Uma Abordagem Semântica para Detecção de Linhas Duplicadas em Banco de Dados

Priscilla Kelly Machado Vieira[1], Carlos Eduardo S. Pires[2]

[1]Mestranda do Programa de Pós-Graduação em Informática – PPGI - UFPB. e-mail: priscillakmv@gmail.com
[2]Prof. Doutor do Depto. de Sistemas e Computação, UFCG, e-mail: cesp@dsc.ufcg.edu.br

Resumo: A descoberta de conhecimento em bancos de dados é um processo não trivial de identificar em dados padrões que sejam válidos, novos, potencialmente úteis e compreensíveis, visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão. Dentre as inúmeras etapas envolvidas neste processo, destacamos a de limpeza de dados, cujo objetivo principal é melhorar a qualidade dos dados de entrada e, assim, aumentar a qualidade do conhecimento obtido. A qualidade de dados pode ser aprimorada por meio de operações como detecção de dados duplicados, remoção de ruídos, manipulação de campos de dados ausentes, formatação de dados, entre outras. Este trabalho tem por objetivo propor uma abordagem semântica para detecção de linhas duplicadas em bancos de dados. Ontologias de domínio são utilizadas como recurso externo para tentar determinar o tipo de relacionamento semântico entre dados. Uma ferramenta que implementa a abordagem semântica proposta foi desenvolvida e experimentos foram realizados.

Palavras–chave: Dados Duplicados, Descoberta de Conhecimento, Limpeza de Dados, Ontologia, Regras Semânticas

1. INTRODUÇÃO

As organizações públicas e privadas começam, finalmente, a perceber o valor dos dados que possuem a sua disposição, e a considerá-los como um bem importante no aumento da produtividade, eficiência e competitividade de mercado (INMON, 1998). Como consequência, a exploração de enormes volumes de dados assume um papel cada vez mais importante na sociedade atual. No entanto, constata-se que boa parte dos dados armazenados apresenta erros ou anomalias (por exemplo, chaves duplicadas, valores ausentes, valores nulos, entre outros), o que dificulta a extração correta de informações (BALLOU, 1999). É neste contexto que surge a importância da limpeza de dados, que visa detectar e remover anomalias nos dados com o objetivo de melhorar sua qualidade.

Os problemas de qualidade podem surgir em conjuntos de dados isolados, como arquivos e bases de dados, sendo ainda mais críticos quando múltiplas fontes de dados necessitam ser integradas (ADRIAANS, 1996). Isto acontece em virtude das diversas fontes conterem dados redundantes sob diferentes representações. De modo a possibilitar um acesso preciso e consistente aos dados, é necessário consolidar as diferentes representações, detectar e eliminar possíveis duplicações.

Para a tentativa de detecção de dados duplicados, é comum a utilização de técnicas linguísticas, ou seja, técnicas que analisam os dados com base em sua grafia (OLIVEIRA, 2008). São exemplos de técnicas linguísticas: (i) comparação numérica: determina se dois valores numéricos são iguais ou não; (ii) distância de edição: considera o número de alterações que devem ser realizadas para transformar um dado em outro; (iii) correspondência de nomes: identifica correspondências entre os dados devido ao uso de abreviaturas; e (iv) soundex: os dados são comparados em função do som das suas pronúncias. Como os dados são providos de semântica, tais técnicas normalmente não apresentam resultados satisfatórios em termos de precisão (OLIVEIRA, 2008). Neste sentido, a combinação de técnicas

ISBN 978-85-62830-10-5

linguísticas e semânticas pode ajudar a melhorar os resultados obtidos na detecção de linhas duplicadas.

O uso de técnicas semânticas ainda é pouco explorado pelas principais ferramentas de detecção de linhas duplicadas. Este trabalho propõe uma abordagem baseada em semântica para detecção de linhas duplicadas em tabelas de bancos de dados. Regras semânticas são propostas para determinar o relacionamento semântico (e.g. equivalência e especialização) entre os dados com base em uma ontologia de domínio. Tais regras são incorporadas a um processo de detecção de linhas duplicadas baseado inicialmente em técnicas linguísticas no sentido de obter melhores resultados. Para validar a proposta, uma ferramenta de detecção de linhas duplicadas foi desenvolvida. Os experimentos mostraram que, com o estudo semântico dos dados, o número total de dados indicados erroneamente como duplicados pode ser atenuado. Também foi observado que alguns dados linguisticamente diferentes puderam ser detectados como duplicados após a associação da semântica.

O restante deste trabalho está estruturado da seguinte forma: Inicialmente, apresentamos os Materiais e Métodos, descrevendo o processo utilizado para a detecção de linhas duplicadas, enfatizando as regras semânticas propostas. Em seguida, são descritos os resultados obtidos e realizadas discussões sobre os mesmos. Por fim, são apresentadas as conclusões e sugestões de trabalhos futuros.

2. MATERIAL E MÉTODOS

Entender e conhecer as anomalias que podem ocorrer nos dados de um banco de dados, assim como suas origens e formas de resolução, é uma tarefa importante no processo de descoberta de conhecimento. Neste sentido, Boscariolli (BOSCARIOLI, 2005) dá uma visão geral da importância e métodos para a descoberta de conhecimento em banco de dados e indica processos de pré-processamento de dados como, por exemplo, estimar parâmetros ausentes ou ignorar dados. Na nossa proposta, assume-se que os dados passaram por uma etapa de pré-processamento.

Inicialmente, foi realizado uma análise sobre as funcionalidades oferecidas pelas principais ferramentas de limpeza de dados existentes. Este foi utilizado para detectar as deficiências destas ferramentas. Nesta etapa, foi possível observar a escassez de soluções semânticas na detecção de dados duplicados em tabelas de bancos de dados (OLIVEIRA, 2008). Das ferramentas analisadas: (i) Arktos (OLIVEIRA & RODRIGUES, 2004), (ii) Intelliclean (OLIVEIRA & RODRIGUES, 2004), (iii) Data Macth 2010 (DATA, 2012) e (iv) Duplicate Record Remover (DUPLICATE, 2012), não foi identificado o uso de técnicas semânticas. Muitas delas utilizam as técnicas de "fuzzificação" (DUPLICATE, 2012) e vizinhança ordenada (OLIVEIRA, 2008), no entanto apenas com análises linguísticas (DATA, 2012). A Tabela 1 sintetiza o comparativo entre as ferramentas analisadas.

Tabela 1 – Quadro comparativo entre as ferramentas analisadas

Nome	Uso de Semântica	Técnica	Privada
Arktos	Não	Vizinhança ordenada	Não
Inteliclean	Não	Vizinhança ordenada	Não
Data Macth	Não	Vizinhança ordenada	Sim
Duplicate Record Remover	Não	Fuzzificação	Sim

ISBN 978-85-62830-10-5

Com base nos trabalhos e ferramentas analisadas, observou-se a importância da detecção de dados duplicados no processo de descoberta de conhecimento. Sendo assim, propomos uma melhoria desta detecção com um estudo semântico dos dados. A seguir, apresentamos a metodologia utilizada para detecção de dados duplicados.

Processo de Detecção de Linhas Duplicadas

Por não considerarem o significado dos dados, técnicas baseadas apenas na grafia dos dados não são suficientes para a detecção satisfatória de dados duplicados (OLIVEIRA, 2008). Em muitos casos, dados linguisticamente diferentes podem ser considerados equivalentes no nível semântico. Para exemplificar, considere a Tabela 1 que mostra os dados da tabela Employee (Empregado). Esta tabela contém informações sobre funcionários de uma organização acadêmica indicando nomes, cargos, nacionalidades e sua colocação dentro da organização. A coluna LINE foi adicionada apenas para facilitar o processo de referência dos dados da tabela ao longo do trabalho. Podemos notar que os dados ("FullProfessor" e "Professor") das linhas 1 e 2 pertencentes à coluna TYPE, não apresentam nenhuma similaridade linguística. Entretanto, semanticamente, esses dados possuem um relacionamento do tipo especialização no qual todo "FullProfessor" é um tipo de "Professor".

Tabela 2 - Dados da tabela *Employee*

LINE	REGISTRATION	NAME	TYPE	ORGANIZATION	NATIONALITY
1	20789809	Full Professor	Full Professor	University	El Paso
2	89098796	John M. Zurembrg	Professor	Organization	Marshal
3	20789808	Paul Lennon	Research Assistant	University	Texas
4	20789807	John Paul	Student	University	Arizona

A Figura 1 oferece uma visão geral do processo de detecção de linhas duplicadas proposto neste trabalho. O processo recebe como entrada as linhas de uma tabela de banco de dados e realiza comparações duas a duas com base em técnicas linguísticas e semânticas. Em cada comparação são produzidos dois graus de similaridade, sendo um linguístico e outro semântico. Cada técnica de comparação recebe um peso fornecido pelo usuário, de acordo com a importância da técnica no processo de detecção. A similaridade global entre duas linhas é obtida por meio da combinação entre os graus de similaridade linguístico e semântico. Caso o valor de similaridade seja maior do que um determinado limiar (também definido pelo usuário), as duas linhas são incluídas no resultado final do processo, indicando ao usuário a possibilidade de serem linhas duplicadas. Os parâmetros de controle incluem o peso de cada técnica e o limiar de similaridade.

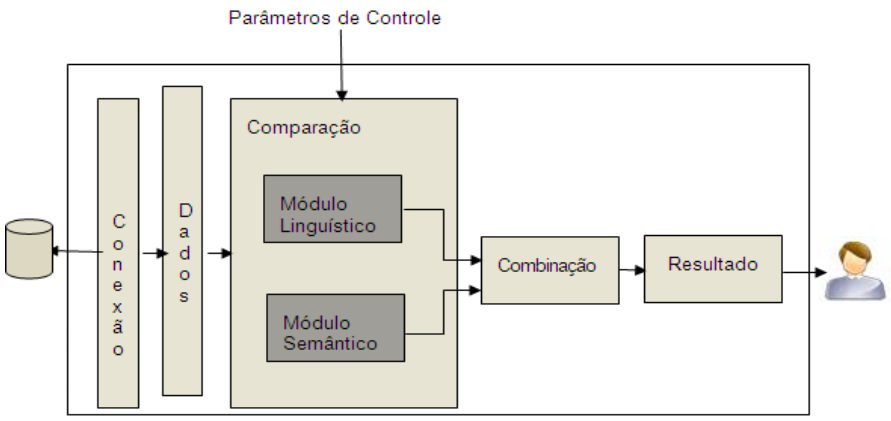

Figura 1- Processo de Detecção de Linhas Duplicadas

ISBN 978-85-62830-10-5

Para descrever o processo de detecção de linhas duplicadas, vamos usar o seguinte formalismo:
- Uma tabela T é uma estrutura formada por linhas e colunas: T = (L, C);
- L é um conjunto finito de linhas: L = {L1...Ln};
- C é um conjunto finito de colunas: C = {C1...Cm};
- Cada coluna Cm ∈ C é identificada por um nome distinto: Cm.nome;
- A interseção de linha Ln com uma coluna Cm corresponde a uma célula (Cl) que contém o dado D(n, m).

Módulo Linguístico

A comparação linguística entre dois dados é feita utilizando a "distância de Levenshtein" (DL) (BUETTCHER et *al*., 2010), que mede a quantidade de modificações necessárias para transformar um dado em outro. Consideramos que os dados são representados no formato textual (string). Por exemplo, calculando a "distância de Levenshtein" entre os dados T(L1, C3.NAME) e T(L2, C3.NAME) da Tabela 1 ("John Mark Zuremberg" e "John M. Zurembrg", respectivamente), a distância de Levenshtein retorna o valor 4 (quatro), indicando que são necessárias quatro modificações para transformar a primeira na segunda. O valor retornado pela Distância de Levensthein pode ser normalizado dividindo-se o resultado pelo tamanho do maior dado. Nesse caso, o resultado é sempre um valor entre o intervalo [0,1]. Para obter o valor de similaridade subtraímos a Distância de Levenstehin do valor 1. A Equação 1 exibe a fórmula para cálculo do grau de similaridade linguística entre dois dados D(x,y) e D(w, y) usada neste trabalho.

O grau de similaridade linguística (Sl) entre duas linhas (Lx e Ly) é determinado pela média aritmética do somatório das similaridades dos seus dados, como indicado na Equação 2. Por exemplo, a similaridade linguística entre as linhas 1 e 2 (Sl(L1.L2)), da Tabela 1, considerando apenas as colunas NAME e TYPE, será a soma da S("John Mark Zuremberg", "John M. Zurembrg"), 0.76, com S("FullProfessor", "Professor"), 0.69, resultando 1.45, dividido pelo número de colunas envolvidas, neste caso m=2, obtendo 0.73. Portanto, as linhas em questão possuem uma similaridade linguística de 73%, considerando apenas estas colunas. A escolha das colunas utilizadas na comparação é feita pelo usuário.

$$S(D_{(x,y)}, D_{(w,y)}) = 1 - \frac{levenshtein D_{(x,y)}, D_{(w,y)}}{\max(size(D_{(x,y)}), size(D_{(w,y)}))} \quad \textbf{(1)}$$

$$Sl(L_x, L_y) = \frac{\sum_{i=1}^{m} S(D_{(x,i)}, D_{(y,i)})}{m} \quad \textbf{(2)}$$

Módulo Semântico

Para a comparação semântica, utilizamos uma Ontologia de Domínio (OD) para inferir os tipos de relacionamento semântico entre os dados. Uma ontologia de domínio modela um domínio específico do mundo real, descrevendo o significado dos termos do domínio em questão (GUARINO, 1989). Por exemplo, o termo "manual" pode ter distintos significados. Uma ontologia do domínio educacional poderia modelar seu significado como um tipo de publicação, outra de domínio comercial para um tipo de produção do produto (manual ou industrial). A Figura 2 ilustra o trecho de uma ontologia do domínio educacional.

ISBN 978-85-62830-10-5

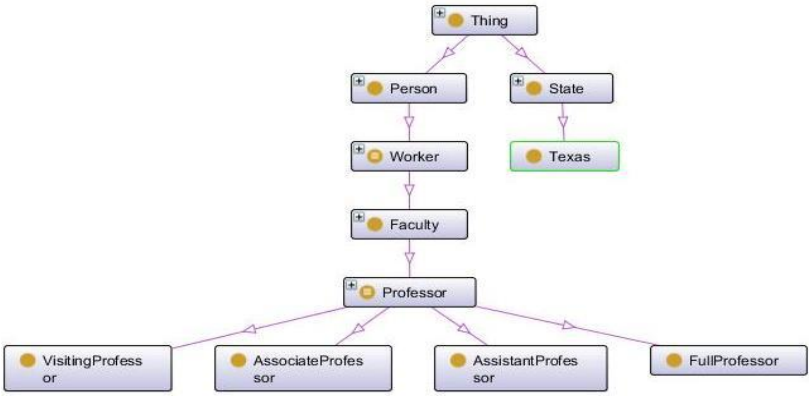

Figura 2 - Trecho de uma ontologia do domínio educacional

O uso de conhecimento prévio contido em ontologias de domínio permite a identificação de relacionamentos semânticos entre os dados (por exemplo, equivalência e especialização). A descoberta do grau de sobreposição semântica é bastante útil em processos de comparação de dados. Os dados da tabela que contém linhas duplicadas são mapeados para classes ou instâncias na ontologia de domínio. Em seguida, os relacionamentos entre classes descritas na OD são mapeados em relacionamentos entre os dados. Com isto, quatro regras semânticas foram definidas: equivalência, especialização, generalização e instâncias de uma mesma classe, respectivamente ilustradas na Figura 3.

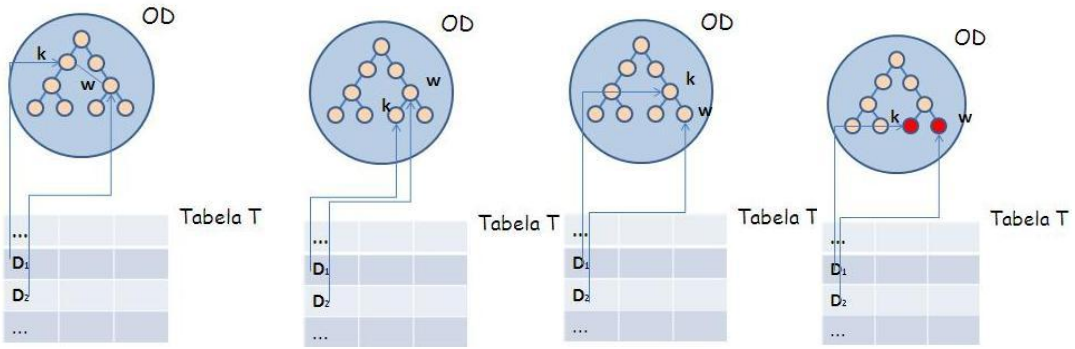

Figura 3 – Regras Semânticas

- Equivalência: Um dado D1:t é equivalente a um dado D2:t se (i) D1:t é idêntico a um conceito k na ontologia de domínio; (ii) D2:t é idêntico a um conceito w na mesma ontologia de domínio; e (iii) k e w estão ligados por um relacionamento equivalentClass na ontologia de domínio.
- Especialização: Um dado D1:t é uma especialização de um dado D2:t se (i) D1:t é idêntico a um conceito k na ontologia de domínio; (ii) D2:t é idêntico a um conceito w na mesma ontologia de domínio; e (iii) k e w estão ligados por um relacionamento subClassOf na ontologia de domínio.
- Generalização: Um dado D1:t é uma generalização de um dado D2:t se: (i) D1:t é idêntico a um conceito k na ontologia de domínio; (ii) D2:t é idêntico a um conceito w na mesma ontologia de domínio; e (iii) k e w estão ligados por um relacionamento superClassOf na ontologia de domínio.
- Instâncias de uma mesma classe: Dois dados D1:t e D2:t são instâncias próximas se: (i) D1:t é idêntico a uma instância k na ontologia de domínio e D2:t é idêntico a uma instância w na mesma ontologia de domínio; e k e w são instâncias do mesmo conceito.

As regras semânticas são aplicadas na comparação semântica de dois dados. A cada regra semântica é associado um peso de acordo com sua importância no processo de detecção de linhas duplicadas (Bilenko & Mooney, 2003). As regras mais relevantes recebem um peso maior. Neste

ISBN 978-85-62830-10-5

trabalho, usamos os seguintes pesos: 1.0 para equivalência, 0.8 para especialização e generalização, e 0.4 para instâncias de mesma classe. A escolha foi feita com base em experimentos. Como mais de um tipo de relacionamento semântico pode ser identificado para dois dados, o grau de similaridade semântica entre dois dados é dado pelo valor do maior peso dentre os diferentes tipos de relacionamentos semântico encontrados, como indicado na Equação 3.

O grau de similaridade semântica entre duas linhas ($Ss(L_x, L_y)$) é determinado pela média aritmética dos graus de similaridade dos seus dados, como indicado na Equação 4.

$$S(D_1, D_2) = \max(weightRule_1, ..., weightRule_n) \quad \textbf{(3)}$$

$$Ss(L_x, L_y) = \frac{\sum_{i=1}^{m} S(D_{(x,i)}, D_{(y,i)})}{m} \quad \textbf{(4)}$$

Para exemplificar, calculamos o grau de similaridade semântica entre as linhas L1 e L2 da Tabela 1 (Ss(L1.L2)), considerando apenas as colunas C3.TYPE e C6.NATIONALITY. A ontologia de domínio utilizada no exemplo é a mesma da Figura 2. De acordo com esta ontologia, o tipo de relacionamento semântico encontrado entre "FullProfessor" e "Professor" é especialização, e entre "El paso" e "Marshal" é instância de mesma classe (nesse caso, Texas) com pesos 0.8 e 0.4, respectivamente. Conforme mostrado na Equação 4, para determinar o grau de similaridade entre as duas linhas, soma-se o valor dos pesos e, em seguida, divide-se pelo número de colunas envolvidas, neste caso m=2. Portanto, o grau de similaridade semântico entre as linhas em questão é de 60%.

Similaridade Entre Linhas

A avaliação da similaridade global entre duas linhas é calculada pela média ponderada das similaridades linguística e semântica entre as mesmas linhas, como indicado na Equação 5.

$$S_{(Lx,Ly)} = WeightLinguistic * Sl_{(lx,ly)} + WeightSemantic * Ss_{(lx,ly)} \quad \textbf{(5)}$$

3. RESULTADOS E DISCUSSÃO

Para realização dos experimentos e desenvolvimento da ferramenta, foi utilizado um notebook Dell Inspiron 1428, 4GB de memória RAM e 250GB de disco rígido. A escolha da base de dados não foi uma tarefa simples, visto que era necessário uma base de dados pública contendo dados duplicados e que possuíssem algum tipo de relacionamento semântico. Algumas bases com duplicações foram encontradas: base de restaurantes (Restaurant) (RECORD LINKAGE, 2012) e base de citações de artigos de computação (DBLP) (RECORD LINKAGE, 2012). No entanto, não foi detectado nenhum tipo de relacionamento semântico associado aos dados contidos nestas bases. Da mesma forma, geradores de dados foram analisados e descartados por não gerarem dados com semântica, a exemplo de: UIS Database Generator (RECORD LINKAGE, 2012) e Data Generator (DATA GENERATOR, 2012). Após a análise de diversas bases de dados, optamos por utilizar a base disponibilizada pela TLU (Texas Lutheran University) (TLU, 2012) e a ontologia pública de domínio educacional a education.owl (EDUCATION, 2012). A priori, a base não apresentava dados duplicados. Estes foram inseridos de forma arbitrária por um especialista do domínio educacional. A base continha 100 (cem) registros.

Os experimentos foram realizados da seguinte maneira: primeiramente o especialista de domínio determinou manualmente o Resultado de Referência (RR) que continha todas as linhas duplicadas da tabela usada para experimentos. Em seguida, o resultado de referência foi comparado com o Resultado da Ferramenta (RF) no sentido de determinar o grau de precisão do processo proposto. Para isso, foram usadas as métricas de precisão, *precision* em inglês, (YUANPENG, 2005), que mede a porcentagem de registros indicados corretamente como duplicados (RC) pela ferramenta

ISBN 978-85-62830-10-5

(Equação 6), e de cobertura, *recall* em inglês, (YUANPENG, 2005), que calcula a porcentagem de registros duplicados esperados pelo usuário que são detectados pela ferramenta (Equação 7).

$$precision = \frac{RC}{RF} \quad (6) \qquad recall = \frac{RF}{RR} \quad (7)$$

Na avaliação foi efetuada uma variação nos pesos das técnicas linguística e semântica. O peso semântico foi gradativamente aumentado em 10%, como indicado na Figura 4, na qual o eixo das abscissas representa a variação do peso semântico e o eixo das ordenadas o valor da métrica em estudo. É importante salientar que aumentar o peso da técnica semântica, implica em reduzir na mesma proporção o peso da técnica linguística, e vice-versa. O limiar foi mantido em 50% durante todas as configurações.

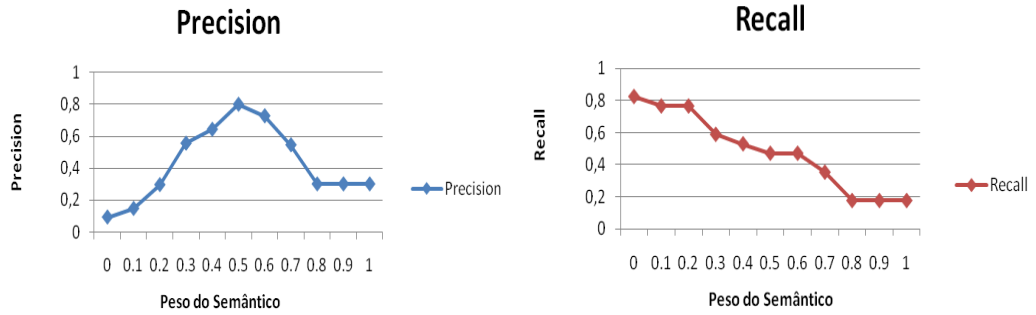

Figura 4 – Métricas utilizadas nos experimentos

De acordo com os experimentos (Figura 4), observou-se que, com o aumento do peso semântico, e a consequente diminuição do peso linguístico, as métricas precision e recall, tendem a se estabilizarem. Este comportamento indica que a partir de certo ponto torna-se indiferente a variação dos pesos. O equilíbrio entre os pesos das técnicas é definitivo para o bom resultado ao fim do processo. Em algumas configurações, considerando a semântica dos dados, pode-se alcançar um aumento de precisão. No entanto, é importante lembrar que esta variação é dependente da ontologia utilizada e dos dados pertencentes à base de dados em estudo.

Após a análise dos experimentos, observou-se uma diminuição em número na detecção de dados duplicados. No entanto, também ocorreu uma diminuição de escopo, ou seja, o total de dados indicados erroneamente como duplicados foi atenuado.

A configuração na qual se utilizou apenas a técnica linguística como forma de detecção, foi a que encontrou maior número de duplicações corretas (Figura 5), no entanto, com esta configuração não foi possível a detecção de linhas duplicadas como as indicadas na Figura 5, que foram detectados apenas com a utilização da semântica. Com isto, é possível perceber a limitação da técnica linguística.

Hettinger, Deborah	Professor	Biology	6030	dhettinger@tlu.edu	Moody Science
Hettinger, D.	VisitingProfessor	Medicine	6030	null	M. Science

Figura 5 – Dados duplicados semanticamente

4. CONCLUSÕES

Este trabalho propôs uma abordagem para detecção de linhas duplicadas em tabelas de banco de dados, combinando técnicas linguísticas e semânticas. O módulo linguístico foi baseado na distância de Levenshtein. Para o módulo semântico foram especificadas e implementadas regras semânticas baseadas em relacionamentos existentes em ontologias. Como resultado, foi desenvolvido uma ferramenta de detecção com abordagem proposta.

ISBN 978-85-62830-10-5

Diante dos resultados provenientes dos experimentos e de sua métrica de avaliação, foi possível concluir que a eficiência do processo de detecção de dados duplicados por meio do uso de ontologias de domínio depende intrinsecamente da ontologia utilizada, e que esta pode tornar a detecção semântica de extremo sucesso ou simplesmente ser indiferente. Foi possível observar deficiências nas técnicas linguística e semântica desenvolvidas. No entanto, foi perceptível a importância do estudo semântico do dado para a detecção de linhas duplicadas. Esta foi capaz de indicar potenciais duplicados que não são detectados linguisticamente.

Como trabalhos futuros, sugerimos a criação de novas regras semânticas (por exemplo, criação de axiomas e utilização de outros tipos de relacionamento semântico existentes em ontologias), assim como a utilização de mais de uma ontologia de domínio obtidas da Web.

5. AGRADECIMENTOS

Ao CNPq pelo financiamento do projeto e pela concessão da bolsa PIBITI. Aos colegas do Laboratório de Sistema de Informação (LSI) do Departamento de Sistemas e Computação, onde este projeto foi desenvolvido e realizado, por todo apoio e contribuição.

6. REFERÊNCIAS

ADRIAANS, P.; Zantinge, D. Data mining. Syllogic, 1996.

BALLOU, D., Tayi, G. K. Enhancing Data Quality in Data Warehouse Environments. Communications of the ACM, 42(1): 73-78, 1999.

BILENKO, M. e MOONEY, R. J. (2003) – Adaptive Duplicate Detection Using Learnable String Similarity Measures. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington (EUA), Agosto de 2003. pp. 39-48.

BOSCARIOLI, C.Pré-processamento de Dados para Descoberta de Conhecimento em Banco de Dados: Uma Visão Geral. In: Unicentro. (Org.). Anais do III CONGED - Congresso de Tecnologias para Gestão de Dados e Metadados do Cone Sul. Guarapuava: Unicentro Editora, 2005, v. I, p. 101-120.

BUETTCHER, S., CLARK, C. L. A., CORMACK. G. V. Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, 2010.

DATA GENERATOR. Disponível em: <http://www.generatedata.com/#about>. Acessado em 09/06/2012.

DATA Match. Disponível em: <http://www.dataladder.com/index.html>. Acessado em: 12/08/2012.

DUPLICATE Record Remover . Disponível em: <http://www.duplicaterecordremover.com/>. Acessado em: 12/08/2012.

EDUCATION. Disponível em: <http://www.cin.ufpe.br/~speed/SemMatch/UnivCsCMO.owl>. Acessado em 28/07/2012.

GUARINO, N. Formal Ontology and Information Systems, 1989. In: Galton, A. and Mizoguchi, R. (eds.), Proceedings of the Sixth International Conference on Formal Ontology and Information Systems (FOIS 2010), p.89-102. IOS Press.

INMON, W. H. Data Warehouse Performance. New York: John Willey, 1998. Edição 1.

OLIVEIRA, P.; RODRIGUES, F. E RODRIGUES, P. (2004) – Limpeza de Dados: Uma Visão Geral. In Proceedings of the Data Gadgets 2004 Workshop – Bringing Up Emerging Solutions for Data Warehousing Systems (em conjunto com a JISBD'04), Málaga (Espanha), Novembro de 2004. pp. 39-51.

OLIVEIRA, Jorge Paulo, Detecção e Correção de Problemas de Qualidade dos Dados: Modelo, Sintaxe e Semântica, 2008. Tese de doutoramento em informática. Universidade do Minho.

RECORD LINKAGE. Disponível em: <http://www.cs.utexas.edu/users/ml/riddle/data.html>. Acessado em: 09/06/2012.

TLU. Disponível em: <www.tlu.edu/i/about_tlu/name_directory.xls>. Acessado em: 10/04/2012.

YAUPENG. J. Huang, R. Powers, and G. T. Montelione (2005) Protein NMR Recall, Precision, and F-measure Scores (RPF scores): Structure Quality Assessment Measures Based on Information Retrieval StatisticsJournal of the American Chemical Society, 127(6), 1665-1674.

ISBN 978-85-62830-10-5