Determinação do escopo geográfico de textos através de uma hierarquia adaptativa de classificadores

Eduardo Marcel Maçan
Escola Politécnica da Universidade de São Paulo
Email: macan@usp.br

Edson Satoshi Gomi
Escola Politécnica da Universidade de São Paulo
Email: gomi@usp.br

***Resumo*—A ambiguidade geográfica de topônimos em textos é o principal obstáculo para a atribuição correta de coordenadas geográficas a textos que mencionam locais e a principal causa de erros em tarefas de anotação geográfica. Este artigo apresenta um novo método para determinação do escopo geográfico de um texto através de uma abordagem adaptativa baseada em uma hierarquia de classificadores de texto. Adicionalmente, este trabalho apresenta uma análise da estrutura da ambiguidade geográfica de topônimos brasileiros. A Wikipédia foi utilizada como fonte de um conjunto de documentos anotados geograficamente para o treinamento de uma hierarquia de SVMs (Support Vector Machines) para a determinação do escopo geográfico e consequente redução da ambiguidade geográfica dos textos.**

I. Introdução

De acordo com o Google [4], em 2011 cerca de 20% de todas as consultas realizadas a partir de computadores pessoais foram efetuadas por usuários em busca de informação local. Este número chega a 40% para consultas realizadas a partir de dispositivos móveis. A busca por informações locais a partir de dispositivos móveis é uma consequência natural do fato de que pessoas em movimento geralmente buscam informações sobre pontos de interesse nas suas imediações. A busca por informações locais é facilitada se os textos disponíveis na internet estiverem anotados geograficamente, isto é, se para cada texto houver um rótulo (tag) que identifique a região à qual o texto se refere. A disponibilidade de textos anotados geograficamente permitirá o surgimento de novas classes de aplicativos móveis, como sistemas automotivos embarcados que selecionem e leiam notícias relevantes para o motorista durante seu trajeto.

A anotação geográfica é o processo de associar uma região geográfica a um documento. Esta associação é feita a partir da identificação de nomes de locais, denominados topônimos, existentes no texto. No entanto, anotar geograficamente documentos da internet é um processo custoso e impraticável se realizado manualmente. Por esta razão é fundamental encontrar um modo automático para extrair informação geográfica precisa de textos.

A criação de métodos de anotação geográfica exige a resolução de diversos problemas. Um dos problemas relevantes é o da ambiguidade de topônimos. Um topônimo é ambíguo quando mais de um local possui o mesmo nome. Por exemplo, existem no Brasil 826 ruas "Getúlio Vargas", localizadas em 26 estados diferentes, segundo dados do Censo de 2010 [7]. Assim, mesmo que seja possível encontrar nomes de locais dentro de um texto, ainda é preciso resolver as eventuais ambiguidades dos topônimos encontrados, antes de associar o texto a uma determinada região.

A probabilidade de existência de topônimos ambíguos é maior numa área geográfica mais ampla. Por esta razão é importante reduzir a área a ser associada a um texto, de forma a permitir a eliminação da ambiguidade dos topônimos nele contidos. A região geográfica resultante, que contém os topônimos do texto, é chamada de foco ou escopo geográfico.

Neste artigo é apresentado um novo método adaptativo para determinação do escopo geográfico. A correta determinação do escopo geográfico do texto reduz a quantidade de alternativas de mapeamento para topônimos, pois locais fora da área estabelecida são descartados para a resolução dos topônimos encontrados no documento.

O restante dste artigo está estruturado da seguinte forma: a Seção "Anotação Geográfica de Textos" apresenta uma breve descrição das principais referências em anotação geográfica relevantes para este trabalho. A seção "Ambiguidade Geográfica" analisa a ambiguidade geográfica dos logradouros brasileiros, apresentando visualizações de sua estrutura. A seção "Dados e Experimentos" descreve como foram obtidos os dados e estruturados os experimentos. A seção "Análise e Conclusões" apresenta considerações sobre os resultados obtidos e "Conclusões" sumariza os resultados obtidos até o presente momento.

II. Anotação Geográfica de Textos

A maioria dos algoritmos para a anotação geográfica de textos utiliza gazetteers para encontrar informação geográfica relativa aos topônimos identificados. Gazetteers, ou diretórios toponímicos, são bancos de dados que armazenam atributos associados a nomes de locais, como a população de uma cidade, suas coordenadas geográficas e dados políticos e econômicos. Estes atributos são utilizados por heurísticas para decidir quais coordenadas ou áreas geográficas melhor representam os nomes de locais mencionados no texto. O Getty Thesaurus of Geographic Names Online [6] e o GeoNames [5] são dois exemplos notórios de gazetteers que podem ser consultados pela internet.

Um dos primeiros sistemas para anotação geográfica foi o GIPSY (Geographical Information Processing System, ou Sistema de Processamento de Informação Geográfica), desenvolvido por Woodruff e Plaunt [23]. O sistema GIPSY utiliza um algoritmo de três passos para a anotação geográfica. O primeiro passo identifica palavras e frases que possuem relevância geográfica, comparando-as com os nomes existentes num gazetteer. O segundo passo busca representações poligonais das áreas mencionadas. O último passo combina os polígonos em uma representação tridimensional, tal como a apresentada na Figura 1. A sobreposição de polígonos das áreas mencionadas no texto cria um relevo tridimensional, cujos picos indicam no mapa as principais regiões às quais o documento se refere.

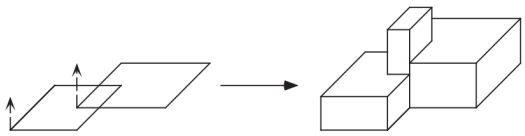


Figura 1: Sobreposição de áreas geográficas. [23]

Leidner [9] cunhou os termos "Extração de Topônimos" e "Resolução de Topônimos" para representar os dois passos principais que a maior parte dos algoritmos de anotação geográfica de textos implementa. No GIPSY, a extração de topônimos é realizada no primeiro passo e a resolução de topônimos, nas demais etapas. Leidner também descreveu 16 heurísticas e regras comuns, utilizadas por diferentes autores para eliminar a ambiguidade de topônimos. Por exemplo, uma dessas heurísticas recomenda escolher os topônimos dentro das áreas com maior população quando em dúvida sobre que local o topônimo identificado representa.

O sistema Web-a-Where [1] foi um dos primeiros a propor a anotação geográfica automática de conteúdo da Web. O primeiro passo executado pelo Web-a-Where é a busca nomes de cidades, estados e países e atribuição de um peso a cada topônimo encontrado no texto. Estes pesos variam entre 0 e 1 e são atribuídos de acordo com uma medida arbitrária de ambiguidade. Para topônimos não ambíguos, como "Chicago, IL" o valor 0.95 é atribuído. Se existem vários topônimos iguais, mas apenas um deles não ambíguo, um valor entre 0.8 e 0.9 é atribuído a todos (por exemplo, múltiplas menções a "Chicago" e apenas uma a "Chicago, IL"). Para todos os outros topônimos ambíguos encontrados é atribuído o valor 0.5 e cada topônimo é associado ao local com a maior população entre aqueles de mesmo nome. O foco geográfico é então calculado através de um sistema de pontuação que utiliza o peso atribuído a cada topônimo no primeiro passo. Ao encontrar um topônimo que representa uma cidade, representada por um foco geográfico no formato Cidade/Estado/País, soma-se para este foco geográfico o valor p^2, onde p é o peso do topônimo encontrado. Este peso é então propagado para o foco geográfico hierarquicamente superior, com um fator de atenuação $d = 0.7$. Ao encontrar uma menção ao foco geográfico Cidade/Estado/País e somar p^2 a sua pontuação, os valores p^2d e p^2d^2 são somados às pontuações de Estado/País e País, respectivamente. Os focos geográficos são finalmente ordenados de acordo com sua pontuação e as quatro maiores entre aquelas com valor maior que 0.9 são selecionadas para representar o(s) foco(s) geográfico(s) do documento. Os valores de d e p utilizados pelo Web-a-Where foram determinados de forma empírica.

O Web-a-Where foi testado com um conjunto de páginas web classificadas manualmente [24] e foi capaz de identificar corretamente a que país um texto fazia referência em 92% dos casos, dos quais apenas 38% identificavam corretamente o estado ou cidade a que o texto se referia.

Overell [11] mapeia termos da Wikipédia para termos da WordNet [10] e usa os verbetes que representam palavras associadas a locais no banco de dados léxico WordNet para determinar quais verbetes da Wikipédia se referem a locais. No passo seguinte, buscam-se pares de nomes de locais que ocorrem dentro de cada texto da Wikipédia. A ideia é que locais geograficamente próximos são frequentemente mencionados dentro no mesmo verbete. A frequência com que pares de locais ocorrem juntos é utilizada para determinar o escopo geográfico correto para os pares encontrados no documento sendo analisado.

Estas abordagens representativas das técnicas de determinação de escopo geográfico e de extração de topônimos baseiam-se em consultas a tabelas de nomes de locais para a identificação de topônimos em um texto. A extensão e correção destas tabelas influencia diretamente na qualidade dos algoritmos, como observador por Amitay [1]. Muitos dos nomes de locais em um gazetteer serão ambíguos. Ao lidar com nomes de locais em uma coleção de textos históricos Smith e Crane [16] observaram que de todos os topônimos em seus documentos, 92% podiam ser traduzidos em mais de uma coordenada geográfica.

Algumas das heurísticas de desambiguação de nomes de locais catalogadas por Leidner [9] forçam que o resultado do algoritmo se adapte a uma característica da distribuição geográfica dos documentos, sem incremento à eficiência do algoritmo em si. Por exemplo, a regra de se privilegiar locais com maior população sempre que um topônimo resultar em referências a mais de uma localidade no mapa.

Overell [11] descreve e valida experimentalmente o que batizou de "Hipótese de Steinberg", que afirma que locais geograficamente próximos ao assunto do texto e a seu autor possuem relevância muito maior do que locais distantes. Desta forma, para se identificar os topônimos de um texto e posicioná-los corretamente em um mapa é necessário estimar previamente a que região geográfica o documento se refere. Um local pouco povoado, porém geograficamente próximo ao assunto do texto é muito mais relevante ao traduzir topônimos em posições no mapa. A heurística de desambiguação por população está em direta oposição a este princípio.

Outras heurísticas comuns se utilizam de regras específicas da língua em que os textos estão escritos, como assumir que sequências de palavras escritas com letras maiúsculas possam

ser candidatas a nomes de locais, ou, a exemplo de Overell, utilizam léxicos da língua inglesa como a WordNet, não disponíveis para outras línguas, para identificar que palavras são nomes de locais.

A Wikipédia em português foi utilizada como fonte de documentos para este trabalho, mas não são assumidas premissas a respeito da língua ou estrutura sintática dos documentos.

III. Ambiguidade Geográfica

O problema da ambiguidade de nomes de locais foi investigado desde os primeiros experimentos para a extração e resolução de topônimos. [23] estiveram entre os primeiros a enumerar as diferentes maneiras com as quais locais falham em ser identificados corretamente por seus nomes. Eles encontram os seguintes fatores que contribuem para a ambiguidade de nomes de locais:

- **Nomes de locais raramente são únicos**. Por exemplo, Cambridge pode se referir a Cambridge, Massachussets ou Cambridge, Inglaterra;
- **Fronteiras políticas mudam com o tempo**, como resultado de tratados ou guerras;
- **Nomes de locais mudam**, o que faz com que seja difícil manter uma lista atualizada de nomes ao mesmo tempo em que se preserva sua utilidade para referências em documentos históricos;
- **Variação de grafia**. Nome de locais são frequentemente escritos de maneiras diversas em diferentes línguas.

Amitay et Al. [1] classificam os principais tipos de ambiguidade em geo/geo e geo/non-geo, pois locais podem ser confundidos com outros locais homônimos ou com palavras que não representem um lugar, respectivamente.

A ambiguidade do tipo geo/non-geo, ou seja, nomes que não representem locais, mas são confundidos com nomes de locais ("Mariana" e "Mariana, MG", por exemplo) depende da interpretação semântica do texto e da área geográfica sendo considerada. Por outro lado, a ambiguidade do tipo geo/geo, locais diferentes que possuem o mesmo nome ("California, Estados Unidos" e "Califórnia, Paraná", por exemplo) pode ser estimada, desde que uma lista de nomes de locais suficientemente completa esteja disponível para uma região. Para este trabalho um gazetteer com mais de 2,5 milhões de nomes de locais brasileiros foi construído e as métricas de ambiguidade absoluta de uma área e de um documento foram definidas.

- **Ambiguidade Absoluta de uma área**: representa a ambiguidade de um conjunto de topônimos em relação aos topônimos da área que os contém, por exemplo: a ambiguidade absoluta da cidade de Londrina em relação ao Estado do Paraná. Dados os conjuntos L_c e E de logradouros tal que L_c contém logradouros da cidade c e E contém logradouros de uma área que inclui a cidade c e a função F definida para pares $[logradouro, cidade]$ l_i de L_c, a ambiguidade absoluta de L_c é definida pela função $A(L_c)$ da equação 1.

$$A(L_c) = \frac{\sum_{i=1}^{n} F(l_i)}{|E|},$$
$$F(l_i) = \begin{cases} 0 & se \quad l_i \notin E - L_c, \\ 1 & se \quad l_i \in E - L_c \end{cases} \quad (1)$$

- **Ambiguidade total de um documento**: A quantidade de topônimos ambíguos identificados em um documento de texto. Sejam n topônimos T_i (i de 1 a n) identificados no texto e para cada topônimo T_i, seja A_i o número de locais diferentes com o mesmo nome de T_i identificados, a Ambiguidade total do documento A_{doc} é definida pela equação 2.

$$A_{doc} = \sum_{i=1}^{n} A_i \quad (2)$$

IV. Dados e Experimentos

A. Dados do Censo IBGE 2010

O IBGE (Instituto Brasileiro de Geografia e Estatística) disponibiliza arquivos contendo dados de todos os endereços visitados por pesquisadores de campo em áreas urbanas e rurais durante o Censo de 2010 [7]. Grande parte desses endereços possui coordenadas geográficas precisas. Neste projeto de pesquisa, nos casos em que latitude e longitude não estão disponíveis são consideradas as coordenadas da cidade da qual os endereços fazem parte.

Estes dados foram utilizados para construir um gazetteer suficientemente completo do território Brasileiro. A Figura 2 apresenta a área coberta por este gazetteer.

Figura 2: Os pontos cinza representam as cidades brasileiras cobertas pelo Gazetteer construído para este projeto de pesquisa.

B. Documentos Georreferenciados da Wikipédia

A base de artigos da Wikipédia em português foi utilizada como fonte para um conjunto de documentos georreferenciados.

A Wikipédia é composta por documentos escritos com uma linguagem própria que inclui marcação para representar coordenadas geográficas no texto. Sempre que um verbete diz respeito a um local estas marcações são traduzidas para o usuário final como links para seviços de mapas online, através do projeto [21]. Todos os documentos da Wikipédia em português que contêm marcações do tipo "geocoordenadas" [19], "satélite" [20] e "coord" [18] foram identificados através do uso de expressões regulares e tiveram suas respectivas posições geográficas extraídas.

O conteúdo do verbete e metainformação como título, URL e coordenada de cada documento identificado foi inserido em uma tabela de um banco de dados geográficos PostGIS [13], resultando em uma base georreferenciada de documentos com 22438 itens. A figura 3 exibe um detalhe da distribuição geográfica destes documentos.

Figura 3: Detalhe dos Artigos Georreferenciados da Wikipédia em Português

O IBGE disponibiliza polígonos [8] em alta resolução dos contornos de todas as cidades do país. Para cidade são selecionados os documentos da Wikipédia com coordenadas internas ao perímetro municipal através de uma query ao banco de dados geográfico. Rótulos são atribuídos a cada documento, correspondentes à cidade, micro e mesorregião, estado e região do país (Norte, Nordeste, Centro-Oeste, Sudeste ou Sul) a que pertencem.

Estas regiões possuem entre si relacionamentos do tipo "contém" e "está contida em", formando uma hierarquia geográfica.

A representação HTML dos verbetes selecionados foi obtida através de requisições web para a Wikipédia e convertidas para texto puro, totalizando 7863 documentos dentro do território brasileiro, georreferenciados e anotados com seus respectivos rótulos geográficos.

Cada um destes rótulos atribui uma classe ao documento, ao mesmo tempo em que o relaciona a uma região geográfica. Um classificador de textos capaz de atribuir corretamente um destes rótulos a um documento determinará também seu escopo geográfico.

C. Cálculo da ambiguidade absoluta de locais

A noção intuitiva de que haverá menos ambiguidade de nomes de locais quanto menor for a área considerada pode ser verificada medindo-se a ambiguidade absoluta dos logradouros de um município em áreas progressivamente menores.

A Tabela I apresenta amostras da ambiguidade absoluta de algumas cidades brasileiras. A ambiguidade absoluta mínima foi observada para a cidade de Brasília, DF com $A_{Brasilia} = 0.0145$ de ambiguidade e a máxima para a cidade de São José do Hortêncio, RS com$A_{SJH} = 0.93$. É interessante observar que Brasília possui um modelo de endereçamento único no território nacional e que na pequena São José do Hortêncio a maior parte de seus 61 endereços é apenas numerada, daí os extremos observados.

Tabela I: Ambiguidade absoluta de cidades medida dentro de diferentes áreas

City	Brazil	Region	State
S. J. Hortêncio	0.9344	0.8852	0.8033
Belo Horizonte	0.4806	0.4336	0.3125
Rio de Janeiro	0.4765	0.3959	0.2113
Campinas	0.3272	0.2930	0.2113
São Paulo	0.3249	0.2345	0.1672
Curitiba	0.2341	0.1748	0.1425
Brasília	0.0145	0.0099	-

Observe que a ambiguidade dos logradouros de uma cidade decresce quando calculada para logradouros dentro de uma área menor. A ambiguidade geo/geo calculada desta forma pode ser tomada como uma estimativa do limite inferior da ambiguidade, pois apenas correspondências perfeitas entre nomes de locais completos e sem abreviações foram levadas em consideração. "Rua Rui Barbosa" e "Avenida Rui Barbosa" não são considerados homônimos neste experimento.

D. Wikipédia: Número de exemplos por região

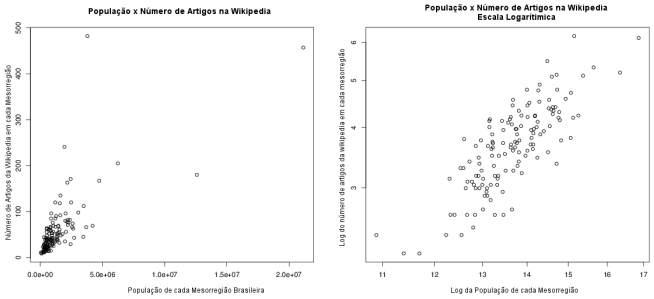


Figura 4: Número de documentos georeferenciados da Wikipédia por mesorregião brasileira.

O IBGE subdivide os estados brasileiros em 137 mesorregiões e 554 microrregiões compostas de cidades vizinhas que compartilham algumas características geopolíticas, econômicas ou sociais [12], como as regiões metropolitanas, por exemplo. A Figura 4 apresenta a relação entre número

de páginas georreferenciadas da Wikipédia dentro de cada mesorregião brasileira, em escala natural e logarítmica.

Estes dados foram obtidos pela extração de coordenadas geográficas presentes em aproximadamente 1% dos artigos da Wikipedia em português, a partir da base de verbetes da Wikipedia disponível em formato XML [22]. A figura 4 sugere uma correlação forte entre o número de artigos em uma determinada área do mapa e a população desta área.

O coeficiente de correlação de Spearmam [17] é utilizado para a análise não-paramétrica de correlação entre duas listas de valores (rankings). O coeficiente ρ, calculado para as variáveis "população" e "número de artigos na Wikipédia" para cada mesorregião é de $\rho = 0.775$, valor que indica uma correlação forte. Esta constatação confirma a noção intuitiva de que existe uma relação entre a população de uma área e o número de artigos na Wikipédia sobre esta área e permite conjecturar que esta relação também se aplique a outros tipos de documentos, como por exemplo, notícias online. Um banco de documentos anotado geograficamente a partir de documentos da web apresentará portanto uma distribuição não uniforme de documentos por toda a área considerada. Esta distribuição não uniforme de documentos deve ser levada em conta para o uso de ferramentas automáticas de classificação de textos.

E. Classificação entre duas cidades

O uso de classificadores de texto que usam a abordagem "Bag of Words"de contagems de palavras do texto é frequente em aplicações como detecção automática de propaganda não solicitada por e-mail (SPAM), como proposto por Sahami et Al. [14]. O uso de mais de uma palavra como token individual, conhecido como n-grama, foi estudado por Berrkerman [2] e por Caropreso [3], entre outros, como features para a classificação de textos, com diferentes resultados dependendo do domínio de aplicação.

Para este experimento inicial de classificação foi utilizado um classificador Naive Bayes [14] com seleção automática de atributos usando TF-IDF [15] em uma base de treinamento limitada a documentos pertencentes às cidades do Rio de Janeiro e São Paulo, extraídos da base de verbetes georreferenciados da wikipedia. Variando-se o número de documentos reservados para o treinamento, pode-se observar que a partir de poucas dezenas de artigos exemplo é possível atingir perto de 90% de acerto de classificação, conforme mostra a figura5

Este resultado permite avaliar que ao menos algumas dezenas de exemplos são necessários em cada região geográfica, para que os classificadores possam ser adequadamente treinados. Os documentos da Wikipédia não são distribuídos uniformemente. Algumas microrregiões, como as regiões metropolitanas de São Paulo e Campinas, possuem sozinhas mais documentos georreferenciados do que alguns estados inteiros do centro-oeste e norte do Brasil.

F. Hierarquia de Classificadores de Aprendizagem de Máquina

Ao invés de treinar um único classificador capaz de atribuir um escopo geográfico mais específico diretamente (um

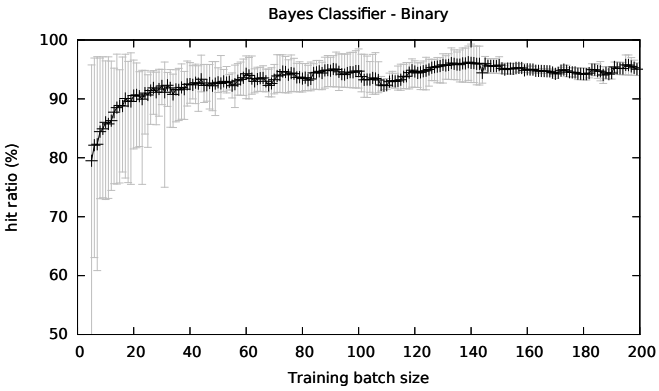


Figura 5: Sucesso de classificação Naive Bayes em verbetes georreferenciados da wikipedia para o Rio e São Paulo, com lotes de treinamento de tamanhos diversos. As barras de erro representam o melhor e pior resultado de cada lote de treinamento e validação cruzada.

único classificador com 5566 classes) foram treinados diversos classificadores para cada região, treinados com os documentos pertencentes a suas subdivisões regionais, de forma a compor uma hierarquia de classificadores, correspondente à hierarquia das regiões geográficas.

Um classificador capaz de atribuir corretamente um destes rótulos a um texto também determinará seu escopo geográfico, pois cada rótulo corresponde diretamente a uma área no mapa. A identificação correta dos locais mencionados no texto se tornará mais simples, pois muitas interpretações incorretas de nomes de locais com coordenadas fora do escopo geográfico poderão ser eliminadas.

A cada execução do experimento foram escolhidos aleatoriamente documentos para compor um conjunto de treinamento e outro de validação, na proporção de 70% e 30% respectivamente. Um classificador capaz de atribuir o documento a categorias correspondentes às 5 regiões brasileiras é treinado para a raiz da hierarquia. Para cada uma destas regiões treinamos um classificador capaz de atribuir o documento a um dos estados da região selecionada como escopo geográfico e assim por diante até o nivel das cidades. A figura 6 exibe um subconjunto dos nós da hierarquia proposta.

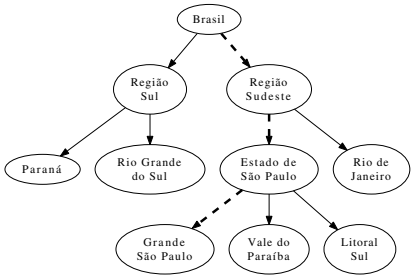
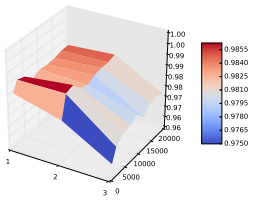


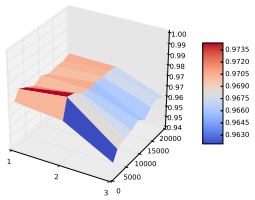
Figura 6: Hierarquia geográfica. Cada classificador corresponderá a um nó, com classes que representam seus nós filhos

Para a determinação do escopo geográfico o documento

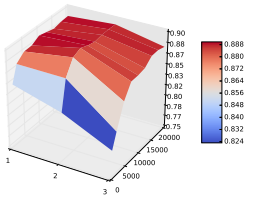
foi avaliado pelo classificador da raiz da hierarquia e foi atribuído um rótulo correspondente a uma região brasileira. Na iteração seguinte, o documento será avaliado pelo classificador correspondente àquela região que por sua vez lhe atribuiu o rótulo de um estado e assim sucessivamente, até chegarmos a atribuição de um rótulo correspondente a uma folha da árvore da hierarquia geográfica. As linhas tracejadas da figura 6 exemplificam um possível "percurso" de um documento pela hierarquia até a atribuição do escopo geográfico mais específico.



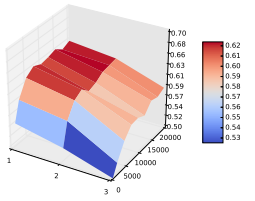
(a) Escopo geográfico: Região.



(b) Escopo geográfico: Estado.



(c) Escopo geográfico: Mesorregião



(d) Escopo geográfico: Microrregião

Figura 7: Taxas de acerto da hierarquia de SVMs em função dos tamanhos do vocabulário e dos n-gramas utilizados

Podemos observar que o desempenho cai, à medida que se busca determinar um escopo geográfico mais específico, é possível determinar que um texto pertence a um determinado estado do país com mais de 95% de correção, mas este índice cai a cerca de 65% para as microrregiões e abaixo de 20% para as cidades, conforme a figura 8

G. Hierarquia de classificadores: Abordagem Adaptativa

A distribuição geográfica não uniforme dos documentos, evidenciada pela figura 3 faz com que muitas cidades não possuam documentos exemplo em número suficiente para o treinamento do classificador de textos e este fator foi ignorado na construção da hierarquia de classificadores. Este é um fator significativo, que não pode ser ignorado.

Outro ponto a ser considerado é que o escopo geográfico de um documento não precisa ser sempre o mais específico. Um documento que tenha seu escopo geográfico determinado como "Nordeste"é correto, embora menos preciso do que o escopo geográfico "Ceará". Ainda assim, é preferível obter como resposta o escopo correto "Nordeste", do que o escopo geográfico errado "Pernambuco", se o documento se refere ao "Ceará".

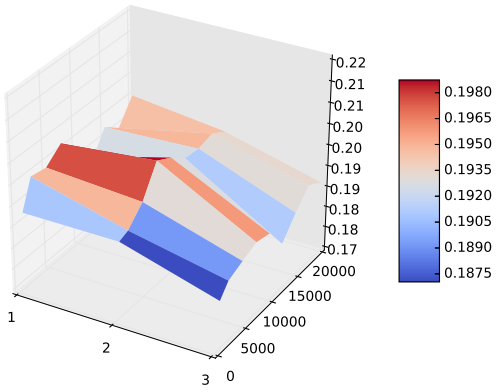


Figura 8: Desempenho para o escopo geográfico Cidade

Uma versão adaptativa da hierarquia de classificadores foi desenvolvida de forma a treinar classificadores apenas para as regiões que já possuem um número mínimo de exemplos disponíveis na base de treinamento. À medida que novos exemplos de treinamento se tornarem disponíveis, o número de classificadores e a topologia da hierarquia serão reconfiguradas, segundo os seguintes passos:

Para cada nó da hierarquia de classificadores, a partir da raiz, percorrido em largura:

- Verficar se o número de exemplos para treinamento deste nó mudou desde a última execução e há exemplos suficientes
- Obter os exemplos de treinamento para relevantes para este nó da hierarquia.
- Treinar o classificador do nó atual da hierarquia.

A inserção de novos exemplos na base de treinamento se deve dar sempre numa folha da árvore da hierarquia geográfica. Neste projeto de pesquisa o escopo mais específico é a cidade. Portanto todo novo documento a compor a base de treinamento deve ter tido uma cidade atribuída.

Porém, esta versão adaptativa da hierarquia de classificadores não atribui sempre uma cidade a um documento. Neste caso, pode-se utilizar o gazetteer para identificar os locais mencionados no texto, aproveitando-se da redução de ambiguidade proporcionada pela determinação do escopo geográfico, ou recorrer à validação de um usuário, conforme figura 9.

Alteração do número de documentos disponíveis para treinamento para um nó se propaga para todos os nós hierarquicamente superiores até a raiz, pois também farão parte de seus conjuntos de treinamento.

A figura 7 mostra a taxa de acertos da determinação do escopo geográfico em cada nível da hierarquia geográfica em função dos tamanhos dos n-gramas e do vocabulário dos classificadores SVM utilizados pela hierarquia adaptativa.

Embora note-se melhora de performance nas classificações de meso e microrregiões, os resultados para classificação

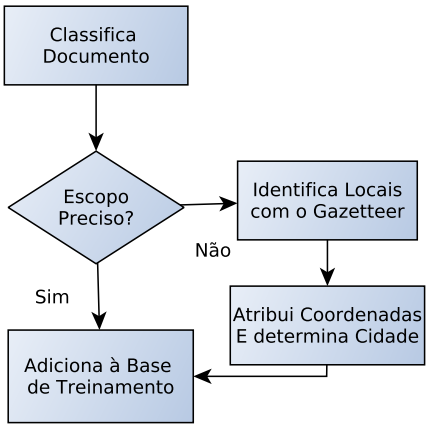


Figura 9: Inserção de novo documento na base de treinamento

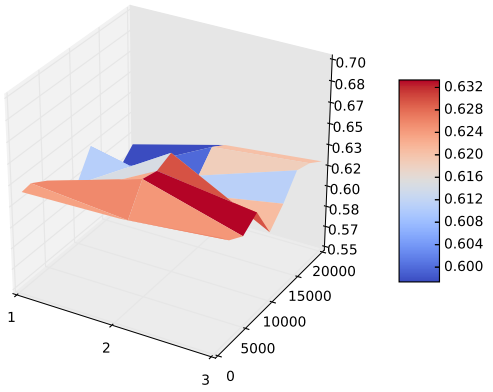
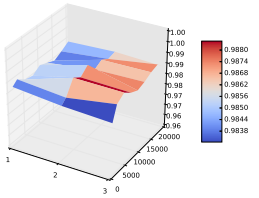
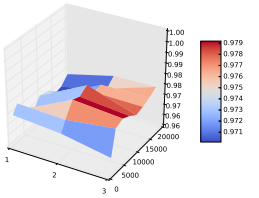


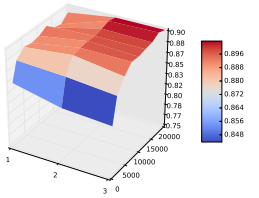
Figura 11: Desempenho para o escopo geográfico Cidade



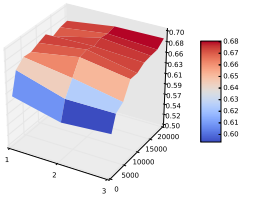
(a) Escopo geográfico: Região.



(b) Escopo geográfico: Estado.



(c) Escopo geográfico: Mesorregião



(d) Escopo geográfico: Microrregião

Figura 10: Taxas de acerto da hierarquia adaptativa de SVMs em função dos tamanhos do vocabulário e dos n-gramas utilizados

de documentos entre cidades é significativamente melhor, na ordem de 60% de classificações corretas, conforme figura 11.

Esta melhora se deve ao fato que esta nova abordagem de treinamento irá interromper o percurso do documento pela árvore de classificadores nas regiões que não possuem informação suficiente para treinamento, atribuindo um escopo geográfico menos preciso, mas que tem mais chances de estar correto.

V. Análise e Conclusões

A hipótese de Steinberg, formulada por Overell [11] diz que "todas as pessoas possuem visões de mundo similares com respeito a sua própria localidade" e que a relevância de um local para uma pessoa decresce com a distância. Ou seja, todos os indivíduos interpretam a relevância de locais da mesma maneira, tomando sua própria posição como referencial. Uma consequência importante desta hipótese é que para desambiguar corretamente os locais mencionados em um documento é preciso primeiro ter uma idéia aproximada de que região considerar para o texto. A hierarquia de classificadores proposta neste trabalho se mostrou uma ferramenta com potencial a ser explorado para esta tarefa.

A popular heurística das áreas mais populosas para desambiguação de textos assume que áreas mais populosas são provavelmente as corretas devido à correlação forte entre a população de uma área e o número de textos disponíveis pertencentes àquela área. No caso de uma alteração desta relação, esta heurística contribuiria para a classificação incorreta destes textos.

Não é prático construir um único classificador com 5565 classes para a tarefa de atribuição de escopo geográfico para todas as cidades brasileiras. Muitas cidades podem não possuir exemplos disponíveis em número adequado para treinamento. Uma abordagem hierárquica em que cada classificador pode se adaptar às peculiaridades de cada região geográfica é mais adequada.

Ao invés de depender de um único gazetteer com informações de todo o mundo em todas as línguas, com uma estrutura de manutenção complexa, é possível utilizar gazetteers menores e especializados, para atribuição de coordenadas precisas aos topônimos de sua região, após a determinação do foco geográfico, quando necessário.

A hierarquia de classificadores permite também, ao se utilizar seleção automática de atributos por tf-idf, que as características mais importantes para cada região específica, emergentes dos documentos exemplo sejam utilizadas, mesmo quando o número de atributos é limitado para evitar a explosão dimensional ocasionada pelo uso de n-gramas. Uma analogia visual pode ser feita em relação às interfaces gráficas para navegação em mapas. Quando o nível de zoom é tal que

abrange todo o território nacional, apenas alguns atributos de cada estado como rios e principais cidades são exibidos. Os atributos mudam e se tornam mais específicos, conforme o escopo é estreitado e passamos a exibir o conteúdo de um estado, ou de uma cidade.

O método apresentado exibe excelente desempenho para documentos que possuam foco geográfico único e bem definido, como notícias online. Um texto comparando Paris a São Paulo, por exemplo, não possui um único foco geográfico definido, mas pode ser dividido em partes menores submetidas individualmente à determinação do foco geográfico.

REFERÊNCIAS

[1] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, page 273, New York, New York, USA, July 2004. ACM Press.

[2] R. Bekkerman and J. Allan. Using bigrams in text categorization, 2003.

[3] M. F. Caropreso, S. Matwin, and F. Sebastiani. Text databases & document management. chapter A Learner-independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization, pages 78–102. IGI Global, Hershey, PA, USA, 2001.

[4] N. Fox. Official google blog: Ads are just answers. http://googleblog.blogspot.com.br/2011/10/ads-are-just-answers.html, Oct. 2011. Acessado em 10/12/2013.

[5] Geonames. Geonames geographical database. http://www.geonames.org/, Oct. 2005. Acessado em 17/06/2014.

[6] Getty. Getty thesaurus of geographic names. http://www.getty.edu/research/tools/vocabularies/tgn/, Oct. 1987. Acessado em 17/06/2014.

[7] IBGE. Cadastro Nacional de Endereços para Fins Estatísticos. ftp://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Cadastro_Nacional_de_Enderecos_Fins_Estatisticos, 2010. Accessado em 19/07/2012.

[8] IBGE. Malhas digitais Brasileiras. ftp://geoftp.ibge.gov.br/malhas_digitais/municipio_2010/, 2010. Acessado em: 19/07/2012.

[9] J. L. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, School of Informatics, University of Edinburgh, 2007.

[10] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[11] S. Overell. Geographic Information Retrieval: Classification, Disambiguation and Modelling. (July):1–181, 2009.

[12] J. R. Portela, editor. *Divisão Regional do Brasil em Mesorregiões e Microrregiões Geográficas*, volume 1. IBGE, 1990.

[13] P. S. C. PostGIS. PostGIS. http://postgis.net, 2012. acessado em: 19/07/2012.

[14] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk e-mail, 1998.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.

[16] D. Smith and G. Crane. Disambiguating geographic names in a historical digital library. . . . *and Advanced Technology for Digital Libraries*, 2001.

[17] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

[18] Wikipedia. Predefinição: Coord. http://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Coord/doc, Oct. 2005. Acessado em 10/12/2013.

[19] Wikipedia. Predefinição: Geocoordenadas. http://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Geocoordenadas/doc, Oct. 2005. Acessado em 10/12/2013.

[20] Wikipedia. Predefinição: Satélite. http://pt.wikipedia.org/wiki/Predefini%C3%A7%C3%A3o:Sat%C3%A9lite, Oct. 2005. Acessado em 10/12/2013.

[21] Wikipedia. GeoHack. https://wiki.toolserver.org/view/GeoHack, 2012. Accessado em 19/07/2013.

[22] Wikipédia. Wikimedia database dumps. http://dumps.wikimedia.org/backup-index.html, 2014. Acessado em 10/05/2014.

[23] A. G. Woodruff and C. Plaunt. GIPSY : Automated Geographic Indexing of Text Documents. pages 1–21, 1994.

[24] Yahoo. Dmoz.org - the open directory project. http://www.dmoz.org, 1999. Acessado em: 10/01/2014.