

**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE  
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA  
COORDENAÇÃO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**RANKING DE RELEVÂNCIA BASEADO EM INFORMAÇÕES GEOGRÁFICAS E  
SOCIAIS**

JÚLIO HENRIQUE ROCHA

Campina Grande - PB

2016

**JÚLIO HENRIQUE ROCHA**

**RANKING DE RELEVÂNCIA BASEADO EM INFORMAÇÕES GEOGRÁFICAS E  
SOCIAIS**

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 obtenção do grau de Mestre em Ciência da Computação.

Professores Orientadores:

Cláudio de Souza Baptista, Ph.D.

Cláudio Elízio Calazans Campelo, Ph.D.

Campina Grande-PB

2016

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

R672r

Rocha, Júlio Henrique.

*Ranking* de relevância baseado em informações geográficas e sociais / Júlio Henrique Rocha. – Campina Grande, 2016.

116 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2016.

"Orientação: Prof. Dr. Cláudio de Souza Baptista; Coorientação: Prof. Dr. Cláudio Elízio Calazans Campelo".

Referências.


1. *Geographic Information Retrieval*. 2. *Ranking* de Relevância. 3. Notícias. 4. Redes Sociais. I. Baptista, Cláudio de Souza. II. Campelo, Cláudio Elízio Calazans. III. Título.

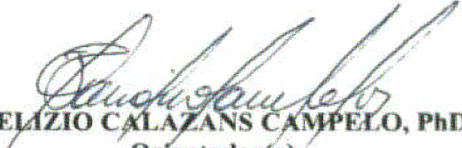
CDU 004.77(043)

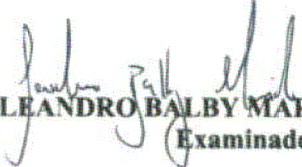
**"RANKING DE RELEVÂNCIA BASEADO EM INFORMAÇÕES GEOGRÁFICAS E  
SOCIAIS"**


**JÚLIO HENRIQUE ROCHA**

**DISSERTAÇÃO APROVADA EM 23/08/2016**

  
**CLÁUDIO DE SOUZA BAPTISTA, PhD., UFCG**  
**Orientador(a)**

  
**CLAUDIO ELIZIO CALAZANS CAMPELO, PhD., UFCG**  
**Orientador(a)**

  
**LEANDRO BALBY MARINHO, Dr., UFCG**  
**Examinador(a)**

  
**ANSELMO CARDOSO DE PAIVA, DR., UFMA**  
**Examinador(a)**

**CAMPINA GRANDE - PB**

## RESUMO

Recuperação de Informação Geográfica (GIR) é uma área de pesquisa que desenvolve e viabiliza a construção de mecanismos de busca por conteúdos distribuídos pela Internet envolvendo algum contexto geográfico. Os motores de busca geográfica, que são artefatos produzidos na área de GIR, podem ser especificados para trabalhar em diversos contextos (e.g., esportes, concursos públicos), buscando um tratamento adequado ao tipo de documento manipulado. Atualmente, a comunidade científica e o meio comercial vêm concentrando esforços na construção de motores de busca geográfica com o foco em encontrar notícias distribuídas na Internet. Contudo, motores de busca (geográfica ou não) com foco em notícias, deveriam considerar o fator de credibilidade da informação contida nas mesmas no momento de ordená-las. Infelizmente, na maior parte das vezes, isso não acontece. Mensurar a credibilidade de notícias é uma atividade onerosa e complexa, por exigir o conhecimento dos fatos relatados. Dessa forma, os motores de busca acabam deixando a cargo do usuário a responsabilidade em confiar no que está sendo lido. Nesse contexto, esta dissertação propõe um método de *ranking* de relevância com foco em notícias e baseado em informações colhidas em redes sociais, para valorar um grau de credibilidade e, assim, ordená-las. O valor de credibilidade da notícia é calculado considerando a afinidade dos usuários, que a compartilharam em sua rede social, com as localidades mencionadas na notícia. Por fim, o *ranking* de relevância proposto é integrado a uma ferramenta de busca e leitura de notícias, denominada GeoSEn News, que viabiliza a consulta por meio de diversas operações espaciais e permite a visualização dos resultados em diferentes perspectivas. Tal ferramenta foi utilizada para avaliar o método proposto através de experimentos utilizando dados colhidos na rede social Twitter e em mídias informativas espalhadas pelo Brasil. A avaliação apresentou resultados promissores e atestou a viabilidade da construção do *ranking* de relevância que se baseia em informações coletadas em redes sociais.

**Palavras chaves:** *Geographic Information Retrieval*, *Ranking* de Relevância, Notícias, Redes Sociais.

## ABSTRACT

Geographic Information Retrieval is a research field that develops and allows the construction of search engines to retrieve information with geographic context that is available on the Internet. Produced in the GIR field, geographic search engines can be specified to work in many different contexts (e.g., as sports, concerts), seeking proper ways to handle the chosen document type. Nowadays, the scientific community and the commerce are focusing efforts on building geographic search engines to find news over the Internet. However, search engines (geographical or otherwise) focused on news should consider the information credibility factor in the moment of ranking them. Unfortunately, in most cases, it is not what happens. Measure the news credibility is a complex and expensive task since it requires knowledge of the stated facts. Thereby, search engines end up giving the user the responsibility to trust or not what is being read. In this context, this work proposes a relevance ranking method focused in news and based on information collected from social networks, to evaluate a credibility factor and thus, rank them. The news credibility value is calculated considering the affinity of users who have shared it on their social network with the locations mentioned in the news. Lastly, the proposed relevance ranking is integrated with a search engine and reading news tool called GeoSEn News, which enables various spatial operations queries and allows result visualization in different perspectives. Through experiments using data collected in the social network Twitter and informational media throughout Brazil, this tool was used to evaluate the proposed method. The evaluation presented promising results and certified the feasibility of building relevance ranking based on information collected from social networks.

**Palavras chaves:** Geographic Information Retrieval, Relevance Ranking, News, Social Network.

## AGRADECIMENTOS

Agradeço primeiramente a Deus pelo dom da vida, por ter guiado minhas decisões e fornecido as forças necessárias para alcançar todos os meus objetivos.

Aos meus pais (Ronildo e Vera), avós (Heleno e Severina), meus irmãos (Jonathan, Helson e Mirtes) e familiares, os quais participaram efetivamente de toda essa trajetória, oferecendo todo o suporte afetivo e emocional necessário para que eu conseguisse vencer os obstáculos encontrados durante todo o percurso. Agradeço de coração.

À minha linda Áquila, que aceitou, durante esses anos de pós-graduação, o papel de noiva e o exerceu perfeitamente. Apesar de separados pela distância, estávamos sempre unidos pelo sentimento e carinho mútuo. Aos gritos de “não vai ter golpe”, ela esperou pacientemente o fim desta dissertação para marcar a data do casamento. Obrigado por esperar. Amo você!

Sou grato aos meus orientadores, o professor Cláudio de Souza Baptista e o professor Cláudio Elízio Calazans Campelo, pela dedicação, orientação, paciência e experiências compartilhadas. Obrigado por transmitirem um pouco do muito que sabem, pela total disponibilidade e conselhos na construção desta pesquisa.

Aos meus amigos e colegas do Laboratório de Sistemas de Informação, que entre alguns copos de cerveja e artigos escritos sempre me apoiaram e ajudaram nos momentos que mais precisei, seja com convites para beber ou na troca de experiências científicas. A vocês, meu muito obrigado.

Agradeço os professores e demais funcionários do curso de Pós-Graduação em Ciência da Computação e da UFCG, que contribuíram direta ou indiretamente para a boa condução dos trabalhos e permitiram o meu crescimento pessoal e profissional.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de nível Superior (CAPES), pelo apoio financeiro.

## SUMÁRIO

<b>CAPÍTULO 1</b> .....	<b>14</b>
1.1. DEFINIÇÃO DO PROBLEMA .....	16
1.2. OBJETIVOS .....	17
1.2.1. Objetivos Gerais .....	17
1.2.2. Objetivos Específicos .....	17
1.3. CONTRIBUIÇÕES .....	18
1.4. TRABALHOS PUBLICADOS .....	19
1.5. ORGANIZAÇÃO ESTRUTURAL .....	19
<b>CAPÍTULO 2</b> .....	<b>20</b>
2.1. RI – RECUPERAÇÃO DA INFORMAÇÃO .....	21
2.1.1. Operação textual.....	21
2.1.2. Indexação .....	22
2.1.3. Ranking de Relevância.....	23
2.1.4. Arquitetura de Sistemas de RI.....	23
2.1.5. Modelos de RI .....	24
2.2. MECANISMOS DE BUSCA PARA WEB .....	26
2.2.1. Spider (Web crawler) .....	27
2.2.2. Indexação .....	28
2.2.3. Ranking de Relevância.....	29
2.3. GIR- RECUPERAÇÃO DA INFORMAÇÃO GEOGRÁFICA .....	29
2.3.1. Detecção de Referências Geográficas .....	29
2.3.2. Modelagem do Escopo Geográfico .....	32
2.3.3. Indexação Espaço-textual.....	34
2.3.4. Ranking de Relevância Geográfica .....	35
2.3.5. Avaliação de Ranking de Relevância Geográfica .....	36
2.3.6. Consultas e Interface Gráfica .....	37
2.4. GEOTEN.....	39
2.4.1. Detecção de Referências Geográficas .....	41
2.4.2. Modelagem do Escopo Geográfico .....	42
2.4.3. Indexação Espaço-textual.....	42



2.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO.....	42
<b>CAPÍTULO 3.....</b>	<b>44</b>
3.1. INFERÊNCIA DE LOCALIZAÇÃO.....	45
3.1.1. Sumarização dos Métodos.....	48
3.1.2. Considerações Sobre o Estado da Arte em Inferência de Localização .....	48
3.2. RANKING DE RELEVÂNCIA GEOGRÁFICA .....	50
3.2.1. Sumarização dos Métodos.....	52
3.2.2. Considerações Sobre o Estado da Arte em <i>Ranking</i> de Relevância Geográfica .	52
3.3. FERRAMENTAS DE LEITURA DE NOTÍCIAS.....	53
3.3.1. Sumarização das Ferramentas .....	57
3.3.2. Considerações Sobre as Ferramentas de Leitura de Notícias.....	58
3.4. CONSIDERAÇÕES FINAIS DO CAPÍTULO.....	60
<b>CAPÍTULO 4.....</b>	<b>61</b>
4.1. COLETOR DE NOTÍCIAS .....	61
4.2. AFINIDADE LOCAL .....	65
4.3. RANKING DE RELEVÂNCIA GEOSOCIAL .....	75
4.4. GEOPEN NEWS .....	78
4.4.1. Camada de Dados.....	80
4.4.2. Camada de Negócio .....	81
4.4.3. Camada de Aplicação.....	83
4.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO.....	86
<b>CAPÍTULO 5.....</b>	<b>88</b>
5.1. CONJUNTO DE DADOS .....	88
5.2. EXPERIMENTO 1: AFINIDADE LOCAL.....	90
5.2.1. Dados do Experimento .....	91
5.2.2. Projeto do Experimento.....	92
5.2.3. Resultados .....	93
5.2.4. Discussão.....	94
5.3. <i>SURVEY</i> 1: <i>RANKING</i> DE RELEVÂNCIA GEOGRÁFICO .....	95
5.3.1. Dados do Survey .....	96
5.3.2. Projeto do Survey .....	97
5.3.3. Coleta das Respostas .....	98
5.3.4. Resultados .....	99

5.3.5. Discussão.....	101
5.4. ESTUDO DE CASO 1: GEOTEN NEWS .....	102
5.4.1. Conjunto de Dados .....	103
5.4.2. Hipóteses de Pesquisa .....	103
5.4.3. Concepção .....	104
5.4.4. Resultados .....	104
5.4.5. Discussão.....	106
5.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO.....	107
<b>CAPÍTULO 6.....</b>	<b>108</b>
6.1. CONTRIBUIÇÕES .....	109
6.2. TRABALHOS FUTUROS .....	110

## LISTA DE ABREVIATURAS E SIGLAS

IR	<i>Information Retrieval</i>
GIR	<i>Geographic Information Retrieval</i>
CEP	Código de Endereçamento Postal
TF	<i>Term Frequency</i>
IDF	<i>Inverse Document Frequency</i>
HTML	<i>Hyper Text Markup Language</i>
PDF	<i>Portable Document Format</i>
HTTP	<i>Hypertext Transfer Protocol</i>
URL	<i>Uniform Resource Locator</i>
TD	<i>Toponym Desambiguation</i>
POI	<i>Point of Interest</i>
SGBD	Sistema de Gerenciamento de Banco de Dados
API	<i>Application Programming Interface</i>
IP	<i>Internet Protocol</i>
RSS	<i>Rich Site Summary</i>
XML	<i>Extensible Markup Language</i>
BPMn	<i>Business Process Modeling</i>
VGI	<i>Volunteered Geographic Information</i>
REST	<i>Representational State Transfer</i>
URI	<i>Uniform Resource Identifier</i>
SIG	Sistema de Informação Geográfica

## LISTA DE FIGURAS

Figura 1: Representação da técnica de arquivos invertidos.....	23
Figura 2: Arquitetura básica de um sistema de RI.....	24
Figura 3: Arquitetura básica de um <i>crawler</i> Web. ....	27
Figura 4: Exemplo do processo de Desambiguação de Topônimo. ....	32
Figura 5: (a) Representação por escopo múltiplo; (b) Representação por escopo simples. ....	33
Figura 6: Arquitetura do GeoSEn.....	39
Figura 7: Instância de uma <i>geotree</i> .....	41
Figura 8: Interface Gráfica do NewsStand. ....	56
Figura 9: Interface Gráfica do TwitterStand.....	57
Figura 10: Processo de coleta de notícias no Twitter. ....	63
Figura 11: Arquitetura do coletor de notícias.....	65
Figura 12: Processo de inferência de afinidade local. ....	67
Figura 13: Processamento do cálculo de afinidade local.....	71
Figura 14: Geoparsing em micro texto.....	72
Figura 15: Fluxo do processo de identificação de termos influentes. ....	73
Figura 16: Exemplo da <i>geotree</i> . ....	74
Figura 17: Exemplo da atividade de medir a credibilidade de uma notícia. ....	76
Figura 18: Arquitetura do GeoSEn News.....	79
Figura 19: Partição do diagrama ER do banco do GeoSEn News.....	80
Figura 20: Exemplo da adição de outro valor ao índice espacial. ....	82
Figura 21: Interface gráfica do GeoSEn News.....	84
Figura 22: Formulário de consulta do GeoSEn News. ....	84
Figura 23: Seleção da terceira notícia e o destaque das localidades. ....	85
Figura 24: Exibição do cluster com infowindow de notícias. ....	85
Figura 25: Interface gráfica em diferentes dispositivos.....	86
Figura 26: Top-10 veículos de comunicação com mais seguidores. ....	89
Figura 27: Montagem do ambiente de experimentação.....	90
Figura 28: Distribuição espacial dos usuários participantes.....	92
Figura 29: Boxplot e sumário dos dados de afinidade local identificados. ....	92
Figura 30: Gráfico com as taxas de acertos obtidas em cada tratamento. ....	94
Figura 31: Ferramenta de auxílio construída para coleta das respostas. ....	98

Figura 32: Ordenação dos documentos sendo avaliada na primeira questão. ....	99
Figura 33: Gráfico de dispersão do coeficiente de correlação (tau) de todas as avaliações... ..	100
Figura 34: Resultado do teste de proporção retornado pela ferramenta R. ....	101

## LISTA DE TABELAS

Tabela 1: Tabela de sumarização dos métodos avaliados. ....	49
Tabela 2: Tabela de sumarização dos métodos de <i>ranking</i> de relevância geográfica. ....	52
Tabela 3: Tabela de sumarização das ferramentas de consulta de notícias. ....	58
Tabela 4: Resultados obtidos em cada tratamento executado. ....	94
Tabela 5: Consultas em linguagem natural avaliadas pelos participantes. ....	97
Tabela 6: Tabela de resultados. ....	105

## LISTA DE CÓDIGOS

Código 1: Pseudo-código para coleta de notícias no Twitter. ....	64
Código 2: Pseudo-código de coleta do comportamento de usuários. ....	68
Código 3: Mapa de referências encontradas. ....	69
Código 4: Pseudo-código para cálculo de afinidade local. ....	70
Código 5: Exemplo de uma requisição ao Web Service REST. ....	83

## LISTA DE EQUAÇÕES

Equação 1 .....	25
Equação 2 .....	26
Equação 3 .....	73
Equação 4 .....	75
Equação 5 .....	77
Equação 6 .....	78



## CAPÍTULO 1

### INTRODUÇÃO

Fundamentada pela acessibilidade irrestrita de informações e a facilidade de produção de conteúdo, a Internet cresce em uma velocidade sem precedentes e atinge um patamar inimaginável desde sua idealização. Atualmente, estima-se que a Internet indexável (i.e., a porção da rede que pode ser armazenada e recuperada por motores de busca), contém aproximadamente 45 bilhões de páginas<sup>1</sup>. De forma análoga, o número de internautas na Web já rompe a barreira dos 3 bilhões<sup>2</sup>, o que a caracteriza como um recurso importante e essencial para a população.

Da ampla variedade dos serviços hoje disponíveis na Internet, destacam-se aqueles que permitem buscar e explorar conteúdos que estão distribuídos, de forma desorganizada e não estruturada, por toda rede. A área que estuda esses serviços recebe o nome de Recuperação da Informação (IR) (Kowalsky, 1997) e envolve a pesquisa em diversos processos associados à atividade de recuperar informação. O motor de busca é responsável por coletar, analisar, indexar e armazenar conteúdos de forma a facilitar a descoberta por meio de palavras-chave representando o desejo do usuário.

Apesar dos motores de busca atuais se encontrarem bastante evoluídos, os anseios e pretensões dos usuários podem superar as funções disponíveis. Os mecanismos de busca tradicionais são baseados na similaridade entre as palavras-chave fornecidas no momento da busca com o conteúdo dos documentos. Isto não é suficiente quando a busca envolve um contexto geográfico, como, por exemplo, a consulta “retorne páginas sobre eventos políticos que estão fora de um raio de 100 km da cidade de Campina Grande/PB”. Para suprir tal carência, surgiu a área de Recuperação da Informação Geográfica (GIR) (Jones, 2008), que envolve, além das funções básicas tradicionais, o reconhecimento de referências geográficas espalhadas pelo conteúdo da página (e.g. nome de lugares, códigos postais, códigos de área telefônicos) permitindo a descoberta de conteúdos relacionados a uma localidade geográfica por meio de operações geográficas e tem seu uso aplicável em diversos domínios (e.g. aplicações com foco em turismo; ferramentas específicas para busca e leitura de notícias).

---

<sup>1</sup> The size of the World Wide Web. Disponível em: <<http://www.worldwidewebsite.com/>>. Acesso em dezembro de 2015.

<sup>2</sup> Internet usage statistics. Disponível em: <<http://www.internetworldstats.com/stats.htm>>. Acesso em dezembro de 2015.

Porém, dos processos envolvidos em ambos os motores de busca tradicionais e o de busca geográfica, várias questões de pesquisa ainda estão em aberto, incluindo a relevância dos resultados em resposta a uma determinada consulta, técnica conhecida como *ranking*. A pesquisa em relevância textual está bem mais avançada do que a pesquisa em relevância geográfica (Kumar e Boll, 2013). Segundo Kumar (2011), um dos desafios presentes na área de GIR é o de como ordenar a resposta de uma busca de forma que essa seja relevante para o usuário.

A dificuldade de ranquear resultados existe para o conteúdo distribuído na Internet e se agrava quando o escopo se reduz apenas às notícias informativas. No cenário que envolve a descoberta de notícias, um fator fundamental para medir a relevância está relacionado diretamente com a credibilidade e veracidade associadas às mesmas. Infelizmente, mensurar este tipo de dado é uma tarefa complexa, pois é fundamentada na investigação fiel do fato, se há confiança no que está escrito e se a fonte é idônea e segura.

Apesar da dificuldade em creditar um valor de confiabilidade às notícias informativas, os usuários (leitores) continuam procurando alternativas para se manterem bem informados e encontraram nas redes sociais uma opção ágil, mas que não traz solução para este problema. Atualmente, já são cerca de 1,96 bilhões de usuários de redes sociais conectados e há projeção para que sejam 2,44 bilhões em 2018<sup>3</sup>. Por esse número expressivo de usuários conectados, do qual ampla parcela faz uso da rede para se informar, grandes empresas do ramo se esforçam para oferecer funcionalidades, agregadas às redes sociais, que permitam que o usuário busque, leia e difunda notícias. No entanto, esta ideia ainda é muito incipiente, havendo espaço para pesquisas visando fornecer uma melhor experiência para o usuário que procura este tipo de serviço.

Deste modo, nesta dissertação é proposta a criação de um motor de busca geográfica, destinado à leitura de notícias distribuídas na Internet, que faz uso de dados proveniente de redes sociais para ordená-las considerando, além da relevância textual e geográfica, o fator de credibilidade atribuído às mesmas. Este fator recebe o nome de *Relevância Geosocial* e é baseado na localização dos usuários que difundem a notícias em suas redes sociais. Assim, viabilizar consultas em um contexto geográfico, o que é, na maioria das vezes, ausente em aplicações destinadas à leitura de notícias, e criar uma forma de integrar informações

---

<sup>3</sup>Number of social network users worldwide from 2010 to 2018  
<<http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>>. Acessado em janeiro de 2016.

provenientes de redes sociais, estabelecerá uma nova percepção do usuário sobre a veracidade e confiabilidade do que se está lendo.

Para medir o fator de credibilidade associado à notícia, assumimos que quando um usuário compartilha com seus amigos uma notícia ocorrida na sua região, este *feedback* representa um indício de credibilidade dada pelo usuário àquela notícia. A concepção do novo método de *ranking* de relevância segue a premissa de que uma notícia, relatando um acontecimento em determinada região, pode ter sua veracidade validada pela população envolvida nesta região descrita, independente se o meio de comunicação seja ou não confiável. Como exemplo, um jornalista pode noticiar constantes engarrafamentos em uma determinada cidade na qual ele nunca visitou. Essa informação pode ser verdadeira, mas ter sido coletada de uma fonte não confiável. Contudo, os habitantes daquela região, implicitamente, podem confirmar essa notícia apenas difundindo-a em suas redes sociais.

### 1.1. DEFINIÇÃO DO PROBLEMA

Atualmente, há uma grande demanda informativa por meio da população de internautas. Segundo a assessoria de imprensa do Twitter<sup>4</sup>, cerca de 60% dos seus usuários utilizam sua ferramenta para se manterem bem informados. Da mesma forma, segundo a Pesquisa Brasileira de Mídia<sup>5</sup>, 67% dos brasileiros acessam a Internet com o propósito de se manter atualizado dos acontecimentos. Em contrapartida, poucas são as ferramentas disponíveis que possibilitem uma busca geográfica por notícias e informações. A distribuição de veículos de comunicação na rede (i.e. cada veículo de comunicação possui seu portal de notícias), a falta de uma estrutura das páginas e a necessidade de apresentar ao leitor uma notícia verdadeira e confiável, são alguns dos obstáculos que retardam o surgimento de ferramentas com este propósito.

Neste âmbito, para fornecer ao usuário uma notícia na qual ele possa confiar no que está sendo dito, é exigido que essa notícia, proveniente de mídias espacialmente espalhadas e produtoras de informações sobre diversas regiões fora de sua alçada, passe por um crivo manual (i.e. averiguação realizada por pessoas capacitadas) verificando a consistência dos dados e das informações ali contidas. Em um mundo onde os acontecimentos são relatados constantemente em tempo real, esse trabalho torna-se complexo de ser realizado por meio manual.

---

4 <https://twitter.com>

5 <http://www.brasil.gov.br/governo/pesquisa-brasileira-de-midia>

Portanto, encontrar uma solução que extraia, analise e integre informações oriundas de redes sociais, torna-se factível na produção de um fator no *ranking* de relevância que mede o grau de confiabilidade de uma determinada notícia.

## 1.2. OBJETIVOS

### 1.2.1. Objetivos Gerais

O objetivo principal desse trabalho é produzir um novo *ranking* de relevância para notícias baseado na integração de dados oriundos das redes sociais e aplicá-lo a uma ferramenta de busca geográfica no contexto de notícias.

### 1.2.2. Objetivos Específicos

1. **Construir aplicação para descoberta de notícias no Twitter:** construir uma aplicação capaz de coletar e armazenar notícias que estão espalhadas na rede social Twitter e os usuários que interagiram com as mesmas. O objetivo é, partindo de um conjunto de contas oficiais de veículos de comunicação no Twitter, capturar e armazenar *posts* e *retweets* sobre alguma notícia. Quando uma notícia é postada por um veículo de comunicação em sua conta oficial no microblog, esta segue o formato “manchete *link*”, onde este *link* redireciona para o conteúdo completo da notícia. Este, por sua vez, será utilizado para a indexação da notícia e o perfil dos usuários que a difundiram (i.e. *retweet*) será usado para inferir a credibilidade desta.
2. **Desenvolver mecanismo de análise de perfil dos usuários do Twitter:** com o propósito de checar se os usuários que difundiram a notícia no microblog vivem ou estão espacialmente próximos da localidade tratada na notícia, é primordial a análise do perfil destes usuários. Esse perfil é composto por diversos campos informativos, fornecidos pelo próprio usuário no momento da sua inscrição no microblog. Um destes campos está relacionado com a localização de moradia, que é informada de maneira textual e sem verificação de consistência. Este campo será analisado de forma a identificar geograficamente o que foi fornecido. Por exemplo, se o usuário preencheu o campo localização do seu perfil como o valor “João Pessoa/PB” esse processo irá converter esse dado textual para um dado espacial (i.e. latitude, longitude), atividade esta conhecida como geocodificação. Por fim, esse dado será utilizado para mensurar o grau de credibilidade daquela notícia na região retratada e assim compor o novo ranking de relevância.

3. **Desenvolver mecanismo de inferência de afinidade com a localização:** a partir da interação do usuário com o microblog, identifica-se uma lista de localidades a qual o mesmo se relaciona ou tem afinidade. Embora a existência do campo de localização de moradia no perfil do usuário ser de extrema importância, nem sempre o mesmo é coerente ou é fornecido. Para viabilizar uma alternativa para essa adversidade, infere-se um grau de afinidade do usuário com uma determinada localidade, permitindo que o cálculo de credibilidade não seja impactado pela ausência de informação no perfil do usuário. Esse grau de afinidade do usuário com uma localidade é determinado pelas atividades por ele exercidas no microblog (e.g. *posts*, relações de amizades).
4. **Estender o motor de busca GeoSEn:** adicionar um módulo específico para o tratamento de notícias provenientes de *links* coletados no microblog Twitter. Essa extensão visa aprimorar o GeoSEn no tocante ao *ranking* de relevância, que será acrescido do novo fator proposto no trabalho.
5. **Construir ferramenta de busca e leitura de notícias:** para utilizar todo o ferramental fornecido pelo GeoSEn, é preciso que a interface gráfica seja remodelada para permitir a leitura de notícias informativas. Essa nova ferramenta de leitura de notícias, chamada de GeoSEn News, é responsável por prover ao usuário final uma nova forma de buscar, encontrar e ler notícias que estão distribuídas por toda rede. Para isso, um novo módulo de entrada de dados e um mapa interativo foram construídos, onde é possível visualizar rapidamente a localidade da qual a notícia trata, agilizando o processo de leitura.

### 1.3. CONTRIBUIÇÕES

Dentre as principais contribuições apresentadas nesta dissertação, pode-se destacar:

1. A proposição de uma nova abordagem no modo em que notícias informativas são ordenadas: considerar a localização dos usuários que difundem uma determinada notícia, em suas redes sociais, de forma a atribuir um valor de credibilidade daquela quando se refere a uma localidade.
2. A modelagem de uma nova técnica para estimar o grau de afinidade entre localidades geográficas e os usuários de microblogs, apenas considerando suas interações nas redes sociais através de *posts* e amizades. Essa técnica é utilizada para medir o valor de credibilidade que um usuário pode agregar a uma determinada notícia que o mesmo compartilhou em sua rede social.

3. A construção de uma plataforma para leitura de notícias em um contexto geográfico, envolvendo todas as funções populares em motores de busca tradicionais acrescidas de funcionalidades características de motores de busca com ênfase geográfico, como, por exemplo, a consulta e visualização de notícias em uma mapa interativo. As notícias disponíveis nesta ferramenta são coletadas das contas oficiais (no microblog Twitter) de vários veículos de comunicação distribuídos pelo país.

#### 1.4. TRABALHOS PUBLICADOS

Esta pesquisa resultou na seguinte publicação:

- *The Geo-social Relevance Ranking: A method based on geographic information and social media data.*
  - Autores: Julio Henrique Rocha, Claudio E. C. Campelo, Claudio de Souza Baptista, Gabriel Joseph Ramos Rafael.
  - Conferência: 13th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2016.
  - Local: Agadir, Marrocos.

#### 1.5. ORGANIZAÇÃO ESTRUTURAL

O restante desta dissertação está organizado da seguinte maneira: no Capítulo 2, são expostos os fundamentos da recuperação da informação geográfica, com foco principal nos motores de busca para Web, bem como o estado da arte dos processos de *ranking* de relevância de documentos da Web. No Capítulo 3, é apresentado um panorama dos trabalhos mais relevantes da área com foco na produção de estudos relacionados ao *ranking* de relevância em documentos. No Capítulo 4, são descritos os artefatos, métodos e técnicas elaboradas para produção deste trabalho, bem como as ferramentas desenvolvidas para assegurar a validade do que foi proposto. No Capítulo 5, descrevem-se os experimentos realizados para validação dos métodos propostos, contemplando a análise e discussão dos resultados. Por fim, no Capítulo 6, apresentam-se as considerações finais, restrições observadas no trabalho e os trabalhos futuros.

## CAPÍTULO 2

### FUNDAMENTAÇÃO TEÓRICA

A proposta fundamental da Recuperação da Informação (*Information Retrieval* – IR), termo este introduzido pelo pesquisador americano Calvin Northrup Mooers no ano de 1952 (Jones, 1997), é fornecer ao usuário a possibilidade de encontrar conteúdo que satisfaça uma necessidade de informação. Na perspectiva do usuário, sua necessidade deve ser convertida em um conjunto de palavras-chave que, por sua vez, são fornecidas ao motor de busca para comparação com os documentos previamente armazenados. A tarefa do motor de busca é retornar ao usuário os documentos ou registros que mais se assemelham com as palavras-chave fornecidas pelo usuário (Manning et al., 2008). Apesar de parecer simples, o processo de recuperação da informação envolve etapas complexas e cruciais para o seu funcionamento adequado.

Com o intuito de expandir as funcionalidades presentes em motores de busca provenientes de IR, surge a área de Recuperação de Informação Geográfica (*Geographic Information Retrieval* – GIR), permitindo que o usuário, além de realizar buscas por documentos que satisfazem um determinado conjunto de palavras-chave, adicione a sua consulta um contexto geográfico. Grande parte dos documentos disponíveis possui alguma referência geográfica em seu conteúdo, como, por exemplo, nome de cidade, CEP, código postal e pontos de interesse. Então, por meio de um tratamento diferenciado a essas referências, torna-se possível a consulta por documentos relacionados a uma localidade ou espaço geográfico. Por exemplo, considere um usuário que deseja obter documentos relacionados aos acidentes automobilísticos no estado da Paraíba. Utilizando um motor de busca tradicional, apenas documentos que possuem a palavra “Paraíba” são retornados. O fato é que esse resultado, apesar de válido, é incompleto. Documentos que não possuem a palavra “Paraíba”, mas sim “Campina Grande”, também deveriam ser retornados, o que não ocorre. Portanto, adicionar uma contextualização geográfica nos processos envolvidos com a recuperação de informação pode propiciar resultados mais satisfatórios ao usuário, ou mesmo permitir que uma simples consulta substitua múltiplas consultas a um sistema de busca tradicional para se obter um resultado similar.

O restante deste capítulo está organizado da seguinte maneira: primeiro, serão apresentados os conceitos básicos que norteiam o RI, seus fundamentos, métodos e técnicas utilizados de forma unânime em motores de busca. Em seguida, descrevem-se as

características de motores de busca direcionados para Web. Por fim, discutem-se as características de um sistema de recuperação de informação geográfica, seus conceitos e desafios envolvidos em sua produção, bem como particularidades consideradas na interface para busca geográfica.

## 2.1. RI – RECUPERAÇÃO DA INFORMAÇÃO

De acordo com a literatura, os sistemas de recuperação da informação se distinguem em dois tipos: sistema de recuperação de dados e sistema de recuperação de informação. Conforme Rijsbergen (1979), os sistemas de recuperação de dados objetivam recuperar todos os objetos, de forma exata e precisa, que satisfazem as condições expressamente definidas por uma linguagem formal. Em contrapartida, um sistema de recuperação da informação traz consigo dificuldades associadas na identificação da real necessidade do usuário, que por sua vez, faz uso de linguagem textual para definir seu desejo. Outro desafio encontrado neste tipo de sistema está relacionado na definição do subconjunto relevante de objetos que satisfazem o real desejo expresso pelo usuário.

Baeza-Yates et al. (1999) definem um modelo de recuperação de informação como uma quádrupla  $[D, Q, F, R(q_i, d_j)]$ , onde:

- $D$  é o conjunto de visões lógicas dos documentos;
- $Q$  é o conjunto de visões lógicas das consultas dos usuários (*queries*);
- $F$  é um framework para modelar  $D$ ,  $Q$  e suas relações;
- $R(q_i, d_j)$  é um método de *ranking* que associa um número real a uma consulta  $q_i \in Q$  e uma visão lógica de um documento  $d_j \in D$ ;

A seguir, demonstram-se alguns conceitos importantes referidos ao processo de recuperação da informação.

### 2.1.1. Operação textual

Um documento pode ser representado por o conjunto de todas as palavras contidas em seu conteúdo. Este método de representação, descrito por Baeza-Yates et al. (1999), recebe o nome de *full text* e consiste em transformar um documento em um conjunto de palavras de forma a facilitar sua recuperação (visão lógica). De acordo com os autores, esta forma de representação é a mais completa. Porém, para grandes coleções de documentos, pode ser necessário diminuir esse conjunto de palavras representativas, com o objetivo de reduzir os custos computacionais envolvidos na recuperação da informação. Esse processo de redução, em geral, é realizado através da eliminação de *stopwords* (e.g. artigos, preposições),



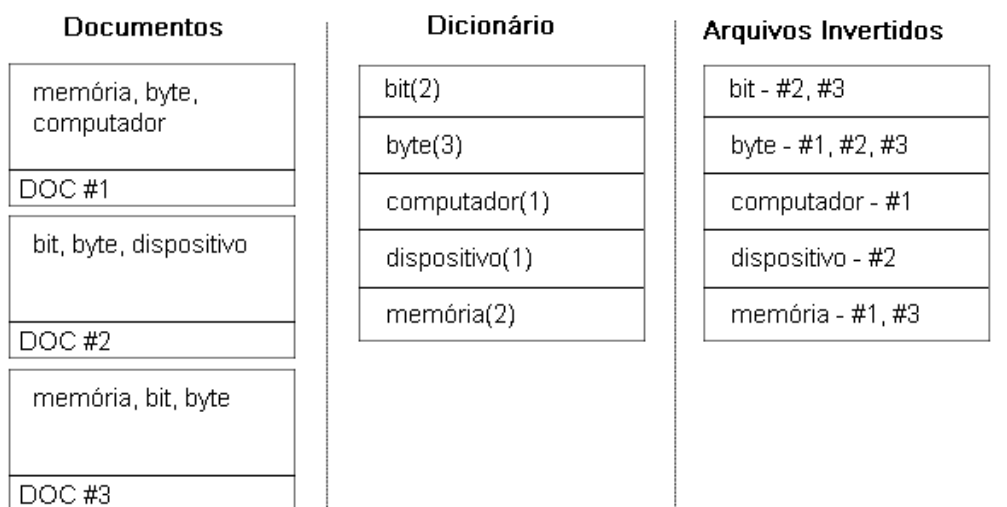
do uso de *stemming* (i.e. redução de palavras distintas a suas raízes gramaticais comuns) e da identificação de grupos de substantivos (e.g. que elimina adjetivos, verbos e advérbios). Estas operações são chamadas de operações textuais ou, simplesmente, transformações. Com estes procedimentos, reduz-se a visão lógica do documento a um conjunto de termos indexáveis. Similarmente, as consultas dos usuários definidas por um conjunto de palavras-chave são submetidas aos mesmos tipos de transformação, para se obter uma visão lógica daquela necessidade.

### 2.1.2. Indexação

Dentre os desafios encontrados na área de recuperação de informação, destaca-se aquele relacionado ao desenvolvimento de algoritmos e estruturas de dados eficientes na armazenagem e recuperação de informação. Este desafio inclui a representação e construção de índices e avaliação de consultas para busca (Zobel e Moffat, 2006). Das principais técnicas desenvolvidas pela comunidade científica, destacam-se as arquivos de assinatura (*signature files*), vetores de sufixo (*suffix arrays*) e arquivos invertidos (*inverted files*), sendo esta última a técnica mais comumente utilizada em ferramentas de RI (Navarro et al., 2001).

Segundo Navarro (1999), um arquivo invertido é um mecanismo de indexação orientado à palavra e sua estrutura é composta de duas partes: o vocabulário e as ocorrências. O vocabulário é um conjunto de palavras distintas, selecionadas por meio de operações textuais (vide Seção 2.1.1). Logo, as ocorrências são listas que fornecem informações sobre cada palavra no vocabulário. Essas informações são determinadas baseadas no contexto da aplicação, como, por exemplo, posição da palavra em determinada estrutura de texto ou *links* para documentos nos quais a palavra (índice) aparece. Na Figura 1 é exemplificada a técnica de arquivo. Na primeira coluna, são exibidos os documentos a serem indexados. Na coluna “Dicionário”, tem-se as palavras que foram identificadas nos documentos, seguido pela quantidade de vezes que são encontradas. Na terceira coluna, apresenta-se o contexto de arquivos invertidos, onde a palavra passa a ser o índice que aponta para os documentos que a mesma pode ser encontrada.

Figura 1: Representação da técnica de arquivos invertidos.



Fonte: elaborado pelo autor.

### 2.1.3. Ranking de Relevância

A principal finalidade de um sistema de RI é fornecer resultados relevantes com qualidade e rapidez que corroboram com as necessidades do usuário transcritas em forma de consulta. O processamento de uma consulta consiste em analisar as palavras ou expressões fornecidas pelo usuário e compará-las com o índice, a fim de encontrar respostas relevantes ao seu interesse. As expressões podem ser compostas por operadores lógicos, como, por exemplo, “e”, “ou” e “não”.

Para estabelecer a ordem ótima de demonstração dos resultados, diversas técnicas de ordenação (do inglês, *ranking*) foram desenvolvidas com o objetivo de construir um grau de similaridade entre o conjunto de identificadores dos documentos e o conjunto de termos transcritos na consulta. Algumas destas técnicas são baseadas na quantidade de vezes que a palavra aparece na página, pela classificação de hubs, dentre outros (Kleinberg e Lawrence, 2001). Outras abordagens, direcionadas para uso em sistemas de RI para Web, consideram a estrutura de *links* da web, como, por exemplo, o PageRank (Page et al., 1999) e o Hyperlink Vector Voting (Li, 1998).

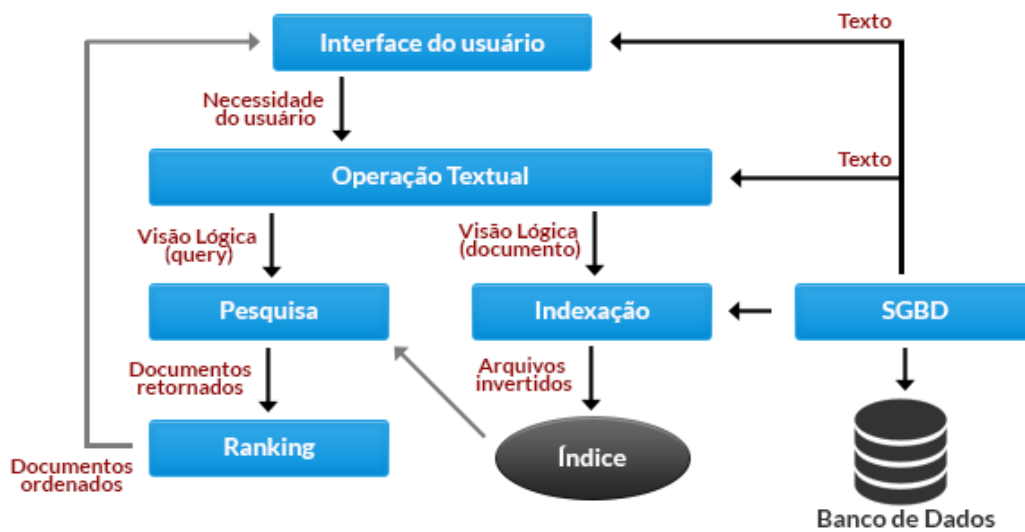
### 2.1.4. Arquitetura de Sistemas de RI

Baeza-Yates et al. (1999) definem uma arquitetura básica em um sistema de RI. Primeiramente, é determinado o mecanismo de armazenamento (banco de dados) no tocante à como os documentos são armazenados, as operações textuais aplicadas aos mesmos, dentre outros. Em seguida, é definida a visão lógica dos documentos proveniente da construção de índices, na maioria dos casos um arquivo invertido. Depois de construído o índice, o processo de busca pode ser iniciado. Na Figura 2, adaptada de Baeza-Yates et al. (1999), é apresentada

uma arquitetura simplificada de um software genérico de RI, onde os retângulos representam os módulos dos sistemas e as arestas, a comunicação entre os mesmos. Seguindo esta arquitetura, os autores propõem as seguintes etapas:

- Etapa de consulta: o usuário define sua consulta, que é analisada e transformada utilizando os mesmos mecanismos de operações textuais aplicados nos documentos;
- Representação formal da consulta: a consulta transformada (i.e. representação do sistema para a consulta) é processada para obter os documentos que a satisfazem;
- Ordenação: os documentos encontrados são ordenados de acordo com uma métrica de relevância. A ordem dos documentos será retornada para o usuário da maior para a menor relevância;
- Resultado: o usuário analisa o resultado retornado pelo sistema RI e, caso a ferramenta dê suporte, inicia o ciclo de *feedback*, onde o sistema utiliza a interação do usuário com os documentos para melhorar uma futura consulta;

Figura 2: Arquitetura básica de um sistema de RI.



Fonte: adaptada de Baeza-Yates et al. (1999).

### 2.1.5. Modelos de RI

A recuperação da informação baseia-se na ideia de palavras-chave para encontrar um documento. Porém, frequentemente os termos de uma consulta podem estar presentes em um documento, mas o mesmo não ser relevante para o usuário. Isto se refere ao teor semântico do termo no documento, o que pode torná-lo relevante ou não. Com isto em vista, para cada termo presente no documento, é atribuído um  $peso(d,s)$  relacionado a sua relevância, ou seja, um valor numérico que indica o grau de relevância do termo  $s$  no documento  $d$ . Em cada

modelo, estes pesos podem ser ou não considerados como mutuamente independentes. Os modelos clássicos mais difundidos são o modelo booleano, o modelo probabilístico e o modelo vetorial (Baeza-Yates et al., 1999).

O modelo booleano é baseado na álgebra booleana. Neste modelo, apenas é determinado se o termo está ou não presente no documento e assim, o peso atribuído a cada termo é binário, ou seja,  $\text{peso}(d, s) \in \{0,1\}$ . Ademais, uma consulta tem formato de uma expressão booleana, a qual é composta por palavras ligadas por conectivos booleanos (i.e. “e”, “ou”, “não”). A grande vantagem do uso deste modelo está em sua simplicidade e sua desvantagem é não permitir uma relevância parcial, ou seja, ou o documento é relevante ou irrelevante.

O modelo probabilístico apresentado por Robertson e Jones (1976) tenta mensurar, para uma dada consulta  $q$  e um documento  $d_i$  na coleção  $D$  de documentos, a probabilidade do usuário considerar o documento  $d_i$  relevante. Para tal, considera-se a existência de um subconjunto  $\epsilon D$  que contém apenas documentos relevantes. Mediante descrição preliminar deste conjunto ótimo, como é chamado, um subconjunto retornado para a consulta é considerado relevante. Então, este subconjunto é reutilizado para aprimorar a descrição do conjunto ótimo. Este processo é feito repetidamente até a obtenção de uma aproximação do conjunto ótimo desejado. A vantagem deste modelo está relacionada à possibilidade de ordenar o resultado de acordo com a probabilidade de relevância dos documentos. Em contrapartida, o fato de não considerar a frequência que um termo ocorre no documento como valor de representatividade, é um ponto negativo deste modelo.

O modelo vetorial propõe a ideia de parcialidade, o que permite variação (não booleanos) aos pesos dos termos. Deste modo, o grau de relevância entre o termo e o documento é considerado no momento da ordenação. Neste modelo, além dos pesos em relação aos documentos, é preciso realizar o cálculo do grau de relevância do termo nas consultas, definindo então, o vetor  $\vec{q} = (s_{1,q}, s_{2,q}, s_{3,q}, \dots, s_{t,q})$ , no qual  $t$  é o número total de termos definidos na consulta. Segundo Baeza-Yates et al. (1999), a similaridade entre um documento e uma consulta pode ser definida como a correlação entre os  $d_j$  e  $q$  e quantificada pelo cosseno do ângulo formado entre eles, como é visto na Equação 1.

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^t w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2} \times \sqrt{\sum_{i=1}^t w_{i,q}^2}} \quad \text{Equação 1}$$

O peso de um termo pode ser calculado de várias formas. Alguns métodos são baseados no número de ocorrências destes no documento, ou seja, a frequência de menções.

Uma das abordagens para o cálculo do peso recebe o nome de TF-IDF. Esta forma tenta balancear características em comum nos documentos e características que distinguem os mesmos, combinando, respectivamente, as medidas de Frequência do Termo (do inglês, *Term Frequency* - TF), que mede a frequência do termo  $t_i$  no documento  $d_j$  e a medida conhecida como Frequência Inversa de Documentos (do inglês, *Inverse Document Frequency* - IDF) que mede a frequência inversa do termo  $t_i$  na coleção de documentos  $D$ . Os cálculos estão definidos seguindo a Equação 2, sendo  $freq_{i,j}$  a frequência do termo  $t_i$  no documento  $d_j$ ,  $N$  o número total de documentos e  $n_i$  a quantidade de documentos em que o termo  $t_i$  aparece.

$$TF_{i,j} = \frac{freq_{i,j}}{\max_l \cdot freq_{l,j}}$$

$$IDF_i = \log \frac{N}{n_i}$$

$$w_{i,j} = TF_{i,j} \cdot IDF_i$$

Equação 2

## 2.2. MECANISMOS DE BUSCA PARA WEB

Documentos disponibilizados na Internet são extremamente heterogêneos. Isto ocorre em diferentes aspectos, seja pela utilização de inúmeras linguagens, pelo formato utilizado para representação de conteúdo (HTML, PDF, texto, vídeo, fotografia, imagem), pelo emprego de diferentes idiomas, volatilidade dos dados, redundância entre os documentos, dentre outros, além de fatores externos ao conteúdo do documento, como exemplo, a confiabilidade da informação, a reputação da fonte e várias outras características (Page et al., 1999). Não obstante, não há nenhum controle sobre a estrutura do documento em relação à organização do conteúdo, a capacidade de comunicação do escritor, nem a forma em que estes são armazenados. Estes obstáculos adicionam um grau de complexidade elevado aos sistemas de informação quando possuem como alvo documentos espalhados na Web.

Um dos ambientes onde mais se aplica o uso da recuperação da informação é a Internet. Por conter uma porção gigantesca de documentos distribuídos, os motores de busca são fundamentais na Internet e se concretizam como o serviço mais utilizado entre os internautas, junto com os sistemas de correio eletrônico. Segundo o Desktop Search Engine Market Share (2016), os motores de busca para Web mais utilizados na atualidade são o Google (67,7%), o Bing (15,6%), o Baidu (7,7%) e o Yahoo (7,1%).

Os sistemas de busca para Web podem ser descritos basicamente em três partes: O *spider* (também conhecido como *Web crawler* ou robô) responsável por localizar e capturar

páginas na Web; o indexador, responsável por indexar documentos encontrados pelo *spider* (ver Seção 2.1.2); e por fim, o processador de consultas, que compara a consulta realizada pelo o usuário, identifica os documentos que a atende e responde para o usuário um conjunto de documentos ordenados do mais relevante para o menos relevante. Estes elementos são descritos em mais detalhes a seguir.

### 2.2.1. Spider (Web crawler)

O *spider* ou *crawler* é um programa que percorre o ciberespaço de forma metódica e automatizada objetivando alcançar as porções mais substanciais do domínio. Este mecanismo faz uso da estrutura de *links* da Internet para visitar outras páginas e encontrar novas que são de interesse. Também é sua finalidade a captura (cópia) do conteúdo para que o mesmo passe por uma etapa de análise futura. Posteriormente à análise submetida às páginas encontradas, as mesmas são indexadas, ficando disponíveis no momento da consulta, que deste modo, será realizada de forma veloz e eficiente. Na Figura 3 é apresentada uma arquitetura básica para um *crawler* Web.

Figura 3: Arquitetura básica de um *crawler* Web.



Fonte: elaborado pelo autor.

Os Coletores são os responsáveis por solicitar, através de requisições HTTP, páginas aos servidores na Web, extrair os *links* das páginas recebidas e enviá-los ao Escalonador e, por fim, requisitar ao Escalonador URLs a serem capturadas. O Escalonador de páginas é responsável por decidir qual a próxima URL será capturada pelos Coletores e suas ações. O Servidor de Dados, por fim, é responsável por armazenar em uma base de dados todas as páginas ou objetos coletados. Também é de sua responsabilidade realizar uma análise a fim de identificar o formato da página e realizar os tratamentos adequados a cada tipo.

O algoritmo executado por um *spider* pode ser descrito da seguinte forma:

1. O administrador do sistema fornece uma lista de URLs como entrada (*seeds*);
2. O Coletor solicita uma URL ao Escalonador;
3. O Escalonador seleciona uma URL e a remove da lista de URLs a serem coletadas;
4. O Escalonador retorna a URL ao Coletor;
5. O Coletor carrega a página ou documento correspondente à URL fornecida pelo Escalonador;
6. O Coletor envia o conteúdo capturado ao Servidor de Dados para que o mesmo seja armazenado;
7. O Coletor extrai os links contidos no documento;
8. O Coletor envia ao Escalonador os links identificados para que sejam adicionados à lista caso ainda não tenham sido visitados anteriormente;
9. O passo 2 é repetido até que não haja mais URLs a serem visitadas;

Os *spiders* possuem diversas competências que permitem a coleta e análise de documentos distribuídos na Internet, contudo, só são capazes de alcançar o conteúdo público, ou seja, apenas a porção disponível na Internet que não seja necessário o preenchimento de formulários ou campos para serem acessadas. Contudo, segundo estimativas realizadas por Raghavan e Garcia-Molina (2001), a Web “oculta” representa cerca de 500 vezes o tamanho da Web pública ou alcançável. Deste modo, os autores propõem uma nova abordagem para coleta de conteúdo oculto na Web seguindo uma nova técnica nomeada de Técnica de Extração de Informação Baseada em Layout. Os autores trazem uma demonstração da eficiência do método aplicando seu uso na extração de conteúdos de páginas em resposta a formulários preenchidos.

### **2.2.2. Indexação**

A etapa de indexação de conteúdo em sistemas de busca para Web é similar ao realizado em sistemas tradicionais de RI. Na maior parte dos casos, os sistemas fazem uso da técnica de arquivos invertidos (ver Seção 2.1.2) para indexar as páginas e documentos que estarão disponíveis para consulta. Por exemplo, o mecanismo de busca do Google (Page et al., 1999) utiliza dois componentes para realizar a indexação: o indexador e o ordenador. O indexador realiza diversas funções, desde a leitura do repositório, a descompressão dos documentos e *parsing* nos documentos. Deste modo, um documento é representado por um conjunto de ocorrências chamadas de *hits*, que por sua vez guardam informações acerca do tamanho da fonte, posicionamento da palavra no documento, utilização de artefatos de

destaque (e.g. negrito, itálico), dentre outros. Também é de responsabilidade do indexador realizar o *parsing*, extraindo das páginas e armazenando informações sobre redirecionamentos (i.e. *links*).

### 2.2.3. Ranking de Relevância

O processo de classificação, ou simplesmente *ranking* de relevância, em sistemas de busca na Web pode adotar técnicas diferentes das utilizadas em sistemas tradicionais de RI. Nos sistemas tradicionais, considera-se apenas a similaridade entre a consulta realizada pelo usuário e os documentos indexados pelo sistema. No entanto, outros fatores podem ser considerados no momento de realizar a classificação, como, por exemplo, o PageRank (Page et al., 1999) da empresa Google. O algoritmo proposto pelos autores Page et al., representa toda a Web em um grafo direcionado, onde os vértices são as páginas, as arestas os *links* entre as mesmas e as direções seguem o sentido do *link*. A partir desta perspectiva, uma página é dita como importante quando a mesma possui um conjunto de *backlinks* (*links* que apontam para esta página) realizados por páginas importantes. Assim, uma página pode ser considerada mais importante que outra apenas observando as páginas que a estão referenciando através de *links*.

## 2.3. GIR- RECUPERAÇÃO DA INFORMAÇÃO GEOGRÁFICA

Com o intuito de expandir as funcionalidades presentes em motores de busca provenientes de IR, surge a área chamada de Recuperação de Informação Geográfica (do inglês, *Geographic Information Retrieval* - GIR) permitindo que o usuário, além de realizar buscas por documentos que satisfazem um determinado conjunto de palavras-chave, possa adicionar a sua consulta um contexto geográfico. No entanto, a expansão das funções de sistemas de RI tradicionais para dar suporte a consultas em contexto geográfico, endereça inúmeros desafios e exigem a adaptação dos principais componentes do sistema tradicional.

Assim, nas seções a seguir, são apresentadas as principais características deste tipo de sistema, bem como apontar diversas soluções propostas para os diversos desafios enfrentados na área.

### 2.3.1. Detecção de Referências Geográficas

Uma das etapas primordiais para o bom funcionamento de um sistema de GIR está relacionada à inferência das localidades geográficas associadas aos documentos Web capturados pelo *crawler*. Existem diversas soluções neste sentido, como, por exemplo, a proposta por Markowetz et al. (2005), onde as informações geográficas podem ser deduzidas



no conteúdo das páginas ou na estrutura dos *links*. A solução proposta pelos autores consiste em subdividir o processo de identificação em duas partes: extração e mapeamento. Na primeira delas, são identificados os elementos capazes de referenciar uma localidade, como, por exemplo, o nome de lugares, códigos postais ou telefones. Na segunda etapa, cada referência identificada é associada a uma localidade geográfica válida.

Diversas outras soluções que utilizam o conceito de *geocoding* na inferência de localidades geográficas no conteúdo da página podem ser encontradas (McCurley, 2001; Markowitz et al., 2005). Porém, apesar do número vasto de soluções para esta atividade, algumas dificuldades ainda são encontradas no processo, como, por exemplo, a ocorrência de ambiguidades (e.g. várias cidades com nome de pessoas ou objetos; vários lugares com o mesmo nome; vários nomes para o mesmo local). O processo que visa resolver o problema de ambiguidade é chamado de Desambiguação de Topônimos (do inglês, *Toponym Disambiguation – TD*) ou Resolução de Toponímia. Algumas soluções de TD fazem uso de mecanismos externos para determinação exata de uma referência geográfica localizada em uma página ou documento. O mecanismo mais utilizado neste sentido é o *gazetteer*, um serviço geográfico do tipo repositório que contém informações acerca de localidades geográficas (e.g. nome; nível hierárquico; *footprint*), bem como dados adicionais (e.g. estatística demográfica dimensão da área; substantivos e adjetivos) sobre a mesma (Hill, 2000). Este mecanismo tem como principal objetivo realizar o *geocoding* (i.e. transformação de uma referência geográfica textual em uma localização geográfica na superfície da terra) em referências geográficas encontradas em textos.

Uma solução alternativa para *TD* é proposta por Wang et al. (2010). Os autores se baseiam no processo de reconhecimento de localidade utilizado por humanos, onde estes fazem uso de algumas regras linguísticas na tentativa de identificar o contexto geográfico que está sendo relatado. Os autores focam em regras semânticas, onde uma localidade é identificada mediante a identificação de outras localidades no contexto que não foi necessário a desambiguação. Por exemplo, considere que um documento possui duas referências geográficas em seu conteúdo. A primeira referência (*A*) foi identificada de forma precisa, pois não havia ambiguidade em sua definição. A segunda referência identificada (*B*) corresponde ao nome encontrado em várias cidades do país. Deste modo, a segunda referência será determinada de acordo com o relacionamento semântico entre a mesma e a localidade (*A*) encontrada de forma precisa. Esta técnica é conhecida como Referência Cruzada (do inglês,

*Cross Reference*) e pode ser encontrada em outros trabalhos da área (e.g. Campelo e Baptista, 2008).

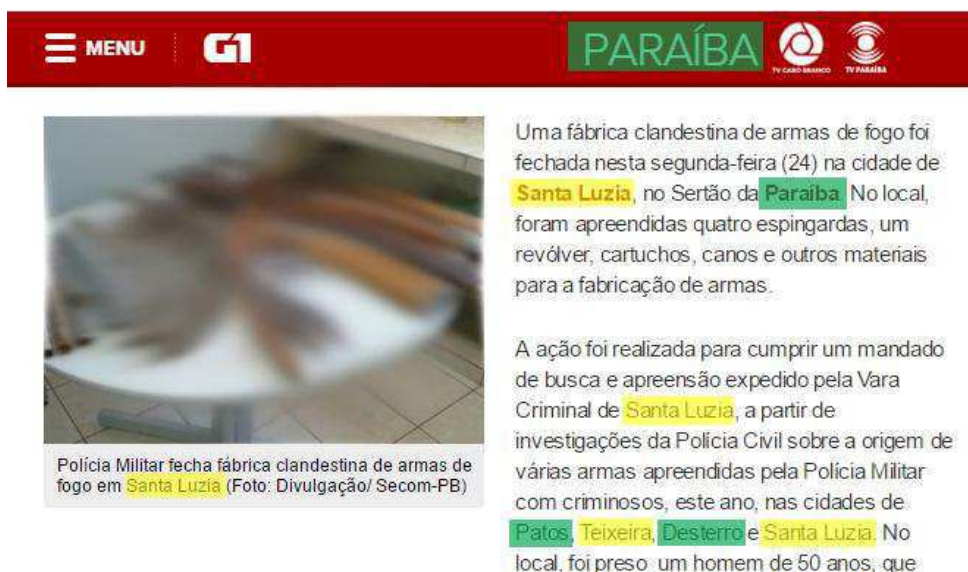
Yu Li et al. (2006) propõem um método para eliminação de ambiguidades de conteúdos em páginas Web. Neste, cada localidade ambígua identificada é submetida a um *gazetteer* para averiguação e então, é realizado o cálculo de um valor que representa a probabilidade daquela referência caracterizar uma localidade. No experimento realizado, foi feito o uso do *gazetteer* TNG<sup>6</sup>, que fornece informações sobre a localidade submetida, bem como dados para que seja possível calcular do valor probabilístico. Em seguida, outras heurísticas são utilizadas para refinar os valores probabilísticos, como, por exemplo, localidades relacionadas cujas referências já foram identificadas, termos geográficos, estatística populacional, dentre outros.

Na Figura 4 é ilustrado o processo de Desambiguação de Topônimos. Nesta pode-se perceber uma página da Web sendo analisada com o objetivo de extrair termos que podem referenciar alguma localidade geográfica. Os termos destacados em verde indicam a existência de uma referência a uma localidade que não foi necessário desambiguar, ou seja, sua identificação foi direta e precisa. Os termos destacados em amarelo indicam a existência de uma possível referência geográfica, mas que por vez, não há exatidão sobre qual localidade está sendo relatada ou se o termo realmente indica uma localização. No exemplo da figura, o termo “Paraíba” foi identificado de forma precisa. No entanto, os termos “Santa Luzia” exigem uma desambiguação, pois podem representar um nome pessoal, substantivo ou inúmeros municípios do Brasil (e.g. Santa Luzia/BA, Santa Luzia/MA, Santa Luzia/PB). A técnica de Referência Cruzada para Desambiguação de Topônimos atua neste tipo de cenário, onde uma referência geográfica encontrada de forma precisa auxilia na identificação de outras localidades, pelo motivo de estarem associadas semanticamente (i.e. “Santa Luzia” é um município do estado da “Paraíba”).

---

<sup>6</sup>TGN. Getty Thesaurus of Geographic Names.  
[http://www.getty.edu/research/conducting\\_research/vocabularies/tgn/](http://www.getty.edu/research/conducting_research/vocabularies/tgn/)

Figura 4: Exemplo do processo de Desambiguação de Topônimo.



Uma fábrica clandestina de armas de fogo foi fechada nesta segunda-feira (24) na cidade de **Santa Luzia**, no Sertão da **Paraíba**. No local, foram apreendidas quatro espingardas, um revólver, cartuchos, canos e outros materiais para a fabricação de armas.

A ação foi realizada para cumprir um mandado de busca e apreensão expedido pela Vara Criminal de **Santa Luzia**, a partir de investigações da Polícia Civil sobre a origem de várias armas apreendidas pela Polícia Militar com criminosos, este ano, nas cidades de **Patos**, **Teixeira**, **Desterro** e **Santa Luzia**. No local, foi preso um homem de 50 anos, que

Polícia Militar fecha fábrica clandestina de armas de fogo em **Santa Luzia** (Foto: Divulgação/Secom-PB)

Fonte: *print screen* do portal de notícias G1.

### 2.3.2. Modelagem do Escopo Geográfico

A etapa seguinte ao processo de detecção de referências geográficas é a modelagem do escopo geográfico dos documentos que serão indexados. Nesta etapa, são definidas quais as localidades geográficas que representam cada um dos documentos, a qual recebe o nome de escopo geográfico. O escopo geográfico pode ser definido como simples ou múltiplo. No primeiro caso, apenas uma localidade representa o documento, que na maioria dos casos, consiste na generalização das localidades identificadas na etapa anterior. No segundo caso, um documento pode ser representado por várias localidades.

Algumas propostas encontradas na literatura diferem em sua forma de representar o escopo geográfico do documento. Campelo e Baptista (2008) sugerem que um documento possa ser representado por inúmeras localidades (escopo múltiplo) e, deste modo, propõem uma nova abordagem para modelagem de escopo geográfico, que considera estatísticas das referências encontradas e a distribuição espacial das mesmas em um documento. Andogah et al. (2012) definem alguns métodos e estratégias, que utilizam o escopo múltiplo de documentos, para melhorar a performance de alguns processos envolvidos em GIR, como, por exemplo, a resolução de toponímia, expansão de consultas e o *ranking* de relevância. Por outro lado, Silva et al. (2006) defendem a suficiência de uma referência geográfica simplificada contendo apenas uma localidade (escopo simples) para definir a abrangência do documento. Segundo os autores, esta localidade poderia ser derivada ou generalizada das referências encontradas no documento. A estratégia de modelagem do escopo geográfica do

documento deve ser escolhida de acordo com os requisitos do sistema, contudo, pode-se perceber que o escopo múltiplo é mais utilizado pela comunidade para representação geográfica, pois permite uma expressão mais precisa do conteúdo encontrado e facilita no momento da recuperação do documento.

Na Figura 5 é exibido o mesmo documento sendo modelado de maneira distinta. No documento, quatro localidades foram identificadas através de referências encontradas em seu conteúdo. Na Figura 5-(a), representando a modelagem de escopo múltiplo, o documento foi representado por cada uma das localidades identificadas (Santa Luzia/PB, Patos/PB, Teixeira/PB e Desterro/PB). Já na Figura 5-(b), que representa a modelagem de escopo simples, o documento é representado por apenas uma localidade que abrange todas as identificadas no conteúdo. Na modelagem de escopo simples, a localidade escolhida para representar o documento não necessariamente precisa abranger todas as localidades que foram identificadas no documento. Esta pode ser escolhida através de heurísticas, como, por exemplo, a localidade de maior relevância para o documento ou a que possui a maior área territorial.

Figura 5: (a) Representação por escopo múltiplo; (b) Representação por escopo simples.



Fonte: montagem elaborada pelo autor.

### 2.3.3. Indexação Espaço-textual

Finalizado o processo de modelagem do escopo geográfico, é dado início à etapa de indexação espacial e textual dos documentos, de forma a viabilizar a recuperação rápida e ao mesmo tempo eficiente.

Diversas são as abordagens para manipular dados espaciais em GIR. Martins et al. (2005) realizam um *survey* em seu trabalho com o objetivo de vislumbrar o estado da arte em estruturas de indexação espacial conhecidas. *R-tree*, *quad-tree*, *grid*, são algumas das estruturas encontradas, sendo a *R-tree* a mais popular entre as demais.

Algumas técnicas podem ser empregadas na indexação espacial de documentos. Uma destas busca utilizar índices puramente textuais na indexação espacial, fazendo uso da mesma estrutura de arquivo invertido empregado em sistemas de busca tradicionais. Contudo, é exigido que o argumento geográfico, utilizado na consulta feita pelo usuário, coincida rigorosamente com o nome associado ao documento armazenado. Por exemplo, suponha que o usuário deseja encontrar “concursos públicos” na localidade A. Sua consulta gerada seria como “curso público em A”. Contudo, sabe-se que na prática o usuário pode estar interessado também em regiões próximas à consultada.

Uma alternativa para esta deficiência é a expansão de consultas. Esta técnica visa expandir a consulta submetida pelo usuário, por meio de heurísticas e relacionamento espacial, para abranger uma área que o usuário também pode estar interessado. Por exemplo, a consulta anteriormente mencionada, quando expandida, poderia gerar um resultado como “curso público em (A ou B ou C)”, considerando que B e C são regiões próximas e relacionadas com a localidade A, esta fornecida pelo usuário no momento da consulta. Por fim, apesar de útil, a expansão de consulta, se não bem modelada, pode ocasionar um crescimento demasiado na consulta interferindo diretamente no desempenho do sistema.

Outra alternativa para indexação espacial de documentos consiste em armazenar o *footprint* espacial documento. Apesar de mais simples desenvolvimento, essa prática esbarra em alguns problemas de desempenho. Armazenar geometrias de cada documento indexado pelo sistema pode demandar muito espaço à medida que o número de documentos presentes no sistema cresce. Outro obstáculo enfrentado neste tipo de abordagem está relacionado à performance do sistema. Realizar consultas espaciais, em tempo de execução, em milhões de geometrias é uma tarefa que exige muitos recursos, como, por exemplo, poder de processamento e memória principal. Em sistemas direcionados para páginas Web, essa

alternativa é pouco usual, por não ser escalável e pela exigência de respostas rápidas e precisas pelo usuário final.

#### **2.3.4. Ranking de Relevância Geográfica**

Presente entre os desafios da GIR está o processo de *ranking* dos resultados para uma determinada busca. De acordo com Andrade e Silva (2006), a maior parte das pesquisas na área está concentrada em solucionar problemas pautados em extração e indexação de informações e poucas são as pesquisas relacionadas com *ranking* de relevância.

Atualmente, mecanismos de busca geográfica estão ordenando os resultados por meio de uma combinação linear entre a similaridade textual e espacial com o documento. Porém, este processo ainda pode ser melhorado, com acréscimo de novos parâmetros, a fim de obter melhores resultados. É evidente que o limiar entre a aceitação ou inutilização da ferramenta de busca pelos usuários está regido pelo método utilizado para realizar o *ranking* de relevância dos resultados. Caso a ferramenta de busca faça uso de um método simples e que não ofereça um *ranking* eficiente de seus resultados, a mesma não será utilizada.

Diversas são as metodologias que podem ser utilizadas para ordenar documentos em resposta a uma consulta em um contexto geográfico. Alguns trabalhos encontrados na literatura aproveitam da ascensão do uso de dispositivos móveis pela população para propor novos métodos capazes de indicar quando um documento ou página é mais relevante para o usuário. Raper (2007) propõe o uso da localização do usuário no momento da busca para realizar o *ranking* de resultados. Segundo o autor, quando a busca está sendo realizada por meio de um dispositivo móvel e a localização do dispositivo está disponível, há fortes indícios que seu interesse principal está ligado à relevância geográfica e em segundo plano na relevância textual.

Outra abordagem proposta por Kumar e Boll (2013) sugere o uso da correlação entre *links* de páginas e suas localizações de provimento (i.e. localização da qual a página foi escrita), para busca por Pontos de Interesse (*Point of Interest* - POI). No momento em que um determinado POI é buscado, deve-se determinar quais são as páginas mais relevantes dentre todas as que mencionam o endereço de tal POI. Os autores pressupõem que há uma tendência para que a página oficial do POI seja referenciada pelas demais páginas e, portanto, esta deve ser a mais relevante dentre as demais.

Por fim, o método GeoRank proposto por Bao e Mokbel (2013) realiza o *ranking* em *feed* de *posts* em redes sociais baseado na temporalidade do *post* e na distância espacial entre os *posts* e o usuário. A localização do usuário é capturada por meio do seu perfil ou através do

georreferenciamento dos últimos *posts*. Como o método é específico para *feed* em redes sociais, houve a preocupação dos autores em permitir que o método consiga processar várias listas de *feed* (um usuário pode “seguir” vários outros e cada um deles possui sua própria lista) de forma eficiente e superficial ao usuário. Experimentos utilizando fontes de dados reais do Twitter indicam a eficiência e escalabilidade do método proposto.

### 2.3.5. Avaliação de Ranking de Relevância Geográfica

As metodologias de avaliação de métodos de *ranking* de relevância geográfica encontrados na literatura distinguem substancialmente dos métodos utilizados para avaliação de sistemas de RI. Para avaliar sistemas clássicos de RI, alguns autores se baseiam em duas métricas: precisão (do inglês, *precision*) e revocação (do inglês, *recall*). A precisão mede a capacidade do sistema de recuperar informações relevantes de acordo com uma consulta, considerando apenas os itens retornados. Por exemplo, quando se diz que um sistema de RI possui 80% de precisão, significa que, dos itens retornados, houve 20% que são irrelevantes para aquela consulta. Por outro lado, a revocação representa a capacidade do sistema retornar itens relevantes dentro do conjunto de itens passíveis de serem recuperados. Por exemplo, quando se afirma que um método possui 60% de revocação, significa que houve 40% de itens relevantes para a consulta que não foram recuperados.

Este tipo de avaliação para sistemas de RI é bastante eficiente em uma escala reduzida de dados, como, por exemplo, em sistemas internos onde o número de documentos é restrito e seus conteúdos conhecidos completamente. No entanto, quando o cenário pressupõe o uso de páginas da Web em grande escala (milhares de páginas indexadas), o uso desta metodologia de avaliação se torna inapropriado, pois exige do experimentador o conhecimento de todo conjunto de itens no sistema, algo inviável para o meio científico.

Deste modo, é possível identificar outras abordagens propostas para avaliação de *ranking* de relevância em sistemas de RI para Web, em especial, para *ranking* de relevância geográfica presentes em sistemas de GIR. Quando é proposto um método de *ranking* de relevância geográfica parametrizada (i.e. vários fatores compõem o método e os mesmos podem ser ajustados para produzir outros resultados), os autores buscam uma forma de compará-los nas diversas configurações paramétricas possíveis. Armenatzoglou et al. (2015) sugerem um método para ordenar usuários de uma rede social e, a partir deste, produzem quatro variações. Para avaliá-los e apontar qual obteve melhor resultado, os autores medem a correlação entre os métodos utilizando o *Kendall's  $\tau_b$  Correlation Coefficient* (Kendall, 1938) e discutem os resultados de forma a identificar a melhor aplicação para cada um deles.

Quando o trabalho busca um aperfeiçoamento de métodos de *ranking* de relevância, os resultados obtidos em ambos são equiparados (método antigo x método aperfeiçoado). Bao e Mokbel (2013) propõem um novo método de *ranking* para *feeds* de redes sociais baseado na associação geográfica entre os *posts* e o usuário. Assim, para verificar se houve ganhos, os autores comparam o método temporal tradicional do Twitter com o método aperfeiçoado com a inserção do contexto geográfico, para então avaliar o proposto.

Das diferentes metodologias para avaliar métodos de *ranking* de relevância geográfica, a mais encontrada na literatura é o *survey*. Isto poderia ser explicado pela subjetividade encontrada em avaliar um *ranking*: algo pode ser relevante para um determinado perfil de usuário, mas irrelevante para outro perfil. Seguindo esta metodologia, Uysal e Croft (2011) selecionaram 10 voluntários para avaliar o método proposto. Este método tem como objetivo ordenar *tweets* de forma a apresentar aos usuários os que possuem maior probabilidade de serem “retuitados” por eles. De forma semelhante, Lee et al. (2007) avaliam seu método de *ranking* de relevância geográfica com um *survey*. Os autores definem 20 exemplos de consultas (e.g. Hotel em NY) que têm seus resultados avaliados pelos participantes do estudo. Cada participante avalia o resultado de 3 a 8 consultas como “não sei”, “não relevante”, “relevante”, “muito relevante”, com média de avaliação de 4.1 participantes por consulta.

### 2.3.6. Consultas e Interface Gráfica

De acordo com Jansen et al. (2008), as consultas realizadas em motores de busca se enquadram em três categorias. São elas:

- Informacional: o usuário deseja obter informação sobre um tópico. É comum o uso de várias páginas para total assimilação da informação procurada;
- Navegacional: o usuário deseja acessar a página oficial de uma entidade ou companhia, porém não tem o conhecimento sobre o endereço URL daquela entidade;
- Transacional: o usuário deseja realizar alguma operação transacional, seja através de compras pela Internet ou operações bancárias;

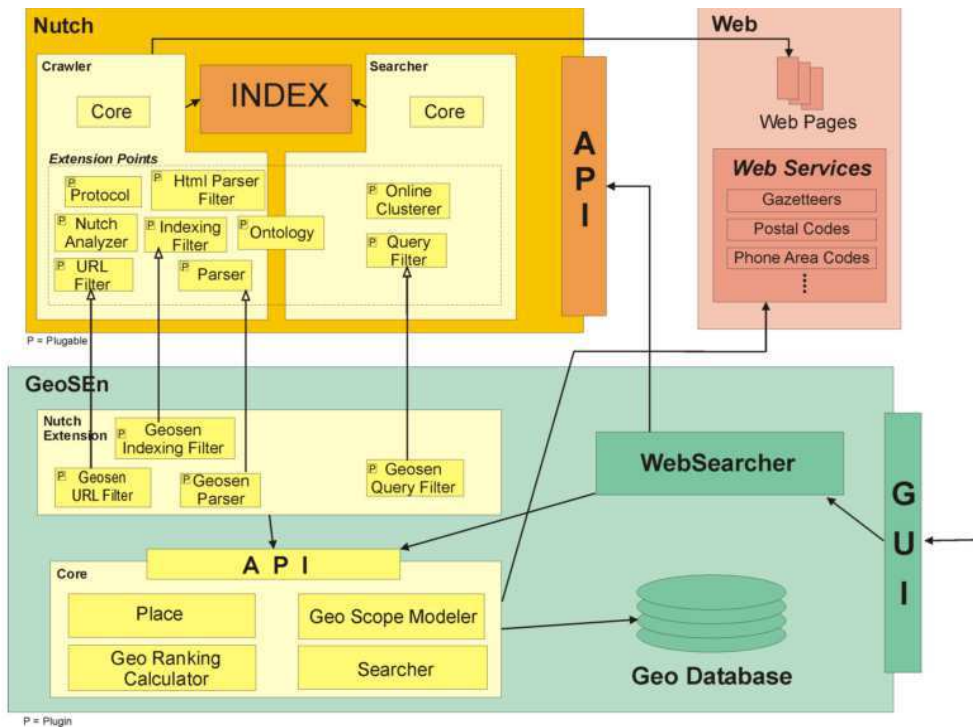
As consultas realizadas em motores de busca tradicional e geográfica se encaixam nas categorias citadas, porém a forma em que é realizada difere. Em motores de busca tradicionais, um campo é fornecido ao usuário para que possa repassar sua necessidade de informação. O usuário deve representar sua necessidade usando palavras-chave no momento da consulta. Melhores resultados estão relacionados diretamente com a qualidade desta representação.



Contudo, motores de busca geográfica fazem uso de outra forma para permitir consultas em contexto geográfico. Na maioria dos casos, três campos são apresentados para o usuário no momento da realização da consulta. O primeiro deles, semelhante ao que é feito em motores de busca tradicionais, recebe as palavras-chave que representam a consulta. No segundo campo, a localidade geográfica deve ser descrita. Por exemplo, nomes de cidades, estados ou regiões podem ser fornecidos. Note que, análogo ao que acontece no momento da análise de conteúdo de um documento, ambiguidades podem ocorrer neste campo. Então, o motor de busca, antes de realizar a consulta, pode tentar resolver essa ambiguidade ou solicitar que o próprio usuário ajude nesta tarefa. Por exemplo, em uma consulta onde o usuário especifica a localidade “Londres”, o sistema pode fazer perguntas ao usuário do tipo: “Você deseja informações sobre London (Reino Unido) ou Londres (Canadá)?”. O terceiro campo é utilizado para indicar qual operação espacial (e.g., continência, interseção, fronteira, raio) o usuário deseja realizar no momento da consulta. Por exemplo, uma consulta plausível em um motor de busca geográfica seria: “Festa de São João contido em um raio de 200 km de Campina Grande/PB”.

O uso de mapas na interface gráfica para realizar consultas também é recorrente. Nesta perspectiva, o usuário pode indicar, diretamente no mapa, a área ou região que deseja realizar a consulta. Kumar et al. (2013) argumentam a necessidade de criar uma interface mais funcional para busca em regiões de interesse. Assim, é proposta uma interface de busca espacial que poderia fornecer análise interativa e ranking de relevância em regiões desejadas. Já Samet et al. (2014) evoluíram o conceito de consultas em sistemas de GIR e apostam no uso massivo de mapas para viabilizar consultas com contexto geográfico. A ferramenta proposta pelos autores tem foco em notícias informativas espalhadas pela Web e permite que o usuário as encontre apenas explorando o mapa através das opções de *zoom* e *pan*. De acordo com a área visualizada no mapa, apenas notícias de mais relevância são exibidas. À medida que o usuário aproxima o mapa em sua região de interesse, outras notícias vão sendo exibidas seguindo o critério de relevância adotado.

Figura 6: Arquitetura do GeoSEn.



Fonte: Campelo (2008).

#### 2.4. GEOSEN

O GeoSEn é um motor de busca com enfoque geográfico que tem como princípio a utilização de *software* livre em sua composição. Por este motivo, o mesmo faz uso do Apache Nutch<sup>7</sup> como motor de busca estrutural, não só por ser baseado em *plugins* (o que o torna facilmente extensível), mas também por possuir uma documentação rica e por ser um projeto contínuo, assegurando sua rotineira evolução. Para armazenamento, o GeoSEn faz uso do SGBD PostgreSQL<sup>8</sup>, um sistema robusto em termos de projeto e que possui uma extensão espacial, o Postgis<sup>9</sup> que atende todas as necessidades exigidas no desenvolvimento das funções de consulta geográfica.

A última versão do GeoSEn (Agosto/2010) foi desenvolvida baseada no motor de busca Nutch (versão 0.9), que, por sua vez, é desenvolvido sobre a arquitetura do Lucene<sup>10</sup>, um *framework* que provê uma API para indexação textual de documentos e uma interface para o sistema de busca textual. O Lucene permite que o desenvolvedor implemente módulos de recuperação de documentos, de fonte e formatos desejados, apenas exigindo a conversão

<sup>7</sup> <http://nutch.apache.org/>

<sup>8</sup> <http://www.postgresql.org/>

<sup>9</sup> <http://postgis.net/>

<sup>10</sup> <https://lucene.apache.org/core/>

desses dados para o formato de texto simples, reconhecido pelo Lucene. Na Figura 6 é exibida a arquitetura do GeoSEn.

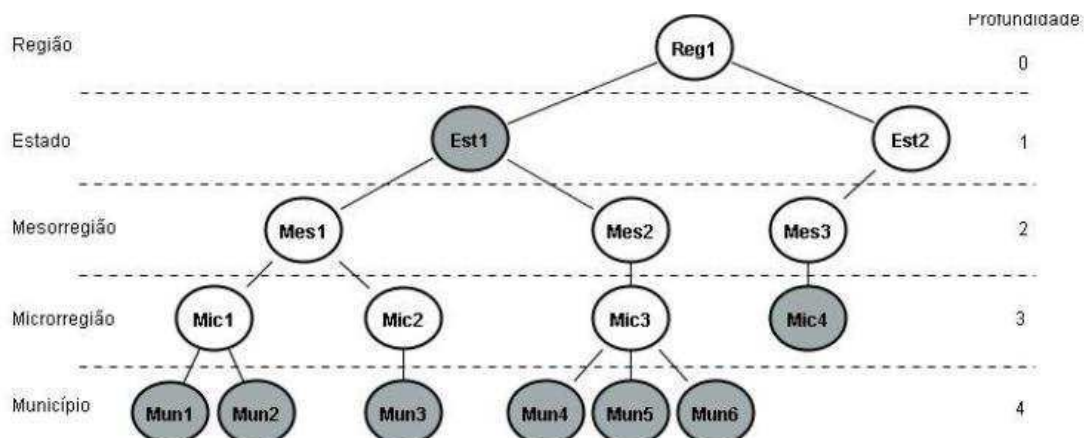
A arquitetura do Nutch pode ser subdividida em dois módulos: o *Crawler* e o *Searcher*. O *Crawler*, também comumente conhecido como “robô” ou “*spider*”, é o responsável por realizar a coleta, análise e a indexação de páginas Web, para futura recuperação. O processo de coleta (i.e. *crawling*) é dividido em várias etapas, que podem ser executadas por meio de um único comando, permitindo uma maior agilidade e praticidade para o administrador do sistema.

Por outro lado, o módulo *Searcher* é responsável por receber e interpretar consultas realizadas pelos usuários, realizar a consulta na base de índices e os retornar como resultados que atendem a consulta. Este módulo pode ser acessado de diversas formas, como, por exemplo, utilizando-se a biblioteca Nutch API, que viabiliza acesso direto ao Nutch por meio da linguagem de programação Java.

Como mencionado anteriormente, o Nutch possui uma arquitetura orientada a *plugins*, permitindo que alguns mecanismos sejam estendidos, de forma a adaptá-los às reais necessidades do motor de busca. Assim, o GeoSEn, por manipular dados espaciais em seu escopo, expande alguns componentes para viabilizar o funcionamento adequado da ferramenta. Mais especificamente, os pontos de extensão são o *Parser*, o *Indexing Filter*, o *Query Filter* e o *URL Filter*, conforme ilustrado na Figura 6, na área identificada como “Extension Points”.

De qualquer forma, estes *plugins* são restritos apenas à comunicação entre o Nutch e o GeoSEn, pois as regras de negócio relacionadas às funcionalidades permitidas estão implementadas no núcleo do GeoSEn, deixando-os assim menos acoplados. Esta baixa acoplagem permite a utilização do GeoSEn por outros sistemas e para outros fins, com um menor esforço de adaptação. O núcleo do GeoSEn compreende os módulos de Detecção de Lugares (*GeoSEn Geoparser*, utilizado na descoberta de referências geográficas em *posts* do Twitter – vide Seção 4.2), Modelagem do Escopo Geográfico (*Geo Scope Modeler*), Ranking de Relevância (*Geo Ranking Calculator*) e o *Searcher*, além dos módulos de acesso à base de dados.

Por fim, o GeoSEn implementa um módulo chamado *Web Searcher*, que faz a comunicação do sistema com o meio externo. Este módulo é responsável por receber as consultas dos usuários através da interface gráfica Web, consultar os índices textuais e espaciais e então retornar os documentos que correspondem à consulta em ambos os aspectos.

Figura 7: Instância de uma *geotree*.

Fonte: Campelo (2008).

Este módulo comunica-se com o núcleo do GeoSEn e com o Nutch através de suas respectivas APIs.

Os principais módulos do GeoSEn estão descritos, de forma sucinta, nas subseções a seguir.

#### 2.4.1. Detecção de Referências Geográficas

O módulo de detecção de referências geográficas tem como objetivo encontrar e extrair, de páginas Web, potenciais dados a se tornarem informações geográficas, como por exemplo, nome de cidades, estados, código postais, dentre outros. Identificados estes termos, é dado início o processo de conversão destes em localidade manipuladas pelo sistema GeoSEn. Por exemplo, um código postal pode ser convertido em uma referência geográfica para a cidade a qual o mesmo pertence.

Para que o processo de identificação e conversão de termos em localidades manipuladas pelo sistema ocorra de forma correta, quando um termo é identificado no conteúdo, título ou URL da página Web, o mesmo é associado a um grau de confiabilidade. Este valor representa a probabilidade do termo ser ou não uma referência válida de localidade e é mensurado respeitando diversas características, como por exemplo, referências cruzadas, sintaxe de escrita do termo, ocorrência de termos influentes, dentre outros. Depois de mensurado o valor de confiabilidade, caso o termo não atinja um valor limite estabelecido, o mesmo será descartado. Ademais, esta etapa do processo tem por finalidade desambiguar as localidades identificadas, que ocorre quando duas ou mais localidades (e.g. cidade, estado, região) possuem o mesmo nome.

### 2.4.2. Modelagem do Escopo Geográfico

O módulo de Modelagem de Escopo Geográfico tem como finalidade mensurar o grau de relevância de cada localidade (identificada na etapa de detecção) em relação ao documento. Um documento pode referenciar várias localidades (escopo múltiplo) explicitamente ou não (i.e. quando outras localidades são derivadas das demais), de modo que, cada uma destas localidades terá seu próprio grau de relevância.

Seguindo a hierarquia geográfica (cidade → microrregião → mesorregião → estado → região), o módulo faz uso da técnica de Expansão de Georreferenciamento, que consiste em calcular a relevância para níveis geográficos mais altos, a partir de referências encontradas em níveis inferiores. O objetivo principal do uso desta técnica está em transferir para tempo de *parsing* e indexação algumas operações espaciais que, de outra forma, seria realizado em tempo de consulta. Esta hierarquia é modelada em uma estrutura de dados do tipo *geotree*, onde cada localidade associada ao documento é representada por um nó da árvore, e este associado a um peso (i.e. relevância daquela localidade em relação ao documento e que será utilizado no *ranking* de relevância). Na Figura 7 é apresentada uma instância da *geotree* onde pode-se perceber os cinco níveis hierárquicos considerados no sistema, as localidades *Mun1*, *Mun2*, *Mun3*, *Mun4*, *Mun5*, *Mun6* e *Mic4* diretas (i.e. localidades explícitas no conteúdo do documento), a localidade *Est1* híbrida (i.e. localidade explícita no conteúdo do documento e derivada da expansão do georreferenciamento) e as demais localidades indiretas (i.e. localidade provida por meio da expansão do georreferenciamento).

### 2.4.3. Indexação Espaço-textual

A abordagem adotada no GeoSEn para o processo de indexação utiliza como base o resultado da etapa de expansão do georreferenciamento. Assim, cada localidade referenciada direta ou indiretamente pelo documento possui uma entrada independente do índice. Cada uma dessas entradas está associada a sua respectiva relevância geográfica, ou seja, cada nó da *geotree* corresponde a um índice espacial e possui um valor de relevância que será considerado no cálculo de relevância geográfica.

## 2.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram apresentados diversos elementos que circundam a área de recuperação da informação tradicional e geográfica. Em cada um destes, foram apresentados os desafios e as principais soluções encontradas na literatura. No próximo capítulo são

descritas as pesquisas e trabalhos no campo da GIR, como foco principal nas soluções para *ranking* de relevância geográfica e interface de consultas disponíveis para o usuário.

### CAPÍTULO 3

#### TRABALHOS RELACIONADOS

Desde seu surgimento, por volta da década de 50, a área de Recuperação da Informação tem ampliado seu ferramental ao ponto de atingir um patamar capaz de fornecer soluções que atendam as necessidades dos usuários. Por sua vez, a área de Recuperação da Informação Geográfica ainda é limitada e, como consequência, pesquisas neste campo são frequentes.

A maior parte dos documentos disponíveis na Web possui alguma referência geográfica em seu conteúdo (e.g., nome de cidade e/ou estado, código postal, endereço). Assim, por meio de um tratamento especializado a este tipo de dado, torna-se possível fornecer funcionalidades que, por sua vez, são inviáveis ou mesmo impossíveis de serem realizadas utilizando ferramentas de busca tradicionais. Por exemplo, considere um usuário que deseja encontrar documentos relacionados às visitas feitas em todo Brasil pelo seu candidato a deputado federal. No entanto, este usuário deseja desconsiderar as visitas feitas ao estado de nascimento deste e estados adjacentes (i.e. estados que fazem fronteira). Formular uma pesquisa deste tipo utilizando os moldes convencionais (i.e. motores de busca tradicionais) é inviável, pois, além da necessidade de definir na consulta todos os estados considerados, apenas páginas que contêm explicitamente o nome do estado serão retornadas como resposta. A área de GIR vem para solucionar esta e diversas outras limitações encontradas em motores de busca tradicionais.

Em contrapartida, alguns desafios ainda precisam ser endereçados na área de GIR. Dentre estes, citam-se: a extração de documentos e referências geográficas; análise e solução de ambiguidades, isto é, quando existe mais de uma localidade com o mesmo nome (e.g. Bom Jesus-PB e Bom Jesus-RN) ou localidades com nome de pessoas ou objetos; modelagem de escopo geográfico de um documento, isto é, como o documento é representado geograficamente (i.e., escopo múltiplo ou escopo simples), dentre outros. Presente entre os desafios está o processo de *ranking* dos resultados para uma determinada busca. De acordo com Andrade e Silva (2006), a maioria das pesquisas na área de GIR estão concentradas em solucionar problemas pautados em extração e indexação de informações e poucas são as pesquisas relacionadas com *ranking* de relevância.

Apesar da existência de alguns sistemas de GIR que trazem consigo soluções válidas para os desafios relatados acima (Buyukkokten et al., 1999; Jones et al., 2002; Cai, 2011),

poucas soluções têm sido observadas na literatura com intuito de fornecer um ambiente completo para leitura de notícias informativas em um contexto geográfico. Em contrapartida, no setor comercial é possível encontrar diversas iniciativas na construção de ferramentas capazes de permitir a busca e a leitura de notícias. Como exemplos, têm-se as ferramentas My Yahoo!<sup>11</sup>, Bing News<sup>12</sup> e Google News<sup>13</sup>, que trazem a proposta de unificar, em um só ambiente, notícias de várias fontes. Outra abordagem encontrada no mercado consiste em implantar, diretamente em redes sociais, mecanismos que permitam a descoberta, por parte do usuário, de notícias do seu interesse. Em linhas gerais, estes mecanismos fazem uso dos dados gerados pelos usuários para aprimorar os resultados apresentados a eles. Destes, pode-se destacar o Twitter Moments<sup>14</sup> e o Instante Articles<sup>15</sup>, do Facebook, que despertam o interesse da população pela sua praticidade, mas que pecam na ausência de funcionalidades no âmbito geográfico.

Assim, uma revisão da literatura foi realizada a fim de identificar valorosas contribuições em modelos de *ranking* de relevância para sistemas de GIR e em ambientes especializados na leitura de notícias informativas. Como o modelo de *ranking* de relevância proposto neste trabalho faz uso de uma técnica capaz de medir o grau de afinidade entre usuário e localidades (i.e. afinidade local) em rede social, trabalhos com este foco também foram avaliados.

As seções seguintes estão organizadas da seguinte forma: na Seção 3.1 é apresentada técnicas de inferência de localização; na Seção 3.2, são descritos trabalhos direcionados para construção de *ranking* de relevância geográfica; na Seção 3.3 são detalhados os trabalhos propostos para criação de um ambiente interativo para consulta e leitura de notícias em um contexto geográfico; e, por fim, na Seção 3.4 as considerações finais do capítulo são apresentadas.

### 3.1. INFERÊNCIA DE LOCALIZAÇÃO

As redes sociais são, nos dias atuais, a maneira mais popular encontrada por usuários para se comunicar, expressar seus pensamentos e compartilhar conteúdo. Estes usuários mantêm um perfil contendo informações pessoais, como interesses, nível de escolaridade, localidade de moradia, dentre outros. Muitas vezes esses atributos são utilizados com o intuito

---

<sup>11</sup> <https://my.yahoo.com/>

<sup>12</sup> <https://www.bing.com/news>

<sup>13</sup> <https://news.google.com/>

<sup>14</sup> <https://about.twitter.com/moments>

<sup>15</sup> <https://instantarticles.fb.com/>



de agrupar usuários para uma melhor exibição de conteúdo, recomendar novos conteúdos e possíveis amigos (Mislove et al., 2010). Diversas outras aplicações podem ser obtidas apenas com o uso destes atributos, como, por exemplo, modelar eleições políticas, desastres ambientais e *blackouts* (Paul e Dredze, 2011).

Dos dados comuns encontrados em perfil de usuários em redes sociais, o que fornece o maior poder para fomentar aplicações como as supracitadas é o atributo de localidade (i.e., localidade de moradia). Infelizmente, pesquisas indicam que apenas 5,34% dos usuários do microblog Twitter informam em seu perfil este dado (Jurgens, 2013), o que dificulta o funcionamento desejado de aplicações que o fazem uso. Para contornar este obstáculo e permitir que a localização dos usuários envolvidos seja empregada em aplicações de diferentes naturezas (e.g. detecção de eventos, desastres ambientais, *blackouts*), algumas ações são encontradas no meio científico com o objetivo de inferir o valor deste campo (i.e., localidade) quando o mesmo não é fornecido pelo usuário.

Cheng, Caverlee e Lee (2010) propõem um *framework* probabilístico para estimar a localização de moradia, em nível geográfico de cidade, de usuários do Twitter. O método baseia-se apenas no conteúdo dos *tweets* inseridos pelo usuário, mesmo que não haja nenhuma “pista geográfica” nos mesmos. Este método baseia-se em três indicadores para estimar a localidade de um dado usuário do microblog. O primeiro deles considera apenas o conteúdo encontrado nos *tweets*, sem a necessidade de checagem de IP ou informações externas à rede social; o segundo indicador é um classificador de palavras, que associa à cada palavra encontrada nos *tweets*, o grau de probabilidade da mesma representar uma localização; o terceiro e último indicador considera, em um grafo de relacionamento, os nós vizinhos para aprimorar a localidade anotada para o usuário. Assim, os três indicadores trabalham em conjunto para inferir K possíveis localizações para o usuário, ordenando-as da mais confiável para a menos confiável. Os experimentos comprovam a utilidade do método, que consegue estimar 51% da localização de moradia dos usuários avaliados, com taxa de erro de aproximadamente 100 milhas. Infelizmente, o método esbarra em obstáculos comuns na área de GIR: ambiguidade, quando o nome pode representar mais de uma localidade ou substantivos; a ausência de uma estrutura de informação; e uso de gírias e variação linguísticas pelos usuários.

Outra abordagem encontrada na literatura consiste em inferir atributos presentes em todo o perfil do usuário, não restringindo apenas à localidade de moradia do usuário. Mislove et al. (2010) propõem uma nova forma de associar atributos ao perfil do usuário, conciliando

outras abordagens encontradas na literatura para este fim. Os autores falam da necessidade dos atributos no perfil do usuário em redes sociais, pois os mesmos podem ser de fundamental importância no agrupamento de usuários e recomendação de conteúdos relevantes. O método de inferência de atributos foi avaliado com um conjunto de 4.000 perfis de estudantes no Facebook, onde foi possível aferir, com 80% de acurácia, os atributos de escolaridade, ano de matrícula e dormitório dos estudantes que não os forneceram em seu perfil. Outro experimento foi realizado com 63.000 usuários da cidade de New Orleans com o mesmo objetivo de identificar atributos ausentes em determinados perfis. Neste, os autores observaram que usuários de determinadas comunidades tendem a compartilhar os mesmos atributos (e.g., interesses, localidades frequentadas e nível escolar). Por se tratar de um mecanismo mais abrangente, ou seja, que envolve a inferência de vários atributos, o método deixa a desejar quando a estimativa envolve a localidade de moradia de um usuário. A inferência deste tipo de atributo deve considerar algumas características (e.g., granularidade da localização e ambiguidades) que foram desconsideradas no momento da avaliação do método (a amostra não considerou a distribuição geográfica dos usuários).

Com o objetivo de aperfeiçoar a análise em tempo real de desastres naturais, os autores Ikawa, Enoki e Tatsubori (2012) sugerem a inferência de localidade em outra perspectiva. Ao invés de tentar estimar a localidade de moradia de usuários, os autores produzem um método capaz de geolocalizar *tweets*. O funcionamento do método pode ser descrito da seguinte forma: são extraídas, de um conjunto de *tweets* geolocalizados (i.e. *tweet* que possui localização em seu metadado), todas as palavras com o intuito de gerar um modelo preditor. Ou seja, o método busca modelar uma localidade a partir das palavras utilizadas em *tweets* postados naquele lugar. O modelo é utilizado para geolocalizar *tweets* que não possuem esse atributo em seu metadado, baseando-se apenas no conteúdo dos mesmos. Os experimentos executados mostram uma boa eficácia do método quando o modelo preditor é individualizado, significando, em linhas gerais, que cada usuário da rede deve ter um preditor único para geolocalizar seus *tweets*. Apesar de promissor, o método encontra limitações devido à ausência de *tweets* geolocalizados para treinamento do modelo. Segundo pesquisas, menos de 1% dos *tweets* são geolocalizados (Jurgens, 2013).

Mahmud, Nichols, e Drews (2012) apresentam em seu trabalho um novo algoritmo para identificar a localização atual de usuários do Twitter, em diferentes granularidades geográficas (i.e. cidade, estado, *time zone*), baseando-se em seu comportamento na rede social, como, por exemplo, os relacionamentos de amizade e *tweets* postados. Para este fim,

são criados classificadores probabilísticos modelados a partir de palavras, *hashtags* e nome de localidades (identificados com auxílio de *gazetteers*). Ademais, uma heurística é proposta para aperfeiçoar o algoritmo apresentado. Esta parte da premissa que um usuário tende a visitar mais vezes os lugares onde vive do que quaisquer outros lugares, assim, possibilitando a descoberta da localidade de moradia apenas verificando o histórico de *tweets* geolocalizados. Os experimentos realizados comprovam melhorias em relação aos métodos encontrados na literatura. Contudo, os autores alertam para a segurança deste dado (i.e., localidade de moradia) que, mesmo que seja ausente ou privado, pode ser facilmente encontrado utilizando artifícios como o proposto nesta dissertação.

No estudo realizado por Jurgens (2013), discute-se a baixa eficácia de técnicas de inferência de localização que consideram apenas a interação do usuário com as redes sociais e seus relacionamentos de amizade. Deste modo, o autor apresenta uma nova abordagem para inferência deste atributo fundamentada na propagação espacial, técnica esta que é fruto do aperfeiçoamento da técnica de *Label Propagation* (Zhu e Ghahramani, 2002). Uma característica importante do método é o de permitir que a propagação espacial seja realizada com poucas localizações, estas identificadas nos *tweets* geolocalizados. Para avaliar a consistência do método de inferência da localização, o autor realizou cinco experimentos com dados oriundos de diversas redes sociais, utilizando como base de treinamento dados precisos e duvidosos. Ademais, uma heurística foi utilizada para melhorar os resultados alcançados na propagação espacial.

### **3.1.1. Sumarização dos Métodos**

O resultado da avaliação dos métodos com o intuito de geolocalizar usuários de redes sociais foi sumarizado na Tabela 1 na forma de métodos (colunas) e características (linhas). A célula preenchida com o símbolo • indica que a característica está presente no método. Por outro lado, quando o símbolo está ausente significa que o método não considera aquela característica em sua abordagem.

### **3.1.2. Considerações Sobre o Estado da Arte em Inferência de Localização**

Dentre os trabalhos publicados com o foco de geolocalizar usuários que não fornecem ou ocultam informações sobre localização em seus perfis de rede social, foi possível perceber uma centralização da abordagem em identificar esta informação em *tweets* ou relacionamentos de amizade dos usuários (Tabela 1, linhas 1 e 2). Os experimentos realizados

Tabela 1: Tabela de sumarização dos métodos avaliados.

ID	Características/Método	Cheng et al., 2010	Mislove et al., 2010	Ikawa et al., 2012	Mahmud et al., 2012	Jurgens, 2013
1	Considera referências geográficas encontradas em <i>tweets</i> .	•	•	•	•	•
2	Considera os relacionamentos de amizade.	•	•		•	•
3	Inferre múltiplas localidades, com grau de confiabilidade, para um único usuário.	•				
4	Propaga os dados espaciais proveniente de outros usuários.					•
5	Utiliza classificador probabilístico baseado em palavras encontradas nos <i>tweets</i> .			•	•	
6	Analisa referências encontradas em <i>links</i> compartilhados.					
7	Identifica localidades através de <i>gazetteers</i> .	•			•	•
8	Faz uso de <i>GeoParser</i> na identificação de referências.					

Fonte: elaborada pelo autor.

por todos os autores comprovam a eficiência das abordagens. Porém, identificar referências geográficas em texto (i.e., resolução de topônimos), apenas com o apoio de *gazetteers*, ainda é um desafio da área e se agrava quando estes textos são reduzidos a poucos caracteres, ocasionando limitações para o uso de técnicas de identificação de localidades, como, por exemplo, referências cruzadas.

Uma solução plausível para minimizar a taxa de erros na identificação de referências geográficas em pequenos textos, seria utilizar um *geoparser* (i.e. mecanismo presente em sistemas de GIR produzido para identificação de termos geográficos presentes em conteúdos de páginas) (Tabela 1, linha 8). O aprimoramento de um *geoparser* para este fim traria ganhos consideráveis, pois o mesmo implementa técnicas úteis na identificação de termos, como, por exemplo, desambiguação, relevância do termo para o texto e taxa de

confiança do termo representar uma localidade. O processo de evolução e manutenção do método também sofreria um impacto positivo, pois reduziria o acoplamento e ajudaria na modularização do mesmo.

Outra característica importante observada foi a ausência de verificação do conteúdo dos *links* compartilhados pelo usuário (Tabela 1, linha 6). Semelhante ao que é feito nos *posts* (e.g. *tweets*), onde o mesmo é submetido ao processo de detecção de referências geográficas, o mesmo deveria ser realizado em conteúdos de *links* compartilhados pelo usuário. A identificação de termos geográficos em páginas de conteúdo, em geral, é mais eficiente do que em pequenos textos, pois aquelas normalmente apresentam uma linguagem mais formal, uma maior estruturação do conteúdo e diversas outras características que auxiliam na identificação de termos geográficos.

### 3.2. RANKING DE RELEVÂNCIA GEOGRÁFICA

Relevância geográfica é definida como um relacionamento entre a necessidade de informação geográfica, expressa pelo usuário através de consultas, e as informações que possuem este atributo geográfico, que podem ser documentos, imagens, mapas, dentre outros (Raper, 2007). Conceitualmente, esta relação é multidimensional, situacional e dinâmica, podendo ser diferente em diversas ocasiões (Cai, 2011). Por exemplo, um sistema de GIR pode considerar a localização atual do usuário para ordenar estabelecimentos procurados considerando a proximidade entre eles e o usuário. Esta ordenação pode variar no momento em que o usuário se desloca para outra região.

Por o *ranking* de relevância geográfica ser um fator determinante para aceitação ou recusa de um sistema de GIR, o tema tem despertado o interesse da comunidade e diversas pesquisas com este foco são encontradas na literatura. As mesmas divergem em suas abordagens e utilizam de diversos elementos, conceitos e dados para melhorar a ordenação dos resultados, mediante consulta submetida a um motor de busca geográfica.

Para ordenar resultados em motores de busca geográfica, o método mais utilizado é o de similaridade geográfica, que consiste em verificar a semelhança entre os *footprints* espaciais da busca e do documento, levando em consideração o grau de similaridade para realizar o *ranking* dos resultados. Vários são os métodos para mensurar a similaridade entre os *footprints* espaciais (Martins et al., 2005; Cai, 2002; Andrade et al., 2006). Dentre estes, destacam-se os que se baseiam na Distância de Hausdorff (Atallah, 1983) para medir a semelhança entre geometrias. Os métodos que usam apenas similaridade espacial entre a

busca e o documento são simples e podem ser aprimorados a fim de prover melhores resultados ao processo de *ranking*.

Semelhante aos métodos que comparam a similaridade entre o *footprint* da consulta com o *footprint* modelado para o documento, Larson e Frontiera (2004) apresentam um modelo de regressão logística para estimar o grau de interseção espacial entre a busca e o documento. O método apresenta resultados promissores, porém, a ausência de características importantes, como, por exemplo, o uso da localização do usuário no momento do *ranking*, dificulta sua adoção.

Vislumbrando a possibilidade de conciliar outros fatores na geração de um modelo de *ranking* de relevância geográfica, Lee, Liu e Miller (2007) propõem um novo método chamado GeoLink, que calcula o valor de relevância com base nas referências entre as páginas e a localização do usuário, além de permitir que ele ajuste alguns fatores para obtenção de melhores resultados. Os autores realizaram uma investigação empírica a fim de comparar métodos de análise de referências sensíveis geograficamente. O método proposto foi comparado com os métodos PageRank (1999) e GeoHITS (uma extensão do algoritmo de *ranking* HITS). O GeoLink apresentou-se como uma solução viável em comparação com os demais métodos analisados pelos autores, no entanto, não se sabe ao certo se o método mantém sua superioridade perante outros métodos de *ranking* de relevância geográfica.

Os métodos de *ranking* de relevância geográfica não se limitam ao uso em motores de busca geográfica. O método GeoRank proposto por Bao e Mokbel (2013) realiza o *ranking* em *feed* de *posts* em redes sociais baseado na temporalidade do *post* e na distância espacial entre os *posts* e o usuário. A localização do usuário é capturada por meio do seu perfil ou através do georreferenciamento dos últimos *posts*. Como o método é específico para *feed* em redes sociais, houve a preocupação dos autores em permitir que o método consiga processar várias listas de *feed* (um usuário pode “seguir” vários outros e cada um deles possui sua própria lista) de forma eficiente e superficial ao usuário. Experimentos utilizando dados do Twitter indicam a eficiência e escalabilidade do método proposto, porém, sua falha está em assumir que todo *post* em uma rede social é georreferenciado, o que nem sempre é verdade. Extrair uma localização do conteúdo do *post* poderia ser uma alternativa promissora na evolução do método GeoRank.

Por fim, percebendo o crescente entusiasmo de usuários em sistemas de busca de pontos de interesse (do inglês, *points of interest* – POI), Kumar e Boll (2013) propõem um método de *ranking* para buscas por POIs baseado na correlação entre *links* de páginas e suas

localizações de provimento (i.e., localização da qual a página foi escrita). No momento em que um determinado POI é buscado, deve-se determinar quais são as páginas mais relevantes dentre todas as que mencionam o endereço de tal POI. Os autores pressupõem que há uma tendência para que a página oficial do POI seja referenciada pelas demais páginas e, portanto, esta deve ser a mais relevante dentre as demais.

### 3.2.1. Sumarização dos Métodos

O resultado da avaliação dos métodos de *ranking* de relevância geográfica foi sumarizado na Tabela 2 na forma de métodos (colunas) e características (linhas). A célula preenchida com o símbolo • indica que a característica está presente no método. Por outro lado, a ausência do símbolo significa que o método não considera aquela característica em sua abordagem.

Tabela 2: Tabela de sumarização dos métodos de *ranking* de relevância geográfica.

ID	Características/Método	Similaridade geográfica	Kunar et al., 2013	Bao et al., 2013	Larson et al., 2004	Lee et al., 2007
1	Considera a localização do usuário.			•		•
2	Analisa os <i>links</i> entre as páginas.		•			•
3	Considera a localização de provimento das páginas.		•			
4	Considera dados oriundos de outras fontes.					
5	Permite balancear os fatores para um melhor resultado.					•
6	Faz uso de um modelo probabilístico.				•	

Fonte: elaborada pelo autor.

### 3.2.2. Considerações Sobre o Estado da Arte em *Ranking* de Relevância Geográfica

Mediante análise realizada em trabalhos científicos com o foco em *ranking* de relevância geográfica, foi possível observar a discrepância entre as abordagens adotadas para ordenar documentos em resposta a uma consulta. Isto se deve ao fato de não haver, em tempo, uma única solução genérica capaz de mostrar eficiência em todos os cenários. Por exemplo, um *ranking* de relevância geográfica proposto para uma rede social poderia considerar

inúmeros fatores (e.g. localização do usuário, interesse, escolaridade, círculos de amizade) no momento da ordenação, estes indisponíveis em aplicações de busca geográfica.

Outro elemento que pode estar relacionado com esta discrepância é o conteúdo que será disponibilizado (i.e., indexado) para consulta. Seguindo o mesmo exemplo anterior, um *ranking* de relevância geográfica que atua diretamente em redes sociais manipula pequenos textos, na maior parte das vezes, sem estrutura e/ou nexos. Em contrapartida, motores de busca geográfica tendem a manipular documentos mais completos, coesos e que seguem, na maior parte das vezes, estrutura de escrita e linguagem adequada, facilitando a identificação de termos geográficos em seu conteúdo (i.e., resolução de topônimo).

Deste modo, uma solução plausível para maximizar a eficiência de um *ranking* de relevância geográfica seria unir ambas as metodologias: coletar e manipular dados oriundos de redes sociais (Tabela 2, linha 4), como, por exemplo, a localização dos usuários que interagiram com o documento; e realizar a indexação das páginas coletadas nas redes sociais, disponibilizando-as em um motor de busca geográfica. Apesar da complexidade em relacionar estas duas vertentes (i.e., dados oriundos de redes sociais e páginas disponíveis na Internet), quando o escopo é reduzido apenas às notícias informativas este esforço é minimizado. Atualmente, é possível perceber um engajamento das mídias informativas em disponibilizar suas notícias diretamente na rede social (e.g., Twitter, Facebook) por conta da proximidade com o leitor. Por este motivo, torna-se viável a construção de um ambiente dedicado à busca e leitura de notícias, que faz uso de dados provenientes das redes sociais, sem a necessidade de integração com as mesmas, para possibilitar o proveito de alguns fatores (e.g., localidade do usuário, data e hora da publicação) no momento da ordenação dos resultados.

### 3.3. FERRAMENTAS DE LEITURA DE NOTÍCIAS

A importância crescente e a penetração das chamadas “novas mídias” na sociedade ficam claras quando os dados sobre os meios de comunicação preferidos pela sociedade são analisados. Segundo a Pesquisa Brasileira de Mídia<sup>16</sup>, realizada no ano de 2015, dos internautas que acessam a Internet todos os dias, 67% estão em busca de se manter bem informado, seja por meio de temas diversos ou informações de um modo geral. Percebendo estes dados, pode-se chegar à conclusão de que há, por parte destes usuários, a necessidade de um ambiente capaz de facilitar e agilizar sua procura por informações de seu interesse.

---

<sup>16</sup> <http://www.secom.gov.br/atuacao/pesquisa/lista-de-pesquisas-quantitativas-e-qualitativas-de-contratos-atuais/pesquisa-brasileira-de-midia-pbm-2015.pdf>



Com foco nesta demanda, o mercado de tecnologia se posiciona no sentido de fornecer ferramentas que encontra, agrupa, estrutura e exhibe notícias informativas de acordo com o desejo do usuário. Por exemplo, podem-se destacar as ferramentas produzidas pela Microsoft<sup>17</sup> (Bing News), pelo Yahoo<sup>18</sup> (Yahoo News) e pela Google<sup>19</sup> (Google News). Todas estas empresas possuem seu próprio motor de busca e procuraram o aperfeiçoamento para manipular notícias informativas proveniente de diversas fontes.

O que se observa entre as iniciativas comerciais propostas é a baixa documentação fornecida acerca de detalhes sobre os algoritmos utilizados, sobre mecanismos de cálculo de relevância e ordenação, bem como materiais e fontes de notícias utilizadas. Esta documentação escassa e limitada dificulta seu total entendimento por parte dos pesquisadores, mas pode ser explicada devido à forte concorrência entre as empresas em busca de espaço comercial. Contudo, do que se pôde perceber das ferramentas, as mesmas fornecem ao usuário funcionalidades bem semelhantes. Em todas as ferramentas comerciais analisadas, há uma categorização de notícias (e.g. cultura, esporte, entretenimento), um campo para consulta textual e botões para restringir o escopo espacial de notícias (i.e., notícias sobre o mundo ou notícias sobre o país de origem do usuário). Uma característica importante encontrada no Google News é a possibilidade do usuário personalizar a quantidade de notícias recebidas de certa fonte, ou seja, caso o usuário não confie em informações oriundas de determinada fonte de notícias, ele pode reduzir ou restringir a exibição de notícias daquela. Ao contrário do que é encontrado nas demais ferramentas citadas, o Yahoo News fornece notícias autorais, o que pode interferir na parcialidade da ordenação dos resultados apresentados.

Nas ferramentas comerciais analisadas, destacam-se negativamente dois fatores importantes em aplicações desta natureza: a ausência de uma consulta temporal por parte do usuário e a falta de funções específicas para uma busca geográfica, seja ela em uma área de interesse ou locais de preferência. Apesar de atrasada em relação às propostas comerciais (e.g., Google News, Yahoo News), algumas soluções acadêmicas consideram estes e outros fatores para fornecer facilidade ao leitor de notícias e uma melhor experiência no uso deste tipo de aplicação.

No meio científico, embora interesses estejam direcionados para a produção de soluções para os inúmeros desafios da área de GIR, poucos são os trabalhos encontrados com a proposta de viabilizar um ambiente para consulta e leitura de notícias informativas com

---

<sup>17</sup> <https://www.microsoft.com/pt-br/>

<sup>18</sup> <https://br.yahoo.com/?p=us>

<sup>19</sup> <https://www.google.com.br/>

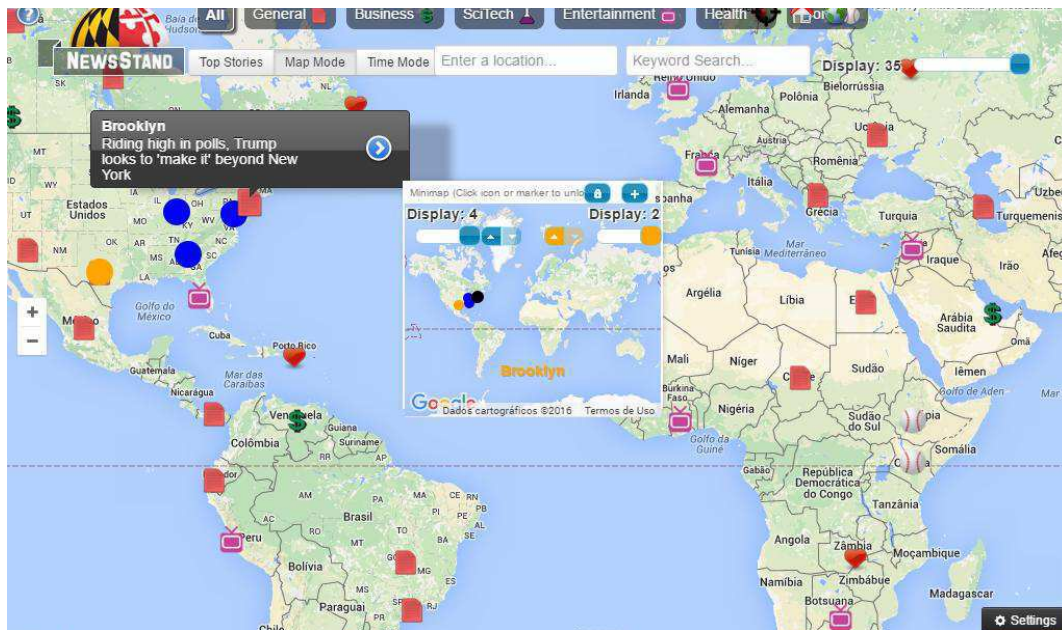
mais ênfase no aspecto geográfico da informação. Dos trabalhos avaliados, alguns merecem destaque. Phelan, McCarthy e Smyth (2009) perceberam a necessidade de um sistema de recomendação de notícias baseadas no perfil do usuário em redes sociais, deste modo, sugerem a construção de um sistema que tem como propósito recomendar notícias informativas, em tempo real, oriundas do Twitter, baseado em atributos presentes no perfil do usuário. Esta abordagem tem funcionamento semelhante ao encontrado em sistemas RSS (vide Seção 4.1), integrando notícias oriundas do Twitter e de fontes RSS. Três estratégias são fornecidas para o usuário: 1) o *ranking* público (notícias ordenadas de acordo com *timelines* públicas encontradas no Twitter); 2) *ranking* de amigos (notícias encontradas apenas nas *timelines* de amigos); 3) *ranking* de conteúdo (não utiliza o Twitter, apenas notícias fornecidas pelo RSS). Um ponto negativo da solução é a ausência de qualquer função geográfica para consulta ou restrição de conteúdo reportado e a interface gráfica pouco amigável.

Em outra abordagem, Sahai e Chan analisam a incorporação de conhecimento geográfico, em notícias informativas, por meio de uma ontologia, tendo em vista que poucas mídias de informação exploram este conceito em suas ferramentas de busca. O objetivo principal desta ferramenta é o de retornar ao usuário, notícias informativas mais relevantes, dada sua necessidade. Algumas técnicas de *ranking* de relevância e de expansão geográfica foram adaptadas para o cenário (i.e., notícias informativas) abordado no trabalho. Os autores relatam a grande dificuldade encontrada em selecionar as fontes de notícias, presentes no sistema, não só pela distribuição e organização estrutural da notícia, mas pela complexidade em determinar o grau de credibilidade associada a cada uma delas. Por tal dificuldade, os autores adotam apenas três fontes de notícias e não sugere nenhum mecanismo para crescimento desta base de forma confiável. O acesso à ferramenta encontra-se indisponível no momento.

Samet et al. (2014) partem do argumento que nem sempre os usuários sabem exatamente o que procuram e, na maior parte das vezes, outras respostas podem ser interessantes para uma determinada necessidade. Por exemplo, um usuário pode buscar por um show em uma determinada cidade, no entanto, cidades vizinhas poderiam ser consideradas no momento da consulta. Deste modo, viabilizar um mecanismo de busca por conteúdos, em um contexto geográfico, pode ser um promissor caminho para novas experiências. Os autores propõem uma aplicação para leitura de notícias informativas, chamada NewsStand, e consideram alguns desafios da área de GIR no momento do desenvolvimento. Para evitar

erros na resolução de topônimos, quando uma notícia está sendo lida pelo usuário, um minimapa é aberto e nele é adicionada uma marcação azul indicando outros lugares com o mesmo nome do que foi encontrado na notícia. Os tópicos (i.e., notícias) são apresentados de acordo com o nível do *zoom* e a sua relevância, porém, pouco se é detalhado como esse valor de relevância é mensurado. Uma característica importante na ferramenta é o agrupamento de notícias correlatas, permitindo a fluidez da leitura de tópicos sobre determinado assunto. A ferramenta se mostra bastante útil para os usuários interessados em notícias, mas o número reduzido de funções espaciais fornecidas (apenas uma consulta por localidade é permitida), a interface gráfica complexa e a ausência de responsividade (i.e. adaptação da interface gráfica em vários tamanhos de tela) são alguns dos pontos negativos encontrados na aplicação (Figura 8).

Figura 8: Interface Gráfica do NewsStand.

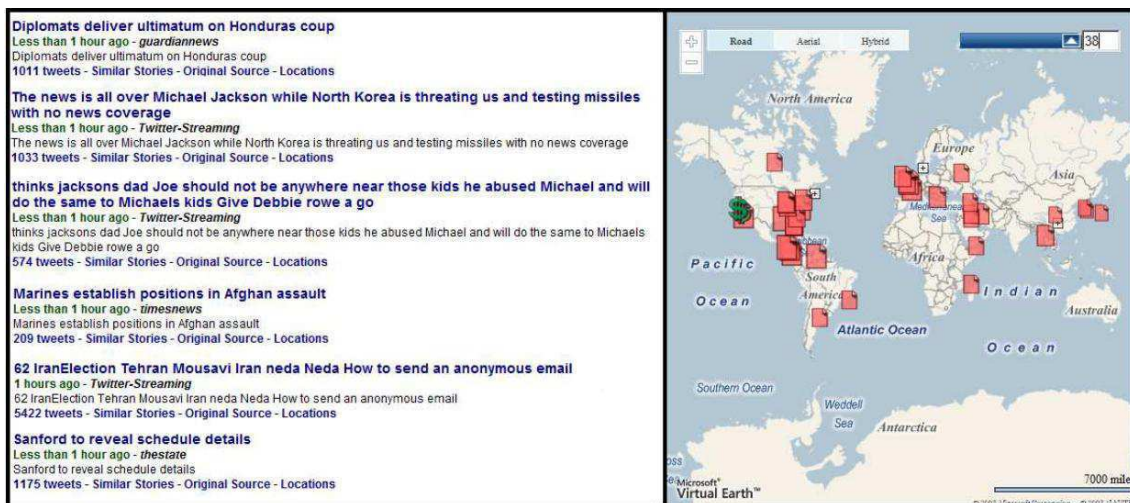


Fonte: *print screen* da aplicação NewsStand.

Por fim, um novo conceito de aplicação para leitura de notícias informativas foi sugerido por Sankaranarayanan et al. (2009). Os autores investigaram o uso do Twitter para construir um sistema de processamento de notícias, chamado TwitterStand. A ideia principal é capturar *tweets* que correspondem à *breaknews* (i.e., notícias em tempo real) permitindo a criação de um sistema central de notícias. O sistema credencia alguns usuários importantes do microblog como “repórteres”. Destes, são analisados seus *tweets* e o número de usuários que estão relatando o mesmo tópico em suas redes sociais. Assim, são identificadas notícias informativas que não necessariamente já foram relatadas ou escritas por um veículo de comunicação. Ao invés de medir o grau de credibilidade do que foi relatado nos *tweets*, o

sistema mede o grau de confiabilidade dos usuários que estão comentando ou interagindo com a notícia, o que nem sempre pode ser levado em consideração. Os *tweets* que estão relatando o mesmo tópico são agrupados e a localização geográfica deste grupo é inferida através das referências encontradas nos mesmos ou pela localidade de moradia dos autores. Ademais, um sistema para consulta destas notícias é disponibilizado, mas a interface gráfica é precária, limitando as possibilidades de consulta, como pode ser visto na Figura 9.

Figura 9: Interface Gráfica do TwitterStand.



Fonte: Sankaranarayanan et al. (2009).

### 3.3.1. Sumarização das Ferramentas

O resultado da avaliação das ferramentas de consulta e leitura de notícias foi sumarizado na Tabela 3 na forma de métodos (colunas) e características (linhas). A célula preenchida com o símbolo • indica que a característica está presente no método. O preenchimento da célula com \* indica a existência da característica, mas com algumas restrições. Por fim, quando a célula está vazia significa que o método não considera aquela característica em sua abordagem.

Tabela 3: Tabela de sumarização das ferramentas de consulta de notícias.

ID	Características/Método	Bing News	Google News	Yahoo News	Phelan et al., 2009	Shai et al.	Samet et al., 2014	Sankaranarayanan et al., 2009
1	Captura notícias em RSS.	.	.	.	.	.	.	
2	Captura notícias em redes sociais.				.			.
3	Agrega múltiplas fontes de notícias.	.	.	.	.	.	.	
4	Permite consultas em uma perspectiva temporal.					.	.	
5	Permite consultas em um contexto geográfico.						.	.
6	Disponibiliza opções de consultas espaciais complexas (e.g. adjacência, distante de, na área definida).							
7	Fornecer uma interface simples e amigável para visualização dos resultados.	.	.	.				
8	Permite perceber instantaneamente as localidades relacionadas nas notícias.						.	.
9	Atribui um grau de credibilidade à notícia.	*	*	*			*	*
10	Considera a responsividade da aplicação.		.	.				

Fonte: elaborado pelo autor.

\* valor de credibilidade associado ao veículo de comunicação e indiretamente à notícia.

### 3.3.2. Considerações Sobre as Ferramentas de Leitura de Notícias

É perceptível que no meio comercial o número de produção de ferramentas para buscar, agrupar e ler notícias informativas é bem maior em relação ao meio acadêmico. Neste

levantamento, foram consideradas apenas três ferramentas do meio comercial com este foco, mas há diversas outras soluções que entregam aos usuários as mesmas funcionalidades encontradas nas analisadas. A motivação para a escolha destas está no fato de representarem as iniciativas das maiores empresas de tecnologia para esta demanda. No entanto, percebe-se que estas ferramentas ainda são carentes de funções capazes de atender todos os tipos de usuários, desde aquele mais simples que deseja se manter informado das notícias do seu país, até aquele usuário mais exigente que almeja encontrar notícias sobre determinada área e/ou em uma faixa restrita de tempo.

Logo, no meio científico, apesar da escassez de soluções com este foco, é possível perceber a preocupação dos autores em fornecer um ambiente mais completo para consultas e leitura de notícias. Dos trabalhos analisados, podem-se destacar o NewsStand (Samet et al., 2014) e o TwitterStand (Sankaranarayanan et al., 2009), que propõem um ambiente de consulta e leitura de notícias, prezando pelo contexto geográfico envolvido nos tópicos e que trazem notícias provenientes de mídias informativas e de redes sociais, respectivamente. Contudo, a baixa qualidade da interface gráfica, a dificuldade de exibição dos resultados, as opções de consultas espaciais reduzidas e a falta de responsividade, em ambas as ferramentas, podem desencorajar sua adoção por parte dos usuários.

O grau de credibilidade associado à notícia é um fator importante, mas desconsiderado nas soluções analisadas. Na análise, foi possível perceber que, para valorar o grau de confiança da notícia, as ferramentas consideram apenas a credibilidade do veículo que a produziu, ignorando a vulnerabilidade desta informação. É sabido que os veículos de comunicação estabelecem suas sedes em pontos estratégicos e relatam notícias de várias localidades, sem haver qualquer regulamentação quanto a isso. Ou seja, um veículo de comunicação pode relatar um acontecimento em uma localidade desconhecida ou até mesmo nunca visitada pelo jornalista. Deste modo, é arriscado associar o valor de credibilidade do veículo de comunicação à notícia informada por este. O valor de confiança da notícia deve ser medido seguindo diversos outros fatores, como, por exemplo, a confirmação dos envolvidos com as localidades relatadas na mesma.

Portanto, construir um ambiente completo para busca e leitura de notícias de diversas fontes de informação, associadas com um grau de confiança calculado mediante análise de dados oriundos de redes sociais, que permite várias opções de consultas em contexto geográfico e temporal, aliado ao uso de uma interface gráfica dinâmica, interativa e

responsiva, configura-se como uma solução promissora para sanar as limitações encontradas nos trabalhos analisados.

#### 3.4. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo, foi possível analisar algumas técnicas, métodos e ferramentas encontradas na literatura com o foco em notícias informativa. Os trabalhos e projetos analisados foram selecionados com base em sua importância e documentação disponíveis. Desta análise, foi possível perceber algumas carências e diretrizes que podem ser seguidas para a evolução do estado da arte em termos de técnicas e ferramental disponíveis.

No capítulo seguinte é apresentada a solução proposta pelo presente trabalho, bem como os detalhes sobre sua construção e funcionamento. Ademais, o protótipo de validação desenvolvido, chamado GeoSEn News, é detalhado em seus aspectos estruturais e comportamentais.

## CAPÍTULO 4

### GEOSEN NEWS

Este capítulo tem por finalidade apresentar as técnicas e metodologias produzidas e implantadas em um motor de busca geográfica com enfoque em notícias informativas distribuídas, o GeoSEn News. O GeoSEn News é uma ferramenta extensora do motor de busca com enfoque geográfico chamado GeoSEn. Essa ferramenta servirá de apoio na implantação e avaliação de um novo conceito de *ranking* de relevância baseado na localização geográfica de usuários que interagem com notícias em redes sociais.

O objetivo principal desta dissertação é construir um conceito de *ranking* de relevância para notícias informativas baseado na integração de dados oriundos das redes sociais (i.e. *posts* de veículos de comunicação e localização geográfica dos usuários) e aplicá-lo em um motor de busca geográfica no contexto de notícias. Para tornar viável sua construção, alguns componentes do motor de busca GeoSEn foram especializados para o tratamento de notícias originadas por diversos veículos de comunicação e coletadas em redes sociais.

Para desenvolvimento dos componentes utilizados no GeoSEn, descritos nas seções subsequentes, foi considerado apenas a rede social Twitter, por possuir uma API de acesso robusta e de vasta documentação. No entanto, este conceito apresentado no trabalho pode ser generalizado para o uso de informações provenientes de qualquer rede social, e caso seja necessário o trabalho em outra plataforma, como, por exemplo, o Facebook<sup>20</sup>, estes devem sofrer alguns ajustes tecnológicos no tocante às bibliotecas, recursos e armazenamento dos dados que serão utilizados.

O restante do capítulo está organizado na seguinte forma. Na Seção 4 descreve-se a etapa de captação e análise de notícias. Na Seção 4.2 é detalhado o mecanismo de estimativa de afinidade local. Na Seção 4.3 é descrito a produção do *ranking* de relevância geosocial. Na seção seguinte (4.4), será apresentada a arquitetura do GeoSEn News, bem como sua interface multi-modo e principais funcionalidades, seguindo pela Seção 4.5 com as considerações finais do capítulo.

#### 4.1. COLETOR DE NOTÍCIAS

Atualmente, inúmeras são as iniciativas de tornar as redes sociais ferramentas de leitura de notícias. Destas destacam-se o *My Yahoo!*, Bing News e Google News, que

---

<sup>20</sup> <https://facebook.com>



fornece como funcionalidade a agregação de notícias, em que para cada usuário na rede, são exibidas as notícias mais lidas por seus amigos (i.e. *trending*) e as publicadas por mídias informativas preferidas por ele (Bao, Mokbel e Chow, 2012).

Antes da ascensão das redes sociais e suas facilidades, os portais de notícias e blogs difundiam suas notícias utilizando o mecanismo, baseado em XML, chamado *Really Simple Syndication* (RSS). O RSS é um formato de consumo (*feed*) para Web, usado com o objetivo de agregar conteúdos de blogs e portais de notícias informando aos usuários assinantes sobre atualizações ocorridas. O mesmo é uma extensão do XML que sumariza itens e *links* para o conteúdo completo, podendo assim ser acessados diretamente no *browser* ou em *software* especializado na agregação desses *feeds* (Murugesan, 2007). O RSS ainda é bastante utilizado por portais de conteúdos, mas não como único recurso de divulgação.

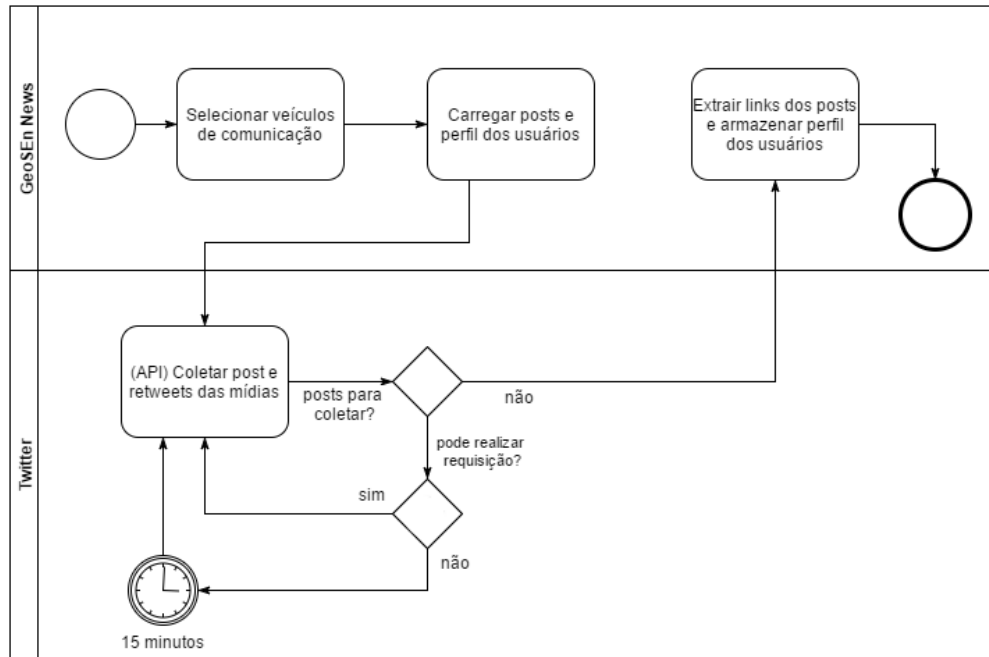
Durante a pesquisa desta dissertação foi possível identificar uma nova abordagem de difusão, por parte dos veículos de comunicação e mídias em geral, que consiste na inserção direta da notícia na rede. Ou seja, os veículos de comunicação normalmente possuem uma conta oficial em uma rede social e, assim que escrevem uma nova notícia em seu portal, a publicam também em sua conta oficial.

Como já relatado na Seção 2.2.1, os primeiros passos na indexação de documentos para um sistema de RI, seja ele tradicional ou não, é a seleção de páginas que servirão como sementes (do inglês, *seeds*) para o coletor. Essa seleção é feita usualmente de forma manual e de acordo com o escopo da ferramenta. Por exemplo, se o motor de busca é dedicado ao mundo esportivo, as sementes escolhidas deverão ser páginas de portais que tratam sobre do assunto.

O objetivo desta dissertação é criar um novo conceito para *ranking* de relevância em notícias informativas, fazendo uso da interação de usuários, em redes sociais, com as mesmas. Este novo conceito consiste em medir o grau de credibilidade da notícia com base no relacionamento entre as localidades retratadas na notícia e a localização dos usuários que a difundiram em sua conta no microblog.

De acordo com Waldo Tobler (1970), a Primeira Lei da Geografia diz que “*Tudo está relacionado com tudo o resto, mas coisas próximas estão mais relacionadas do que coisas distantes*”. Inspirando-se nesta lei, pode-se afirmar que um usuário residente em uma determinada região tem maior propriedade para atestar uma notícia vinculada sobre esta região em comparação com um usuário que não reside nesta. Ou seja, quando este usuário

Figura 10: Processo de coleta de notícias no Twitter.



Fonte: elaborado pelo autor.

difunde (e.g. *retweet*) uma notícia sobre sua região, considera-se que este está implicitamente afirmando que há certa confiança no que foi escrito.

Deste modo, percebendo esta nova forma de divulgação de notícias por parte dos veículos de comunicação e necessitando selecionar sementes (i.e. *links* de notícias) para fornecer ao *crawler* (i.e. mecanismo presente em sistemas de RI responsável pela coleta, análise e indexação de conteúdo que estará disponível para ser consultado), a forma mais astuta é concentrar a coleta diretamente no Twitter, permitindo a associação entre os *links* coletados e as interações dos usuários com os mesmos. Ao invés de usar páginas de portais de notícias como sementes, são utilizadas as contas oficiais de veículos de comunicação no microblog como ponto de partida para a coleta de notícias. Na Figura 10 é apresentado o processo de coleta de notícias, modelado seguindo o padrão de Modelagem de Processos de Negócios BPMn (do inglês, *Business Process Modeling*).

O processo, diagramado na Figura 10, pode ser descrito da seguinte forma: no início, o administrador do sistema seleciona os veículos de comunicação (i.e. contas oficiais no Twitter) que se deseja coletar as notícias. Posteriormente, são coletados os *posts* (referente às notícias) e o perfil dos usuários que compartilharam (i.e. *retweet*). Esta coleta é feita de forma interativa, por veículo de comunicação, sempre respeitando a janela de requisições (i.e. tempo de espera, neste caso de 15 minutos) quando o número de requisições extrapolar o permitido

pela API do Twitter. Por fim, quando não houver *posts* a coletar, os mesmos, junto com os perfis coletados, são submetidos à atividade de extração de *links* e armazenamento.

O algoritmo, formalizado no Código 1, pode ser descrito da seguinte forma:

1. Fornece uma lista de contas oficiais associadas aos veículos de comunicação e o número máximo de *posts* que serão coletados de cada uma delas (linha 1);
2. Para cada conta oficial pertencente à lista de contas (linha 2);
3. Coleta os *posts* na conta oficial do veículo selecionado, respeitando o número máximo (*maxPosts*) de *posts* que serão coletados por conta oficial (linha 3);
4. Itera sobre todos os *posts* coletado (linha 4);
5. Persiste o *post* na base de dados (linha 5);
6. Coleta o perfil dos usuários que realizaram *retweet* do *post* selecionado (linha 7);
7. Itera sobre cada usuário, executando o *geocoding* no campo “*location*” do perfil com o objetivo de identificar a localidade de moradia do mesmo (linha 8-11);
8. Persiste o perfil do usuário na base de dados (linha 13);
9. Repete o passo 2 até que não haja mais contas oficiais a serem analisadas;

Para desenvolvimento da arquitetura do componente coleta de notícias, foi adotado o modelo de comunicação entre duas camadas: camada de negócio e camada de dados. Na Figura 11 é ilustrada a arquitetura do coletor. A camada de negócio é responsável pelo

Código 1: Pseudo-código para coleta de notícias no Twitter.

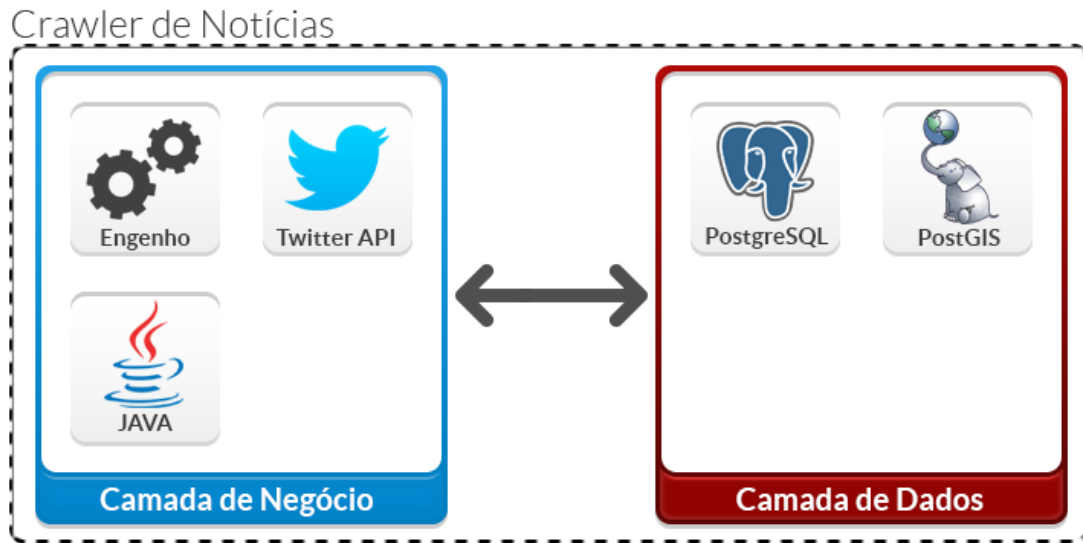
```

1  function newsCrawler(newsSourceList, maxPosts) {
2      foreach(newsSource in newsSourceList) {
3          posts = collectTweets(newsSource, maxPost);
4          foreach(post in posts) {
5              success = persistPost(post);
6              if(success) {
7                  users = retweetUsers(post);
8                  foreach (user in users) {
9                      geoloc = geocoding(user.location);
10                     if(geoloc) {
11                         user.geolocation = geoloc;
12                     }
13                     persistUser(user);
14                 }
15             }
16         }
17     }
18 }

```

Fonte: elaborado pelo autor.

Figura 11: Arquitetura do coletor de notícias.



Fonte: elaborado pelo autor.

processamento de coleta de notícias no microblog, desenvolvido utilizando a linguagem de programação JAVA. Esta camada comporta um mecanismo específico para coleta de *posts* e suas respectivas informações (e.g. *retweets*, *replies*, *favorites*), fazendo uso da API<sup>21</sup> disponibilizada pelo Twitter para este propósito. Por fim, a camada de dados é utilizada para armazenar todos os dados (i.e. *links*, perfil dos usuários) oriundos da camada de negócio e que serão analisados futuramente.

#### 4.2. AFINIDADE LOCAL

A efetividade do método de *ranking* de relevância geosocial depende exclusivamente das informações geográficas contidas no conteúdo da página e no perfil do usuário que a difundiu em sua rede social. De posse dessas informações, será realizada uma análise a fim de mensurar um grau de credibilidade daquela página, apenas comparando as referências geográficas encontradas em seu conteúdo com as localidades de moradia declarada no perfil dos usuários.

Nesta etapa de verificação, alguns obstáculos são enfrentados no momento da identificação da localidade de moradia fornecida no perfil dos usuários. É comum que redes sociais, como, por exemplo, o Twitter, forneçam ao usuário a possibilidade de informar uma localização de moradia em seu perfil. No entanto, esse campo segue apenas o formato textual, deixando a critério do usuário, a melhor forma de declarar onde reside. O preenchimento deste campo de forma textual é bastante sensível a erros de grafia, inconsistências e

<sup>21</sup> <https://dev.twitter.com/>

ambiguidades. Por exemplo, um usuário pode não se preocupar com a granularidade geográfica da informação e apenas informar que mora no “Brasil”. Outro poderia informar apenas o nome da cidade “Santa Luzia” sem se preocupar com o estado “Paraíba”, causando ambiguidade no momento da inferência de sua real localidade, utilizando algum mecanismo de geocodificação.

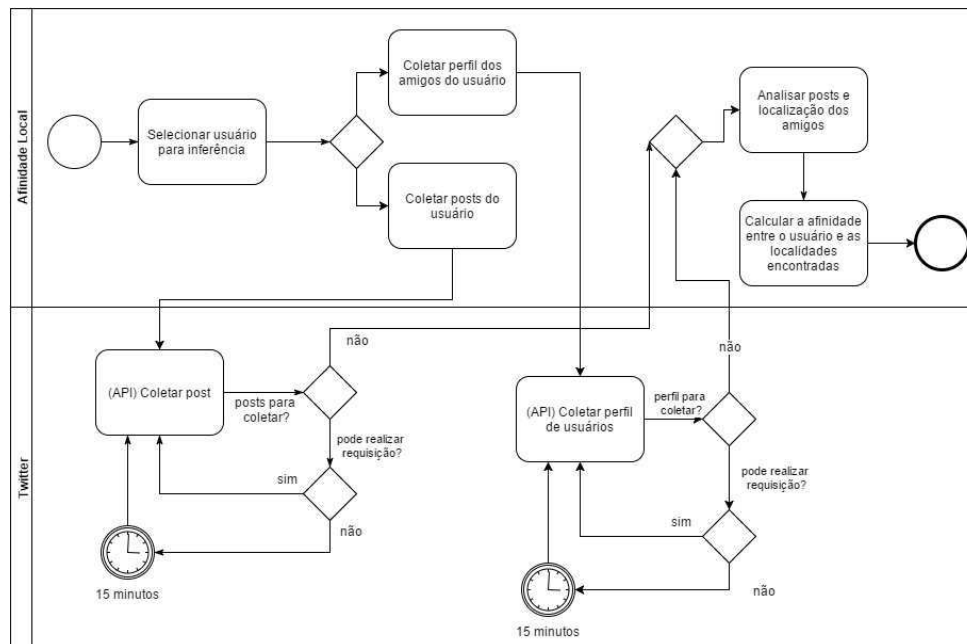
Não obstante, pesquisas apontam que apenas 5,34% dos usuários do Twitter informam em seu perfil sua localidade de moradia (Jurgens, 2013). Este número baixo dificulta diretamente na eficácia do método, impossibilitando seu uso de forma propícia. Então, torna-se importante prover uma solução capaz de inferir a localidade de moradia desses usuários que não a fornecem no momento do cadastro do perfil.

Assim, uma investigação foi realizada a fim de avaliar trabalhos que surgem com o propósito de inferir a localidade de moradia de usuários, mediante análise de comportamento em redes sociais. Jurgens (2013) propõe uma nova abordagem para inferir a localidade de um usuário mediante propagação espacial, considerando o relacionamento com amigos e distribuição espacial dos seus *posts*. O autor aponta o relacionamento entre pessoas como uma forte fonte de evidências da residência do usuário. Os *posts* geolocalizados, apesar de representarem apenas 1% dos *posts*, também se configuram como fator importante na identificação da localização de residência do usuário investigado. Os experimentos realizados avaliam positivamente o método de inferência, com taxas de erros de aproximadamente 10 km.

Em outro trabalho, Mahmud, Nichols e Drews (2012) partem de duas premissas básicas. A primeira delas diz que um usuário cita o nome de sua cidade em seus *posts* mais que qualquer outra. A segunda afirma que um usuário visita lugares na cidade onde mora mais que em qualquer outro lugar. Assim, partindo dessas deduções, os autores apresentam um algoritmo para identificar a localização atual do usuário do microblog Twitter, em diferentes granularidades (e.g. cidade, estado, *timezone*), baseado apenas em seu comportamento na rede. Experimentos realizados demonstram melhorias nos resultados encontrados em algoritmos de inferência anteriores.

Nos trabalhos mencionados anteriormente, a abordagem consistia em identificar a localidade de vivência do usuário baseado em seu comportamento em redes sociais, porém essa função pode ser generalizada para outros cenários. Bergren et al. (2015) sugerem um método que utiliza a distribuição de Gauss para mensurar o grau de localidade que tem uma determinada palavra presente em um *post* de qualquer fonte (e.g. blog, fórum de discussão). O

Figura 12: Processo de inferência de afinidade local.



Fonte: elaborado pelo autor.

princípio do método consiste na identificação de termos de uso regional, como, por exemplo, gírias ou variação linguística. Uma coleção grande de textos, de várias regiões e espacialmente anotada, é utilizada como conjunto de treinamento na geração de um modelo de predição. Desta forma, para identificar a localidade do autor do *post* basta verificar os termos utilizados e identificar a região que são mais usuais. O autor descreve uma série de experimentos para determinar como *posts* de blogs espacialmente anotados podem ser usados para aprender a localidade de outros *posts* e de seus respectivos autores. Os experimentos avaliaram a eficiência do método, que conseguiu um resultado satisfatório em 33% com precisão de 100 km de distância.

Apesar de eficientes no propósito de inferir a localização de residência de usuários, os trabalhos anteriormente citados não satisfazem todas as necessidades exigidas no *ranking* de relevância geosocial. Ao invés de trabalhar apenas com a localização de moradia atual do usuário, convém utilizar uma cadeia de localidades com as quais ele tem alguma relação e, conseqüentemente, domínio para falar sobre tal. Por exemplo, um usuário pode ter vivido muitos anos em uma determinada região e ter se transferido para outra. Esse usuário terá propriedade para falar e confirmar informações tanto do lugar onde viveu quanto do lugar onde está vivendo atualmente. O mesmo acontece com lugares onde o mesmo possui um número considerável de amigos ou lugares que visitou recentemente.

Por tal necessidade, optou-se por construir um método capaz de medir o grau de afinidade que um usuário tem com determinada localidade. A abordagem assemelha-se ao que

é feito por Bergren (2013), analisando o comportamento do usuário no microblog e inferindo várias localidades relacionadas a este. Na Figura 12 é ilustrado o processo de inferência da afinidade local de usuários.

O processo de inferência, diagramado na Figura 12, pode ser descrito da seguinte forma: inicialmente, seleciona um usuário da lista de usuários armazenados no sistema; logo depois, são coletados os *posts* do usuário bem como uma lista de amigos na sua rede, sempre respeitando as restrições de recursos impostos pela API do Twitter; posteriormente, quando não houver mais *posts* e perfil de amigos daquele usuário a serem coletados, os mesmos são submetidos à etapa de extração de referências geográficas nos *posts* e identificação da localidade de moradia no perfil dos amigos; por fim, são calculados os valores de afinidade entre o usuário e as localidades encontradas em seu comportamento na rede social.

O processo de inferência pode ser dividido em duas etapas. A primeira delas, responsável apenas pela coleta de informação, é formalizada no Código 2, e segue o algoritmo:

1. Fornece uma lista de usuários para a ferramenta de detecção de afinidade local, um número máximo de *posts* coletados e um número máximo de perfis dos amigos capturados (linha 1);
2. Seleciona um usuário da lista de usuários (linha 2);
3. Captura os *posts* na linha de tempo do usuário, respeitando o limite estabelecido no início do processo (linha 3).
4. Atribui os *posts* coletados ao usuário em questão e os armazena para análise futura (linhas 4-5);
5. Coleta o perfil dos amigos, respeitando o limite estabelecido (linha 6);
6. Executa, em cada perfil de amigo coletado, o mecanismo de geocodificação com o

Código 2: Pseudo-código de coleta do comportamento de usuários.

```

1  function behaviorCrawler(usersList, maxPosts, maxFriends) {
2      foreach(user in usersList) {
3          posts = collectTweets(user, maxPosts);
4          user.posts = posts;
5          persistPosts(user.posts);
6          friends = collectFriends(user, maxFriends);
7          user.friends = friends;
8          persistFriends(user.friends);
9      }
10 }

```

Fonte: elaborado pelo autor.

Código 3: Mapa de referências encontradas.

```

1  var referencias_encontradas = {
2      "ref_id_0" : (mp, lp, fi) ,
3      "ref_id_1" : (mp, lp, fi) ,
4      "ref_id_n" : (mp, lp, fi)
5  }

```

Fonte: elaborado pelo autor.

intuito de identificar a localidade de moradia do mesmo;

7. Atribui os perfis coletados ao usuário em questão e os armazena para futura análise (linhas 7-8);
8. Repete o passo 2 até que não haja mais usuários à serem analisados;

Depois de realizada toda coleta de *posts* e relacionamentos na rede social, é dado início à etapa de análise destes. Nesta etapa do processo, os *posts* do usuário são submetidos a um mecanismo capaz de identificar referências geográficas em pequenos textos, pois um *post* no Twitter é restrito a 140 caracteres. O artefato construído neste processo pode ser descrito como um mapa, onde a chave é o identificador único da referência geográfica (*ref\_id*) e o valor é uma lista contendo a quantidade de menções em *posts* (*mp*), a quantidade de *posts* geolocalizados naquela localidade (*lp*) e o número de amigos que residem naquela localidade (*fi*), respectivamente. O Código 3 representa o mapa de referências avaliadas no cálculo de afinidade local.

O algoritmo de análise, formalizado no Código 4, pode ser descrito da seguinte forma:

1. Fornece uma lista de usuários, com seus respectivos *posts* e relacionamentos de amizade (linha 1);
2. Seleciona um usuário da lista de usuários (linha 2);
3. Itera sobre cada *post* do usuário selecionado com o objetivo de identificar referências geográficas no texto e/ou geolocalização (i.e. localização atribuída por meio do GPS diretamente no *post*) (linha 5):
  - a. Caso encontre uma referência no *post*:
    - i. Examinar se a referência já está na lista. Se sim, incrementa o valor de “mp” daquela referência. Se não, acrescenta a nova referência na lista com “mp” igual a 1 (linhas 7-12);
  - b. Caso encontre um *post* geolocalizado:



Código 4: Pseudo-código para cálculo de afinidade local.

```

1  function localAffinity(usersList){
2  foreach(user in usersList){
3      mapLocations = {};
4      posts = user.posts;
5      foreach(post in posts){
6          references = extractReferences(post);
7          foreach(ref in references){
8              if(mapLocations.keys.contains(ref.id)){
9                  mapLocations[ref.id][0]++; //mp
10             }else{
11                 mapLocations[ref.id] = [1,0,0];
12             }
13         }
14         geotag = extractGeotag(post);
15         if(geotag && mapLocations.keys.contains(geotag.id)){
16             mapLocations[ref.id][1]++; //lp
17         }else if(geotag){
18             mapLocations[ref.id] = [0,1,0];
19         }
20     }
21     friends = user.friends;
22     foreach(friend in friends){
23         geolocation = geocoding(friend.location);
24         if(geolocation && mapLocations.keys.contains(geolocation.id)){
25             mapLocations[ref.id][2]++; //fi
26         }else if(geolocation){
27             mapLocations[ref.id] = [0,0,2];
28         }
29     }
30     locations = localAffCalculator(mapLocations);
31     geoTreeAffinity = geoTreeGenerator(locations);
32 }
33 }

```

Fonte: elaborado pelo autor.

- i. Examina se a referência já está na lista. Se sim, incrementa o valor de “lp” daquela referência. Se não, acrescenta a nova referência na lista com “lp” igual a 1 (linhas 14-19);
- c. Prossegue a iteração até que não haja *posts* a serem analisados;
4. Itera sobre todos os perfis dos amigos do usuário selecionado (linhas 22):
  - a. Caso o usuário consiga ser geolocalizado:
    - i. Examina se a localidade já está na lista. Se sim, incrementa o valor de “fi” daquela referência. Se não, acrescenta a nova referência encontrada na lista com o valor de “fi” igual a 1 (linhas 23-29);
    - b. Prossegue até que não haja mais perfis a serem analisados;
5. Calcula o grau de afinidade do usuário com as localidades encontradas, descrito na seção subsequente (linha 30);
6. Monta a árvore geográfica de afinidades do usuário e a persiste na base de dados. A técnica utilizada na geração da *geotree* é descrita na sequência (linha 31);
7. Repete o passo 2 até que não haja mais usuários a serem avaliados;

Figura 13: Processamento do cálculo de afinidade local.



Fonte: elaborado pelo autor.

Na Figura 13 é apresentado um exemplo do processo de análise e extração de referências em *posts*.

A detecção de referências em pequenos textos é tarefa complexa e muitas vezes difícil de ser realizada. Alguns trabalhos da literatura compartilham da mesma solução para esse problema. Os trabalhos de Jurgens (2013) e Bergren et al. (2015) fazem uso de *gazetteer* na identificação de referências geográficas em pequenos textos, como, por exemplo, um *post* no microblog *Twitter*. *Gazetteer* é um dicionário geográfico que tipicamente é usado na identificação de referências geográficas, que, por sua vez, não são necessariamente nomes de cidades ou região. Por exemplo, o termo “Rainha da Borborema” é um adjetivo dado à cidade de Campina Grande/PB que pode ser identificado, em uma resolução de topônimos, com o auxílio de um *gazetter*.

Porém, para a identificação de referências geográficas em pequenos textos, que serão utilizadas para inferência de um valor de afinidade entre uma localidade e o usuário, optou-se por utilizar o *GeoSEn Geoparser* (Campelo e Baptista, 2009), módulo responsável por detectar referências geográficas em textos escritos em português. Este módulo foi construído exclusivamente para atender o processo de *parser* em documentos indexados no GeoSEn, no entanto, com alguns ajustes pontuais, foi possível utilizá-lo em outros cenários (e.g. detecção de referências geográficas em *posts* do *Twitter*). Neste estágio, os termos candidatos são localizados no conteúdo e depois passaram por um crivo, responsável por eliminar os termos candidatos que não satisfazem determinadas métricas. Na Figura 14 é demonstrado o *geoparsing* realizado em um texto curto, destacando os termos candidatos. Oliveira et al. (2014) expõem uma abordagem para automatização de *posts* (i.e. inserção automática de conteúdo) em sistemas de Informação Geográfica Voluntária (do inglês, *Volunteered Geographic Information – VGI*) baseado no georreferenciamento de textos publicados na Web. Para avaliar tal, os autores utilizaram o *GeoSEn Geoparser* em textos do microblog

Twitter, que, por sua vez, mostrou-se bastante eficiente na identificação de referências geográficas em pequenos textos.

Contudo, apesar dos resultados do *GeoSEn Geoparser* serem satisfatórios para este cenário, evoluções na técnica de detecção de topônimos foram feitas em paralelo ao desenvolvimento desta dissertação com o intuito de atingir uma melhor eficácia no processo. O método de detecção de topônimos utilizado pelo *GeoSEn Geoparser* é baseado em várias heurísticas e técnicas de processamento de linguagem natural. Uma destas técnicas baseia-se na presença de termos influentes, que são termos que podem indicar a existência de uma referência geográfica. Por exemplo, os termos “na cidade de” e “nas imediações de” são considerados termos influentes, pois, na maior parte dos casos, antecedem o nome de alguma cidade ou localização, respectivamente, influenciando o grau de confiança destas referências geográficas no processo de desambiguação.

Assim, com a finalidade de enriquecer o conjunto de termos influentes utilizados pelo *GeoSEn Geoparser* na identificação de referências geográficas, Jerônimo, Campelo e Baptista (2015) apresentam uma abordagem para identificar, de forma automática, termos influentes relevantes e um conjunto de atributos que relacionam esses termos aos topônimos (i.e. termos candidatos a referências geográficas). Foram realizados experimentos e os resultados indicam que a técnica é eficaz na identificação automática de termos influentes e que houve ganho significativo, por parte do *Geoparser*, na capacidade de detecção de topônimos quando usada a lista de termos de influência gerada. Na Figura 15 é descrito o fluxo do processo da identificação de novos termos influentes.

No entanto, o conjunto de termos influentes pode variar de acordo com o tipo de

Figura 14: Geoparsing em micro texto.



Fonte: elaborado pelo autor.

Figura 15: Fluxo do processo de identificação de termos influentes.



Fonte: adaptado de Jerônimo, Campelo e Baptista (2015).

texto a ser analisado, como, por exemplo, se é um texto que emprega linguagem formal ou informal. Como o objetivo deste trabalho é identificar topônimos em pequenos textos oriundos de redes sociais, a técnica de geração automática de termos influentes foi executada com um conjunto de treinamento de aproximado de 25.000 *posts* do microblog Twitter. Desta execução foram extraídos 1.173 termos influentes que são utilizados como apoio no momento do *parsing*, realizado pelo módulo *GeoSEn Geoparser*.

Finalizados os estágios de coleta e análise do comportamento do usuário, o valor de afinidade entre o usuário e as localidades encontradas em seus *posts* e amigos é calculado. Esse valor é obtido conforme a Equação 3.

$$al(u, r) = \left( \frac{u_{r,mp}}{\#pc} \times pmp \right) + \left( \frac{u_{r,lp}}{\#pc} \times plp \right) + \left( \frac{u_{r,fi}}{\#ac} \times pfi \right), r \in u_r \quad \text{Equação 3}$$

Onde:

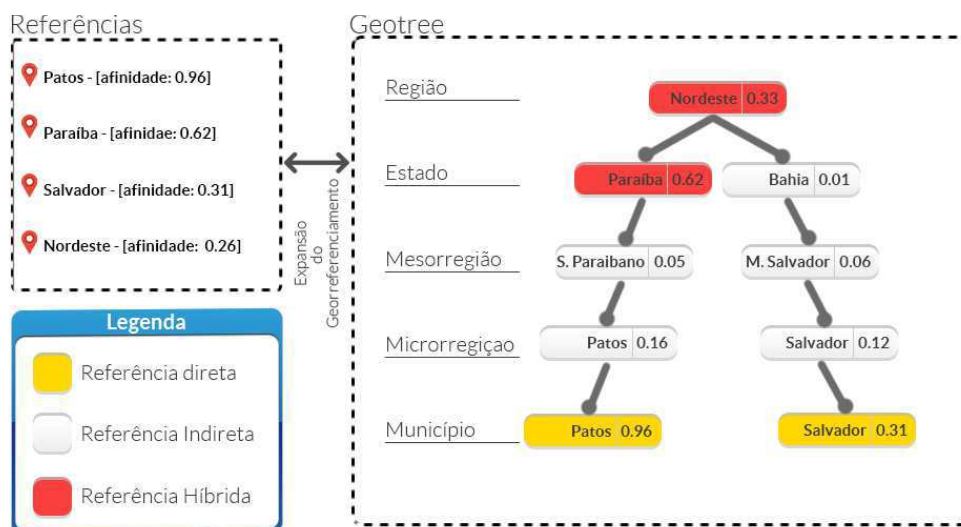
- $al(u, r)$ : método responsável por calcular a afinidade local entre o usuário  $u$  e a localização  $r$ ;
- $u_{r,mp}$ : valor de referências para a localidade  $r$  encontradas nos *posts* do usuário  $u$ ;
- $u_{r,lp}$ : valor de *posts* geolocalizados no local  $r$  pelo usuário  $u$ ;
- $u_{r,fi}$ : valor da quantidade de amigos do usuário  $u$  que residem na localização  $r$ ;
- $\#pc$ : quantidade de *post* coletados para inferência do valor de afinidade local;
- $\#ac$ : quantidade de perfil de amigos coletados para inferência do valor de afinidade local;
- $pmp$ : peso dado para referências geográficas encontradas em *posts*. O valor é empiricamente determinado e varia entre 0 e 1;
- $plp$ : peso dado para *posts* geolocalizados. O valor é empiricamente determinado e varia entre 0 e 1;
- $pfi$ : peso dado para o relacionamento de amizade. O valor é empiricamente determinado e varia entre 0 e 1;

Finalizado o procedimento de cálculo dos valores de afinidade local entre o usuário e as referências geográficas encontradas em seu comportamento no microblog, é dado início ao estágio de construção de uma árvore geográfica (do inglês, *geotree*). *Geotree* é uma estrutura de dados em formato de árvore produzida para dar suporte à implementação da técnica de expansão do georreferenciamento (Campelo, 2008). Cada nó da *geotree* representa uma localidade que faz parte do contexto geográfico do cenário, neste propósito, para o comportamento do usuário. Sua utilização permite uma maior flexibilidade no momento do cálculo do *ranking* de relevância geosocial, fazendo com que qualquer localidade avaliada, apesar de não ter sido referenciada diretamente, possa ter sua afinidade valorada.

Para esclarecer o processo de expansão, seguindo a instância na Figura 16, suponha que *R* é a lista de referências geográficas encontradas no comportamento do usuário *u* e que já foi submetida para o processo de cálculo de afinidade local. Esta lista contém *Patos* e *Salvador* no nível hierárquico de município, *Paraíba* no nível de estado e *Nordeste* no nível hierárquico de região. Os municípios *Patos* e *Salvador* pertencem ao estado da *Paraíba* e *Bahia*, respectivamente. Note que estes municípios pertencem às determinadas microrregião (*Patos*; *Salvador*) e mesorregião (*Sertão Paraibano*; *Metropolitana de Salvador*) que não foram diretamente referenciadas, mas que deveriam ser utilizadas para valorar o fator geosocial de um documento. Assim, as referências diretamente mencionadas devem servir para expandir as referências de nível hierárquico superior, mas que não foram explicitamente mencionadas.

O cálculo do valor de afinidade local com uma localização que foi expandida geograficamente segue a fórmula da Equação 4.

Figura 16: Exemplo da *geotree*.



Fonte: elaborado pelo autor.

$$peso(x) = \frac{\sum_{i=1}^{\#f(x)} peso(f(x)[i])}{tf(x)} \quad \text{Equação 4}$$

Onde,

- $peso(x)$ : é o valor da afinidade local do nó  $x$  na *geotree*;
- $\#f(x)$ : é a quantidade de itens no conjunto de filhos de  $x$ ;
- $f(x)[i]$ : é o  $i$ -ésimo filho do nó  $x$ ;
- $tf(x)$ : o total máximo de filhos para o nó  $x$ ;

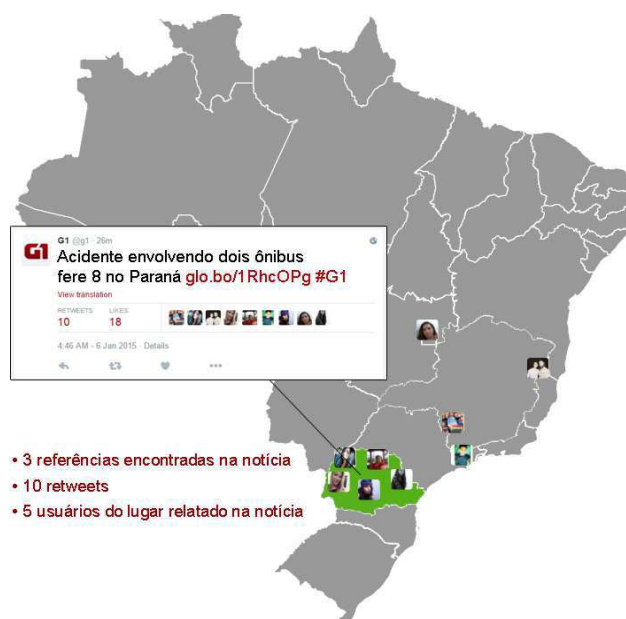
Por fim, a *geotree* de afinidade local de cada usuário é armazenada na base de dados. Estas informações serão acessadas novamente na etapa de cálculo do *ranking* de relevância geosocial.

#### 4.3. RANKING DE RELEVÂNCIA GEOSOCIAL

Esta seção é responsável por apresentar o processo relacionado ao cálculo do *ranking* de relevância geosocial, implantado no motor de busca geográfica com enfoque em notícias informativas, o GeoSEn News. Este *ranking* consiste na agregação de um novo fator ao método de *ranking* de relevância já existente na ferramenta GeoSEn. O cálculo deste fator e o modo como é adaptado ao *ranking* de relevância já existente, são descritos de forma detalhada no decorrer da seção.

*Ranking* de relevância é um dos grandes desafios da área de Recuperação da Informação e consiste em ordenar, de forma eficiente e satisfatória, de acordo com algum critério estabelecido, os documentos retornados com base em uma determinada consulta realizada pelo usuário. O *ranking* de relevância geográfica, por sua vez, estabelece uma relação entre a necessidade humana de informação geográfica e objetos com informações georreferenciadas (e.g. documentos, vídeos, fotos, textos) (Raper e Jonathan, 2007).

Figura 17: Exemplo da atividade de medir a credibilidade de uma notícia.



Fonte: elaborado pelo autor.

O *ranking* de relevância geográfica é tradicionalmente dividido em dois tipos: o dependente de consulta e o independente de consulta. O primeiro deles consiste na combinação da similaridade textual e espacial da página com relação ao texto e ao *footprint* da consulta. Uma consulta que deseja encontrar “concursos públicos em um raio de 200 km da atual localização”, é um exemplo de busca dependente de consulta. O segundo tipo consiste em calcular, no momento da indexação da página, valores como “popularidade” ou “significância” de determinadas páginas, seguindo alguma métrica predeterminada e apenas ordenar o conjunto de resultados de acordo com esses valores (Kumar e Boll, 2013).

A proposta desta dissertação consiste em promover um novo *ranking* de relevância geográfica, dependente de consulta, que consiste no acréscimo de um terceiro fator, que recebe o nome de *Relevância Geosocial*. Este coeficiente é medido baseado na localização geográfica do usuário que compartilha determinada notícia em sua rede social. Por exemplo, considere um veículo de comunicação que publicou, em sua conta oficial no Twitter, uma notícia sobre a capital de Minas Gerais, Belo Horizonte. Esta notícia foi compartilhada (i.e. *retweet*) por diversas pessoas vivem em Belo Horizonte ou nas proximidades. Conseqüentemente, esta notícia terá um alto grau de relevância social, pois as pessoas que estão realizando a ação de *retweet* estão implicitamente ratificando aquela informação. Na Figura 17 é descrito um caso similar.

Diante disto, é possível perceber que uma notícia que foi “retuitada” por usuários próximos ou pertencentes àquela região sobre a qual a notícia se refere tende a ter um alto

grau de credibilidade da sua informação. Em contrapartida, uma notícia que foi compartilhada por usuários que estão distantes ou que não possuem nenhuma relação com a região sobre a qual se refere a notícia, não terá uma credibilidade tão alta.

O novo método de *ranking* de relevância inclui, além dos fatores nativos de relevância textual (cálculo realizado pelo indexador de documentos Solr<sup>22</sup>) e relevância espacial do GeoSEn (detalhado na Seção 2.4), um terceiro fator que representa o grau de credibilidade da notícia, inferido com base na interação de usuários com o microblog Twitter. O coeficiente de relevância geosocial (*gsr*) é calculado da seguinte forma: seja *L* o conjunto de *links* de notícias (capturados em *posts* no Twitter) e *R<sub>i</sub>* o conjunto de usuários que realizaram *retweet* do *link*  $i \in L$ . Assim,

$$gsr(i, l) = \sum_{n=1}^{\#R_i} al(R_i[n], l), i \in L \quad \text{Equação 5}$$

Onde,

- $gsr(i, l)$  é a função que determina a relevância geosocial do *link* (i.e. notícia) na localização *l*;
- $R_i[n]$  é o *n*-ésimo usuário em  $R_i$ ;
- $al(R_i[n], l)$  é a afinidade local entre o usuário  $R_i[n]$  com a localização *l* (Equação 3).

Por meio desta Equação 5, calcula-se o valor de relevância geosocial para cada *link* e para cada localidade associada ao artigo (i.e. localidades detectadas pelo *geoparser* no conteúdo da notícia). Por exemplo, se uma notícia retrata um ocorrido em duas cidades, cada uma terá seu valor de relevância geosocial calculado. Note que, como uma notícia pode ser publicada por um ou mais veículos de comunicação, os *links* são agrupados de forma a somar os *retweets* em todas as publicações. Este cálculo de relevância geosocial é realizado no momento em que está sendo realizado o *parser* da notícia, o que evita atrasos no momento do processamento da consulta submetida pelo usuário.

Por fim, o cálculo da relevância geosocial (*rgs*) pode ser definido pela seguinte fórmula:

---

<sup>22</sup> [https://lucene.apache.org/core/5\\_5\\_0/core/org/apache/lucene/search/package-summary.html](https://lucene.apache.org/core/5_5_0/core/org/apache/lucene/search/package-summary.html)



$$rgs(q, d) = rt(d, q_{texto}) + \sum_{n=1}^{\#q_l} rg(d, q_l[n]) + \sum_{n=1}^{\#q_l} gsr(d, q_l[n]) \quad \text{Equação 6}$$

Onde,

- $rgs(q, d)$  é a função que realiza o cálculo da relevância do documento  $d \in D$  perante a consulta  $q$ ;
- $q_{texto}$  é o texto correspondente a consulta realizada;
- $\#q_l$  é o total de elementos no conjunto de localizações associadas a consulta  $q$ ;
- $q_l[n]$  é o  $n$ -ésimo elemento pertencente ao conjunto de localizações associadas a consulta  $q$ ;
- $rt(d, q_{texto})$  é o valor do coeficiente de relevância textual do documento  $d$  em relação à consulta  $q$ ;
- $rg(d, q_l[n])$  é o valor do coeficiente de relevância geográfica do documento  $d$  em relação ao  $n$ -ésimo elemento do conjunto de localizações associadas a consulta  $q$ ;
- $gsr(d, q_l[n])$  é o valor do coeficiente de relevância geosocial do documento  $d$  em relação ao  $n$ -ésimo elemento do conjunto de localizações associadas a consulta  $q$ ;

O valor do coeficiente de relevância textual ( $rt$ ) é calculado pelo *framework* SOLR. A fórmula de *ranking* faz uso do Modelo Booleano para “filtrar” os documentos que satisfazem a consulta submetida e do Modelo Vetorial (vide Seção 2.1.5) para então ordenar os resultados obtidos pela consulta. Maiores detalhes sobre o cálculo do referido coeficiente podem ser encontrados na documentação do *framework* SOLR<sup>23</sup>.

Enfim, o conjunto de documentos que satisfaz a consulta submetida será ordenado, em ordem decrescente, pelo valor de  $rgs$  de cada documento e retornado como resposta para o usuário.

#### 4.4. GEOPEN NEWS

Nesta dissertação, o foco será apenas em páginas Web de notícias. Atualmente, é perceptível um engajamento da grande mídia em disponibilizar mecanismos eficientes na busca de notícias. Em paralelo a isto, importantes ferramentas de redes sociais empregam esforços na adição de novas funcionalidades neste contexto. Por exemplo, recentemente, o Twitter lançou o aplicativo Twitter Moments<sup>24</sup>, um recurso que agrupa, de forma histórica,

<sup>23</sup> [https://lucene.apache.org/core/2\\_9\\_4/api/core/org/apache/lucene/search/Similarity.html](https://lucene.apache.org/core/2_9_4/api/core/org/apache/lucene/search/Similarity.html)

<sup>24</sup> <https://about.twitter.com/moments>

Figura 18: Arquitetura do GeoSEn News.



Fonte: elaborado pelo autor.

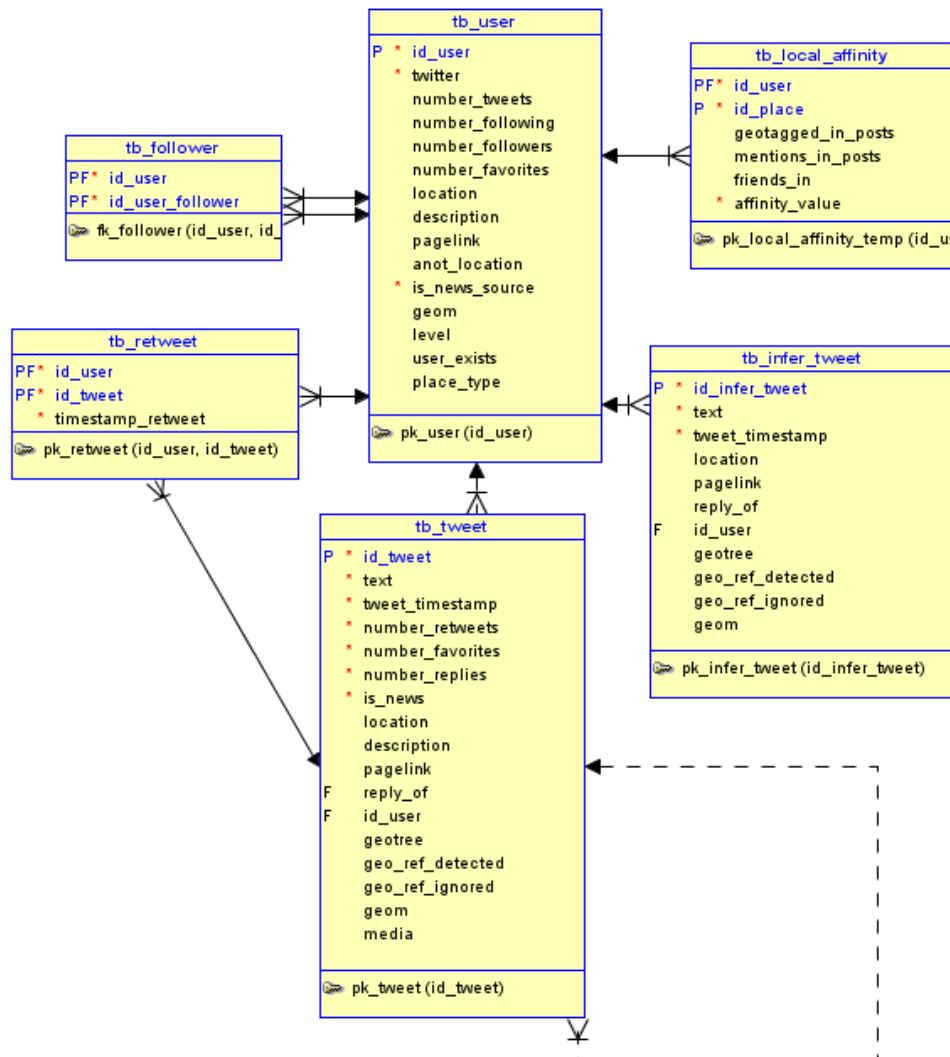
notícias sobre assuntos relevantes publicadas em seu microblog. Similarmente, a plataforma Facebook Notify<sup>25</sup> foi lançada, uma ferramenta com foco em notícias integrada à rede social.

Perante este cenário de engajamento na construção de plataformas que viabilizam a busca e o encontro de notícias informativas, e entendendo que pouco se tem feito na construção dessas plataformas com enfoque geográfico, o motor de busca GeoSEn foi estendido para produzir um ambiente capaz de buscar, encontrar e ordenar, de forma eficiente, notícias distribuídas em diversas mídias jornalísticas, além de permitir a consulta em uma perspectiva geográfica, pouco explorada nas demais ferramentas existentes. GeoSEn News foi o nome dado a nova plataforma de leitura de notícias informativas.

O GeoSEn News foi construído possuindo como estrutura o GeoSEn. Na Figura 18 é ilustrado uma visão macro da arquitetura do sistema. A aplicação foi projetada seguindo o modelo 3-camadas, objetivando facilitar a sua portabilidade, a atualização dos dados e a modularização total do sistema. As três camadas que compõem o sistema são: Camada de aplicação, que é responsável por toda manipulação gráfica da interface; a Camada de Negócio, camada esta que compõe o *crawler* de notícias, a extensão do GeoSEn e REST Web Service; e a Camada de Dados, responsável pelo controle de transações entre o Web Service e os dados. A seguir, as três camadas do GeoSEn News serão detalhadas de modo a auxiliar no seu entendimento.

<sup>25</sup> <http://newsroom.fb.com/news/2015/11/introducing-notify-a-notifications-app-from-facebook/>

Figura 19: Partição do diagrama ER do banco do GeoSEn News.



Fonte: elaborado pelo autor.

#### 4.4.1. Camada de Dados

A camada de dados é composta por um banco de dados objeto relacional com suporte espacial (PostgreSQL/Postgis) e o framework Apache Solr<sup>26</sup>. É de responsabilidade desta camada o acesso, gestão, armazenamento das informações e prover mecanismos de acesso aos dados utilizados na aplicação.

Como havia a necessidade de manipular dados provenientes do Twitter, o esquema do banco de dados do GeoSEn foi melhorado, passando a agregar este tipo de dado. Assim sendo, foi possível permitir que as informações sobre contas oficiais de veículos de comunicação, *tweets*, *retweets*, perfil de usuários, dentre outros, fossem armazenadas na base

<sup>26</sup> <http://lucene.apache.org/solr/>

de dados. Este segue o mesmo relacionamento encontrado no Twitter: um usuário possui vários *tweets* e vários relacionamentos de amizade; um *tweet* possui uma lista de usuários que realizaram *retweet*; e, para o funcionamento do *ranking* de relevância geosocial, um usuário passa a ter uma lista de localidades que o mesmo possui afinidade. O esquema que permitiu esta modelagem pode ser visualizado na Figura 19.

Na perspectiva de melhorias no desempenho da ferramenta e permitir o acesso por outras fontes, o método de armazenamento dos documentos indexados foi modificado. Anteriormente, o GeoSEn fazia uso de um método de indexação nativo do Lucene, o que inviabilizava seu acesso por outras ferramentas que não fosse do próprio Nutch. No entanto, vislumbrando tal melhoria, optou-se por indexar as páginas Web utilizando uma plataforma de busca textual, *open-source*, chamada Apache Solr. O Solr, como é comumente chamado, inclui como funcionalidade a consulta textual, destaque de resultados, agrupamento dinâmico e diversas outras funções que tornam a plataforma bastante popular para esta finalidade.

A escolha do Solr abriu caminho para a construção de um serviço central de notícias distribuídas, na forma de um Web Service REST, que pode ser utilizado por qualquer veículo de comunicação, portal de notícias ou *website* interessado em expor notícias da sua região de abrangência, valorizando a agilidade e eficácia que a plataforma pode oferecer e a diversidade de consultas em um contexto geográfico disponível no GeoSEn. Este serviço é detalhado na Seção 4.4.2.

#### **4.4.2. Camada de Negócio**

A camada de negócio foi construída de forma a permitir que vários sistemas possam consumir informação de maneira simples (i.e. por meio de uma URL), utilizando o estilo de arquitetura REST, largamente difundido em Web Services na Internet. Ademais, nesta camada estão presentes o módulo de captura de notícias provenientes do Twitter (vide Seção 4) e o núcleo do motor de busca GeoSEn.

Figura 20: Exemplo da adição de outro valor ao índice espacial.

Espacial

Índice Espacial	Relevância Espacial	Relevância Geosocial
51	0.01	0.01
414	0.63	0.29
312	0.12	0.10
2153	0.21	0.23
2154	0.22	0.13
15236	0.81	0.53
15228	0.86	0.00
15365	0.92	0.81

GeoSEn     GeoSEn News

Fonte: elaborado pelo autor.

A construção do GeoSEn News implicou em mudanças em alguns módulos do núcleo do GeoSEn. Nestas modificações, o objetivo foi permitir a adição da técnica de relevância geosocial no *ranking* de relevância da ferramenta. A primeira modificação ficou por conta do módulo de Modelagem do Escopo Geográfico, descrito na Seção 2.4.2, onde se é calculado a relevância espacial entre a página indexada e as localidades encontradas em seu conteúdo. Além deste, o cálculo do fator de credibilidade da página (i.e. relevância geosocial) também é realizado (Equação 5) neste momento.

A outra modificação necessária foi realizada no módulo de Indexação Espaço-Textual, detalhada na Seção 2.4.3. A abordagem adotada pelo GeoSEn para o processo de indexação espacial foi mantida, onde é feito uso do resultado da expansão do georreferenciamento. Assim, cada localidade referenciada de forma direta ou indireta pelo documento possui uma entrada independente do índice, associada a sua respectiva relevância geográfica. No GeoSEn News, o valor de relevância geosocial será uma nova entrada associada à localidade, portanto, ao invés de apenas uma, serão dois valores atribuídos à cada localidade no documento: relevância espacial e relevância geosocial, como pode ser visto na Figura 20.

No anseio de proporcionar um ambiente para leitura de notícias informativas, eficiente, escalável e reusável, um serviço REST (i.e. RESTful Web Service) foi construído. Este tipo de serviço REST (do inglês, *Representational State Transfer*) (Richardson e Ruby, 2007) é um estilo de arquitetura com restrições específicas, interface de acesso uniforme, que permite alta performance e escalabilidade. Neste serviço os recursos são acessados usando URIs (do inglês, *Uniform Resource Identifiers*), tipicamente *links* para Web. O exemplo de

Código 5: Exemplo de uma requisição ao Web Service REST.

```
http://[url]/rest/news.php?func=findNews&textQuery='acidente
automotivo' &isLocated=true&bbox='-48.196095 -15.498145, -47.312694 -
16.044810'
```

Fonte: elaborado pelo autor.

uma URI para busca por notícias sobre “acidente automotivo” no Distrito Federal pode ser visto no Código 5.

O serviço REST foi construído usando a linguagem de programação PHP, o que permite ser adicionado em qualquer servidor de aplicação que ofereça suporte à linguagem. O mesmo faz uso de uma biblioteca para acesso aos documentos no Solr, o *Solr PHP*<sup>27</sup>, fornecida pela *PECL Extensions*<sup>28</sup>.

#### 4.4.3. Camada de Aplicação

A camada de aplicação é responsável pelo domínio gráfico da ferramenta de busca com enfoque em notícias informativas. Nesta, estão dispostos todos os componentes gráficos de mapa e visualização de dados, bem como mecanismos que facilitam a busca e a visualização de notícias, respeitando as diretrizes de responsividade.

A interface gráfica do GeoSEn News foi construída utilizando AngularJS<sup>29</sup>, um *framework* Javascript que estende atributos HTML com diretivas e manipula dados de forma simplificada e ágil. Para permitir a responsividade e prover um visual mais amigável ao usuário final, a interface gráfica faz uso de um dos projetos *open source* mais popular da Web, o Bootstrap<sup>30</sup>.

Algo que foi incorporado ao GeoSEn News e que pouco é explorado em motores de busca geográfica, é a possibilidade de perceber, de forma paralela, as localidades relatadas na página desejada. Esta funcionalidade permite que o usuário perceba, por meio de um mapa interativo, no momento em que visualiza a notícia, se aquela relata acontecimentos de sua área de interesse. Para implementação do mapa interativo, foi utilizado o *framework* OpenLayers versão 2.0. Na Figura 21 é ilustrada a interface gráfica do GeoSEn News.

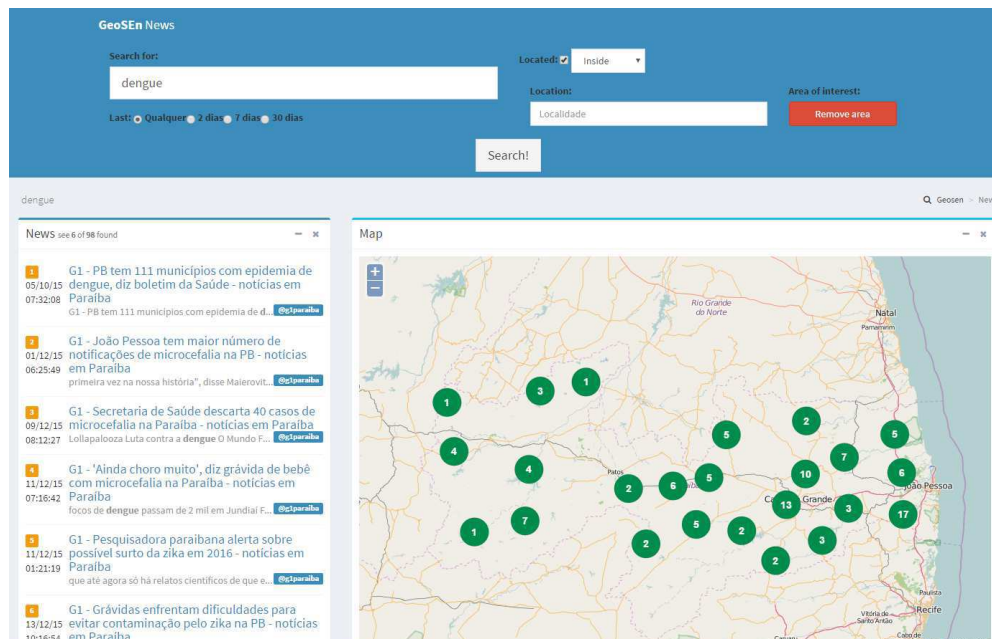
<sup>27</sup> <http://php.net/manual/en/book.solr.php>

<sup>28</sup> <https://pecl.php.net/package/solr>

<sup>29</sup> <https://angularjs.org/>

<sup>30</sup> <http://getbootstrap.com/>

Figura 21: Interface gráfica do GeoSEn News.



Fonte: *print screen* da ferramenta GeoSEn News.

Dentre as contribuições da interface gráfica, destacam-se:

**Mecanismo de consulta:** meio que fornece de forma simplificada a possibilidade de realizar consultas no âmbito textual, espacial e temporal. Em uma consulta textual, um campo é disponibilizado semelhante ao que é feito em motores de busca tradicionais. No domínio espacial, é possível realizar consultas por documentos cujos contextos geográficos estão localizados ou não (i.e. operador de negação) dentro, adjacentes ou em um raio de uma determinada localização selecionada. Esta localização pode ser determinada através de um campo auto completar (no formulário) ou através de uma ferramenta de desenho de área de interesse (i.e. *bounding box*) no próprio mapa. Na Figura 22 é apresentada o formulário de consulta.

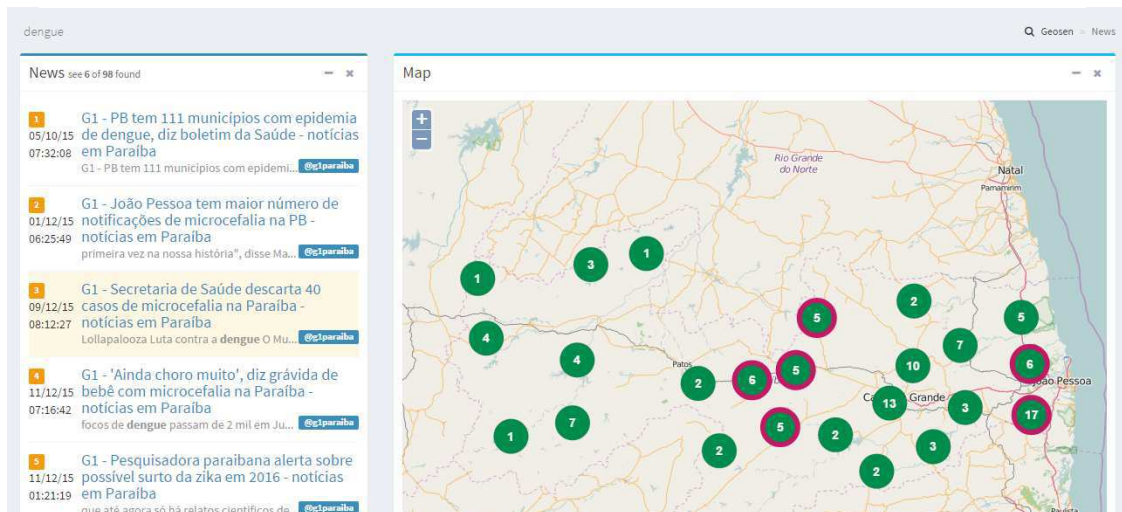
**Visualização do resultado:** a visualização do resultado da consulta é possível em uma perspectiva integrada multi-modo: listagem e mapa interativo. A listagem, tradicional em motores de busca, permite a visualização das páginas que correspondem àquela consulta de

Figura 22: Formulário de consulta do GeoSEn News.



Fonte: *print screen* da ferramenta GeoSEn News.

Figura 23: Seleção da terceira notícia e o destaque das localidades.

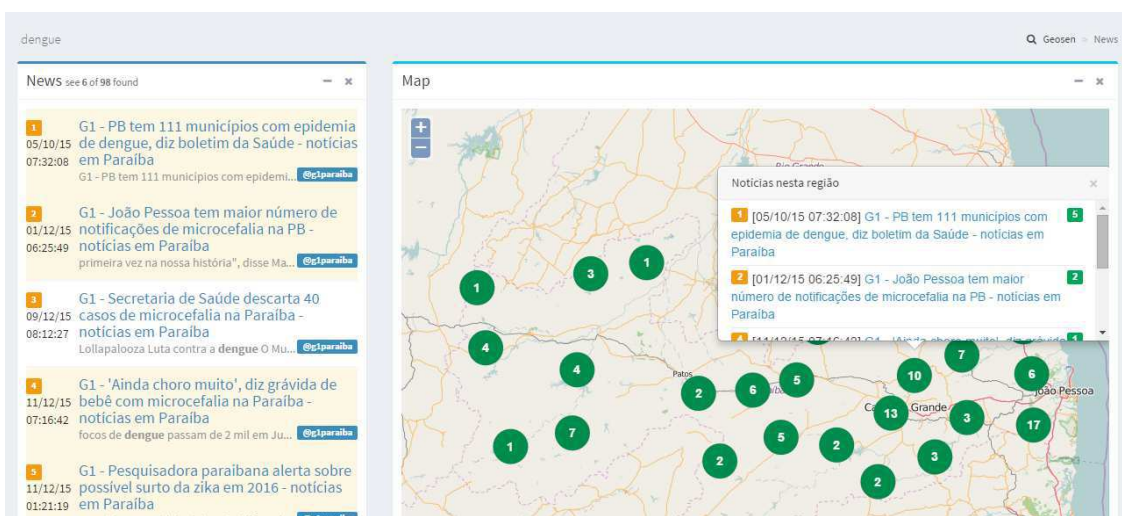


Fonte: *print screen* da ferramenta GeoSEn News.

forma paginada e com sua posição no ranking explícita. O mapa interativo permite o contato direto do usuário com a notícia em um Sistema de Informação Geográfica (SIG). Estes dois ambientes de visualização trabalham de forma unificada, pois, no momento que o usuário aponta seu cursor do *mouse* para uma notícia, as localizações encontradas na mesma são destacadas no mapa interativo. O inverso também acontece. Quando o usuário seleciona uma localidade no mapa, imediatamente, as notícias que se referem àquele lugar são destacadas, como mostra na Figura 23.

**Agrupamento:** o mapa interativo apresenta todas as localidades mencionadas nas notícias retornadas para uma determinada consulta. No entanto, esse número pode ser demasiadamente grande e/ou o usuário optar por visualizar o mapa em um *zoom* mais

Figura 24: Exibição do cluster com infowindow de notícias.



Fonte: *print screen* da ferramenta GeoSEn News.



afastado da região de interesse. Neste caso, a poluição visual é um problema e o mesmo foi sanado implantando a técnica de agrupamento (do inglês, *clustering*) de localidades. Assim, independente do número de localidades reportadas nas notícias ou o nível de *zoom* que o usuário opte, o mapa interativo se adapta agrupando os marcadores próximos, como pode ser visto na Figura 24. Ademais, quando o usuário clica em um grupo (*cluster*), uma janela de informações é aberta com informações sobre as notícias naquela região, suas respectivas posições no *ranking* bem como a quantidade de lugares reportados naquele grupo por cada notícia.

**Responsividade:** preocupação em toda plataforma Web atual, a responsividade é um fator crucial na adoção da ferramenta por usuários finais. Por tal, o GeoSen News é preparado para se adequar à qualquer tamanho de tela, como, por exemplo, *smartphones* e *tablets*. Na Figura 25 é ilustrado o exemplo da ferramenta sendo acessada por diversos meios.

#### 4.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo, descreveu-se toda a metodologia, as técnicas e as funcionalidades do motor de busca geográfica com enfoque em notícias proposto nesta dissertação. Vale destacar que a ideia consiste em gerar e implantar, em uma plataforma de leitura de notícias, um novo método de *ranking* de relevância que considera informações provenientes de redes sociais. O Twitter foi usado para este trabalho, porém outras redes sociais podem ser empregadas, necessitando apenas de alguns ajustes na ferramenta, conforme discutido anteriormente.

Em todas as etapas do desenvolvimento, alguns desafios foram enfrentados para tornar a ideia factível. O primeiro desafio estava relacionado com o modo de aquisição das

Figura 25: Interface gráfica em diferentes dispositivos.



Fonte: montagem elaborada pelo autor.

notícias distribuídas em diversos portais e mantidas por várias mídias informativas. A solução encontrada foi desenvolver uma ferramenta capaz de coletar notícias, de fontes distribuídas, diretamente na rede social, e informações sobre usuários que interagiram com as mesmas, já antecipando a necessidade dessa informação no momento do cálculo do *Ranking* de Relevância Geosocial. Outro desafio transposto durante o trabalho condiz com a ausência de informações sobre a localidade de moradia de usuários de redes sociais. Para tal, um algoritmo que estima a afinidade entre localidades e usuário da rede social, baseando-se em sua interação e relacionamentos de amizade na rede, foi produzido. Este algoritmo tornou possível a participação de mais usuários de redes sociais no cálculo do fator de credibilidade das notícias, tornando viável o *Ranking* de Relevância Geosocial. Por fim, usuários de motores de busca geográfica exigem respostas relevantes e ágeis. Por esse motivo, a indexação dos documentos foi feita no *Solr* e um *Webservice* REST foi produzido para consumo destes dados, tornando as respostas ágeis, bem como viabilizando portabilidade entre aplicações que desejam fazer uso dos mesmos.

No próximo capítulo são expostas as avaliações experimentais utilizadas, em detalhes, e os resultados alcançados com o protótipo apresentado.

## CAPÍTULO 5

### AVALIAÇÃO EXPERIMENTAL

Neste capítulo, são apresentados os métodos avaliativos utilizados para validar experimentalmente as contribuições propostas nesta dissertação, descritas no Capítulo 4. Inicialmente, um experimento empírico é realizado para avaliar o grau de eficácia do algoritmo em estimar a afinidade local dos usuários do microblog Twitter. No segundo experimento, almeja-se avaliar a sensibilidade do método de *ranking* de relevância geográfico, no contexto de notícias informativas, através de um *survey* realizado com voluntários de diferentes perfis. O terceiro e último experimento busca examinar a facilidade e usabilidade da ferramenta de busca geográfica com foco em notícias informativas, chamada GeoSEn News, por meio de um estudo de caso.

O restante do capítulo está organizado da seguinte forma: na Seção 5.1 são descritos os dados utilizados nos experimentos, suas características e deficiências; na Seção 5.2, descreve-se o arcabouço utilizado na execução do experimento empírico para avaliar o grau de eficiência do algoritmo de afinidade local; na Seção 5.3 é detalhado o *survey* de avaliação do método de *ranking* de relevância geográfico no contexto de notícias; na Seção 5.4, apresenta-se o estudo de caso para avaliação da usabilidade do GeoSEn News; e, por fim, na Seção 5.5 são apresentadas as discussões e as considerações finais sobre o capítulo.

#### 5.1. CONJUNTO DE DADOS

Para a execução dos experimentos propostos neste trabalho, um conjunto de contas oficiais no Twitter, associadas a fontes de notícias brasileiras, foram selecionadas manualmente. Estas contas oficiais serviram como ponto de partida para coleta de notícias, como foi descrito no Capítulo 4.

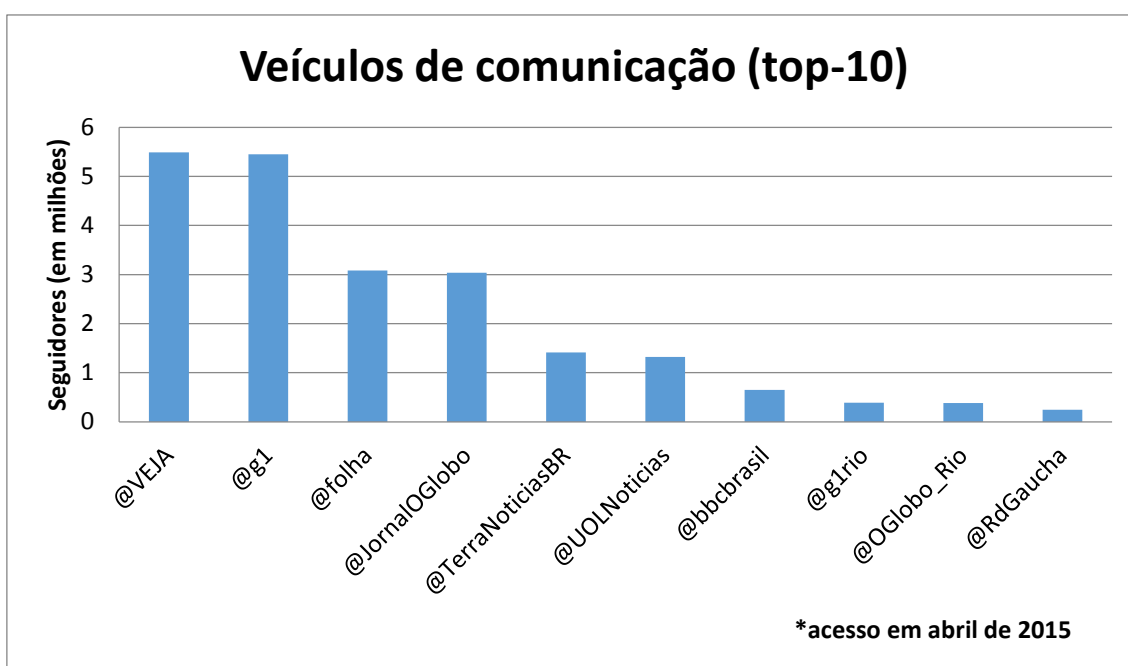
Este conjunto compreende 59 contas oficiais, sendo 18 da região Nordeste, 12 da região Norte, 6 da região Centro-Oeste, 8 da região Sudeste, 8 da região Sul e 7 nacionais, ou seja, abrangem informações de todo o Brasil. Todos estes veículos de comunicação foram geograficamente anotados de forma manual, no nível hierárquico de estado ou país. Dentre as contas, apenas 33 (56%) identificam a localização de sua sede no perfil, ou seja, informam no campo “*location*”, presente no perfil, a localização de onde sua sede é estabelecida.

Este conjunto foi utilizado como entrada para o Coletor de Notícias (vide Seção 4.1). Desta forma, foram coletados 26.160 *posts* no Twitter. Deste conjunto, 25.092 (95,91%) seguia o formato “*manchete link*”, caracterizando uma notícia inserida na rede social. Destes

25.092 *posts*, apenas 398 (0,01%) foram georreferenciados, ou seja, possuem em seus metadados uma localização geográfica.

Os *posts* que não seguem o padrão “manchete *link*” foram desconsiderados. Dos *posts* que representam notícias, 16.640 (66%) são oriundos de veículos de comunicação do grupo G1<sup>31</sup> e 16.811 (67%) dos *posts* possui ao menos um *retweet*. Ainda, se for considerado apenas os 10 veículos de comunicação (Figura 26) que possuem mais seguidores, esse número atinge 99%, ou seja, dos 11.990 (47,7%) *posts* gerados pelo top-10, 11.877 têm ao menos um *retweet*.

Figura 26: Top-10 veículos de comunicação com mais seguidores.

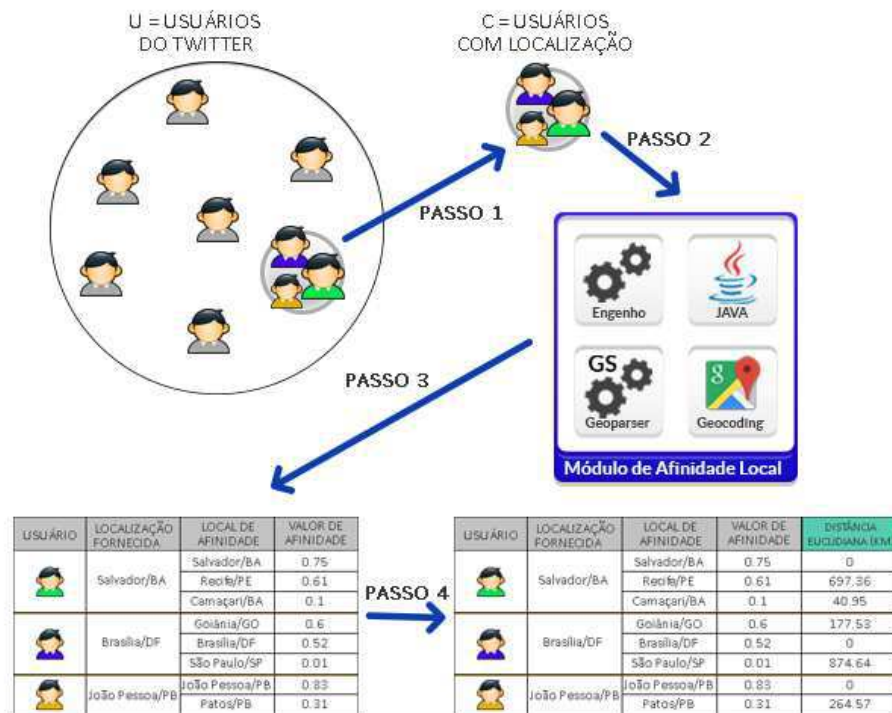


Fonte: elaborado pelo autor.

Do conjunto de *posts* que representam notícias, foram coletados 272.268 *retweets*. No entanto, conseguiu-se geolocalizar através do campo “*location*” (presente no perfil) apenas 19.951 dos usuários que os realizaram (7,33%).

<sup>31</sup> <http://g1.globo.com/>

Figura 27: Montagem do ambiente de experimentação.



Fonte: elaborado pelo autor.

## 5.2. EXPERIMENTO 1: AFINIDADE LOCAL

O objetivo deste experimento é avaliar o algoritmo de estimativa de afinidade local, proposto neste trabalho, com o intuito de valorar sua eficácia em diversos cenários e permitir perceber sua sensibilidade em afirmar as relações entre usuário e localidades, utilizando como fonte de informação dados oriundos de rede social.

A eficácia do algoritmo é avaliada verificando se a localização fornecida pelo usuário em seu perfil está presente na lista de localidade que o mesmo possui afinidade (estas identificadas pelo algoritmo de afinidade local). Desta forma, para ser possível realizar tal experimento, apenas usuários que forneceram este atributo em seu perfil e que foi possível geocodificá-lo para o Brasil (i.e., a localidade informada foi identificada como pertencente ao território brasileiro) podem participar do experimento.

Formalmente, tem-se:

$U = \text{conjunto de usuários do Twitter.}$

$L = \text{localidades do Brasil.}$

$C = \{u \mid u \in U \wedge u.\text{location} \in L\}.$

Cada usuário contido em  $C$  é submetido ao algoritmo de estimativa de afinidade local com os pesos de  $pmp$ ,  $plp$  e  $pfi$  atribuídos para 0,33, 0,34 e 0,33, respectivamente. O resultado de cada uma das execuções é uma lista ( $AL$ ) contendo uma quádrupla com o usuário, a localização de afinidade, o valor de afinidade e a distância entre a localidade do usuário e a localidade identificada na afinidade.

Formalmente, tem-se:

$$AL = \{(u, l, v, d) \mid u \in U \wedge l \in L \wedge v, d \in R+\}$$

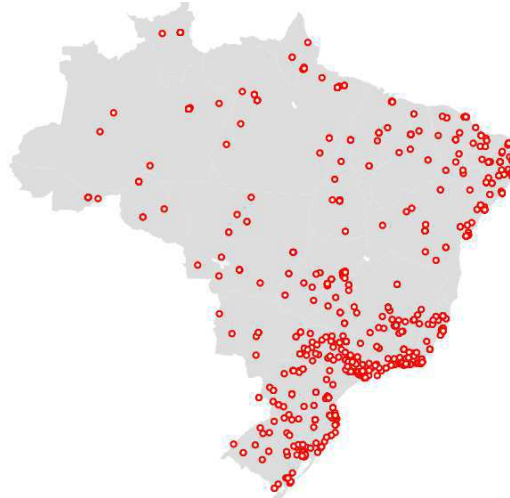
Onde  $d$  é a distância euclidiana entre  $u.location$  e  $l$ .

A distância encontrada entre a localização fornecida pelo usuário e as localidades encontradas pelo algoritmo serve de limiar para afirmar se houve ou não sucesso na identificação das localidades por parte do algoritmo. Na Figura 27 é exibido o fluxo da avaliação do algoritmo. No passo 1, do conjunto de usuários coletados, são escolhidos apenas usuários que possuem localização no seu perfil e que foi possível geolocalizar no Brasil. No passo 2, estes usuários são submetidos ao algoritmo de afinidade local. Por fim, no passo 3 e 4, são calculadas as distâncias euclidianas entre a localização informada no perfil do usuário e as localidades retornadas pelo algoritmo.

### 5.2.1. Dados do Experimento

Para execução do experimento, foram coletados, de forma aleatória, os perfis de 1.090 usuários do microblog Twitter, todos seguindo os requisitos mencionados, como visto na Figura 28. Cada um dos perfis foi submetido ao algoritmo de estimativa de afinidade local, que analisou 200 *tweets* e 200 relacionamentos de amigos (de cada usuário), resultando em 17.575 registros de afinidade no total, sendo 12.514 identificadas de forma direta e 5.061 de forma indireta (média de 16,12 por usuário). A Figura 29 apresenta o gráfico boxplot e o sumário dos dados.

Figura 28: Distribuição espacial dos usuários participantes.



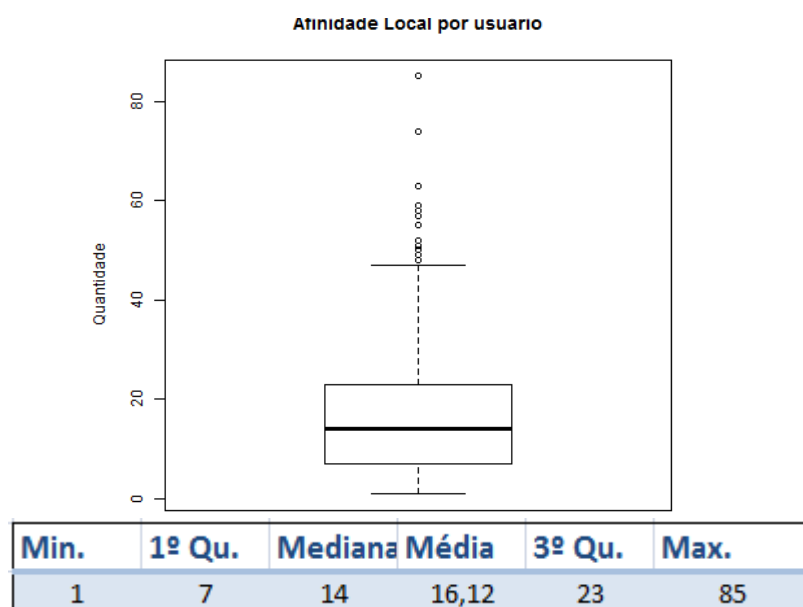
Fonte: elaborado pelo autor.

### 5.2.2. Projeto do Experimento

O experimento realizado no estudo é do tipo investigativo. Tal metodologia é utilizada quando se deseja identificar cenários e fatores ótimos, que contribuem diretamente na eficiência do objeto avaliado. A presença de fatores desconhecidos pode influenciar na variável resposta (taxa de acerto), porém, o estudo visa qualificar os fatores que apresentam maior influência positiva nos resultados (Wohlin et al., 2012).

Com o objetivo de visualizar a interação entre os fatores e a interferência que cada

Figura 29: Boxplot e sumário dos dados de afinidade local identificados.



Fonte: elaborado pelo autor.

um exerce sobre a variável resposta, o modelo utilizado foi o Fatorial  $N^k$  sem repetição, onde  $N$  é a quantidade de opções em cada fator e  $k$  o número de fatores presentes no experimento. Como o algoritmo fornece respostas para sua execução de forma determinística, realizar repetições para cada tratamento não iria contribuir para o experimento. Então, o experimento foi realizado com 9 ensaios totalizando 9 execuções.

Na avaliação do algoritmo de estimativa de afinidade local, uma métrica foi levantada em cada tratamento com intuito de descrever o grau de eficiência (taxa de acerto) em cada cenário experimentado. Para tal, dois fatores foram considerados para este experimento. O primeiro fator representa o número de elementos de afinidade identificados pelo algoritmo (i.e., top-k), ou seja, se o algoritmo identificou um conjunto de  $x$  localidades que o usuário possui afinidade, o *top-k* elementos é um subconjunto de tamanho  $k$ , ordenado pelo valor de afinidade (onde  $k, x \in \mathbb{N}^*$  e  $k \leq x$ ). Este fator compreende três valores: *top-1* (a localidade identificada que o usuário possui maior afinidade); *top-5* (as cinco localidades identificadas que o usuário possui mais afinidade); e *top-10* (as dez localidades identificadas que o usuário possui mais afinidade).

O segundo fator considerado representa um limiar de aceitação da distância entre a localidade identificada pelo algoritmo e a localidade fornecida no perfil do usuário. Por exemplo, considere que o algoritmo foi executado considerando um perfil de usuário residente na cidade de Catolé do Rocha/PB e o algoritmo identificou certa afinidade entre o usuário e a cidade de Brejo do Cruz/PB. Se o teste for realizado considerando o limiar de aceitação de 30 km, o algoritmo obteve sucesso na inferência, pois, a distância euclidiana entre as duas cidades é de 27,39 km e, portanto, abaixo do limiar estabelecido. Para esse fator, três níveis foram escolhidos: limiar de 0 km, que indica que a inferência foi realizada de forma precisa; 10 km, representando o sucesso do algoritmo quando a distância entre as localidades inferidas e a localidade do usuário não ultrapassa este valor; e 100 km, quando a distância entre as localidades inferidas e a localidade fornecida pelo usuário não ultrapassa este valor.

### 5.2.3. Resultados

O processo de execução do experimento consistiu em executar cada tratamento (i.e., combinação entre os fatores) e colher a taxa de sucesso que o algoritmo obteve em identificar as localidades que do usuário de forma correta. As taxas de acertos obtidas são apresentadas na Tabela 4 e no gráfico da Figura 30, sendo discutidas a seguir, na Seção 5.2.4.



Tabela 4: Resultados obtidos em cada tratamento executado.

Top 1			Top 5			Top 10		
0 km	10 km	100 km	0 km	10 km	100 km	0 km	10 km	100 km
57,80%	61,01%	66,79%	79,45%	80,55%	83,49%	83,49%	84,40%	86,70%

Fonte: elaborada pelo autor.

Figura 30: Gráfico com as taxas de acertos obtidas em cada tratamento.



Fonte: elaborado pelo autor.

#### 5.2.4. Discussão

Apesar da influência de todos os fatores na métrica observada, o fator que melhor pode explicar essa variação é o tamanho da lista de afinidades considerada. É possível perceber uma mudança brusca no momento em que o teste passa a considerar uma lista com tamanho 5 (top-5) ao invés de apenas um elemento (top-1). Isto se deve ao fato que o usuário pode ter mais afinidade com localidades onde ele residiu anteriormente do que com a localidade de moradia atual.

Foi possível perceber um bom resultado do algoritmo em seu caso crítico (57,8%), quando o mesmo deve acertar precisamente a localidade de moradia do usuário (tratamento 1: top-1 com limiar 0km). Alguns algoritmos encontrados na literatura e descritos na Seção 3.1 obtêm resultados próximos, mas considerando um limiar entre 10 e 100 km, tornando o algoritmo proposto neste trabalho superior em tais aspectos. Outro destaque perceptível nos resultados é a relação entre a precisão do algoritmo e o resultado obtido (i.e., taxa de sucesso): à medida que a precisão vai sendo desprezada, a taxa de acerto do algoritmo cresce.

Ainda que os resultados tenham se apresentado de forma satisfatória, alguns fatores espúrios podem ter interferido negativamente na variável resposta como, por exemplo: a

participação de usuários que não possuem o hábito de atualizar seu perfil nas redes sociais (e.g., nível de escolaridade, assuntos de preferência, localidade de moradia); o fato que o módulo de identificação de referências geográficas (GeoSEn GeoParser) foi construído especialmente para análise de páginas Web, mas nesta dissertação foi utilizado também para identificação de referências geográficas em *tweets* (140 caracteres), com poucos ajustes; dentre outros.

De qualquer forma, algumas sofisticções podem ser realizadas com o objetivo de ampliar os resultados obtidos. A primeira delas consiste em amplificar as funções de identificação de relação entre usuário e localidades. Na versão atual do algoritmo, são consideradas as localidades encontradas nos *tweets* e no perfil dos amigos do usuário, porém, segundo Jurgens (2013), apenas 5% dos usuários preenchem essa informação em seu perfil (fator motivador para construção do algoritmo proposto), deste modo, aplicar o algoritmo de estimativa de afinidade local também no perfil dos amigos, pode propiciar melhorias nos resultados. Outra eventual melhoria para o algoritmo está no uso de um GeoParser especializado em analisar referências geográficas em pequenos textos, como, por exemplo, um *tweet*. Por fim, além de considerar os *tweets* do usuário na busca por referências geográficas, tentar coletar tais informações no conteúdo de páginas da Web (através *links* compartilhados por ele) pode ser um ato promissor na melhoria do algoritmo proposto.

Finalmente, cabe ao desenvolvedor (ou administrador do sistema) escolher a melhor configuração para diferentes cenários de aplicação. Se em um determinado cenário a ferramenta exige precisão na inferência com distâncias aceitáveis, pode-se usar a configuração *top-1* (i.e., apenas a localidade que o usuário tem mais afinidade é considerada) que obteve uma taxa de sucesso entre 57% e 67%. Caso a ferramenta seja aplicada em um cenário onde se busque perceber a afinidade de usuários em diferentes regiões geográficas, a configuração ideal assemelha-se a *top-5*, que obteve sucesso entre 79% e 84%. Eventualmente, se o objetivo for capturar todo o histórico de moradia do usuário, lugares onde já frequentou e/ou possui amigos e família, o desenvolvedor pode optar pelo *top-10* ou superior, que garante sucesso na estimativa entre 84% e 87%. O *ranking* de relevância geográfico com foco em notícias informativas, proposto nesta dissertação, fez uso da configuração *top-10* (i.e. as dez localidades que o usuário possui maior afinidade).

### 5.3. SURVEY 1: RANKING DE RELEVÂNCIA GEOGRÁFICO

O *survey* realizado neste trabalho é do tipo exploratório e tem como finalidade estimar o impacto da inserção do fator de relevância geosocial, proposto nesta dissertação, no

*ranking* de relevância geográfico implantado no GeoSEn News. Este fator adicional busca representar o grau de credibilidade que uma determinada notícia possui, mediante a avaliação da afinidade que usuários que a compartilharam em sua rede social possuem com as localidades mencionadas na notícia. Este tipo de investigação fornece embasamento científico para avaliar soluções e provê pistas para concepção de novas interações e trabalhos futuros.

Na literatura, inúmeros são os autores que adotam o *survey* exploratório na avaliação de suas soluções de *ranking* de relevância no âmbito geográfico. Outros tipos de abordagens de avaliação são encontrados, mas ocorrem quando há um aperfeiçoamento de uma técnica de *ranking* e, portanto, é possível verificar ganhos ou perdas apenas comparando os resultados obtidos nas duas versões. Neste caso, a inclusão de um novo fator no *ranking* de relevância geográfico do GeoSEn não se configura como um aperfeiçoamento ou melhoria do mesmo, pois o foco difere: o *ranking* de relevância do GeoSEn é mais abrangente em manipular documentos distribuídos em toda a Web, enquanto o *ranking* de relevância implantado no GeoSEn News manipula e ordena apenas notícias informativas.

### **5.3.1. Dados do Survey**

Dos 25.092 *tweets* coletados que seguiam o padrão “manchete *link*” (vide Seção 0), 21.135 (84%) representavam notícias informativas únicas e, portanto, foram indexadas para uso na avaliação do método de *ranking* de relevância geográfico. Deste total de *tweets*, foram capturados 230.450 *retweets*, realizados por 71.175 usuários distintos. Cada um destes perfis de usuários foi submetido ao algoritmo de estimativa de afinidade local, que serviu para mensurar o grau de credibilidade das notícias encontradas no sistema GeoSEn News. Dos perfis submetidos ao algoritmo de afinidade local, foi possível identificar ao menos um registro de afinidade em 19.829 usuários. O número baixo de usuários que o algoritmo conseguiu estimar alguma afinidade pode ser explicado pelo baixo número de *tweets* e relações de amizades coletadas (200 e 200, respectivamente); pela participação de perfis de estrangeiros; pela dificuldade do *geoparser* em trabalhar com pequenos textos; dentre outros.

As notícias indexadas para este *survey* foram reproduzidas por 59 mídias informativas populares espalhadas pelo Brasil, que publicam notícias sobre todo o território nacional e internacional. Notícias que envolvem apenas localidades internacionais foram desconsideradas neste estudo.

### 5.3.2. Projeto do Survey

Com o propósito de avaliar melhor todas as causas e reações dos voluntários nas respostas aos questionários, o modelo do *survey* escolhido foi o supervisionado, onde há o acompanhamento presencial do pesquisador no instante da resolução. Os questionários submetidos aos voluntários contam com duas questões, de baixa complexidade, a respeito do resultado apresentado pelo método de *ranking* de relevância geográfico em resposta a uma consulta pré-definida.

Para o *survey*, dez consultas, recorrentes em sistemas de GIR, foram pré-definidas pelos pesquisadores (Tabela 5). Essas consultas foram avaliadas por 36 voluntários de diferentes perfis, onde cada um dos participantes avaliou o resultado de 5 consultas escolhidas aleatoriamente dentre as 10 disponíveis. A atividade realizada pelos voluntários consistiu em duas tarefas. Na primeira, o voluntário deveria verificar a ordem das notícias retornadas e, caso a ordem não estivesse correta (na opinião do voluntário), ele deveria reordenar os itens apresentados pela ferramenta utilizando a interface gráfica (descrita na Seção 5.3.3). Na segunda tarefa, o participante deveria responder se o resultado (conjunto de itens) apresentado para aquela consulta era, em geral, “Pouquíssimo relevante”, “Pouco relevante”, “Relevante”, “Muito relevante” ou “Muitíssimo relevante”.

Tabela 5: Consultas em linguagem natural avaliadas pelos participantes.

Consultas
<b>Desemprego em um raio de 600 km do estado de Minas Gerais</b>
<b>Doença e epidemia em um raio de 600 km do estado de São Paulo</b>
<b>Concurso público nas adjacências do estado de Minas Gerais</b>
<b>Pesquisa e estudo em um raio de 500 km do estado de São Paulo</b>
<b>Assalto e roubo no estado do Rio Grande do Sul</b>
<b>Concurso público nas adjacências do estado do Ceará</b>
<b>Doença e epidemia nas adjacências do estado de São Paulo</b>
<b>Pesquisa e estudo em um raio de 600 km do estado do Ceará</b>
<b>Assalto e roubo no estado do Ceará</b>
<b>Concurso público nas adjacências do estado de São Paulo</b>

Fonte: elaborado pelo autor.

Os participantes do *survey* são pessoas de ambos os sexos, de idade que varia entre 18 e 51 anos, de profissões distintas (e.g., estudante, advogado, professor universitário,

bibliotecário) e que compreendem um subconjunto da população de interessados em aplicações desta natureza.

### 5.3.3. Coleta das Respostas

Com o intuito de auxiliar a resolução dos questionários, padronizar o formato das respostas e facilitar sua coleta, uma ferramenta de auxílio foi construída para este *survey*. Esta foi desenvolvida para Web e faz uso de tecnologias atuais (e.g., JSP, JSF, JavaScript) suportadas pelos principais *browsers* do mercado. Quando o participante do *survey* acessa o sistema para responder aos questionários a ele submetidos, uma página inicial é apresentada, contendo 7 principais elementos, exibidos em destaque na Figura 31.

Figura 31: Ferramenta de auxílio construída para coleta das respostas.

1) A ordem do mais relevante (topo) para o menos relevante está correta? Se não, reordene usando as setas. (ii)

Resultado da consulta: "Saúde no Nordeste" (iii)

G1 São Paulo: Taxa de homicídios por distrito de São Paulo  
<http://l1.co/MLhSpaITa>  
 G1 São Paulo: Taxa de homicídios por distrito de São Paulo Taxa de homicídios por distrito de São Paulo

G1 - Fílmoteca Acreana exibe mostra em homenagem ao Dia do Trabalho - notícias em Acre  
<http://l1.co/0LabWOC1Uo>  
 G1 - Fílmoteca Acreana exibe mostra em homenagem ao Dia do Trabalho - notícias em Acre MENU G1 Acre

(iv) G1 - Ex-agente penitenciário mata PM a tiros no bairro do Antares, em Maceió - notícias em Alagoas  
<http://l1.co/yMf0m6s1>  
 G1 - Ex-agente penitenciário mata PM a tiros no bairro do Antares, em Maceió - notícias em Alagoas

G1 - Câmara escondida flagra precariedade em unidade de saúde de Maragogi, AL - notícias em Alagoas  
<http://l1.co/EPFmMvWm>  
 G1 - Câmara escondida flagra precariedade em unidade de saúde de Maragogi, AL - notícias em Alagoas

2) Qual o grau de relevância do resultado? (v)

Pouquíssimo relevante  
 Pouco relevante  
 Relevante (vi)  
 Muito relevante  
 Multíssimo relevante

Avaliar (vii)

Fonte: *print screen* da ferramenta de auxílio.

Os elementos presentes são: (i) elemento que representa o total de avaliações feitas e a quantidade de restantes; (ii) descrição da primeira tarefa; (iii) consulta avaliada; (iv) a primeira questão em relação à ordenação do resultado; (v) a descrição da segunda tarefa; (vi) as opções da segunda questão respondida em relação à resposta fornecida pela ferramenta; (vii) botão para enviar a avaliação e partir para a avaliação seguinte, caso ainda reste alguma a ser feita;

Para responder a primeira questão, caso o participante não concorde com a ordem definida pelo *ranking* de relevância geográfico, ele deve utilizar as setas, localizadas no lado direito de cada item do resultado, para reordená-los. Na segunda pergunta, o usuário deve escolher uma das respostas apresentadas sobre o grau de relevância do resultado apresentado (de forma geral).

O mapa disposto ao lado direito da primeira questão serve para facilitar a percepção do usuário acerca das localidades identificadas nos documentos apresentados. Cada localização é exibida no mapa seguindo a mesma cor atribuída ao documento. Inicialmente, o mapa não apresenta nenhuma localidade, ficando a cargo do participante a escolha de quais documentos são exibidos no mapa, clicando no *checkbox* localizados à esquerda de cada um dos documentos listados. Na Figura 32 é exibido um exemplo da primeira questão sendo respondida.

Figura 32: Ordenação dos documentos sendo avaliada na primeira questão.

2 de 5 avaliações

1) A ordem do mais relevante (topo) para o menos relevante está correta? Se não, reordene usando as setas.

Resultado da consulta: **Acontecimentos no estado de São Paulo**

- G1 - Acordos entre domésticos e patrões impedem cumprimento de PEC no AC - notícias em Acre  
<http://t.co/xK9Pni3CvL>  
 G1 - Acordos entre domésticos e patrões impedem cumprimento de PEC no AC - notícias em Acre MENU G1
- G1 - 'Cultura foi bem representada', diz pajé Yawanawá sobre desfile no SPFW - notícias em Acre  
<http://t.co/qSLVhacSs>  
 G1 - 'Cultura foi bem representada', diz pajé Yawanawá sobre desfile no SPFW - notícias em Acre
- G1 São Paulo: Taxa de homicídios por distrito de São Paulo  
<http://t.co/MUvSgaNTJs>  
 G1 São Paulo: Taxa de homicídios por distrito de São Paulo Taxa de homicídios por distrito de São
- G1 - 'Desespero', diz haitiana após ser separada da filha na fronteira do AC - notícias em Acre  
<http://t.co/nafUT1QT5a>  
 G1 - 'Desespero', diz haitiana após ser separada da filha na fronteira do AC - notícias em Acre
- G1 - Empresa de ônibus dá desconto de até 20% para imigrantes saírem do AC - notícias em Acre  
<http://t.co/ZB2UBJ1Lv6>  
 G1 - Empresa de ônibus dá desconto de até 20% para imigrantes saírem do AC - notícias em Acre MENU
- G1 - Homem furta casa, é agredido pela população e acaba preso em Maceió - notícias em Alagoas  
<http://t.co/j61W9ym929>  
 G1 - Homem furta casa, é agredido pela população e acaba preso em Maceió - notícias em Alagoas MENU

Mapa do Brasil com marcadores coloridos em São Paulo, Maceió e outras localidades.

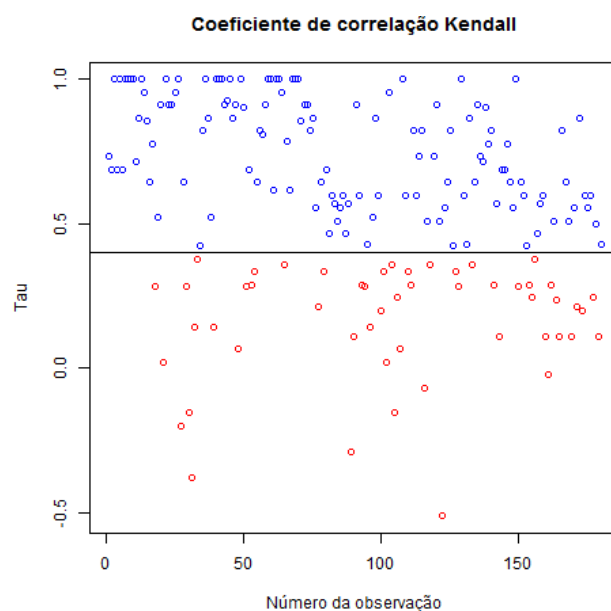
Fonte: *print screen* da ferramenta de auxílio.

### 5.3.4. Resultados

Para a primeira questão do formulário, onde o voluntário deveria reordenar o resultado retornado pela ferramenta quando julgasse necessário, optou-se por utilizar uma medida estatística de correlação de *ranking*. Esta medida busca representar a associação de ordem entre diferentes variáveis ordinais ou diferentes *rankings* de mesma variável em diferentes observações desta. Este coeficiente de correlação de *ranking* mede o grau de similaridade entre dois *rankings* e pode ser usado para obter o grau de significância da relação entre eles. Existem vários métodos não paramétricos de significância que usam a correlação de *ranking*, como, por exemplo, Wilcoxon signed-rank test (Wilcoxon, 1945) e o Kendall's tau correlation coefficient (Kendall, 1970). Este último foi escolhido para validação estatística deste questionário.

O coeficiente de correlação de Kendall é usado como um teste estatístico para estabelecer quando a ordem de duas variáveis pode ser estatisticamente dependente. A hipótese nula ( $t_0$ ) afirma que há independência entre as variáveis enquanto a hipótese alternativa ( $t_1$ ) assegura a dependência entre elas. Deste modo, para verificar se há similaridade entre a ordem fornecida pelo algoritmo e a ordem ótima apontada pelo voluntário, as 180 avaliações de *ranking* foram submetidas ao método de coeficiente de correlação de Kendall. O resultado pode ser apreciado no gráfico da Figura 33, onde os pontos em azul indicam que houve correlação moderada ou muito forte e os pontos em vermelho indicam a existência de uma correlação baixa ou muito fraca.

Figura 33: Gráfico de dispersão do coeficiente de correlação (tau) de todas as avaliações.



Fonte: elaborado pelo autor.

Já no segundo quesito do formulário, o voluntário deveria apontar o grau de relevância do resultado apresentado pelo método de *ranking* de relevância proposto. Ele deveria escolher entre 5 opções (escala Likert) o quão relevante é aquele resultado à consulta definida. Assim, as 180 respostas obtidas para esse quesito foram avaliadas seguindo o método estatístico chamado de Teste de Proporção. Deseja-se poder estatístico para assegurar que mais de 80% dos voluntários afirmou que o resultado retornado pelo método de *ranking* está entre “Relevante” e “Muitíssimo Relevante”.

- Teste de proporção:
  - **Questão:** Há evidências estatísticas para afirmar que 80% da população considerou o resultado “Relevante”, “Muito Relevante” ou “Muitíssimo Relevante”?

- **Hipótese nula:** Menos de 80% considerou o resultado “Relevante”, “Muito Relevante” ou “Muitíssimo Relevante”?
- **Hipótese alternativa:** 80% ou mais da população considerou “Relevante”, “Muito Relevante” ou “Muitíssimo Relevante”;

O teste foi executado na ferramenta R. A saída da execução pode ser vista na Figura 34.

Figura 34: Resultado do teste de proporção retornado pela ferramenta R.

```
> prop.test(148,180,alternative = "greater", conf.level = .95)

1-sample proportions test with continuity correction

data: 148 out of 180, null probability 0.5
x-squared = 73.472, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.767668 1.000000
sample estimates:
      p
0.8222222
```

Fonte: elaborado pelo autor.

### 5.3.5. Discussão

Os resultados obtidos nas avaliações estatísticas foram conclusivos e satisfatórios. De uma maneira geral, os voluntários foram orientados a considerar três elementos no momento de avaliar o *ranking* provido pelo algoritmo: 1) se os documentos estavam de acordo com a consulta textual. Por exemplo, se a consulta fosse por “concurso público”, qualquer documento que contivesse ao menos uma destas palavras seria considerado válido; 2) se alguma referência geográfica encontrada no documento pertencesse à área buscada; e 3) se a fonte de notícia tinha, de fato, propriedade para falar daquele lugar. Foi pedido para os voluntários uma atenção especial ao terceiro elemento por representar a fonte de estudo desta pesquisa.

Na primeira questão do formulário, 10 documentos eram apresentados para o usuário como resposta à consulta definida. Tais documentos foram ordenados pelo algoritmo proposto, porém, não representavam os 10 primeiros resultados. Na verdade, estes documentos representavam os 5 primeiros resultados, seguido pelo 10º, 15º, 20º, 25º e 30º documento da ordenação. Este particionamento buscava identificar problemas mais críticos do algoritmo de *ranking* de relevância, onde, por exemplo, o 30º documento seria posto, pelo avaliador, nas primeiras posições.



Assim, na primeira questão, onde o voluntário deveria reordenar o resultado da consulta caso julgasse necessário, foi possível perceber que a maior parte das avaliações pouco modificou o resultado obtido pelo algoritmo e, quando o fez, ocorreram em casos de baixo impacto (i.e., quando ordenação ocorreu entre os 4 primeiros resultados). Das 180 avaliações, 128 (71%) obtiveram o coeficiente de correlação superior a 0.4 (i.e., correlação entre moderada e muito forte). As causas das demais avaliações não alcançarem este valor pode ser explicada por diversos motivos: dificuldades encontradas no momento da identificação de referências geográficas nos documentos indexados (resolução de topônimos); modelagem imprecisa do escopo geográfico; poluição de informações nas páginas de notícias (e.g., quando dados são exibidos de outras notícias ou notas informativas); dentre outros. A interpretação equivocada dos três elementos supracitados (i.e., relevância textual, geográfica e de credibilidade) pode ter sido motivo de mudanças mais rigorosas (i.e., quando os últimos documentos foram postos nos primeiros resultados).

Deve-se ressaltar que, em avaliações desta natureza, há sempre algum viés do voluntário no momento de responder os questionários. Infelizmente, o voluntário não conhece a veracidade dos fatos (i.e. notícias) relatados e acaba, em alguns casos, avaliando positivamente uma notícia que não possui credibilidade alguma. Para minimizar esse viés, apenas notícias de grandes veículos de comunicação do país foram utilizadas neste *survey*.

Logo, para as avaliações da segunda questão do formulário, o teste de proporção de Wilcoxon pôde confirmar, com confiança de 5% ( $p\text{-value} = 2.2e^{-16}$ ), que mais de 80% das avaliações consideraram o resultado apresentado pelo *ranking* como “Relevante”. Apesar de satisfatório, este número também pode ter sido influenciado negativamente por falhas em processos anteriores, como, por exemplo, resolução de topônimos, modelagem do escopo geográfico e vários outros. Assim, acredita-se que a evolução de outros componentes do sistema pode melhorar substancialmente os resultados obtidos pelo *ranking* de relevância proposto neste trabalho.

#### 5.4. ESTUDO DE CASO 1: GEOTEN NEWS

A avaliação do ambiente de consulta e leitura de notícias informativas, em um âmbito geográfico, que envolve dados coletados de redes sociais para estimar um grau de confiabilidade da notícia e assim usá-lo no *ranking* de relevância, foi conduzida experimentalmente por um estudo de caso confirmatório. Apesar do estudo de caso não estabelecer um modelo experimental rigoroso, este tipo de abordagem é útil para confirmar algumas hipóteses primárias baseada na redução do conjunto de elementos, indicando

algumas evidências de eficácia da ferramenta avaliada, bem como apontar algumas possíveis melhorias que podem ser feitas.

Deste modo, tem-se como questão de pesquisa: *como a aplicação avaliada se comporta, em termos de facilidade, agilidade, usabilidade e resposta, em consultas genéricas na procura por notícias informativas?*

#### **5.4.1. Conjunto de Dados**

Os dados utilizados em todos os experimentos presentes nesta seção e descritos em detalhes na Seção 5.1 foram reduzidos a fim de produzir um ambiente experimental mais controlado. Para isto, um cenário temático foi fixado, de modo que, apenas notícias que continham pelo menos uma das palavras “dengue”, “zika” ou “microcefalia” participaram do estudo de caso. Atualmente, o Brasil se depara com uma crise de saúde pública, caracterizada pelo crescimento alarmante de recém-nascidos com condições neurológicas raras em que a cabeça da criança e cérebro são significativamente menores que o habitual. Esta condição é conhecida como microcefalia e, recentemente, despertou a atenção da ciência por uma possível correlação entre a condição e o zika vírus. O zika vírus é transmitido pelo mosquito *Aedes aegypti*, responsável também pela transmissão da dengue e chikungunya.

Os casos de microcefalia começaram a ser descobertos na região do Nordeste do Brasil, mas ganharam grande repercussão em todo o território nacional, sendo notificados por diversas mídias espalhadas pelo país. Assim, acredita-se que perceber a eficiência do fator nesse cenário crítico, onde diversos meios de comunicação distribuídos por todo o país escrevem sobre uma determinada região longe de sua área de abrangência, configura-se uma oportunidade de investigação em um contexto real.

Ao todo, foram coletados 959 *tweets*, publicados pelos veículos de comunicação supracitados (vide Seção 0). Destes, 824 *links* (*tweets* que seguiam o padrão “manchete *link*”) e seus respectivos *retweets* foram passados para as etapas de *parsing* e indexação (nestas etapas analisa-se o conteúdo das notícias apontadas pelos *links* contidos nos *tweets*). Dos *links* encontrados, 458 (55%) possuem ao menos uma interação do usuário. Consequentemente, os demais *links* obtiveram o valor zero no fator de relevância geosocial, pois este valor depende diretamente do número de interações de usuários na rede social.

#### **5.4.2. Hipóteses de Pesquisa**

As hipóteses de pesquisa são definidas como:

*H1: A aplicação fornece elementos que facilitam e favorecem a usabilidade da aplicação em consultas por notícias informativas?*

*H2: A aplicação se mostrou ágil no momento da realização de consultas por notícias informativas?*

*H3: As notícias recuperadas pelo GeoSEn News representam a real necessidade do usuário e são ordenadas de acordo com o critério de credibilidade exigido?*

### **5.4.3. Concepção**

Com o intuito de avaliar as hipóteses de pesquisa, foram formuladas três consultas com contextos geográficos, submetidas à aplicação GeoSEn News. As consultas (descritas em linguagem natural) são as seguintes:

*Q1: Retorne todas as notícias sobre “dengue” no estado de Minas Gerais;*

*Q2: Retorne todas as notícias sobre “zika” dentro da área que compreende os estados do Ceará, Rio Grande do Norte, Paraíba, Pernambuco e Alagoas;*

*Q3: Retorne todas as notícias sobre “microcefalia”, nos últimos 7 dias, nas adjacências do estado de Pernambuco;*

### **5.4.4. Resultados**

Todas as consultas foram executadas na aplicação GeoSEn News respeitando todos os parâmetros mencionados anteriormente. Como há um grande número de resultados para cada uma das consultas, optou-se por investigar e apresentar apenas a primeira página de resultados. Cada página de resultado contém 6 itens. Para que a hipótese *H3* seja avaliada corretamente, foram coletados os resultados reportados pela aplicação com e sem a participação do fator de relevância geosocial, contribuição deste trabalho. Os resultados são sumarizados na Tabela 6.

Tabela 6: Tabela de resultados.

Results			
Q1: Retorne todas as notícias sobre “dengue” no estado de Minas Gerais			
Com relevância geosocial		Sem relevância geosocial	
Título da notícia	Mídia	Título da notícia	Mídia
1	Instituto Butantan se prepara para estudar vacina contra o zika vírus - Notícias - Saúde	@UOLNoticias	G1 - Cidades estão em estado de alerta contra a dengue no Sul de Minas - notícias em Sul de Minas
2	Vacina contra dengue será testada em 17 mil pessoas, em 12 Estados - Notícias - Saúde	@UOLNoticias	G1 - Seminário debate enfrentamento de dengue, zika vírus e chikungunya - notícias em Minas Gerais
3	G1 - Secretaria de Saúde investiga 11 casos de microcefalia em Minas - notícias em Minas Gerais	@g1mg	G1 - Mais de 30 casos de microcefalia por zika vírus são investigados em MG - notícias em Minas Gera
4	Paciente foi infectado com vírus zika por transfusão de sangue - Jornal O Globo	@OGlobo_Rio	G1 - Saúde de BH notifica 10 casos de microcefalia por suspeita de zika vírus - notícias em Minas Ge
5	G1 - Sobe para 26 número de mortes por dengue este ano em Goiás - notícias em Goiás	@g1goias	G1 - Secretaria de Saúde investiga 11 casos de microcefalia em Minas - notícias em Minas Gerais
6	Dilma diz que governo investirá em vacina contra o vírus zika - Notícias - Saúde	@UOLNoticias	G1 - Caso de microcefalia em recém-nascido de Uberlândia é investigado - notícias em Triângulo Minei
Q2: Retorne todas as notícias sobre “zika” dentro da área que compreende os estados do CE, RN, PB, PE, AL			
1	Zika vírus pode ser transmitido por relação sexual e aleitamento materno   Local: Diário de Pernambuco	@DiarioPE	G1 - Ministério da Saúde envia equipe para investigar microcefalia no RN - notícias em Rio Grande do
2	Instituto Evandro Chagas confirma primeira morte por vírus Zika no país   Brasil: Diário de Pernambuco	@DiarioPE	G1 - Número de casos confirmados do Zika vírus aumenta em AL, diz secretaria - notícias em Alagoas
3	Casos de microcefalia já são 1.248 no Brasil e 646 em Pernambuco   Local: Diário de Pernambuco	@DiarioPE	G1 - AL atinge 64% da meta mínima da vacinação contra o vírus Influenza - notícias em Alagoas
4	G1 - No RN, 27 municípios registram alta incidência de casos de dengue - notícias em Rio Grande do Norte	@g1rn	G1 - Secretaria da Saúde de AL confirma 42 casos de microcefalia no estado - notícias em Alagoas
5	G1 - Bebê de 8 meses morre com suspeita de dengue hemorrágica em Natal - notícias em Rio Grande do Norte	@g1rn	G1 - AL confirma 25 casos da síndrome de Guillain-Barré; zika pode ser a causa - notícias em Alagoas
6	G1 - RN tem maior taxa de incidência de dengue do Nordeste, diz ministério - notícias em Rio Grande do Norte	@g1rn	G1 - No RN, 27 municípios registram alta incidência de casos de dengue - notícias em Rio Grande do N
Q3: Retorne-me notícias sobre “microcefalia”, nos últimos 7 dias, nas adjacências do estado de Pernambuco			
1	Encontro Estadual de Vigilância Sanitária abordará o zika vírus   Paraíba Online	@paraiba_online	Secretaria de Estado da Saúde divulga novo Boletim Epidemiológico sobre microcefalia   Paraíba Onlin
2	Zika vírus pode ser transmitido por relação sexual e aleitamento materno   Local: Diário de Pernambuco	@DiarioPE	G1 - João Pessoa tem maior número de notificações de microcefalia na PB - notícias em Paraíba
3	Instituto Evandro Chagas confirma primeira morte por vírus Zika no país   Brasil: Diário de Pernambuco	@DiarioPE	G1 - Secretaria de Saúde descarta 40 casos de microcefalia na Paraíba - notícias em Paraíba
4	Casos de microcefalia já são 1.248 no Brasil e 646 em Pernambuco   Local: Diário de Pernambuco	@DiarioPE	Divulgado onde foram registrados os casos de microcefalia na PB   Paraíba Online
5	Governo avalia distribuir repelente a gestantes para tentar conter microcefalia   Local: Diário de P	@DiarioPE	Secretaria de Estado da Saúde divulga Boletim sobre casos suspeitos de microcefalia   Paraíba Online
6	Secretaria de Estado da Saúde divulga novo Boletim Epidemiológico sobre microcefalia   Paraíba Onlin	@paraiba_online	G1 - 'Ainda choro muito', diz grávida de bebê com microcefalia na Paraíba - notícias em Paraíba

Fonte: ferramenta GeoSEN News.

#### 5.4.5. Discussão

As hipóteses de pesquisa H1 e H3 são respondidas de forma satisfatória. O GeoSEn News fornece elementos simples e responsáveis (i.e., se adaptam ao tamanho da tela) que têm como objetivo viabilizar uma interface funcional para o usuário (vide Seção 4.4.3). Sua usabilidade é aperfeiçoada com componentes de paginação dos resultados, de autocomplemento no formulário de consulta, com ferramenta de delimitação de área de interesse, bem como através de um mapa interativo que permite um entendimento rápido e panorâmico sobre as localizações relatadas nos resultados (H1).

Na Tabela 6 pode-se perceber que as respostas retornadas pelo GeoSEn News estão de acordo com a necessidade espacial e textual do usuário (H3). Em alguns casos, o título pode diferir da temática pesquisada, como, por exemplo, na terceira notícia retornada, como fator de relevância geosocial ativado, para a consulta *Q3*. No entanto, o termo “microcefalia” pesquisado está presente no conteúdo do documento, o que satisfaz à consulta realizada. Outro detalhe importante é o impacto positivo do fator de relevância geosocial nas respostas retornadas pela aplicação. Na *Q1* com o fator habilitado, das 6 notícias retornadas, 3 foram produzidas pelo veículo de comunicação “@UOLNoticias”, este de abrangência nacional e de grande reputação no país. Nota-se que o mesmo resultado reportado, sem a adição do fator de relevância geosocial, não foi satisfatório. Embora o “@g1mg” tenha alta credibilidade no estado, o “@UOLNoticia” é mais aceito pela população, por ser uma mídia de abrangência nacional, com mais recursos e de alta credibilidade no país.

Para a H2, a resposta é ainda incerta. As consultas realizadas na aplicação obtiveram resposta em menos de 1 segundo. Porém, essa mesma eficiência pode não ser obtida em cenários de operação. O estudo de caso foi realizado em um ambiente local e controlado, com um número reduzido de documentos indexados e com uma conexão de banda larga que facilita na renderização da camada vetorial do mapa (i.e. Google Street Map<sup>32</sup>). Assim, para responder de forma precisa esta hipótese, uma avaliação mais aguçada deve ser realizada, em um ambiente simulado mais próximo do operacional, com um número maior de notícias indexadas, acesso simultâneo de usuários e consultas complexas. Deste modo, é possível fornecer um parecer seguro sobre a questão de agilidade no tempo de resposta da aplicação.

---

<sup>32</sup> maps.google.com.br

## 5.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram descritos as metodologias utilizadas e os resultados obtidos nos experimentos realizados para validação das contribuições que circundam esta dissertação. Por meio dos experimentos, foi possível constatar a eficácia do algoritmo de estimativa de afinidade local para usuários de redes sociais, a boa aceitação do *ranking* de relevância geográfico por parte dos interessados e a usabilidade da aplicação GeoSEn News. No entanto, lacunas existentes no módulo de detecção de referências geográficas (GeoSEn), podem ter sido refletidas nas etapas subsequentes e conseqüentemente interferido na obtenção de resultados superiores.

No próximo capítulo, são apresentadas as considerações finais sobre o trabalho desenvolvido nesta dissertação, bem como suas contribuições e os trabalhos futuros.

## CAPÍTULO 6

### CONCLUSÕES E TRABALHOS FUTUROS

Dos desafios presentes na área de Recuperação da Informação Geográfica, merece destaque aquele relacionado à função de ordenar os documentos em resposta a uma busca realizada pelo usuário. Estudos recentes na área apontam para a necessidade de novas abordagens em *ranking* de relevância para mecanismos de busca Web com enfoque geográfico (Kumar, 2011; Bao e Mokbel, 2013). Neste aspecto, há discrepância na eficiência entre as soluções de *ranking* de relevância propostas em RI e as encontradas em motores de busca geográfica. Assim, pesquisas na área de GIR buscam propor soluções que se equiparam ou superam as soluções presentes em RI.

Ademais, a dificuldade em ordenar os resultados de uma busca se intensifica quando o escopo se reduz às notícias espalhadas pela Internet. No cenário que envolve a descoberta de notícias, um fator importante para medir a relevância está relacionado com a confiabilidade do conteúdo das mesmas. Contudo, há diversas implicações em medir esse fator, como, por exemplo, a necessidade do conhecimento do fato descrito na notícia e a verificação da fidelidade das fontes de informação. Por tal dificuldade, os motores de busca, que atuam no contexto de notícia, acabam desconsiderando esse fator em seus modelos de *ranking* de relevância, deixando sobre o leitor a responsabilidade em confiar no que está sendo lido.

Apesar desta carência, os usuários (i.e. leitores) continuam em busca de alternativas para se manterem bem informados. Contudo, as soluções encontradas no meio comercial e científico não trazem desfecho para este problema. Assim, vários fatores, como, por exemplo, o crescimento acelerado do número de internautas com necessidade de informação confiável, segura e consistente, o advento das redes sociais como meio de informação e as limitações encontradas nas soluções propostas pelo meio comercial e científico, vêm contribuindo para a configuração de um campo oportuno de pesquisas, visto que elementos e parâmetros, desconsiderados até então, podem ser a chave para produzir um ambiente eficaz no contexto de notícias.

Portanto, o principal objetivo deste trabalho foi a criação de um método de *ranking* de relevância geográfico, com foco em notícias informativas, baseado na relação entre a localização dos usuários que as difundiram em redes sociais, e as localidades encontradas na própria notícia. O valor medido desta relação entre localidades atua como um fator de credibilidade associado à notícia, partindo da premissa que usuários podem atestar notícias

que são relatadas dentro da sua área de moradia e/ou que possui alguma relação de convívio, seja de amigos ou familiar.

Por fim, para tornar essa estimativa de credibilidade possível e avaliar seu funcionamento, foi exigida a construção de outros arcabouços tecnológicos: o algoritmo de afinidade local, que estima o grau de afinidade entre localidades e usuário baseando-se em suas ações na rede social; e um motor de busca geográfica, com foco em notícias informativas e descendente do motor de busca GeoSEn, para implantar o *ranking* de relevância e permitir que o mesmo seja utilizado em um ambiente especializado neste sentido.

O restante deste capítulo está organizado da seguinte maneira: na Seção 6.1 são descritas as contribuições do trabalho proposto e, na Seção 6.2, são listadas algumas sugestões de trabalhos futuros.

## 6.1. CONTRIBUIÇÕES

No capítulo 3 foram apresentados trabalhos relacionados na área de GIR, que tratam de melhorias em mecanismos de *ranking* de relevância geográfico e aplicações de motores de busca geográfica com foco em notícias informativas. Além disso, foram analisados trabalhos que tinham como proposta a inferência da localidade de moradia de usuários de redes sociais que não informavam esse campo em seu perfil.

Na revisão bibliográfica realizada, foram elencadas diversas características encontradas em cada contexto analisado (i.e., inferência da localidade de moradia de usuários, *ranking* de relevância geográfico e motores de busca geográfica com foco em notícias informativas). Diante desta revisão, foi possível identificar as abordagens, técnicas utilizadas e limitações, permitindo compreender as carências e propor soluções para tais. As soluções que circundam esta dissertação buscam agregar fatores e parâmetros desconsiderados até então, como, por exemplo, o grau de credibilidade da notícia medido através de informações colhidas em redes sociais, e implantá-los em um motor de busca geográfica que permite a consulta por meio de diversas operações espaciais e viabiliza a visualização dos resultados em diferentes perspectivas.

Como principais resultados do trabalho descrito nesta dissertação, citam-se:

- **Algoritmo de estimativa de afinidade local:** algoritmo que objetiva identificar de um usuário, dada sua interação no microblog Twitter, uma lista de localidades da qual ele possui alguma afinidade, seja ela de moradia (i.e., quando o usuário reside na localidade) ou de relação de convívio com amigos ou familiares;



- **Ranking de relevância geográfico:** o *ranking* de relevância geográfico proposto comporta, além dos fatores comuns (relevância textual e relevância geográfica), um terceiro fator que representa o grau de confiabilidade da notícia, chamado de Relevância Geosocial, estimado pela correlação entre as localidades identificadas na notícia e a afinidade dos usuários que a difundiram com estas localidades.
- **Extensão do motor de busca GeoSEn:** um módulo foi construído para realizar o tratamento adequado em notícias provenientes de *links* coletados no microblog Twitter. Este visa aprimorar o GeoSEn no tocante ao *ranking* de relevância, que será acrescido do novo fator de relevância geosocial;
- **Aplicação de busca e leitura de notícias:** uma ferramenta capaz de coletar e armazenar notícias informativas, de autoria de diversas mídias distribuídas pelo país, por meio de contas oficiais no Twitter, foi construída para coleta de notícias. Esta ferramenta foi incorporada em uma interface gráfica que permite a busca e leitura de notícias informativas por parte de usuários comuns. Esta aplicação, chamada GeoSEn News, foi fruto de uma extensão do motor de busca GeoSEn e provê ao usuário uma nova experiência no encontro de notícias espalhadas pela Internet. Esta comporta a exibição de dados em várias perspectivas (i.e., listagem de resultados; mapa interativo de resultados) prezando a simplicidade, usabilidade e responsividade exigidos em sistemas desta natureza.

## 6.2. TRABALHOS FUTUROS

Para continuidade das pesquisas iniciadas neste trabalho, algumas limitações funcionais são encontradas e estas podem guiar os trabalhos futuros. As sugestões são:

- Para o algoritmo de estimativa de afinidade local:
  - Identificar outros elementos que possam indicar a afinidade entre usuário e localidade;
  - Analisar *links* compartilhados pelo usuário em sua rede social, em busca de referência geográficas no conteúdo das páginas por estes apontados, para serem utilizados na estimativa de afinidade;
  - Realizar um experimento com o intuito de identificar a melhor configuração dos pesos *pmp*, *plp* e *pfi* possível. Foi realizada uma tentativa com regressão linear, mas que não surtiu o efeito esperado.
- Para o método de *ranking* de relevância geográfico:

- Agregar diversas redes sociais para coleta de notícias e interações de usuários;
- Realizar um experimento comparativo com outras técnicas de *ranking* de relevância no contexto de notícias;
- Permitir traçar um perfil do usuário que busca por notícias e, assim, adicionar outros fatores no *ranking* de relevância (e.g., localização do usuário, assuntos de interesse);
- Realizar um experimento comparativo para identificar ganhos ou perdas em comparação ao modelo de *ranking* de relevância do GeoSEn. Este experimento também deve identificar o impacto na adição do novo fator de relevância geosocial.
- Coletar dados sobre a interação do usuário com ferramenta de consulta para perceber o grau de aceitação das notícias e, assim, ordená-las de forma mais eficiente;
- Para o GeoSEn News:
  - Desenvolver uma heurística que permita a descoberta de novas fontes de notícias. Pesquisas primárias indicam que um número considerável de *retweets* em notícias pode apontar a atuação de jornalistas, que também poderiam ser considerados fontes de notícias.
  - Avaliar experimentalmente, de forma mais robusta, os aspectos de usabilidade na aplicação GeoSEn News.
  - Alguns elementos do *layout* da página podem atrapalhar no momento da indexação de documentos, principalmente em páginas relacionadas às notícias informativas. Assim, investigar uma abordagem para identificar e considerar apenas o conteúdo da notícia no momento da indexação pode trazer melhorias nos resultados alcançados.
  - Permitir o uso da aplicação e indexação de documentos em diversas línguas. Atualmente, apenas o português é considerado.
  - Incorporar no GeoSEn News outros tipos de credibilidade e combiná-los para comparar com o fator de relevância geosocial. Uma combinação que poderia ser considerada é a do engajamento de usuários na rede (e.g. curtidas, *replay*, dentre outros).

## BIBLIOGRAFIA

Andogah, Geoffrey, Bouma, G., & Nerbonne., J. (2012). Every document has a geographical scope. *Data & Knowledge Engineering 81* , pp. 1-20.

Andrade, L., & Silva, M. J. (2006). Relevance Ranking for Geographic IR. *The Workshop on Geographic Information Retrieval, SIGIR 06, Seattle, USA* .

Armenatzoglou, N., Ahuja, R., & Papadias, D. (2015). Geo-Social Ranking : functions and query processing. *The VLDB Journal* .

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: ACM Press Book.

Bao, J., & Mokbel, M. F. (2013). GeoRank: an efficient location-aware news feed ranking system. *The 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* , pp. 184-193.

Bao, J., Mokbel, M. F., & Chow, C. Y. (Abril de 2012). GeoFeed: A location aware news feed system. *International Conference on Data Engineering* , pp. 54-65.

Bergren, M., & al., e. (2015). Inferring the location of authors from words in their texts. *NoDaLiDa* .

Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., & Shivakumar, N. (1999). Exploiting Geographical Location Information of Web Pages. *WebDB (Informal Proceedings)* .

Cai, G. (2002). GeoVSM: An Integrated Retrieval Model For Geographical Information Lecture Notes. . *Second International Conference on GIScience, Baltimore, MD, USA* , pp. 65-79.

Cai, G. (2011). Relevance ranking in Geographical Information Retrieval. *SIGSPATIAL Special, 3(2)* , pp. 33–36.

Calazans Campelo, C. E., & de Souza Baptista, C. (Outubro de 2008). Geographic scope modeling for web documents. *2nd international workshop on Geographic information retrieval. ACM* , pp. 11-18.

Campelo, C. E. (2008). *GeoSEn : um Motor de Busca com Enfoque Geográfico*. Campina Grande: Editora Universidade Federal de Campina Grande - EDUFPG.

Campelo, C. E., & Baptista, C. S. (2009). A Model for Geographic Knowledge Extraction on Web Documents. *Advances in Conceptual Modeling - Challenging Perspectives, LNCS 5833* , pp. 317-326.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. . *The 19th ACM International Conference on Information and Knowledge Management* , pp. 759–768.

Comer, D. (1979). Ubiquitous B-tree. *ACM Computing Surveys (CSUR) 11.2* , pp. 121-137.

*Desktop Search Engine Market Share.* (s.d.). Acesso em 27 de 03 de 2016, disponível em NetMarketShare: <https://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0>

Hill, L. L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. *International Conference on Theory and Practice of Digital Libraries. Springer Berlin Heidelberg* .

Ikawa, Y., Enoki, M., & Tatsubori, M. (2012). Location inference using microblog messages. *Proceedings of the 21st International Conference Companion on World Wide Web* , pp. 687–690.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management 44.3* , pp. 1251-1266.

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *Geographical information retrieval. International Journal of Geographical Information Science* , 219-228.

Jones, C. B., Purves, R., Ruas, A., Sanderson, M., Sester, M., Kreveld, M. J., et al. (2002). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. *SIGIR 2002: The 25th Annual Intern ational ACM SIGIR Conference on Research and Development in Information Retrieva, Tampere, Finland* , pp. 387-388.

Jones, K. S., & Willett, P. (1997). *Readings in Information Retrieval.* Morgan Kaufmann.

Jurgens, D. (2013). That’s what friends are for: Inferring location in online social media platforms based on social relationships. *AAAI Conference on Weblogs and Social Media* , pp. 273–282.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30 , pp. 81–93.

Kendall, M. G. (1948). *Rank correlation methods.*

Kleinberg, J., & Lawrence, S. (30 November 2001). The Structure of the Web. *Science*, v. 294 , 1849-1850.

Kowalski, G. (1998). Information retrieval systems: theory and implementation. *Computers and Mathematics with Applications* , p. 133.

Kumar, C., & Boll, S. (Novembro de 2013). Criteria of query-independent page significance in geospatial web search. *7th Workshop on Geographic Information Retrieval* , pp. 35-42.

Kumar, C., Heuten, W., & Boll, S. (2013). Geographical queries beyond conventional boundaries. *The 7th Workshop on Geographic Information Retrieval - GIR '13. New York, New York, USA: ACM Press.* , pp. 84–85 .

Larson, R. R., & Frontiera, P. (2004). Spatial ranking methods for geographic information retrieval (gir) in digital libraries. *In Research and Advanced Technology for Digital Libraries* , pp. 45-56.

Lee, H. C., & Haifeng Liu, R. J. (2007). Geographically-Sensitive Link Analysis. *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)* , pp. 676–682.

Li, Y. ( Jul/Aug de 1998). Toward a qualitative search engine. *IEEE Internet Computing*, v. 2, n. 4 , pp. 24-29.

Li, Y., Moffat, A., Stokes, N., & Cavedon, L. (2006). Exploring probabilistic toponym resolution for geographical information retrieval. *SIGIR, Workshop on Geographical Information Retrieval* , pp. 17-22.

Libânio, C., Jerônimo, M., Campelo, C. E., & Baptista, C. D. (2015). Mining influential terms for toponym recognition and resolution. *GeoInfo* , pp. 143-154.

Mahmud, J., Nichols, J., & Drews, C. (2012). Where Is This Tweet From? Inferring Home Locations of Twitter Users. *International AAAI Conference on Web and Social Media* , pp. 511-514.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval. Vol. 1.* Cambridge: Cambridge university press.

Markowetz, Chen, Y.-Y., Suel, T., Long, X., & Seeger, B. (2005). Design and Implementation of a Geographic Search Engine. *WebDB* .

Martins, B., Silva, M. J., & Andrade, L. (Outubro de 2005). Indexing and Ranking in Geo-IR Systems. *Workshop on Geographic Information Retrieval at CIKM 2005* .

McCurley, K. S. (Maio de 2001). Geospatial Mapping and Navigation of the Web. *Tenth International World Wide Web Conference* , pp. 221-229.

Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You are who you know. *The third ACM international conference on Web search and data mining - WSDM '10*. New York, New York, USA: ACM Press. , p. 251.

Murugesan, S. (2007). Understanding Web 2.0. *IT professional* , 34-41.

Navarro, G. (1999). Indexing and Searching. In: Baeza-Yates, *Modern Information Retrieval* (pp. 191-228). New York: ACM Press Book.

Navarro, Gonzalo, Baeza-Yates, R. A., Sutinen, E., & Tarhio, J. (2001). Indexing methods for approximate string matching. *IEEE Data Eng. Bull.*24, no. 4 , pp. 19-27.

Oliveira, M. G., Baptista, C. D., Cláudio, E. C., Amilton, J., Acioli, M., Gabrielle, A., et al. (2014). Automated Production of Volunteered Geographic Information from Social Media. *Geoinfo Synposium* .

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web. *Technical Report SIDL-WP-1999-0120, Stanford Digital Library* .

Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *ICWSM* .

Phelan, O., McCarthy, K., & Smyth, B. (2009). Using twitter to recommend real-time topical news. *Proceedings of the Third ACM Conference on Recommender Systems - RecSys '09* , p. 385.

Raghavan, S., & Garcia-Molina, H. (2001). Crawling the HiddenWeb. *27th international Conference on Very Large data Bases (VLDB)* , pp. 129-138.

Raper, J. (2007). Geographic relevance. *Journal of Documentation* 63.6 , 836-852.

Richardson, L., & Ruby, S. (2008). *RESTful web services*. O'Reilly Media, Inc.

Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths.

Robertson, S. E., & Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Sciences*, 27(3) , pp. 129-146.

Sahai, G., & Chan, C. (n.d.). Building a Geographically Intelligent News Search Utility.

Samet, H., Sankaranarayanan, J., Lieberman, M. D., Adelfio, M. D., Fruin, B. C., Lotkowski, J. M., et al. (2014). Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*, 57(10) , pp. 64–77.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). TwitterStand. *Proceedings of the 17th ACM SIGSPATIAL International Conference*

on *Advances in Geographic Information Systems - GIS '09*. New York, New York, USA: ACM Press , p. 42.

Silva, M. J., Martins, B., Chaves, M. S., Afonso, A. P., & Cardoso, N. (Julho de 2006). Adding Geographic Scopes to Web Resources. *CEUS – Computers, Environment and Urban Systems* .

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., & Sperling, J. (2008). NewsStand. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems - GIS '08 (Vol. 2008)* , p. 1.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography* 46.sup1 , 234-240.

Uysal, I., & Croft, W. B. (2011). User oriented tweet ranking: A filtering approach to microblogs. *International Conference on Information and Knowledge Management, Proceedings* , pp. 2261–2264.

Wang, X., Zhang, Y., Chen, M., Lin, X., Yu, H., & Liu, Y. (2010). An evidence-based approach for toponym disambiguation. *18th International Conference on Geoinformatics* .

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*. 1(6), 80-83.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107. Carnegie Mellon University.

Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys (CSUR)* .