pelo (*grupo 1*), com a abordagem proposta (*grupo 2*). Assim, o *p-valor* obtido foi de $9{,}78\times10^{-10}$, portanto, menor do que o nível de significância de 5%. Logo, rejeita-se a hipótese nula de que as médias são iguais e se aceita a hipótese alternativa de que a média da abordagem com TF-IDF e a expansão de consulta é menor do que a média da medida F da abordagem proposta. Portanto, assim como aconteceu com a abordagem com TF-IDF, em um nível de confiança de 95%, pode-se afirmar que o presente trabalho apresenta uma melhora na recuperação de documentos relevantes por parte do usuário em relação à abordagem com TF-IDF e expansão de consulta. Na Tabela 10, são apresentados os resultados do teste-t realizado.

Tabela 10 - Resultados do teste-t entre a abordagem proposta e a abordagem com TF-IDF + Expansão de Consultas.

Informação	Valor
T	-11,02308909
Graus de Liberdade	18
P-valor	9,78E-10
Média no grupo 1:	0,53484
Média no grupo 2:	0,63684
Desvio padrão amostral do grupo 1:	0,015036415
Desvio padrão amostral do grupo 2:	0,025102643
Desvio padrão agrupado:	0,020691018
Intervalo de Confiança	0
Limite Superior	95%

5.5 Georreferenciamento do Tesauro

Como foi visto na seção 4.2, a técnica de expansão de consultas proposta neste trabalho emprega um tesauro cujos termos são georreferenciados. Na abordagem proposta, para cada termo, associa-se uma lista de localidades geográficas de tamanho previamente definido. Desse modo, foi avaliado o impacto do tamanho da lista de localizações geográficas aos termos do tesauro no processo de recuperação da informação.

Assim como na seção anterior, avaliaram-se a precisão, o *recall* e a medida F do sistema, utilizando-se a técnica de expansão de consulta proposta nesse trabalho, para os diferentes tamanhos de lista de localidades geográficas associadas aos termos do tesauro. Na Tabela 11, é possível observar que existe um padrão de decaimento à medida que se aumenta o número de localidades geográficas por termo do tesauro. Isso quer dizer que os melhores resultados foram obtidos através de listas com uma ou duas localidades geográficas, por termo, do tesauro.

Além disso, é possível comprovar que o processo de marcação geográfica dos termos do tesauro é um fator importante na abordagem proposta, uma vez que, quando

se amplia o número de localizações geográficas por termo, a eficiência do sistema diminui e, em alguns casos, pode ser inferior às abordagens tradicionais.

Tabela 11 - Análise experimental da abordagem proposta em relação ao número de localidades geográficas por termo do tesauro.

TAMANHO DA LISTA DE LOCALIDADES GEOGRÁFICAS	PRECISÃO	*RECALL*	MEDIDA F
1	**60,65 %**	**67,99 %**	**64,11%**
2	59,96%	65,44 %	62,58%
3	55,98 %	60,19 %	58,01%
4	51,37 %	58,51 %	54,71%
5	50,89 %	55,76 %	53,21%
6	49,06 %	52,27 %	50,61%
7	47,18 %	51,90 %	49,43%
8	46,84 %	50,83 %	48,75%
9	45,79 %	50,11 %	47,85%
10	45,87 %	50,03 %	47,36%
MÉDIA	**51,36%**	**56,30%**	**53,71%**

5.6 Considerações Finais

Nesse capítulo foram descritos os métodos adotados e os resultados obtidos nos experimentos realizados para validação da abordagem desenvolvida. Através dos experimentos apresentados, é possível constatar que o uso de informações contextuais, em especial o geográfico, pode melhorar a eficiência do processo de aquisição de informação nos SRI. Além disso, foi mostrado que o processo de georreferenciamento dos termos tem um papel fundamental no desempenho da abordagem proposta. No próximo capítulo, serão apresentadas as conclusões desse trabalho e alguns trabalhos futuros que poderão ser desenvolvidos para o prosseguimento dessa pesquisa.

Capítulo 6

Considerações Finais

O emprego de novas abordagens com o fim de melhorar os mecanismos de recuperação da informação tem sido muito explorado pelos recentes trabalhos na área de RI. Entre alguns dos aspectos considerados, está o uso do contexto geográfico e de técnicas de expansão de consulta.

Neste trabalho, foram apresentados os principais fundamentos da recuperação da informação e dos SRI, descrevendo-se os aspectos estruturais desses sistemas e os modelos clássicos de RI que são utilizados como base para os modelos adotados nos principais sistemas de recuperação atuais. Do mesmo modo, foram descritos os principais conceitos e as características do processo de recuperação de informação geográfica, incluindo os sistemas de RIG. Também foi apresentado o estado da arte em técnicas de expansão de consulta que utilizam o contexto geográfico onde foram mencionadas as principais contribuições relacionadas ao tema.

Foram descritos os métodos para georreferenciamento automático dos termos do tesauro e a expansão semântica de consultas com a utilização do contexto geográfico, incluindo o protótipo desenvolvido para a validação dos métodos propostos. Por fim, foram descritos os experimentos realizados para a avaliação funcional do protótipo.

6.1 Contribuições

Foi proposto um sistema de recuperação da informação, que faz uso de um tesauro georreferenciado para expandir semanticamente as consultas dos usuários com base em informações geográficas. Nesse sistema, foram descritos, em detalhes, o processo de indexação e de busca, bem como a forma de representar os documentos e as consultas.

Como parte fundamental do sistema proposto, foi descrito um método de georreferenciamento automático dos termos do tesauro a partir do número de resultados retornados por uma consulta submetida a um motor de busca tradicional contendo o termo a ser georreferenciado e a localidade geográfica. Com essa técnica, é possível atribuir um ou mais contextos geográficos aos termos do tesauro.

Além disso, foi apresentada uma nova técnica de expansão de consulta, cujo objetivo é de expandir os termos através de sinônimos presentes no tesauro que tenham coordenadas geográficas próximas às localidades geográficas presentes na consulta. Dessa forma, num cenário onde o termo buscado não seja comum na localidade geográfica, os resultados relevantes podem ser apresentados mediante um sinônimo conhecido nessa localidade. Por fim, a abordagem proposta se mostrou mais eficiente quando comparada com sistemas tradicionais de recuperação da informação.

6.2 Trabalhos Futuros

Com o objetivo de dar prosseguimento às pesquisas iniciadas neste trabalho, apresentamos as seguintes sugestões para trabalhos futuros:

- No processo de georreferenciamento, ao invés de se utilizar um motor de busca tradicional, como o Google, deve-se empregar um sistema de recuperação por meio do qual seja possível identificar o contexto dos documentos e, a partir dessa informação, avaliar o número de ocorrências do termo em documentos de um mesmo contexto geográfico;

- Utilização de uma ontologia geográfica, a fim de expandir os termos geográficos e avaliar essa expansão em conjunto com a expansão semântica proposta;

- Aumentar o número de casos de teste, a fim de se avaliar melhor a técnica proposta.

- Utilizar outras relações, além da sinonímia, para expandir os termos da consulta, como por exemplo, as relações de hipernímia e hiponímia;

- Adição da capacidade de manipular documentos escritos em múltiplos idiomas, bem como a expansão do tesauro para relacionamento com termos de outros idiomas.

- Investigar técnicas de construção automática de tesauro para que o sistema possa evoluir com o passar do tempo, uma vez que novas palavras podem surigir ou ainda se relacionar com outros termos ao longo do tempo. Dessa forma, é preciso que o georreferenciamento dos termos também aconteça à medida que esses novos termos são acrescidos ao tesauro.

Referências Bibliográficas

ANDOGAH, G. **Geographically Constrained Information Retrieval**. 2010. 189f. Tese (Phd Computer Science), University of Groningen, Holanda.

ARGUELLO, J.; ELSAS, J. L.; CALLAN, J.; CARBONELL, J. G. Document representation and query expansion models for blog recommendation. In: **Proceedings of the 2nd International Conference on Weblogs and Social Media**, p. 10–18, 2008.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. **Modern Information Retrieval**. First Edition. Addison-Wesley Longman Publishing Co., Inc., 1999.

BATISTA, D. S; SILVA, M. J.; COUTO, F. M.; BEHERA, B. **Geographic signatures for semantic retrieval**. Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10, p. 191-198, 2010.

BAZIRE, M.; BRÉZILLON, O. **Understanding context before using it**. Proceedings of the 5th international conference on modeling and using context, p. 29-40, 2005.

BHOGAL, J.; MACFARLANE, A.; SMITH, P. A review of ontology based query expansion. **Journal Information Processing and Management**, v. 43, n. 4, p. 866-886, 2007.

BRIN, S.; PAGE, L. **The anatomy of a large-scale hypertextual web search engine**. Computer Networks 30, p. 107–117, 1998.

BUCHHOLZ, S.; HAMANN, T.; HUBSCH, G. **Comprehensive Structured Context Profiles (CSCP): Design and Experiences**. Pervasive Computing and Communications Workshops, IEEE International Conference, Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, p. 43, 2004.

BUSCALDI, D.; ROSSO, P.; ARNAL, E. S. Using thewordnet ontology in the geoclef geographical information retrieval task. In: **Accessing multilingual information Repositories**, v. 4022/2006, p. 939–946, Springer (Lecture Notes in Computer Science LNCS), 2006.

BUSCALDI, D.; ROSSO, P. **Using geowordnet for geographical information retrieval**. Proceedings of the 9th cross-language evaluation forum conference on evaluating systems for multilingual and multimodal information access, p. 863-866, 2009.

BUYUKKOKTEN, O.; CHO, J.; GARCIA-MOLINA H.; GRAVANO L.; SHIVAKUMAR N. **Exploiting geographical location information of web pages**: Proceedings of Workshop on Web Databases (WebDB 99), 1999.

CAMPELO, C.E. GeoSen, **Um motor de busca com enfoque geográfico**. 2008. 150f.

Dissertação (Mestrado em Computação) - Universidade Federal de Campina Grande, Campina Grande.

CAO, G.; GAO, J.; NIE, J. Y.; BAI, J. Extending query translation to cross-language query expansion with markov chain models. In: **Proceedings of the 16th Conference on Information and Knowledge Management** (CIKM'07). 2007.

CARDOSO, N.; SILVA, M. J. Query expansion through geographical feature types. In: **GIR '07**: Proceedings of the 4th ACM workshop on Geographical information retrieval, p. 55–60, 2007.

CARPINETO, C.; ROMANO, G. **A** (CSUR), v. 44, 2012.

CHOWDHURY, G. **Introduction to Modern Information Retrieval**. 3rd Edition, Facet Publishing, 2010.

CLEVERDON, C. W. **The significance of the cranfield tests on index languages**. Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, p. 3–12, 1991.

CRASWELL, N. **GOV2 Test Collection**. Disponível em: < http://ir.dcs.gla.ac.uk/test_collections/gov2-summary.htm>. Acesso em 27/06/2012.

DELBONI, T. M.; DORGES, K. A. V.; LAENDER, A. H. F.; DAVIS, C. A. Semantic expansion of geographic web queries based on natural language positioning expressions. In**: Transactions in GIS**, v. 11, p. 377–397, 2007 .

DEY, A. K. Understanding and using context. **Journal Personal and Ubiquitous Computing**, v. 5, p. 4-7, 2001.

DING, J.; GRAVANO, L.; e SHIVAKUMAR N. **Computing geographical scopes of web resources**. 26th International Conference on Very Large Databases, pp. 445-456, 2000.

DING,Y.; GHOWDHURY, G. G.; FOO, S. Incorporating the results of co-word analyses to increase search variety for information retrieval. **Journal of Information Science**, v. 26, p. 429-451, 2000.

FERNANDES, R. M. **GeoSen_tags**:um motor de busca geográfico com suporte a tags. 2010. 108f. Dissertação (Mestrado em Computação) - Universidade Federal de Campina Grande, Campina Grande.

FU G.; JONES, C. B.; ABDELMONTY A. I. Ontology-based spatial query expansion in information retrieval. In: **on the move to Meaningful Internet Systems 2005**: CoopIS, DOA, and ODBASE, v. 3761/2005, p. 1466–1482. Springer (Lecture Notes in Computer Science LNCS), 2005.

FUHR, N. Probabilistic Models in information retrieval. **The Computer Journal**, v. 35, p. 243-255, 1992.

UFCG/BIBLIOTECA/BC

GAN, Q.; ATTENBERG, J.; MARKOWETZ, A.; SUEL, T. **Analysis of geographic queries in a search engine log**. Proceedings of the first international workshop on Location and the web, p. 49-56, 2008.

GHIDINI, C; GIUNCHIGLIA, F. Local models semantics, or contextual reasoning = locality + compatibility. **Journal Artificial Intelligence**, v. 127, p. 221-259, 2001.

HENRICKSEN, K; INDULSKA, J; RAKOTONIRAINY, A. Modeling context information in pervasive computing systems. In: **Pervasive '02**: proceedings of the first international conference on pervasive computing, p. 167-180, 2002.

HILL, L. L. **Georeferencing**: The Geographic Associations of Information. MIT Press, 2007

HILL, L. L.; GOODCHILD, M.; JANEE G. Research directions in georeferenced IR based on the Alexandria digital library. In: **The SIGIR Workshop on Geographic Information Retrieval**, 2004.

HORNG, Y.; CHEN, S.; CHANG, Y.; LEE C. **A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques.** IEEE Transactions on fuzzy systems, v. 13, p. 216-228, 2005.

IMRAN, H.; SHARAN, A. Thesaurus and query expansion. **International Journal of Computer Science & Information Technology** (IJCSIT), v. 1, n. 2, p. 89-97, 2009.

JENSEN, F. V. **Bayesian networks and decision graphs**. information science and statistics. springer, 2002.

JONES, C.; ABDELMONTY A.; FINCH D., FINCH G.; VAID S. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In: **Geographic Information Science**, p. 125–139, 2004

KOWALSKI, G. **Information Retrieval Systems**: theory and implementation. Kluwer Academic Publishers, 1997.

LARSON, R. R., **Geographical information retrieval and spatial browsing**. Geographical Information Systems and Libraries: patrons, maps, and spatial information, pp. 81-124, 1996.

LARSON, R. R.; GEY, F. C.; PETRAS, V. **Berkeley at GeoCLEF**: Logistic Regression and Fusion for Geographic Information Retrieval. In Accessing Multilingual Information Repositories, v. 4022/2006, p. 963–976. Springer (Lecture Notes in Computer Science LNCS), 2006.

LEITE, M. A. A.; RICARTE, I. L. M. **Document retrieval using fuzzy related geographic ontologies**. Proceedings of the 2nd international workshop on Geographic information retrieval, p. 47-54, 2008.

MANNING, C. D.; RAGHAVAN, P.; SCHTZE, H. Introduction to Information Retrieval. **Cambridge University Press**, 2008.

MARKOWETZ, A.; YEN-YU C.; TORSTEN S.; XIAOHUI L.; BERNHARD S.

SCHMIDT, A; BEIGL, M; GELLERSEN, H.-W. **There is more to context than location**. Computers and Graphics, v. 23, p. 893-901, 1999.

SILVA, M. J.; MARTINS, B.; CHAVES, M. S.; AFONSO A. P.; CARDOSO, N. Adding **geographic scopes to web resources**. CEUS – Computers, Environment and Urban Systems. v. 30, p. 378-399, 2006.

SMITH D.; CRANE G. Disambiguating geographic names in a historical digital library. In: **Research and advanced technology for digital libraries**, p. 127–136, 2001.

TRUONG, K. N.; ABOWD, G. D.; BROTHERTON, J. A. **Who, What, When, Where, How**: Design issues of capture access applications. In: Ubicomp 2001: Ubiquitous Computing, Third International Conference, p. 209-224, 2001.

VAN RIJSBERGEN, C. J. **Information Retrieval**. Butterworths, 2nd edition, 1979.

VIEIRA, V.; SOUZA, D.; SALGADO, A. C.; TEDESCO, P. **Uso e Representação de Contexto em Sistemas Computacionais**, *Mini-curso apresentado no Simpósio de Fatores Humanos em Sistemas Computacionais* (IHC 2006), Natal, Brasil, 2006.

VOORHEES, E. M.; HARMAN, D. K. **TREC**: experiment and evaluation in information retrieval (Digital Libraries and Electronic Publishing). The MIT Press, 2005.

WALLER, W. G.; KRAFT, D. H. A mathematical model of a weighted boolean retrieval system. **Information Processing and Management Journal**, v. 15, n. 5, p. 235-245, 1979.

WANG, X. H.; ZHANG, D. Q.; GU, T.; PUNG, H. K. **Ontology based context modeling and reasoning using OWL**. Pervasive Computing and Communications Workshops - Proceedings of the Second IEEE Annual Conference, p. 18- 22, 2004

XIANG, B.; JIANG, D.; PEI, J.; SUN, X. CHEN, E.; LI, H. **Context-aware ranking in web search**. Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, p. 451-458, 2010.

XU, J.; CROFT, W. B. Query expansion using local and global document analysis. In: **Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval**, p. 4-11, 1996.

XU, J.; CROFT, W. B. **Improving the effectiveness of information retrieval with local context analysis.** ACM Transactions on Information Systems, v. 18, p. 79-112, 2000.

YUE, Y.; FINLEY, T.; RADLINSKI, F.; JOACHIMS, T. **A Support vector method for optimizing average precision**. Proceeding SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, p. 271-278, 2007.

ZOBEL, J.; ALISTAIR M. **Inverted files for text search engines**. ACM Computing Surveys v. 38, n. 2, 2006.