Detecção de Autores Duplicados Utilizando Estrutura de Comunidades em Redes de Cooperação Científica

Breno Júnio V. da Silva[1], **Robson Motta**[2], **Alneu de Andrade Lopes**[3]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP) – São Carlos – SP – Brazil

Abstract. *In collections of scientific papers is frequent to find different quotational names of a same author. For many applications, these duplicate records need to be identified. This is an instance of the problem known as identification of duplicates, for which good results have not been achieved yet. This study investigated the use of scientific cooperation networks and community detection techniques to deal with the problem of identifying duplicates. The results indicate that such strategy not only improves the accuracy of the identification of duplicates, but also reduces the computational cost associated with this task when compared with approaches in which a record is compared against all the others.*

Resumo. *Nas coleções de artigos científicos é frequente encontrar nomes diferentes de citação de um mesmo autor. Para muitas aplicações, estes registros duplicados devem ser identificados. Esta é uma instância do problema conhecido como a identificação de duplicados, para o qual bons resultados não foram alcançados ainda. Este estudo investigou a utilização de redes de cooperação científica e técnicas de detecção da comunidade para lidar com o problema de identificação de duplicados. Os resultados indicam que essa estratégia não só melhora a precisão da identificação de duplicatas, mas também reduz o custo computacional associado a esta tarefa quando comparado com as abordagens em que um registro é comparado com todos os outros.*

1. Introdução

A grande quantidade de dados armazenados em meio digital cresceu consideravelmente nas últimas décadas, dificultando a possibilidade de exploração de conhecimento específico e relevante. Desta forma, diversos trabalhos em diferentes áreas da ciência desenvolvem métodos automáticos para extrair conhecimento a partir de grandes volumes de dados [Fayyad et al. 1996].

Aprendizado de Máquina, subárea de Inteligência Artificial, é uma das áreas de pesquisa que estuda algoritmos e técnicas relacionadas a obtenção automática de conhecimento a partir de dados. O aprendizado de máquina é uma área multidisciplinar envolvendo conhecimentos de estatística, computação, matemática, biologia entre outros [Mitchell 1997]. Para o uso das técnicas de aprendizado de máquina é fundamental definir como os dados e o conhecimento serão representados, bem como o mecanismo de inferência apropriado para a tarefa em questão. A representação dos dados de entrada para a maioria dos algoritmos pode ser dividida em dois modelos: a representação proposicional, na qual os objetos são representados somente pelas suas características; e a

representação relacional, na qual considera as características individuais e as relações existentes entre os objetos [Raedt 2008].

Um grafo (ou rede) pode ser utilizado para representação dos dados de forma relacional. Redes são estruturas compostas por um conjunto de vértices e um conjunto de arestas que ligam os pares de vértices conforme as relações existentes no mundo real ou no sistema em questão [Newman 2003]. Redes Complexas refere-se a uma rede que apresenta uma estrutura de conexões não trivial, como por exemplo, a *World Wide Web* ou as redes sociais como *Orkut*[1], *Facebook*[2] e outras [Reka Albert and Barabasi 1999].

A representação relacional, neste trabalho, é usada para agregar informações sobre cooperação científica entre pesquisadores e usar essa informação para melhorar a detecção de registros duplicados em base de dados. Bases de dados podem conter registros que se referem à mesma entidade no mundo real, porém, não são sintaticamente idênticos. Variações na representação dos dados podem surgir de erros de digitação, erros ortográficos, abreviações, não padronização na entrada de dados e integração de múltiplas bases de dados. Dados extraídos de documentos não estruturados, semi-estruturados ou páginas da *web* tendem a propagar mais ruídos às bases de dados [Bilenko and Mooney 2003]. Neste trabalho, quaisquer pares de registros que sejam sintaticamente diferentes, mas representem o mesmo objeto no mundo real, serão identificados como duplicados.

O problema de registros duplicados se estende a bibliotecas digitais, onde a não utilização de padrões de citações e/ou união de diversas bibliotecas, fazem com que surjam registros duplicados de autores. É comum o mesmo autor aparecer em portais de periódicos com nomes de citação diferentes, por exemplo, “M. E. J. Newman”, “Mark Newman” e “Newman, Mark E.”. Esse problema é tradicionalmente tratado por uma tediosa análise manual dos dados e com técnicas automáticas e semi-automáticas que ainda não alcançam uma precisão desejável. Além disso, detectar um autor duplicado, grafado de formas distintas, é um problema computacionalmente caro. Uma implementação básica, em que são confrontados todos os autores de uma base de dados, seria no mínimo de complexidade computacional de ordem quadrática no número de registros.

Considerando o problema de autores duplicados , o objetivo deste trabalho é usar o formalismo de redes complexas de cooperação científica e de detecção de comunidades para identificar registros duplicados de autores em uma coleção de artigos científicos. A abordagem investigada detecta grupos de vértices fortemente conectados em uma rede de cooperação científica, e os utiliza para reduzir a complexidade do problema, minimizando a quantidade de pares de autores a serem comparados, e melhorar a precisão da detecção de duplicados.

O trabalho está organizado da seguinte forma. Na Seção 2, é apresentada uma revisão dos trabalhos relacionados ao problema. Na Seção 3, a modelagem proposta para simular o problema é apresentada. Na Seção 4, os resultados de cobertura e precisão da técnica são apresentados, bem como sua avaliação da complexidade de tempo. Por fim, na Seção 5 a conclusão e proposta de trabalhos futuros.

[1]http://www.orkut.com
[2]http://www.facebook.com

2. Detecção de Registros Duplicados

Em grandes bases de dados, são facilmente encontrados dados duplicados ou referências erradas. No mundo real, o problema se torna ainda mais complexo quando várias bases de dados são agrupadas. O problema se estende para bibliotecas digitas, pois, a falta de convenções para referências agrava o problema, dificultando obter medidas que caracterizam a base de dados. Medidas tradicionais de similaridade entre cadeias de caracteres podem ser usadas para identificar erros tipográficos e outros tipos de ruídos. No entanto, é difícil identificar quando referências sintaticamente distintas se referem ao mesmo objeto no mundo.

Bhattacharya e Getoor [Bhattacharya and Getoor 2004, Bhattacharya and Getoor 2007] utilizaram um algoritmo iterativo para eliminar registros duplicados em bases de dados a partir dos relacionamentos existentes entre as entidades. A modelagem utilizada procura por evidências que determinam se dois registros idênticos são na verdade distintos, por exemplo, dois registros de nome de pessoas idênticos em uma base de dados cadastrais de pessoas. Essas evidências são informações adicionais a um registro, as quais não são consideradas em abordagens tradicionais que utilizam medidas de aproximação de cadeia de caracteres. Como evidência adicional há, por exemplo, informação de filiação nos registros ou nome do cônjuge.

Algoritmos genéticos foram aplicados para descobrir uma função de similaridade que é capaz de dizer se dois registros são o mesmo objeto no mundo real [de Carvalho et al. 2006]. A motivação para usar tal técnica é a capacidade existente nos algoritmos genéticos de se adaptar aos dados, identificar padrões de informações que não são óbvias e diminuir a interferência do especialista na definição dos parâmetros de entrada dos modelos de detecção dos registros duplicados.

Dong e colegas [Dong et al. 2005] usaram um algoritmo que se baseia na construção de um grafo de dependência. As dependências são informações contidas no próprio trabalho científico, como instituição do autor, endereço eletrônico, grupo de pesquisa entre outros. Os vértices representam as referências e as arestas representam as dependências entre as referências. A construção do grafo visa o uso extensivo de informações de contexto para fornecer evidências necessárias para identificar registros duplicados. Por exemplo, dados duas pessoas, considera-se as relações de co-autoria e o endereço eletrônico para auxiliar na decisão de identificar se são a mesma pessoa.

No contexto da *web*, observou-se que um determinado registro A pode ser um registro duplicado de B, se A aparece frequentemente associado a um tipo de informação, em que B também aparece. Utilizando motores de busca e um conjunto de entidades fornecidas, o algoritmo proposto por Elmacioglu e colegas [Elmacioglu et al. 2007] utiliza uma função de densidade baseada nos resultados obtidos dos motores de busca para identificar os registros duplicados. Diferentemente dos outros modelos propostos, este por sua vez, utiliza informações da *web*, então os resultados variam no tempo conforme ocorrem mudanças na *web*.

Conforme a quantidade de registros aumenta, a natureza quadrática torna algumas abordagens proibitiva para o problema [Kim and Lee 2007]. Diante deste contexto, são propostas diversas formas de abordarem ao problema, considerando-se o relacionamento entre os registros da base de dados, com resultados consideráveis.

De maneira geral, as propostas de soluções para o problema consideram o relacionamento entre os registros, medida de similaridade e grupos ou classes de registros, porém, os métodos necessitam de interferência humana. A proposta de substituir os parâmetros estimados por especialista para métodos dinâmicos que alteram e adaptam os parâmetros aos dados em tempo de execução aumentam as chances de classificar corretamente um registro, porém, não ajuda no problema de custo computacional [Paskalev and Antonov 2006, Paskalev and Antonov 2007].

3. Detecção de Comunidades na Identificação de Duplicados

As bases de dados de artigos científicos são, de uma maneira geral, compostas por documentos não estruturados, ou seja, documentos nos quais não existem etiquetas separando as informações no texto. Para a construção das redes de cooperação científica é necessário extrair informações pertinentes dos documentos. Para isto, a ferramenta IESystem, desenvolvida no ICMC para extração de informações em artigos, se mostrou eficiente [Rossi et al. 2010]. A ferramenta IESystem foi desenvolvida especificamente para extrair metadados (título, autores, resumo, referências e etc) de artigos científicos em diferentes idiomas.

Utilizando um esquema de modelos proposto pelo autor, a ferramenta IESystem transforma o texto não estruturado em um modelo estruturado. Por exemplo, se utilizar o modelo contendo título, autores e referências, o texto é separado por *tags* em que o título é inserido entre as etiquetas <Título> e < \Título>, a lista de autores é inserida em <Autores> e < \Autores> e por último, as referências são inseridas entre <Referências> e < \Referências>.

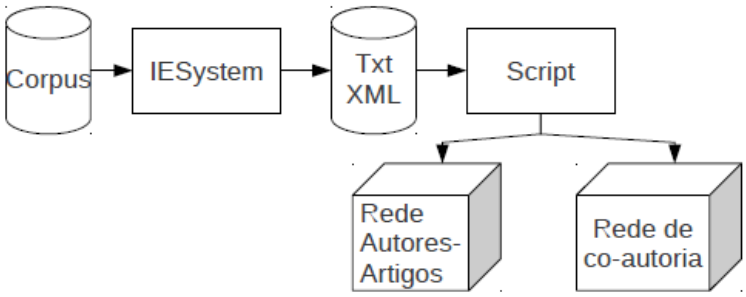

Figura 1. Sequência de atividades para extrair as informações dos artigos científicos e então construir a rede de cooperação científica.

Com as informações já identificadas pela ferramenta, uma nova base de dados parcialmente pré-processada e que mantêm as características de relacionamento original é construída. Para geração da rede de cooperação científica somente as informações referentes aos autores é necessária, sendo que neste trabalho somente os artigos completos estão sendo considerados, eliminando os autores das referências. A Figura 1 sintetiza as etapas do processo utilizado para construção das redes, utilizando *scripts* para processar os dados extraídos pela ferramenta IESystem e construir a rede de cooperação científica, conectando-se dois autores que estão presentes em um mesmo artigo.

A técnica proposta consiste em construir um sistema capaz de identificar autores duplicados em uma rede de cooperação científica, tendo como entrada a coleção de artigos

científicos em formato PDF. Para avaliar a qualidade da ferramenta na execução da tarefa foram utilizadas as medidas cobertura, precisão e *f-measure* [Witten and Frank 2005]. A ferramenta foi construída com a linguagem de programação *Python* [3].

O processo de detecção de autores duplicados é dinâmico e não dispensa a avaliação do especialista. A Figura 2 ilustra o modelo simplificado do funcionamento geral do sistema. Na entrada de dados, o sistema recebe uma rede de cooperação científica (ou co-autoria). Então processa essa rede e sugere ao especialista os possíveis autores duplicados. O especialista analisa os resultados e altera os parâmetros, de forma que se observe melhoria na qualidade dos resultados. Por fim, é possível atualizar a base de dados eliminando os itens duplicados.

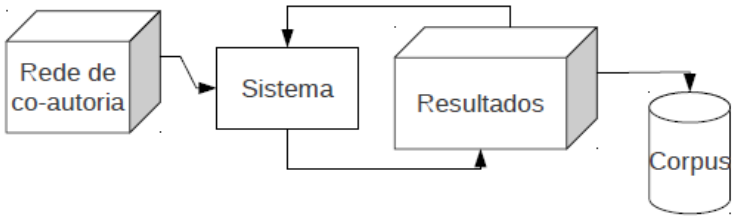

Figura 2. Sequência de atividades executadas pelo sistema para identificar autores duplicados na rede cooperação científica fornecida.

O sistema é dividido em quatro partes. Primeiro são as tarefas de pré-processamento da entrada de dados, em seguida detecção de comunidades na rede fornecida. Terceiro o processamento das comunidades com a medida de similaridade proposta e por fim o pós-processamento dos resultados.

A tarefa de pré-processamento da entrada de dados consiste em atribuir um número identificador (ID) para da cada elemento de entrada. Eliminar caracteres especiais como “pontos” (usados em abreviações) e espaços em branco entre nomes e sobrenomes. A eliminação desses caracteres formam as chaves de busca usadas na identificação dos duplicados, e o ID representa o vértice da rede.

Feito o pré-processamento da entrada de dados, inicia-se a fase de detecção de comunidades da rede. Estrutura de comunidades é uma característica comumente presente em redes complexas. As comunidades são formadas por vértices que são densamente conectados entre si e fracamente conectados a outros grupos [Newman 2004]. Normalmente, essas comunidades conectam vértices com alguma similaridade entre si.

Para detecção de comunidades exige-se como parâmetros uma rede e o número desejado de comunidades. Neste trabalho, utilizou-se o método de detecção de comunidades baseado na remoção de *betweenness* [Girvan and Newman 2002]. Neste método gera-se o caminho mínimo entre todos os pares de vértices na rede, e assume-se que arestas que fazem parte de muitos caminhos mínimos ligam grupos fortemente conectados, ou seja, as comunidades. A remoção iterativa destas arestas gera cada vez mais componentes, que são as comunidades identificadas na rede.

Após a detecção das comunidades, estas podem ser processadas individualmente. Comparando objeto por objeto a fim de identificar elementos duplicados usando a medida

[3]http://www.python.org/

de similaridade entre cadeia de caracteres (Medida de Levenshtein, também conhecida como Distância de Edição). A similaridade é definida em porcentagem sobre o comprimento da cadeia de caracteres. Por exemplo, dada uma cadeia de caracteres A de comprimento 10 e definido um limiar de diferença de 30% (ou limiar de similaridade em 70%), então diz-se que uma outra cadeia de caracteres B é similar a A se for necessário mudar 30% (ou manter 70%) dos caracteres de A para que esse se iguale a B.

Por fim, para cada comunidade é construída uma lista de itens duplicados, considerando o valor definido para o limiar de diferença utilizado na medida de similaridade. A fase de pós-processamento, exige que um especialista avalie os resultados, efetuando, se necessário, alterações no valor do limiar de diferença e no número de comunidades, refazendo o procedimento de identificação de duplicados. Um especialista pode também avaliar os itens duplicados, já os unificando na rede, permitindo, possivelmente, uma melhoria nos resultados futuros.

4. Resultados Obtidos

Para avaliação do modelo proposto, foi realizado um estudo de caso com um conjunto de artigos científicos. O conjunto de dados é formado por 992 artigos, sendo 654 documentos publicados entre 1994 e 2008 na área de *Case-Based Reasoning*, e 338 documentos publicados entre 1997 e 2008, na área de *Inductive Logic Programming*.

Para avaliar o modelo proposto a partir dos resultados obtidos foram utilizados 7 diferentes limiares de diferença para a medida de Levenshtein (20%, 25%, 30%, 35%, 40%, 45% e 50%), e o intervalo de 1 à 20 comunidades para cada limiar de diferença. A Tabela 1 contêm informações complementares da base de dados que se aplicam à rede de cooperação científica, no caso, cada autor é um vértice da rede e cada aresta é uma relação de co-autoria.

Base de Dados	Total de Artigos	Total de Autores
CBR-ILP	992	856

Tabela 1. Informações complementares da base de artigo científicos CBR-ILP.

Devido a dificuldade de saber exatamente quais são os duplicados existente na coleção, para a avaliação da metodologia, nesta base de dados foram inseridos 268 erros controlados, alterando então o número de objetos referentes ao mundo real de 856 para 1124. Os erros inseridos foram de alteração no formato de citação de um mesmo autor e erros de digitação. Erros de citação se referem a aproximadamente 70% dos erros inseridos, por exemplo, ao autor ”Mobyen Uddin Ahmed” são inseridas abreviações para o nome como “Mobyen U. A.” e “M. U. Ahmed”. Erros de digitação correspondem aos outros 30%, por exemplo, ao autor “José Ramon Méndez“, são inseridos erros de digitação ”José Ramom Méndez“ e ”José Ramon Méndes“. A avaliação foi feita com base na precisão e cobertura alcançada pela proposta, considerando o total de erros inseridos.

A primeira configuração para processar a base de dados CBR-ILP foi o limiar de diferença em 20% e quantidade de comunidades de 1 à 20. A média de objetos identificados como duplicados foi 170 e a média dos objetos identificados corretamente como duplicado foi 166, de um total de 268 erros inseridos.

Na Figura 3 observa-se o que no primeiro experimento a precisão é alta nos dados recuperados, porém, a cobertura dos objetos recuperados é baixa, e a variação na quantidade de comunidades não tem influência direta nos resultados. Tais características ocorrem também para os limiares de diferença em 25% e 30%, com uma pequena variação quando considerada uma comunidade ou valores acima de 15 comunidades, porém, conforme incrementa-se o limiar de diferença, consequentemente melhora-se a cobertura e prejudica-se a precisão.

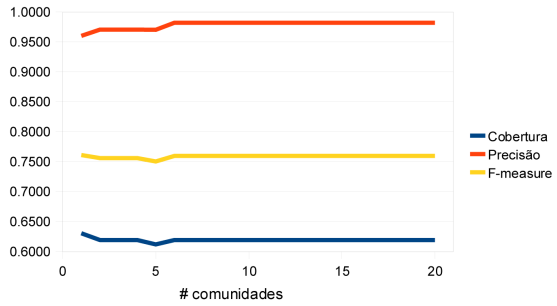

Figura 3. Cobertura, precisão e F-*measure* para o limiar de diferença em 20%.

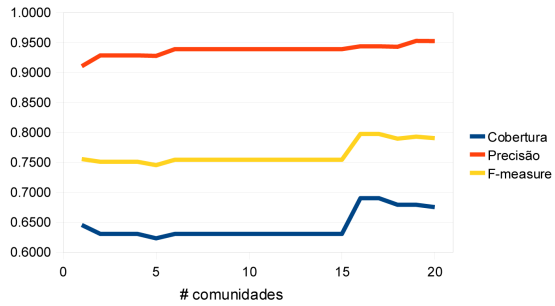

Figura 4. Cobertura, precisão e F-*measure* para o limiar de diferença em 25%.

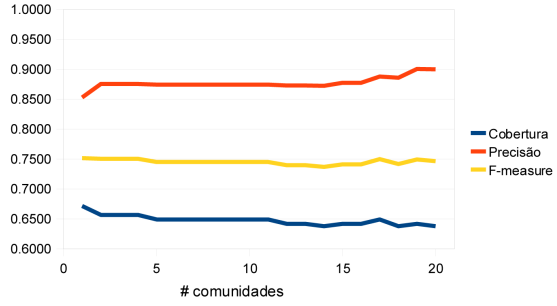

Figura 5. Cobertura, precisão e F-*measure* para o limiar de diferença em 30%.

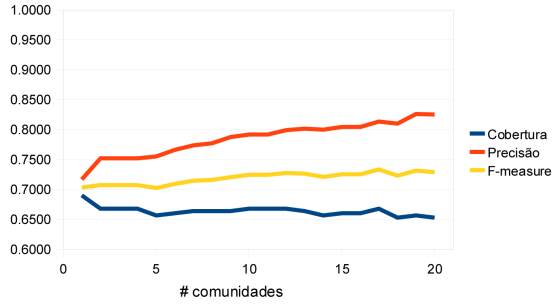

Figura 6. Cobertura, precisão e F-*measure* para o limiar de diferença em 35%.

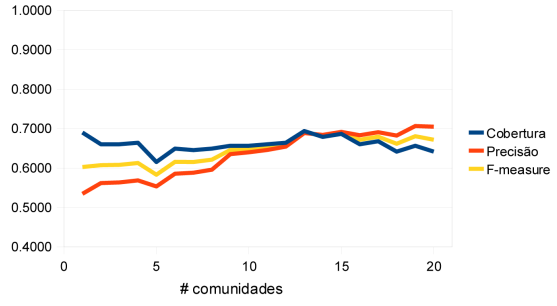

Figura 7. Cobertura, precisão e F-*measure* para o limiar de diferença em 40%.

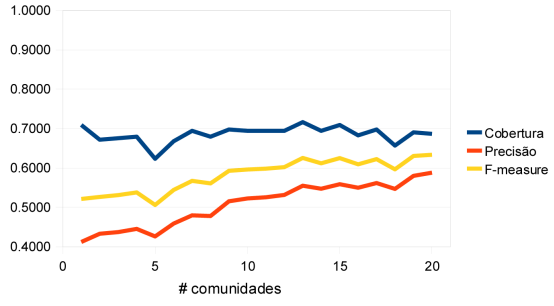

Figura 8. Cobertura, precisão e F-*measure* para o limiar de diferença em 45%.

Considerando os limiares de diferença 35% e 40%, observa-se que conforme aumenta-se o número de comunidades melhora-se a precisão e piora-se a cobertura, porém, a medida *f-measure* demonstra que a melhoria na precisão tem maior relevância

do que a perda da cobertura. Isso demonstra que a comparação somente entre objetos das comunidades apresenta uma melhoria comparada ao método que verifica todos os pares de objetos (equivalente a observar as Figuras 6 e 7 com 1 comunidade).

Mesma característica ocorre para os limiares de diferença 45% e 50%, com a cobertura atingindo altos valores, mas a precisão é prejudicada, ocorrendo muitos falsos positivos, i.é, registros, de fato distintos, são apresentados como possíveis duplicados. Ou seja, foram recuperados uma quantidade de duplicados superior a quantidade de erro inserido, resumida na Tabela 2.

Limiar de diferença	45%	50%
Média de objetos recuperados	366	553

Tabela 2. Objetos recuperados em média para limiares de diferença de 45% e 50%.

Portanto, foi possível observar que os resultados se dividiram em três grupos. O primeiro, contendo os valores para os limiares de diferença iguais a 20%, 25% e 30%, não apresenta ganho de cobertura e precisão no uso das comunidades, comparados com a verificação de todos os pares de objetos nos dados (ou seja, 1 comunidade), mas apresenta ganho no processamento, pois menos pares são comparados. O segundo grupo é para um limiar intermediário, no caso os valores de 35% e 40%, demonstrando um ganho nos resultados com o uso das comunidades. E, por fim, o terceiro grupo contém os valores dos limiares de diferença acima de 40%, o qual apresenta grande cobertura, mas contendo muitos falsos positivos.

Como já comentado, a complexidade computacional para determinar os autores duplicados no modelo tradicional é quadrática, sendo necessário comparar todos com todos, e considerando as comunidades encontradas este tempo é reduzido significativamente, comparando apenas elementos de uma mesma comunidade (Figura 9). Porém, para o método proposto há o custo computacional para identificação de comunidades. O método utilizado, baseado na remoção de *betweenness* [Girvan and Newman 2002], possui custo computacional quadrático no número de vértices, mas há outros métodos, que serão implementados e possuem custo computacional reduzido, por exemplo o método *FastGreedy* [Girvan and Newman 2002] com custo computacional $O(Nlog^2N)$, com N sendo o número de vértices.

Por fim, analisando os resultados obtidos, a capacidade de identificar elementos duplicados na base de dados está relacionada ao erro dominante na base, ou seja, se a base contém erros de digitação no qual geralmente se caracteriza por troca ou falta de poucos caracteres na referência ao autor, o limiar de diferença deve ser pequena. Combinando bases de dados, geralmente, o erro dominante está relacionado a modelos de citação diferentes, então o limiar de diferença deve ser maior.

5. Conclusões

A utilização de detecção de comunidades em redes de cooperação científica para reduzir a complexidade computacional se mostrou útil no estudo de caso realizado. O aumento no número de comunidades reduz o tempo necessário para identificar autores duplicados, porém, aumentar de maneira demasiada diminui a precisão na detecção de duplicados.

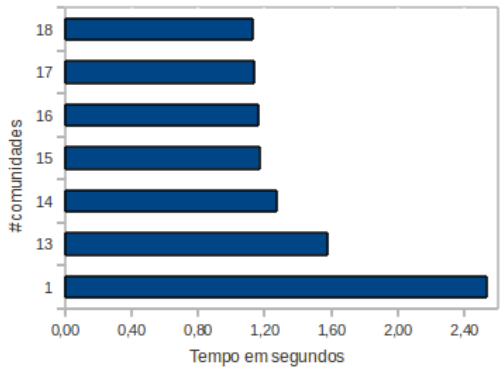

Figura 9. Tempo médio em segundos de execução dos métodos para identificar os autores duplicados na base de dados.

Assim sendo, as redes de cooperação científica tendem a agrupar autores da mesma linha de pesquisa na mesma comunidade. Então, os registros duplicados acompanham essa tendência. Aplicar a medida de similaridade nas comunidades se mostrou eficiente comparado ao método tradicional que compara todos pares de objetos, pois, mesmo conseguindo uma cobertura menor, a precisão é alta nos registros identificados.

As principais contribuições deste trabalho foram a abertura para uma nova abordagem ainda não explorada, até onde foi pesquisado em trabalhos relacionados, de uso de redes de cooperação científica e desenvolvimento do método de identificação de comunidades na detecção de duplicados. E, por fim, a aplicação da técnica de detecção de duplicados baseada nas comunidades, embora mantenha a complexidade computacional quadrática, ela passa a ser quadrática no número médio de elementos das comunidades.

Como proposta de trabalhos futuros, primeiramente torna-se necessário uma avaliação considerando uma grande quantidade de conjunto de dados. A utilização de outros modelos de detecção de comunidades [Clauset et al. 2004], principalmente para redução da complexidade do problema, permitindo escalabilidade do algoritmo. Além disso, verificou-se que é possível realizar uma comparação mais eficiente entre os pares de autores de uma mesma comunidade, considerando não somente a medida de similaridade utilizada, mas também comparando, por exemplo, as abreviações presentes no nome de um autor com os nomes completos de outro autor.

Além disso, as redes de cooperação científica foram construídas baseadas no relacionamento entre autores em trabalhos científicos. Outros modelos de rede podem ser utilizados, como modelos que consideram similaridade entre tópicos abordados nos artigos. Por fim, pode-se considerar aplicar processamento pararelo entre as comunidade, pois não existe relacionamento direto durante o processamento dos duplicados entre comunidades.

Referências

Bhattacharya, I. and Getoor, L. (2004). Iterative record linkage for cleaning and integration. In *DMKD '04: Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 11 – 18, New York, NY, USA. ACM.

Bhattacharya, I. and Getoor, L. (2007). Collective entity resolution in relational data. *ACM Trans. Knowl. Discov. Data*, 1:5.

Bilenko, M. and Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 39–48, New York, NY, USA. ACM.

Clauset, A., Newman, M., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(1):066111.

de Carvalho, M. G., Gonçalves, M. A., Laender, A. H. F., and da Silva, A. S. (2006). Learning to deduplicate. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 41–50, New York, NY, USA. ACM.

Dong, X., Halevy, A., and Madhavan (2005). Reference reconciliation in complex information spaces. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 85–96, New York, NY, US. ACM.

Elmacioglu, E., Kan, M.-Y., Lee, D., and Zhang, Y. (2007). Web based linkage. In *WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 121–128, New York, NY, USA. ACM.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communication of the ACM*, 39(11):27–34.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS*, 99(12):7821–7826.

Kim, H.-s. and Lee, D. (2007). Parallel linkage. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 283–292, New York, NY, USA. ACM.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38:321–330.

Paskalev, P. and Antonov, A. (2006). Intelligent application for duplication detection.

Paskalev, P. and Antonov, A. (2007). Increasing the performance of an application for duplication detection. In *CompSysTech '07: Proceedings of the 2007 international conference on Computer systems and technologies*, pages 1–8, New York, NY, USA. ACM.

Raedt, L. D. (2008). *Logical and Relational Learning: From ILP to MRDM (Cognitive Technologies)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Reka Albert, H. J. and Barabasi, A. L. (1999). Diameter of the world-wide web. *Nature*, 401:130–131.

Rossi, R. G., Rezende, S. O., and de Andrade Lopes, A. (2010). Sistema para extração de informações de artigos científicos - iesystem. Technical Report 354, ICMC, São Carlos - SP.

Witten, I. H. and Frank, E. (2005). *Data Mining Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA.