Um Método de Classificação Automática Para a Detecção de Duplicados em Gazetteers

Bruno Martins, Helena Galhardas e Nelson Gonçalves
Technical University of Lisbon and INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Resumo—**Este artigo apresenta uma abordagem para a detecção de registos duplicados no contexto de bases de dados de locais, utilizando uma técnica do estado-da-arte em aprendizagem supervisionada. Especificamente, este artigo relata experiências com o uso de Florestas Aleatórias como forma de classificar pares de registos como duplicados ou não-duplicados, comparando diferentes combinações de métricas de similaridade na representação dos pares de registos. Os resultados demonstram que o uso de vectores de características combinando diferentes métricas de similaridade, baseadas nos nomes dos locais, nas suas localizações geoespaciais e nos seus tipos, conduz a melhores resultados, tendo-se obtido uma exactidão de 97.45%.**

Palavras-Chave: Bases de Dados de Locais, Detecção de Duplicados, Aprendizagem Supervisionada, Florestas Aleatórias

I. INTRODUÇÃO

As bases de dados de locais, armazenando dados descritivos sobre localizações específicas na superfície da Terra, são designadas por *gazetteers* [1]. Estes *gazetteers* são muitas vezes construídos com base na consolidação de múltiplas fontes de informação. Um desafio fundamental é a detecção de registos duplicados, por forma a preservar a identidade dos locais.

Este artigo formula o problema da detecção de registos duplicados, em bases de dados de locais, como uma tarefa de classificação, onde o desafio consiste em usar dados de treino para aprender um modelo capaz de classificar pares de registos como duplicados ou não-duplicados. É usada uma técnica de classificação supervisionada conhecida pelo nome de Florestas Aleatórias (i.e., *Random Forests*), e os pares de registos são representados através de vectores de características que combinam múltiplas métricas de similaridade (e.g., a similaridade entre as localizações geoespaciais, entre as categorias de tipo, entre as relações semânticas, e entre os nomes dos locais). Numa publicação anterior [2], foram apresentados resultados sobre a detecção de duplicados em gazetteers, através de aprendizagem supervisionada. Este artigo introduz o uso de Florestas Aleatórias como uma abordagem de classificação mais eficaz e apresenta resultados experimentais com um conjunto de registos maior e mais representativo da realidade.

Trabalho parcialmente suportado pela Fundação para a Ciência e Tecnologia (FCT), através do projecto PTDC/EIA-EIA/109840/2009 (SInteliGIS).

II. CONCEITOS E TRABALHOS RELACIONADOS

O problema de identificação de registos que são sintacticamente diferentes e ainda assim descrevem a mesma entidade física tem sido amplamente estudado. Os métodos típicos envolvem o cálculo da similaridade entre os pares de registos, pressupondo que os registos muito semelhantes são provavelmente duplicados [3]. Para cada par candidato de registos, a similaridade é calculada usando uma métrica de distância ou um método probabilístico. Os pares candidatos que têm valores de similaridade superiores a um determinado limiar são então associados como duplicados. O fecho transitivo dos pares associados permite formar os grupos finais de equivalência entre os registos duplicados.

Trabalhos anteriores exploraram métodos de aprendizagem supervisionada na detecção de duplicados, recorrendo a dados de treino manualmente marcados como pares de registos duplicados ou não-duplicados [8, 10, 13]. Estes métodos têm sido explorados a dois níveis distintos, individualmente ou em combinação. Estes dois níveis correspondem (i) ao uso de métricas de similaridade treináveis, tais como variantes da distância de edição onde se adaptam os pesos das operações de forma a capturar a similaridade entre atributos específicos, ou (ii) à construção de classificadores para discriminar entre os pares de registos duplicados e não duplicados, utilizando valores de similaridade para os diferentes atributos como características. O trabalho relatado neste artigo cai na segunda categoria, usando classificadores binários que aprendem a combinar diferentes caracteristicas para discriminar registos duplicados em *gazetteers*.

Alguns trabalhos anteriores definiram ainda a resolução de entidades geoespaciais como o processo da obtenção, a partir de fontes de dados distintas referentes a locais específicos na superfície Terrestre, de uma única colecção de locais reais consolidados [14]. Este problema é diferente de outras tarefas de detecção de duplicados mais tradicionais, devido à presença de uma componente geoespacial contínua nos dados georeferenciados (i.e., as coordenadas geoespaciais). Foram, por exemplo, propostas técnicas para combinar funções de similaridade geo-espaciais e não geo-espaciais, embora esta combinação apresente problemas não-triviais devido à necessidade de combinar informação semanticamente distinta. Uma forma de combinar diferentes métricas de similaridade envolve colocar um limiar sobre uma, de seguida utilizando outra métrica como um filtro secundário (isto é, ajudando na rejeição de locais semelhantes que, de acordo com a primeira

métrica, não seriam duplicados), e assim por diante [11]. Hastings descreveu um abordagem baseada na geo-cognição humana, concentrando-se primeiro na informação geoespacial (usando sobreposições de áreas ou proximidade entre os locais), de seguida usando as categorias de tipo dos locais (categorias conceptualmente próximas podem indicar o mesmo lugar) e, finalmente, usando os nomes (i.e., nomes de locais com pequenas variações na grafia ou com abreviações podem indicar o mesmo local) [11]. Contudo, a abordagem acima não permite capturar os locais duplicados que não são muito similares de acordo com cada uma das métricas individuais. Neste caso, é preciso uma medida de similaridade única que combine todas as métricas individuais num único valor. Alguns trabalhos anteriores propuseram usar a similaridade geral entre pares de registos, calculada pela média ponderada das similaridades entre os seus atributos individuais [12]. As médias ponderadas têm a flexibilidade de dar a alguns atributos mais importância do que a outros. No entanto, o ajuste manual dos pesos pode ser difícil e métodos de aprendizagem automática são geralmente mais robustos.

Autores como Zheng et al. ou Sehgal et al. descreveram abordagens de aprendizagem automática para detectar registos duplicados em *gazetteers*, combinando a similaridade dos nomes, da localização geoespacial e das categorias [10, 13]. O trabalho relatado neste artigo segue de perto estes trabalhos anteriores, mas apresenta experiências com classificadores baseados em Florestas Aleatórias e usando representações mais ricas para os pares de registos.

III. DETECÇÃO DE DUPLICADOS EM REGISTOS DE *GAZETTEER*

Neste artigo, argumentamos que a detecção de registos de *gazetteer* duplicados pode ser encarada como um problema de classificação, em que o objectivo é classificar pares de registos como duplicados ou não. Esta tarefa é realizada em duas fases, de forma sequencial, tal como ilustrado na Figura 1. Na fase de *treino*, constrói-se um classificador usando um conjunto de instâncias, obtido através da selecção de um conjunto representativo de pares de registos, anotados como sendo duplicados ou não. As múltiplas características que representam esses pares são extraídas dos elementos que descrevem os locais e um algoritmo de aprendizagem é usado para construir o modelo de classificação. Na fase de *execução*, a detecção de duplicados envolve a geração de todos os possíveis pares de registos emparelhados, a extracção de características desses pares e a utilização do modelo de classificação gerado pela fase anterior para etiquetar cada par como correspondendo a registos duplicados ou não.

O cenário de aplicação genérico para a técnica reportada neste artigo é aquele em que se têm dois conjuntos de registos *A* e *B* provenientes de fontes independentes. Cada registo corresponde a um local na superfície da Terra. Um local é definido por (i) um ou mais nomes pelos quais é normalmente conhecido, (ii) um ou mais tipos de local, que o situam num esquema de classificação estabelecido e que também fornece a base conceptual para os locais listados no *gazetteer*, (iii) zero, uma ou mais localizações geospaciais (*footprints*), que correspondem às coordenadas do local na superfície da Terra e, opcionalmente, designam a sua configuração em termos de área e (iv) zero, um ou mais intervalos temporais de validade para o local. Cada *footprint* geoespacial pode ser dado por um ponto, linha, caixa delimitadora, ou outra forma poligonal que seja suportada pela norma GML (*Geography Markup Language*). A informação de validade temporal é dada por instantes ou intervalos do calendário, também de acordo com a norma GML.

Muitas vezes, um local é conhecido por múltiplos nomes, tipos e *footprints*, atribuídos por entidades diferentes, com distintos objectivos ao longo do tempo. Contudo, no contexto de um *gazetteer*, deve existir uma relação de um-para-um entre um local e o seu respectivo registo. Assim, a construção de novos *gazetteers* a partir de múltiplas fontes requer o tratamento de registos duplicados. O objectivo deste trabalho é assim encontrar pares de registos $<r1,r2>$, tal que $r1 \in A$, $r2 \in B$ e ambos $r1$ e $r2$ correspondam ao mesmo local do mundo real.

É de salientar que os nomes de locais têm frequentemente embebida informação sobre os seus tipos (e.g., *Lisbon Street* ou *Luxembourg City*) e alguns possuem múltiplas interpretações (e.g., *Mississippi* é um local populado assim como um rio). Diferentes locais podem partilhar as mesmas ou semelhantes coordenadas (e.g., a cidade ou o estado do *Mónaco*) e os locais podem ver mudadas as suas fronteiras (*Berlim* antes e depois da queda do Muro), os seus tipos (e.g., o *Rio de Janeiro* era a capital do *Brasil* antes de 1960) ou mesmo os seus nomes (*Congo* era conhecido como *Zaire* entre 1971 e 1997) ao longo do tempo. As mesmas regiões do globo podem também corresponder a diferentes locais ao longo do tempo (e.g., a região da antiga *União Soviética*).

Os vectores de características utilizados no esquema proposto combinam informação de diferentes elementos disponíveis nos registos de *gazetteer*. Um total de 50 características distintas é usado. Estas são agrupadas nas seguintes cinco classes diferentes: (i) similaridade entre nomes, (ii) similaridade entre *footprints* geoespaciais, (iii) similaridade entre tipos, (iv) similaridade entre relações semânticas, e (v) similaridade entre intervalos temporais.

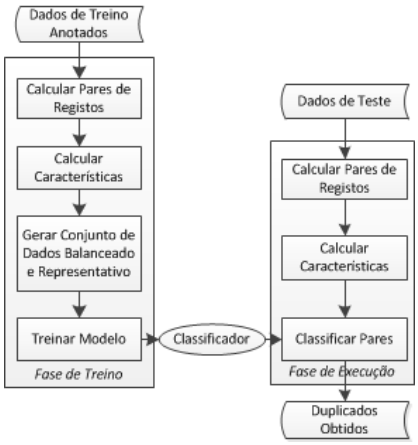

Figura 1: Processo de classificação para detecção de duplicados.

No que diz respeito às características de similaridade entre nomes de locais, a ideia é capturar a intuição de que nomes semelhantes tipicamente descrevem o mesmo local. Porém, o modo como a similaridade é calculada deve suportar palavras com grafias ligeiramente diferente, abreviaturas, etc. Diferentes tipos de métricas de similaridade entre cadeias de caracteres

têm vantagens e desvantagens que se complementam. Em consequência, é útil considerar múltiplas métricas na avaliação de potenciais duplicados. Devido a restrições de espaço, omitimos o conjunto completo de características usado nas experiências. No caso da similaridade entre nomes de locais (entre os nomes principais ou alternativos), são usadas características tais como as medidas de similaridade de Levenshtein e de Jaro-Winkler [4], ou a distância fonética de acordo com o algoritmo *Double Metaphone* [5].

Em termos da similaridade entre *footprints* geoespaciais, a ideia é suportar a intuição de que locais que não estão próximos não podem ser o mesmo. Duas das características consideradas são a distância entre os pontos centróides dos *footprints*, ou a área de sobreposição entre os *footprints*. No que diz respeito à similaridade entre tipos de local, a intuição é que locais com o mesmo tipo têm maior probabilidade de serem duplicados. Os registos de *gazetteer* são representados de acordo com o tipo do *Alexandria Feature Type Thesaurus*[1] *(FTT)*, que define uma hierarquia para a classificação de locais com 6 termos no primeiro nível e 210 termos no total. Tendo em conta a natureza hierárquica do FTT, podemos ter uma noção da proximidade entre os diferentes tipos que são considerados (por exemplo, os tipos de local que correspondem a cidades e capitais são conceptualmente semelhantes e potencialmente descrevem o mesmo local). Usámos então a distância no *Alexandria FTT* como uma característica deste tipo, de acordo com métricas como as propostas por Lin [6] e Resnik [7].

Em termos de medidas de similaridade entre relações semânticas, a intuição é que os locais relacionados com os mesmos outros locais têm maior probabilidade de serem duplicados. Dois exemplos de métricas usadas são os coeficientes de Dice e de Jaccard [4]. Finalmente, no que diz respeito à similaridade entre intervalos temporais, a intuição é que as definições de locais que se referem ao mesmo período temporal são potencialmente duplicadas. Um exemplo de uma métrica é a duração da intersecção entre os períodos de tempo associados aos registos. É, no entanto, de notar que só uma quantidade muito pequena dos registos considerados tem de facto definido um período temporal para a sua validade. Assim, o impacto destas características nos resultados obtidos é negligenciável.

Este artigo reporta experiências com um único método de classificação supervisionada, as Florestas Aleatórias [16]. Estes classificadores correspondem a árvores de decisão, usando uma aproximação de comité baseada em *bagging* e selecção aleatória de características. Várias árvores de decisão são geradas e o classificador retorna a classe que é a mais votada de todas as classes que são produzidas pelas várias árvores individuais. Os classificadores de Florestas Aleatórias são considerados como um dos métodos de classificação mais eficiente e eficaz. Nas experiências, usámos a implementação fornecida pelo pacote de aprendizagem automática Weka [17].

IV. AVALIAÇÃO EXPERIMENTAL

As experiências relatadas neste artigo utilizam um conjunto de registos contendo exemplos de pares duplicados e não duplicados, sendo estes representados de acordo com o esquema XML do serviço de *gazetteer* da ADL (*Alexandria Digital Library*). O conjunto de dados contém um total de 2,864 registos que representam locais na Terra. Um total de 4,401 pares de registos foram anotados manualmente como duplicados. Uma análise dos casos duplicados revela que existem diferentes problemas no conjunto de dados, incluindo nomes com grafias diferentes (e.g., *Wien* e *Vienna*), com *footprints* geoespaciais distintos (e.g., através de coordenadas centróides, rectângulos delimitadores ou geometrias poligonais complexas), tipos de local distintos (e.g., *Lisbon* tanto é uma *cidade* como uma *capital*), e contendo informação com nível de detalhe variável nas relações com outros locais.

Número de registos	2,864
Número de nomes de locais	5,952
Nomes de locais únicos	3,713
Média de nomes por registo	2.08
Número de tipos de local	14
Média de tipos de local por registo	1.00
Número de registos com dados temporais	11
Número de registos representados por centróide	589
Número de pares	4,102,680
Número de pares duplicados	4,401
Número de pares considerados	8,802
Número de pares duplicados considerados	4,401

Tabela 1: Caracterização da colecção de teste.

A abordagem *naïve* de usar todos os pares de registos disponíveis (neste caso 4,102,680 pares de registos, com apenas 4,401 duplicados) resulta num conjunto de treino que contém uma quantidade muito maior de pares não duplicados do que de pares duplicados. Este desvio não só afecta a utilidade de uma abordagem de classificação (e.g., a precisão seria alta mesmo que todos os pares fossem classificados como não duplicados), mas também a eficiência do processo de aprendizagem. Tomando inspiração nas abordagens propostas por Sehgal et al. [10] e por Bilenko e Mooney [9], decidimos balancear o nosso conjunto de dados, de forma a termos a mesma quantidade de pares duplicados e não duplicados. Em resumo, tomamos todos os pares de registos, previamente assinalados como duplicados e uma igual quantidade de pares não duplicados que representam tanto pares fáceis (i.e., bastante dissimilares) ou dificeis de classificar. O número total de pares de registos utilizado nas experiências é assim 8,802. A Tabela 1 apresenta uma caracterização da colecção de teste.

Características	***Precisão***	***Recall***	***F1***	***Exactidão***	***Erro***
Nomes de Locais (12)	0.852	**0.981**	0.912	0.924	07.600
Footprints (7)	0.935	0.951	0.943	0.954	04.577
Nomes + Footprints(19)	0.947	0.953	0.950	0.960	03.995
Todas(50)	**0.967**	0.969	**0.975**	**0.974**	**02.548**

Tabela 2: Comparação dos diversos vectores de características.

Treinamos classificadores de Florestas Aleatórias usando combinações de características propostas. As combinações utilizadas foram as seguintes: *(i)* utilizar apenas características relativas ao nome, *(ii)* utilizar apenas características relativas à localização, *(iii)* utilizar características relativas ao nome e à localização, *(iv)* utilizar todas as características propostas.

Em todos os casos acima mencionados, foi realizada uma validação cruzada com 10 rondas. A Tabela 2 dá-nos uma visão global sobre os resultados obtidos para cada uma das diferentes combinações. O classificador que utiliza todas as

[1] http://www.alexandria.uscb.edu/gazetteer/FeatureTypes/FTT2HTM/

características de similariedade propostas atinge o melhor desempenho (i.e., uma exactidão de 0.974), apesar das características relacionadas apenas com o nome de local ou com as *footprints* geoespaciais oferecerem resultados bastante competitivos (i.e., uma exactidão de 0.924 e de 0.954, respectivamente). Utilizando as características relacionadas com o nome em conjunto com as características relacionadas com a representação espacial leva a uma exactidão de 0.960. As restantes características (i.e., similaridade de tipo de local) têm um impacto bastante reduzido nos resultados finais.

Usámos ainda um *t-test* de forma a medir a significância estatística dos resultados, comparando os diferentes grupos de características. Os resultados mostram que as diferenças entre os grupos são significativas, apesar dos resultados serem bastante semelhantes. É também importante notar que a colecção utilizada contém apenas lugares de alto nível (e.g., cidades). Visto que provavelmente a ambiguidade em nomes de locais se manifesta mais em locais de nível mais baixo (e.g., pequenas vilas e nomes de ruas), o desempenho da similaridade entre nomes de locais pode baixar em registos diferentes.

Através de selecção de características, é muitas vezes possível obter classificadores que são mais eficientes (i.e., usam menos características) e são mais exactos. Como tal, fizemos uma experiência utilizando um método de selecção de características denominado *Greedy Forward Selection*. Este inicia com um conjunto de características pre-seleccionadas *S* que é tipicamente vazio. Para cada característica f_i que ainda não esteja em *S*, um modelo é treinado e avaliado utilizando o conjunto de características $S \cup f_i$. A característica que tenha demonstrado o maior ganho é então adicionada ao conjunto *S* e o processo é repetido até nenhuma característica melhorar a exactidão. Treinámos um classificador com este procedimento, resultando numa exactidão de 0.951, usando um total de 4 características, das quais 2 estão relacionadas com *footprints*, enquanto as outras estão relacionadas com nomes.

V. Conclusões e Trabalho Futuro

Este artigo apresenta uma aproximação baseada em aprendizagem supervisionada para encontrar registos de *gazetteer* duplicados. Reporta uma avaliação da aproximação de classificação *Random Forests*, usando vectores de características que combinam diferentes aspectos de similaridade entre pares de registos. A utilização de diferentes tipos de características de similaridade conduziu a um aumento significativo na correcção dos resultados obtidos, embora a similaridade entre os *footprints* geoespaciais ou entre os nomes de locais, por si só, já sejam aproximações muito eficazes.

Embora os resultados obtidos sejam promisores, existem ainda muitos desafios para trabalho futuro. Um desses desafios diz respeito à eficiência do processo. Durante a detecção de duplicados em grandes conjuntos de dados, a avaliação de todos os possiveis pares de registos usando o modelo de classificação seria altamente ineficiente. Porém, dado que a maior parte dos pares não são semelhantes, podemos desenhar técnicas que só seleccionam pares que sejam potencialmente duplicados (e.g., que partilham termos comuns nos seus nomes) como candidatos para o emparelhamento. Trabalhos anteriores exploraram já medidas de similaridade pouco custosas de modo a limitar o número de comparações necessário na utilização de aproximações caras, tais como os modelos de classificação aqui propostos. Um caso particular que tencionamos explorar é o método *Sorted Neighborhood* (SN) [6], cujo funcionamento básico consiste em ordenar os registos usando uma chave, movendo depois uma janela sequencialmente sobre os registos ordenados. Apenas os registos que pertencem à mesma janela são depois emparelhados e incluidos na lista de pares de registos candidatos, melhorando-se assim a eficiência do processo. No caso particular dos nossos registos, onde temos elementos que codificam propriedades geoespaciais, técnicas como a curva de Hilbert podem ser usadas como critério de ordenação [15].

Referências

[1] L. L. Hill, Core elements of digital gazetteers: Placenames, categories, and footprints, in: Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries, 2000.

[2] B. Martins, A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records, Proceedings of the International Conference on GeoSpatial Semantics, 2011.

[3] A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios, Duplicate record detection: A survey, IEEE Transactions on Knowledge and Data Engineering 19 (1), 2007.

[4] W. Cohen, P. Ravikumar, S. Fienberg, A comparison of string distance metrics for name-matching tasks, Proceedings of the ACM Conference Knowledge on Discovery and Data Mining, 2003.

[5] P. Lawrence, The double metaphone search algorithm, C/C++ Users Journal 18 (6), 2000.

[6] D. Lin, An information-theoretic definition of similarity, Proceedings of the International Conference on Machine Learning, 1998.

[7] P. Resnik, Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research 11, 1999.

[8] M. Bilenko, R. J. Mooney, Adaptive duplicate detection using learnable string similarity measures, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, 2006.

[9] M. Bilenko, R. J. Mooney, On evaluation and training-set construction for duplicate detection, Proceedings of the KDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 2003.

[10] V. Sehgal, L. Getoor, P. D. Viechnicki, Entity resolution in geospatial data integration, Proceedings of the International Symposium on Advances on Geographical Information Systems, 2006.

[11] J. T. Hastings, Automated conflation of digital gazetteer data, International Journal Geographic Information Science 22 (10),2007.

[12] A. Samal, S. Seth, K. Cueto, A feature-based approach to conflation of geospatial sources, International Journal of Geographical Information Science 18, 2002.

[13] Y. Zheng, X. Fen, X. Xie, S. Peng, J. Fu, Detecting nearly duplicated records in location datasets, Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010.

[14] R. J. Bayardo, Y. Ma, R. Srikant, Scaling up all pairs similarity search, Proceedings of the International Conference on World Wide Web, 2007.

[15] R. Gutman, Space-filling curves in geospatial applications, Dr. Dobb's Journal of Software Tools 24 (7), 1999.

[16] L. Breiman, Random forests, Machine Learning 45 (1), 2001.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, SIGKDD Explorations Newsletter 11, 2009.