Método não supervisionado para monitoramento de assuntos de governo nos países de língua portuguesa

Fabio Ferman[1], Luan B. Garrido[1], Tiago S. da Silva[1], Sérgio A. Rodrigues[1], Jano M. de Souza[1]

[1]Instituto Alberto Luiz Coimbra de Pós Graduação e Pesquisa de Engenharia – Universidade Federal do Rio de Janeiro (COPPE/UFRJ) – Cidade Universitária – RJ – Brasil

{fferman,lbgarrido,tiagoss,sergio,jano}@cos.ufrj.br

Abstract. *The rulers around the world need to monitor various aspects of its territory and the planet Earth with the intention of preventing and solving problems. This monitoring in most cases is through the analysis of statistical data which can be difficult and costly to acquire. Through news analysis we can create a simple and inexpensive alternative to achieve the required data. For this the holding of two mining tasks are necessary: (1) recognition of the entities of interest and (2) geolocation of the issues found. This work is focused on the second one, with special attention on countries that have Portuguese as one of its official languages.*

Resumo. *Os governantes ao redor do mundo necessitam monitorar diversos aspectos de seu território e do planeta Terra com intuito de prevenir e solucionar problemas. Esse monitoramento, na maioria dos casos, se dá através da análise de dados estatísticos, os quais podem ser difíceis e custosos de se adquirir. Através da análise de notícias, pode-se criar uma alternativa simples e barata para se conseguir os dados necessários. Para isso, é necessário a realização de duas tarefas de mineração: (1) reconhecimento das entidades de interesse e (2) geolocalização dos assuntos encontrados. Neste trabalho é focado o segundo problema, com enfoque especial para os países que tem o português como uma de suas línguas oficiais.*

1. Introdução

Durante toda a história, cada região do mundo teve que lidar com diversos problemas que surgiram e tiveram que ser superados. No cenário atual, em que temos o planeta divido em países, cada governo precisa lidar com diversas complicações as quais envolvem cada vez mais fatores. Monitorar os problemas existentes e conseguir evitar novos exige dos governos estar ciente da situação e dos acontecimentos, criar estratégias e colocá-las em execução. Estar ciente da situação em muitos casos é gozar de dados estatísticos que revelem com precisão os pontos fortes e fracos acerca de diversas questões, e que podem ser cruzados para assim prospectar medidas a serem tomadas.

Uma vez que ir em busca de estatísticas existentes em diversas bases ou gerar as próprias estatísticas (através de censos por exemplo) é uma tarefa custosa, uma abordagem diferente necessita ser utilizada. Uma alternativa para se adquirir esse conhecimento é através das notícias, no qual grande parte dos acontecimentos são

relatados nela, e assim precisam apenas de um tratamento para transformá-las em estatísticas. Usando os portais de notícias Web junto com a aplicação da técnica de extração de informação (*Information Extraction* - IE), os assuntos de interesse e suas respectivas localizações podem ser extraídas. Para a realização dos dois tipos de extração, duas técnicas são respectivamente aplicadas: (1) reconhecimento de entidades nomeadas (*Named Entity Recognition* - NER), a qual extrai dos textos as ocorrências de entidades pré-estabelecidas, e (2) recuperação de informações geográficas (*Geographical Information Retrieval* - GIR), a qual extrai dos textos todos os lugares mencionados.

Neste trabalho será investigado a segunda técnica (GIR). Assim como nos demais trabalhos, será utilizado um dicionário geográfico (*Gazetteer*) para ajudar na identificação dos topônimos (nome dado aos *tokens* que representam uma localização). Todavia, ao contrário de outras abordagens que tentam resolver o problema (GIR) com apenas uma abordagem, não importando o escopo e a língua utilizada, precisando recorrer em muitas das vezes na necessidade de uma base com grande quantidade de detalhes ou um algoritmo complexo, neste trabalho o algoritmo propõe um método com foco na língua portuguesa, a qual ainda é um idioma pouco explorado para esse tipo de problema [Machado et. al., 2011][Tjong Kim Sang and Meulder, 2003]. Assim, será inicialmente realizado uma análise em cima dos dicionários geográficos dos países que tem o português como uma de suas línguas oficiais de acordo com a *Central Intelligence Agency* (CIA) [UNITED STATES, c2014], e assim explorado as peculiaridades do idioma para que sejam desenvolvidos um algoritmo de *geoparsing* e *geocoding*.

O artigo inicia, na seção II, mostrando os trabalhos relacionados ao tema presentes na literatura. Em seguida, são analisados os dicionários geográficos (seção III). Na seção IV é mostrado o algoritmo proposto, seguido da análise de seus resultados (seção V). A última seção contempla as conclusões.

2. Trabalhos Relacionados

Todos os trabalhos estudados que lidam com GIR discriminam a tarefa em duas [Qin et. al., 2010][Machado et. al., 2011][Leidner, 2006]: (1) *geoparsing*, responsável por identificar todas as possíveis ocorrências em um texto que possa representar uma localização válida; (2) *geocoding*, responsável por mapear cada ocorrência identificada na etapa anterior a uma instância do dicionário geográfico, resolvendo possíveis ambiguidades. Todo esse processo é conhecido como Resolução de Escopo Geográfico (*Geographical Scope Resolution*) [Andogah, Bouma and Nerbonne, 2012].

Para a tarefa de *geoparsing*, Chasin et. al. (2013) utilizou três diferentes abordagens: (1) utilização do SVM, o qual utiliza de determinadas características como número de vogais e consoantes na palavra, presença de letra maiúscula, tamanho da palavra, entre outras, para realização do treinamento; (2) HMM usando LingPipe, o qual é um outro tipo de classificador capaz de aprender instâncias não conhecidas; (3) CRF que provê um framework de aprendizado supervisionado. Algumas técnicas auxiliares são utilizadas para ajudar no processo, dentre elas destaca-se o POS (*part-of-speech*) tagging [Lieberman and Samet, 2012], a qual permite extrair a classe gramatical das palavras.

O *geocoding*, segunda tarefa, lida principalmente com o problema de resolução de topônimos, o qual identifica quais das *n* instâncias do dicionário geográfico com o mesmo nome é a real referência que o texto faz. Para realizar a identificação de forma correta, o problema lida com dois tipos de ambiguidade [Machado et. al., 2011]: (1) geo/geo, na qual um nome pode representar mais de uma localização; (2) não-geo/geo, na qual um nome pode tanto representar uma entidade geográfica como alguma referência não geográfica. Hosokawa (2012) ainda divide o primeiro tipo de ambiguidade em dois: (1) geo/geo horizontal, na qual os locais ambíguos são do mesmo nível na hierarquia (por exemplo, ambos são cidades); (2) geo/geo vertical, na qual os locais ambíguos são de diferentes níveis na hierarquia (por exemplo, um é uma cidade e o outro um bairro). Algumas suposições comuns são utilizadas por grande parte dos algoritmos neste problema [Qin et. al., 2010]: (1) um nome ambíguo que aparece mais de uma vez no texto muito provavelmente representa o mesmo lugar, (2) nomes que aparecem no mesmo contexto tendem a ser locais próximos. Algumas outras heurísticas também são utilizadas para a desambiguação de topônimos [Rauch, Bukatin and Baker, 2003][Andogah, Bouma and Nerbonne, 2012]: tamanho da área das regiões, distância hierárquica, prioridade do tipo do lugar (país, cidade, etc), tamanho da população, popularidade no corpus, menor polígono, restrição de escopo, etc. Alguns *Web Services* já provêem o algoritmo de *geocoding* para uso público, contudo podem não ser a melhor alternativa visto que estes não se beneficiam de alguns detalhes específicos do problema [Magagna, Hess and Sutanto, 2012][Qin et. al., 2010].

3. Dicionários Geográficos

Os dicionários geográficos são bases de dados os quais contêm uma lista de nome dos locais, e cada instância possui uma chave que a identifica unicamente [Leidner, 2006] no mapa, além de possíveis outros atributos e relações. A escolha do *gazetteer* deve ser baseada no seu tipo de representação e outras características como, por exemplo, sua cobertura e corretude [Leidner, 2006].

Tabela 1. Número de topônimos por divisão administrativa (DA)

Gazetteer	1ª DA	2ª DA	3ª DA	Total
Brasil	27 estados	5,570 cidades	0 bairros	5,597
Portugal	20 distritos e regiões autônomas	308 cidades	4,260 freguesias	4,588
Angola	18 províncias	160 cidades	0 comunas	178
Moçambique	11 províncias	153 distritos	1 postos	165
Correios	27 estados	10,085 cidades	41,144 bairros	51,256

Dentre os diversos dicionários geográficos disponíveis *on-line*, Smith and Crane (2001) mostraram que 95% dos topônimos são ambíguos (em uma análise em cima do *gazetteer* TGN). Visto que o objetivo é desenvolver um algoritmo para a língua portuguesa e com restrição de escopo aos países que a falam como uma de suas línguas oficiais, a quantidade de ambiguidade deve ser reavaliada considerando apenas o escopo definido. Utilizando dos dicionários presentes no GeoNames, será realizado a análise nos principais países (de acordo com o PIB de 2013) que tem o português como língua oficial: Brasil, Portugal, Angola e Moçambique. Além desses *gazetteers*, também será

avaliado um dicionário geográfico com maior cobertura para o Brasil, gerado pelos Correios, com intuito de validar o algoritmo proposto adiante. A Tabela 1 mostra para cada dicionário geográfico as quantidades de instâncias por divisão administrativa.

As análises dos *gazetteers* são apresentadas na Tabela 2. Os dois grupos de resultados de ambiguidade horizontal representam as quantidades de grupos de X topônimos que são ambíguos entre si presentes na Yª DA. Percebe-se duas características interessantes: (1) em todos os casos, mais da metade das ambiguidades ocorrem entre apenas dois topônimos e, (2) o único caso em que tem-se mais de 50% de ambiguidade entre os topônimos foi na 3ª DA. Em ambos os casos, tem-se o cenário mais favorável no quesito de desambiguação, uma vez que na primeira situação a dúvida está entre poucas possibilidades, e na segunda situação, por um senso comum de escrita, tem-se no texto presente referências da 1ª ou 2ª DA para melhor identificar o leitor o local da 3ª DA e assim também auxiliar o algoritmo. Já para a ambiguidade vertical, são cruzados os topônimos entre cada par de DAs. O primeiro número em cada cédula representa a quantidade de topônimos da Xª DA e Yª DA (com X e Y diferentes) que sejam homônimos perfeitos, já o valor entre parênteses indica a parcela destes que pertencem a mesma hierarquia, ou seja, um topônimo é pai ou avô do outro. Para a ambiguidade vertical, duas análises realizadas: (1) a quantidade de topônimos ambíguos em relação a quantidade de instâncias em cada uma das DAs é baixa e, (2) a quantidade de ambiguidade na mesma hierarquia, a qual é a mais difícil de se desambiguar, diminui drasticamente para o *gazetteer* dos Correios (o qual apresentava o maior valor absoluto).

Tabela 2. Ambiguidade geo/geo horizontal e vertical

Gazetteer	**Horizontal**						**Vertical**		
Y=	**2ª DA**			**3ª DA**			**1ª com 2ª DA**	**1ª com 3ª DA**	**2ª com 3ª DA**
X=	**2**	**> 2**	**% amb.**	**2**	**> 2**	**% amb.**			
Brasil	205	29	9,12%	-	-	-	8 (4)	0 (0)	0 (0)
Portugal	1	0	0,65%	270	140	25,33%	0 (0)	0 (0)	216 (208)
Angola	1	0	1,25%	-	-	-	0 (0)	0 (0)	0 (0)
Moçambique	1	0	1,31%	-	-	-	1 (1)	0 (0)	0 (0)
Correios	500	257	19,54%	1550	1709	55,11%	8 (3)	14 (6)	1731 (272)

4. Algoritmo proposto

O objetivo do algoritmo a ser desenvolvido é capturar de um texto todas as referências explícitas de localização, sendo estas instâncias de alguma DA analisada nos dicionários acima, agregadas ao resto de sua hierarquia. Para isso, o algoritmo conta com dois métodos (*geoparsing* e *geocoding*) e um *gazetteer* de representação hierárquica. Uma vez que a meta é monitorar assuntos para o governo, assume-se que notícias de determinados temas, por exemplo esporte, não estarão sendo processadas por este algoritmo, evitando assim algumas ambiguidades não-geo/geo (por exemplo, times de futebol com nome de localizações).

O método de *geoparsing* a ser desenvolvido explora as características do idioma e com isso consegue funcionar sem a necessidade de treinamento. Durante a realização

da identificação e coleta dos topônimos, utiliza-se como heurísticas que estes devem ter seus tokens (excetuando preposições e artigos) capitalizados e; caso haja topônimos que compartilhem tokens dentro do texto, é escolhido como desempate aquele que tem o maior número de tokens. Além dos topônimos, são guardados também os níveis das DAs que cada um destes podem se referir. Iniciando o tratamento das ambiguidades geo/geo vertical, se houver alguma palavra chave antecedendo algum topônimo que identifique uma DA, são descartadas todas as demais DAs relacionadas a este. Topônimos que estiverem precedidos de uma palavra chave como, por exemplo, "cidade do Rio de Janeiro", ganham uma identificação adicional **P** que corresponde que o topônimo tem probabilidade alta de se referir de fato a uma instância geográfica.

Os topônimos coletados e não descartados abastecem o método seguinte, o *geocoding*. Este, tem como abordagem tratar os topônimos em quatro grupos distintos: os dois primeiros grupos são topônimos que contenham **P**, sendo o primeiro referente a topônimos que se refiram a apenas uma instância no *gazetteer* e o segundo que se referem a pelo menos duas; os dois últimos grupos são iguais aos primeiros porém não contendo **P**. Os grupos são tratados em ordem, primeiramente tratando os topônimos mais fáceis e que precisem de menos esforço para desambiguação para depois tratar os mais complexos, utilizando da ajuda dos topônimos já resolvidos dos grupos anteriores. Para desambiguar os topônimos utiliza-se as heurísticas: (1) caso haja dúvida entre mais de uma instância para um topônimo, escolhe-se a instância a qual seu pai na hierarquia apareceu mais vezes entre os topônimos já resolvidos, (2) topônimos que não conseguem ser resolvidos são considerados instâncias não geográficas e portanto desconsideradas. Quando um determinado topônimo tem-se identificada a correta instância do *gazetteer*, esta é adicionada no conjunto de resposta juntamente com todos as instâncias acima na hierarquia.

5. Resultados

A validação dos algoritmos propostos foi realizada utilizando o dicionário geográfico dos Correios em cima de notícias retiradas do portal de notícias do governo brasileiro (http://noticias.gov.br/noticias/). Foram sorteadas cem notícias divulgadas durante o mês de março de 2015 que tinham referência geográfica, e o resultado é exposto na Tabela 3. Nota-se que para ambas as métricas o algoritmo apresentou uma taxa alta de acerto, na qual a abrangência não varia muito entre as DAs e a precisão sofre um decaimento.

Tabela 3. Validações

Métricas	Estado	Cidade	Bairro	Todos
Precisão	100,00%	93,85%	80,36%	94,75%
Abrangência	89,19%	88,09%	93,75%	89,03%

6. Conclusão

Neste trabalho desenvolveu-se uma metodologia alternativa para captação de dados estatísticos através da análise de notícias com um algoritmo não supervisionado específico para textos escritos em português e referente aos países que o falam como um de seus idiomas oficiais, com foco no problema de geolocalização. Como analisado na validação, o método de geolocalização se mostrou eficaz e simples de implementação

(como mostrado no fluxo do algoritmo na Seção IV) em comparação aos métodos tradicionais presentes na literatura. A sequência do trabalho tem como objetivo melhorar o desempenho para os bairros, o qual apresentou menor precisão.

Referências

Andogah, G., Bouma, G. and Nerbonne, J. (2012). Every Document has a Geographical Scope. In *Data & Knowledge Engineering*, vol. 81–82, pages 1–20.

Chasin, R., Woodward, D., Witmer, J. and Kalita, J. (2013). Extracting and Displaying Temporal and Geospatial Entities from Articles on Historical Events. In *The Computer Journal*, vol. 57, issue 3, pages 403-426.

Hosokawa, Y. (2012). Improving Vertical Geo/Geo Disambiguation by Increasing Geographical Feature Weights of Places. In *Proceedings of the 2012 ACM Research in Applied Computation Symposium RACS '12*, New York, NY, USA: ACM, pages 92–99.

Leidner, J. L. (2006). An Evaluation Dataset for the Toponym Resolution Task. In *Computers, Environment and Urban Systems*, vol. 30, issue 4, pages 400–417.

Lieberman, M. D. and Samet, H. (2012) Adaptive Context Features for Toponym Resolution in Streaming News. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pages 731–740.

Machado, I. M. R., Alencar, R. O. de, Campos Jr., R. de O. and Davis Jr., C. A. (2011). An Ontological Gazetteer and Its Application for Place Name Disambiguation in Text. In *Journal of the Brazilian Computer Society*, vol. 17, issue 4, pages. 267–279.

Magagna, F., Hess, B. and Sutanto, J. (2012). Building Location-Aware Web with SALT and Webnear.Me. In *Procedia Computer Science ANT 2012 and MobiWIS 2012*, vol. 10, pages 601–608.

Qin, T., Xiao, R., Fang, L., Xie, X. and Zhang, L. (2010). An Efficient Location Extraction Algorithm by Leveraging Web Contextual Information. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM*, pages 53–60.

Rauch, E., Bukatin, M. and Baker, K. (2003). A Confidence-Based Framework for Disambiguating Geographic Terms. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, vol. 1, pages 50–54.

Smith, D. A. and Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In *Research and Advanced Technology for Digital Libraries*, Springer, vol. 2163, pages. 127–136.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, vol. 4, pages 142–147.

UNITED STATES. Central Intelligence Agency [CIA]. (c2014). "The World Factbook", https://www.cia.gov/library/publications/the-world-factbook/fields/2098.html.