UNIVERSIDADE ESTADUAL DE MONTES CLAROS

Centro de Ciências Exatas e Tecnológicas

Programa de Pós Graduação Modelagem Computacional e Sistemas

Herberth Giuliano Amaral Silva

Tratamento de dados faltantes em uma solução de record linkage com datasets heterogêneos

Montes Claros - MG

2015

Sumário

1 Introdução

Introdução

2 Proposta da Pesquisa

2.1 Justificativa

O processo de *record linkage* é um dos desafios nas atividades de limpeza de dados (*data cleasing*) em Mineração de Dados. No entanto, as técnicas clássicas de *record linkage* não lidam eficientemente com dados faltantes (Toan C. Ong, Michael V. Mannino, Lisa M. Schilling, Michael G. Kahn, 2014). Essa característica é uma grande desvantagem porque diferentes bases de dados podem ter diferentes características no que diz respeito à presença dados necessários no processo de *record linkage*. Para não comprometer o processo de *record linkage*, técnicas para lidar com falta de dados precisam ser empregadas.

2.2 Objetivo Geral

Desenvolver um modelo de record linkage que suporte integração de datasets heterogêneos com dados faltantes (aleatoriamente ou não) que apresente valores mínimos de falsos-negativos;

2.3 Objetivos Específicos

a) Um algoritmo eficiente do ponto de vista de precisão/revocação;

b) Uma solução de ajustes de parâmetros automatizada para o modelo em questão;

c) Extensões de algoritmos para comparacao de tuplas em meio a dados faltantes.

3 Referencial Teórico

3.1 *Record linkage*

3.1.1 Introdução

Record linkage é o processo de encontrar registros duplicados em uma ou mais de dados que se referem à uma mesma entidade (Christen P., 2012). O processo é conhecido por deduplicação quando aplicado em um único banco de dados.

A teoria de record linkage foi desenvolvida inicialmente por (Fellegi I., Sunter A., 1969). A parte principal do processo pode ser compreendido por uma função $\mu(t_1, t_2)$ que analisa a similaridade de um par de tuplas. Esta função tem como retorno uma probabilidade, que pode ser classificada como "equivalente"(quando é possível afirmar que o par de tuplas representa a mesma entidade), "possivelmente equivalente"(quando não há informações suficientes para afirmar se é ou não equivalente) e "não equivalente"(quando há informações suficientes para afirmar que o par de tuplas **não** representam a mesma entidade).

O processo de *record linkage* conta com sub-tarefas para sua realização: tais como limpeza e padronização de dados, indexação, comparação e classificação e revisão manual. Cada uma dessas sub-tarefas serão detalhadas nas subseções a seguir.

Como ilustração de funcionamento de uma solução de *record linkage*, considere o conjunto de dados de exemplo abaixo:

Tabela 1 – Dados demográficos

ID	Nome	Data de nascimento	Endereço	Telefone
1	João Pedro da Silva	01/04/1985	Rua das Camélias, 325	92368080
2	João Pedro Silva	01/04/1985	Rua das Tulipas, 180	92338080
3	Joana Paula Silva	02/09/1992	Rua das Tulipas, 180	32225478
4	Joana P. Silva	02/09/1962	Av. das Bromélias, 98	32225478

Para deduplicar o conjunto de dados mostrado anteriormente, é necessário fazer comparações de todos os possíveis pares de tuplas ($C = n!/2(n-2)!$) utlizando a função $\mu(t_1, t_2)$. Desta forma, o conjunto gerado pela comparação dos registros previamente mencionados é:

As comparações de campo-a-campo foram feitas utilizando o algoritmo de Jaro-Winkler (Winkler, W., 1990). No final, define-se um *score* que será utilizado para classi-

Tabela 2 – Comparações de tuplas da tabela de dados demográficos

ID 1	**ID 2**	**Nome**	**Data de nascimento**	**Endereço**	**Telefone**	***Score***
1	2	0.94	1.00	0.85	0.91	3.71
1	3	0.62	0.67	0.85	0.47	2.62
1	4	0.78	0.67	0.62	0.47	2.55
2	3	0.63	0.67	1.00	0.47	2.78
2	4	0.82	0.67	0.67	0.47	2.65
3	4	0.89	0.93	0.67	1.00	3.50

ficação em "equivalente", "não equivalente"e "possivelmente equivalente". Neste exemplo, os pares de tuplas $(1, 2)$ e $(3, 4)$ são equivalentes. Como parâmetros desta solução em particular, pode-se definir os limiares 3.0 para "possivelmente equivalentes"e 3.4 para "equivalentes", uma vez que todos os valores abaixo de 3.0 não são equivalentes e todos os valores acima de 3.4 são equivalentes.

3.1.2 Limpeza e padronização dos dados

Pelo fato que boa parte dos dados do mundo real são "sujos"e contém informações ruidosas, incompletas e mal-fortatadas, a tarefa de limpeza dos dados é crucial para o restante do processo (T. Churches, P. Christen, K. Lim, J.X. Zhu, 2002). Já é reconhecido que a falta de dados de boa qualidade pode ser um dos grandes obstáculos para uma solução de *record linkage* de sucesso (Clark D., 2004).

O objetivo principal da tarefa de limpeza e padronização dos dados é a conversão de dados brutos de entrada em um conjunto de dados bem definido e consistente (T. Churches, P. Christen, K. Lim, J.X. Zhu, 2002).

3.1.3 Indexação

Em sua natureza, o processo de *record linkage* é um processo de complexidade combinatorial, pois precisa analisar todos as combinações de pares de tuplas em uma ou mais bases de dados. Porém, a vasta maioria das comparações é feita em registros que não são equivalentes (Christen P., 2012), o que leva a um desperdício de recursos computacionais.

A abordagem tradicional no contexto de *record linkage* para mitigar o problema previamente descrito é conhecido como *blocking*. Essa estratécia consiste em dividir os dados em blocos não-sobrepostos de forma que somente dados dentro de cada bloco são comparados entre si. Um critério de *blocking*, comumente chamado de chave de *blocking*, é baseado em um único campo ou na concatenação de valores de vários campos (Christen

P., 2012).

3.1.4 Comparação

4 Cronograma

Item	Descrição	Tempo
01	Conclusão das Disciplinas Regulares	1º semestre 2015
02	Levantamento Bibliográfico	1º semestre 2015
03	Definição de Conceitos	2º trimestre 2015
04	Comparação dos algoritmos	2º semestre 2015
05	Desenvolvimento do algoritmo	1º semestre 2016
06	Redação da Dissertação	2º trimestre 2016
08	Defesa	3º trimestre 2016

Tabela 3 – Cronograma

REFERÊNCIAS

Christen P. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, v. 24, p. 1537–1555, 2012. Citado 2 vezes nas páginas 4 e 5.

Clark D. Practical Introduction to Record Linkage for Injury Research. *Injury prevention*, v. 10, p. 186–191, 2004. Citado na página 5.

Fellegi I., Sunter A. A theory for record linkage. *Journal of the American Statistical Association*, v. 64, p. 1183–1210, 1969. Citado na página 4.

T. Churches, P. Christen, K. Lim, J.X. Zhu. Preparation of Name and Address Data for Record Linkage Using Hidden Markov Models. *BioMed Central Medical Informatics and Decision Making*, v. 2, 2002. Citado na página 5.

Toan C. Ong, Michael V. Mannino, Lisa M. Schilling, Michael G. Kahn. Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, v. 52, p. 43–54, 2014. Citado na página 3.

Winkler, W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, p. 354–359, 1990. Citado na página 4.