Uma Estratégia para Seleção de Atributos Relevantes no Processo de Resolução de Entidades

Gabrielle K. Canalle, Bernadette F. Lóscio, Ana Carolina Salgado

[1]Programa de Pós Graduação em Ciências da Computação - Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)

{gkc, bfl, acs}@cin.ufpe.br

Abstract. *Data integration is an essential task for achieving a unified view of data stored in heterogeneous and distributed sources. A key step in this process is the Entity Resolution, which consists of identifying instances that refer to the same real-world entity. Functions that evaluate the similarity between values of attributes are used to identify equivalent instances. This work proposes a strategy for selection of relevant attributes to consider in the instance matching phase in the process of entity resolution. This strategy employs characteristics from attributes, such as the quantity of duplicated and null values, aiming to identify the most relevant to the instance matching process.*

Resumo. *Integração de Dados é um processo essencial quando deseja-se obter uma visão unificada de dados armazenados em fontes de dados heterogêneas e distribuídas. Uma etapa crucial desse processo é a Resolução de Entidades, que consiste em identificar instâncias que se referem à mesma entidade do mundo real. Para a descoberta de instâncias equivalentes, são usadas funções que avaliam a similaridade entre os valores dos atributos que as descrevem. Este trabalho propõe uma estratégia para seleção de atributos relevantes a serem considerados na fase de comparação de instâncias do processo de Resolução de Entidades. A estratégia proposta utiliza características dos atributos, como a quantidade de valores repetidos e valores nulos, a fim de identificar os mais relevantes para o processo de comparação de instâncias.*

1. Introdução

As soluções de Integração de Dados visam combinar dados residentes em diferentes fontes provendo aos usuários uma visão unificada desses dados. Uma etapa importante no processo de Integração de Dados é a de Resolução de Entidades [Christen 2012], que busca identificar a equivalência entre instâncias que representam uma mesma entidade do mundo real [Dong and Srivastava 2015]. A Resolução de Entidades é composta de várias fases, incluindo uma etapa de comparação entre pares de instâncias. Durante essa etapa, é avaliada a similaridade entre os valores dos atributos que descrevem as instâncias que estão sendo comparadas. Nesse contexto, um dos principais desafios a serem enfrentados pelo processo de Resolução de Entidades diz respeito à escolha dos atributos que serão utilizados na fase de comparação.

Para ilustrar a necessidade de selecionar os melhores atributos a serem considerados na fase de comparação do processo de Resolução de Entidades, e como esses atributos impactam na qualidade do processo, considere o exemplo a seguir: *Um serviço de biblioteca digital online, no domínio de Ciências da Computação, que integra dados de*

múltiplas fontes, tais como CiteSeerX[1] e DBLP[2], e possibilita a realização de pesquisas por título, autor, ou palavra-chave. Suponha que um usuário esteja interessado em buscar artigos sobre "Integração de Dados". O serviço de integração submete a consulta para as fontes de dados CiteSeerX e DBLP, e obtém um conjunto de artigos. Uma pequena fração desse resultado pode ser vista na Tabela 1.

Tabela 1. Resultado de uma consulta nas fontes CiteSeerX e DBLP.

ID Paper	ID	Fonte	Author	Title	Year	Venue	Pages
1	1	CiteSeerX	M. Lanzerini	Data Integration: A Theorical Perspective (2002)	2002	Symposium on Principles of Database Systems	NULL
	2	DBLP	Maurizio Lanzerini	Data Integration: A Theorical Perspective	2002	PODS 2002	233-246
2	3	DBLP	Guy Pierra	The PLIB ontology-based approach to data integration	2004	IFIP Congress Topical Sessions	13-18
3	4	CiteSeerX	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - all problems solved	NULL	In 18th IFIP Computer Congress (WCC)	NULL
	5	DBLP	Patrick Ziegler and Klaus R. Dittrich	Three decades of data integration - All problems solved?	2004	IFIP Congress Topical Sessions	NULL

Suponha que a Resolução de Entidades será realizada considerando todos os atributos que descrevem as instâncias. Possivelmente, as instâncias 1 e 2 seriam dadas como não duplicadas, já que dos cinco atributos considerados, dois (*Venue* e *Pages*) possuem uma similaridade igual a 0. O mesmo aconteceria com as instâncias 4 e 5, em que os atributos *Year* e *Pages* também possuem similaridade igual a 0. Com isso, podemos observar que atributos contendo valores nulos podem afetar negativamente o processo de Resolução de Entidades. Isso acontece porque um valor nulo na comparação ocasiona em uma similaridade igual a 0, podendo fazer com que duas instâncias sejam dadas como distintas mesmo sendo correspondentes.

Agora, considere que o subconjunto de atributos *Year* e *Venue* foi selecionado aleatoriamente, sem considerar os valores dos atributos. Provavelmente, as instâncias 3 e 5 seriam consideradas duplicadas, já que os valores desses atributos possuem uma alta similaridade. Dessa forma, pode-se observar que atributos com valores repetidos também podem afetar negativamente o processo de Resolução de Entidades. Isso acontece porque um valor repetido pode contribuir para aumentar o valor de similaridade, o que pode fazer com que duas instâncias sejam dadas como correspondentes mesmo sendo distintas.

Na literatura existem alguns trabalhos que tratam este problema, dentre eles, destacam-se [Chen et al. 2012] e [Su et al. 2010]. Chen et al. propõem um método para Resolução de Entidades baseado em Aprendizagem de Máquina. Esse método busca, por meio de um conjunto de treinamento, encontrar o melhor grupo de atributos para ser utilizado na fase de comparação do processo de Resolução de Entidades. Em Su et al. é feita a detecção de duplicados para resultados de consultas em fontes de dados *Web*. Neste cenário, o trabalho foca em um algoritmo de Aprendizagem de Máquina que tem como objetivo ajustar os pesos dos atributos para o cálculo de similaridade. O algoritmo é capaz de aprender a ajustar os pesos dos atributos utilizando uma amostra de dados que contém instâncias não correspondentes provenientes de diversas fontes de dados. No entanto, utilizar algoritmos de Aprendizagem de Máquina sobre um conjunto de treinamento pode tornar o processo custoso. Além do mais, a definição de um conjunto de treinamento não é uma tarefa trivial, uma vez que o usuário nem sempre tem o conhecimento prévio do domínio.

[1]http://citeseerx.ist.psu.edu
[2]http://dblp.uni-trier.de

Diferentemente de Chen et al. e Su et al., este trabalho propõe uma estratégia de Seleção de Atributos em que a avaliação de relevância dos atributos é realizada por meio de critérios relacionados aos dados e por meio de medatados de qualidade relacionados às fontes. Um atributo é considerado relevante se contribui positivamente para a identificação de correspondências verdadeiras, e irrelevante se contribui na identificação de correspondências erradas (falsos positivos e falsos negativos). Com o uso da estratégia proposta, foi possível alcançar bons resultados na comparação de instâncias do processo de Resolução de Entidades, ou seja, os atributos dados como relevantes foram aqueles que mais contribuíram para encontrar o maior número de correspondências verdadeiras, com o menor número de correspondências erradas. O restante deste trabalho está organizado como segue. Na Seção 2, especificamos a estratégia de Seleção de Atributos proposta neste trabalho. A Seção 3 apresenta os resultados obtidos por meio de experimentos realizados. E, por fim, a Seção 4 apresenta as considerações finais e indicações para trabalhos futuros.

2. Estratégia para Seleção de Atributos Relevantes no Processo de Resolução de Entidades

Para a descrição da estratégia proposta, considere um conjunto de fontes de dados $F = \{f_1, ..., f_n\}$, que oferece um conjunto de dados C, tal que as instâncias de C podem ser provenientes das múltiplas fontes que compõem F. C contém um conjunto de entidades $C.E = \{E_1, ..., E_m\}$, que representam conceitos do mundo real. Cada entidade E é descrita por um conjunto de atributos $E.A = \{A_{i1}, ..., A_{in}\}$, e contém um conjunto de instâncias $\{i_1, i_2..., i_k\}$ denotado por $E.I$, tal que cada i_j é definida por um conjunto de pares $\{(A_{i1}, v_{i1}), ..., (A_{in}, v_{in})\}$, onde $A_i \in E.A$, e v_i é o valor de A_i para a entidade E_i.

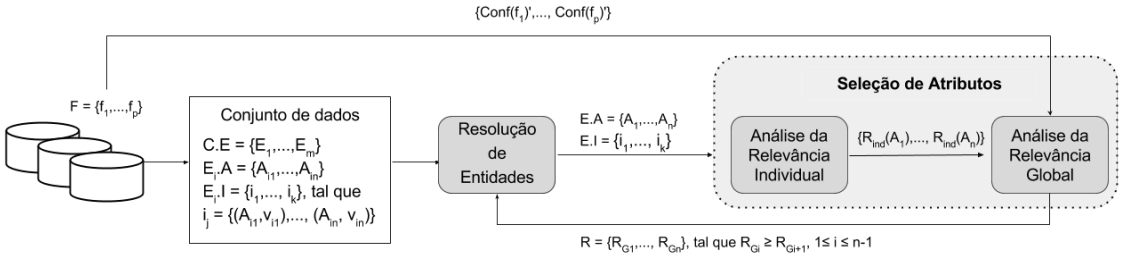


Figura 1. Visão geral da Estratégia de seleção de atributos.

A entrada da seleção de atributos é um conjunto de instâncias $E.I$ correspondentes a uma entidade. Para cada atributo $A_i \in E.A$, será calculada a sua relevância. Nesse sentido, a estratégia que estamos propondo consiste de duas etapas: A **(i)** Análise da Relevância Individual dos Atributos; **(ii)** Análise da Relevância Global dos Atributos, conforme mostra a Figura 1.

Para avaliar a Relevância Individual dos Atributos, são consideradas algumas características relacionadas aos dados, como **Repetição** e **Densidade**. Para avaliar a Relevância Global dos Atributos, são levados em consideração aspectos relacionados às fontes de dados contidas em F, como **Confiabilidade** e **Cobertura**.

A Repetição de um atributo A_i, denotada por $Rep(A_i)$, é dada pela quantidade de vezes que um mesmo valor para um atributo aparece no conjunto de dados. Neste artigo, o valor de Repetição é calculado por meio da divisão entre o total de valores distintos

de A_i, e o número total de valores de A_i. Para identificar se os valores de um atributo são distintos, é utilizada uma função de similaridade. Existem inúmeras funções propostas na literatura [Christen 2012]. Neste trabalho, foi adotada a função de similaridade de *Levenshtein*, ou *edit distance*, por ser uma das funções mais utilizadas quando se deseja comparar *strings* relativamente pequenas e que não precisam necessariamente ter o mesmo tamanho.

A Densidade de um atributo A_i, denotada por $Den(A_i)$, é dada pelo percentual de valores não nulos contidos no conjunto de valores que descreve esse atributo [Naumann and Freytag 2000]. A ausência de valor em um atributo pode fazer com que duas instâncias que são correspondentes sejam dadas como distintas, uma vez que ao se comparar atributos com valores ausentes a similaridade é igual a 0, contribuindo para a identificação de falsos negativos, ou seja, instâncias similares classificadas como distintas. Em nossa abordagem, o valor de $Den(A_i)$ é calculado por meio da divisão entre o total de valores não nulos de A_i, e o total de valores de A_i.

A Relevância Individual, denotada por $R_{ind}(A_i)$ é calculada com base nos critérios de Repetição e Densidade, de acordo com a Equação 1. Para o cálculo, é necessário atribuir um peso, denotado por p, para cada um dos critérios avaliados, que deve ser feito conforme o grau de importância do critério para o cálculo da Relevância Individual, de tal forma que a soma dos pesos deve ser igual a 1.

$$R_{ind}(A_i) = Den(A_i) * p(Den) + (1 - Rep(A_i) * p(Rep) \tag{1}$$

A Confiabilidade de uma fonte, denotado por $Conf(f_k)$, diz respeito ao grau em que os dados fornecidos por ela são verídicos e confiáveis. Dessa forma, semelhante a [Mihaila et al. 2000], assumimos que as fontes possuem metadados de qualidade associados a elas, onde o valor de $Conf(f_k)$, tal que $0 \leq Conf(f_k) \leq 1$, pode ser extraído por meio desses metadados.

A Cobertura de uma fonte de dados, denotada por $Cob(f_k)$, é definida pelo percentual de instâncias que ela fornece para C [Naumann and Freytag 2000]. Para calcular $Cob(f_k)$, é necessário dividir o total de instâncias que uma fonte fornece para C, pelo total de instâncias contidas em C.

Uma vez que temos os valores dos critérios de Confiabilidade e Cobertura para cada fonte de dados $f_k \in F$, a qualidade do conjunto de fontes de dados F, é dada pelo somatório da multiplicação entre os valores de $Conf(f_k)$ e $Cob(f_k)$. Após calcular o valor de qualidade do conjunto de fontes de dados F, a Relevância Global, denotada por $R_{glob}(A_i)$, pode ser calculada conforme a Equação 2.

$$R_{glob}\left(A_i\right) = R_{ind}\left(A_i\right) * \left(\sum_{k=1}^{|f|} Conf(f_k) * Cob(f_k) \right) \tag{2}$$

3. Avaliação Experimental

Nossa estratégia foi avaliada em quatro cenários, criados a partir da base de dados Cora[3]. Essa base de dados contém 1.879 instâncias de diferentes fontes de dados, que fazem referência a produções literárias. Os cenários foram criados para avaliar o comportamento

[3]http://hpi.de/fileadmin/user_upload/fachgebiete/naumann/projekte/dude/CORA.xml

da estratégia em diferentes porcentagens de duplicação, uma vez que a base de dados original possui aproximadamente 90% de duplicação. Os cenários foram divididos da seguinte maneira: **Cenário 1.** 5% - 10%; **Cenário 2.** 15% - 30%; **Cenário 3.** 34% - 50%; **Cenário 4.** 55% - 70%.

O objetivo desse experimento foi validar as seguintes hipóteses: **H1.** Considerar todos os atributos na fase de comparação do processo de Resolução de Entidades ocasiona em um resultado com um baixo *F-measure*[4]; **H2.** Considerar os atributos mais relevantes, de acordo com a classificação realizada pela estratégia proposta, na fase de comparação do processo de Resolução de Entidades faz com que o resultado obtido tenha um alto *F-measure*; **H3.** À medida que atributos menos relevantes são adicionados ao grupo de atributos considerados na fase de comparação do processo de Resolução de Entidades, o número de correspondências erradas aumenta, diminuindo o *F-measure* do resultado.

Para validar as hipóteses elencadas acima, consideramos o resultado obtido por meio da Resolução de Entidades realizada nos diferentes cenários, a partir dos atributos escolhidos por meio da classificação de relevância proposta. Pasa isso, escolhemos a ferramenta DuDe [5] e o algoritmo de resolução *Naive Duplicate Detection*. Para o cálculo da similaridade, adotamos a função de Levenshtein, com um limiar de 0.8. A base de dados Cora possui um *Gold Standard*, que integrado ao DuDe possibilitou que o *F-measure* fosse calculado automaticamente ao fim do processo.

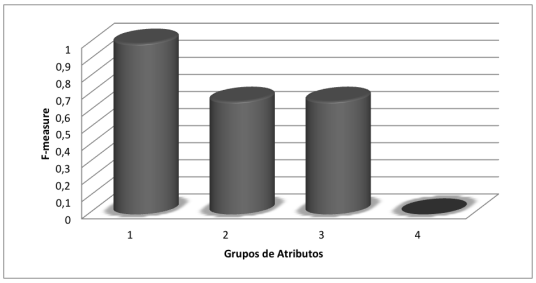
De acordo com [Chen et al. 2012], um único atributo pode não ser suficiente para a fase de comparação. Dessa forma, deve-se utilizar um conjunto de atributos para comparar os pares de instâncias. Sendo assim, com base na classificação obtida por meio da estratégia proposta, consideramos quatro grupos de atributos. O **grupo 1** contendo os dois atributos mais relevantes, o **grupo 2** contendo os três atributos mais relevantes, o **grupo 3** contendo os quatro atributos mais relevantes, e por fim o **grupo 4**, contendo os 8 atributos mais relevantes. Para cada cenário, o processo de Resolução de Entidades foi executado utilizando os quatro grupos descritos acima. Os resultados obtidos [6] são apresentados nos gráficos da Figura 2. O eixo x de cada gráfico representa os grupos de atributos e o eixo y representa os valores de *F-measure*.

A partir da análise dos resultados, podemos concluir que a nossa estratégia se mostrou eficiente em todos os cenários. Além disso, confirmamos que utilizar uma grande quantidade de atributos no processo de Resolução de Entidades não é viável, como mostram os resultados obtidos utilizando o conjunto de atributos do grupo 4. Também foi possível observar que o maior *F-measure* do processo de Resolução de Entidades foi obtido utilizando o grupo 1, contendo apenas os dois atributos mais relevantes. Verificamos também que, à medida que atributos com menor valor de relevância são considerados na comparação, o *F-measure* diminui. Assim sendo, os resultados obtidos por meio dos experimentos validaram nossas hipóteses.

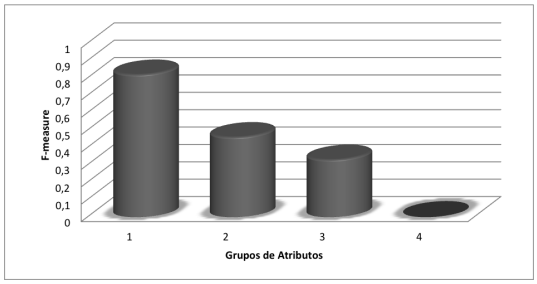
[4]De acordo com [Christen 2012], essa medida é considerada como a mais recomendada para avaliar a qualidade do processo de Resolução de Entidades

[5]http://hpi.de/naumann/projects/data-quality-and-cleansing/dude-duplicate-detection.html#c115302

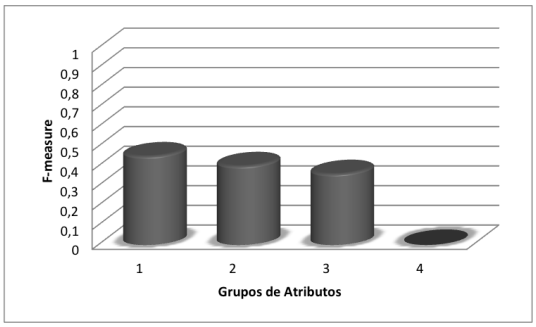
[6]Na avaliação experimental, apenas o resultado de relevância individual dos atributos é apresentado. Os resultados da avaliação da relevância global não foram apresentados por motivo de espaço.



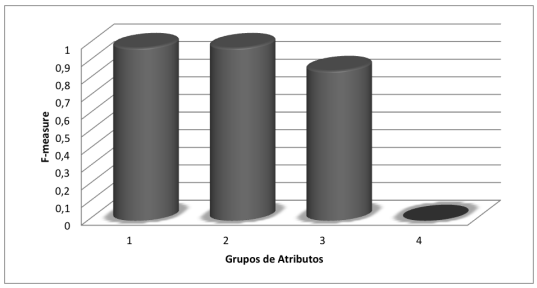
(a) Cenário 1



(b) Cenário 2



(c) Cenário 3



(d) Cenário 4

Figura 2. Resultado da Resolução de Entidades considerando os grupos de atributos elencados.

4. Conclusão

Neste trabalho, propomos uma estratégia para Seleção de Atributos relevantes no processo de Resolução de Entidades. Para avaliar nossa estratégia, realizamos experimentos com a base de dados Cora. Os experimentos comprovaram que, aplicando nossa estratégia foi possível classificar corretamente os atributos de acordo com sua relevância. Como trabalhos futuros, pretendemos realizar experimentos com outras bases de dados para analisar como a estratégia se comportará. Também pretendemos incluir outros critérios no processo de seleção de atributos.

Referências

Chen, J., Jin, C., Zhang, R., and Zhou, A. (2012). A learning method for entity matching. In *In Proceedings of 10th International Workshop on Quality in Databases*, East China Normal University, China.

Christen, P. (2012). *Data Matching*. Springer, Heidelberg.

Dong, X. L. and Srivastava, D. (2015). *Big Data Integration*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.

Mihaila, G. A., Raschid, L., and Vidal, M.-E. (2000). Using quality of data metadata for source selection and ranking. In *WebDB (Informal Proceedings)*, pages 93–98.

Naumann, F. and Freytag, J. C. (2000). Completeness of information sources. Technical report, Humboldt University of Berlin.

Su, W., Wang, J., Lochovsky, F. H., and Society, I. C. (2010). Record Matching over Query Results from Multiple Web Databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(4):578–589.