Resolução da Heterogeneidade na Identificação de Pacientes*

Fábio Filocomo[1] **, Marcelo Finger**[2†]**, Diogo F. C. Patrão**[1]

[1] Laboratório de Informática Médica – CIPE
Fundação Antonio Prudente – Hospital A.C. Camargo

[2]Departamento de Ciência da Computação
Instituto de Matemática e Estatística – Universidade de São Paulo (USP)

{fabio.filocomo,djogo}@cipe.accamargo.org.br, mfinger@ime.usp.br

Abstract. *This work deals with the identification of pacients in heterogeneous databases. We deal with the problem of identifier matching as well as with name matching. For that, a method for matching several identification formats was developed; we also developed the QFS measure, which takes in consideration the statistical distribution of names and cultural factors. We show that QFS-measure is superior to several existing methods for matching names.*

Resumo. *Este trabalho trata da identificação de pacientes em bancos de dados heterogêneos. Lidamos com o problema do emparelhamento de identificadores de pacientes, bem como de nomes. Para tanto, um método de unificação de diversos padrões de identificação e uma medidas de similaridade entre nomes, chamada de medida QFS, foi criada. Mostramos que esta medida é superior a outros métodos no caso da identificação de nomes de pessoas.*

1. Introdução

Como identificar pacientes em duas dezenas de bancos de dados?

O Hospital A. C. Camargo, localizado em São Paulo e fundado em 1953, é um hospital oncológico, sem fins lucrativos, com uma média mensal de 2500 pacientes atendidos. Cada paciente recebe um identificador denominado Registro Geral Hospitalar (RGH), o qual, ao longo dos anos, sofreu modificações em seu formato e no processo de atribuição. Diversos sistemas informatizados foram implantados, cada um com o RGH no formato próprio. Este trabalho descreve nossa estratégia e resultados na tarefa de resolução da heterogeneidade na identificação de pacientes (RHIP).

Para este fim, desenvolvemos dois métodos de identificação de pacientes, a saber, o *emparelhamento de identificadores* e o *emparelhamento de nomes*. Falamos de *emparelhamento de identificadores* quando as pessoas são representadas por um RGH padrão em algum formato. Por *emparelhamento de nome* entende-se a identificação de registros contendo dados referentes a mesma pessoa, independente da forma como esta pessoa está representada nos diversos bancos de dados.

O primeiro método promove o emparelhamento de RGHs; estes identificadores sofreram evolução ao longo de 5 décadas, e formatos diferentes são usados em bancos

*Trabalho desenvolvido e apoiado pelo projeto INCITO CNPq 573589/2008-9.
†Bolsista CNPq PQ 302553/2010-0.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 199–203,
Cuiabá, MT, Brazil, October 24–26, 2011.

de dados diferentes. Para unificá-los foi proposto um identificador único denominado CPFH [Filocomo et al. 2011]. O segundo método busca o emparelhamento dos nomes de pacientes. Esta é uma tarefa significativamente mais complexa, valendo-se de dados estatísticos como a distribuição de probabilidade de nomes de pacientes, fatores culturais e erros típicos de digitação.

Este artigo se desenvolve da seguinte maneira. Na Seção 2, apresentamos os métodos desenvolvidos para emparelhar identificadores e nomes, seguido pela avaliação das medidas e sua comparação com outras existentes na Seção 2.2. Por fim, apresentamos nossas conclusões na Seção 3.

2. Metodologia

A tarefa de identificação de pacientes nos diversos sistemas do hospital seguiu as seguintes fases:

- Unificação dos diversos sistemas de identificação de pacientes nos diversos sistemas de bancos de dados federados (CPFH).
- Avaliação do emparelhamento por CPFH por amostragem
- Criação de uma medida para identificação de pacientes, **QFS** (Qwerty Flex Score).
- Comparação da medida QFS com outros métodos de identificação de nomes.

2.1. Emparelhamento de Nomes

Para a identificação única de pacientes, foi proposto um identificador denominado CPFH [Filocomo et al. 2011]. Entretanto, devido aos métodos próprios de identificação das diversas bases de dados, é comum encontrar diferentes pacientes compartilhando um mesmo RGH ou pacientes que sequer o possuam.

Para solucionar este problema, foi proposto um método de emparelhamento de nomes (EN). O EN leva em conta uma série de características dependentes do contexto geográfico e cultural em que se encontram os pacientes do hospital:

1. a distância Levenshtein com pesos para diferentes erros — este erros foram calibrados por erros comuns de digitação num teclado brasileiro no padrão *ABNT-2*;
2. distribuição estatística dos nomes nas bases de dados do Hospital;
3. datas de nascimento, filtrando erros de preenchimentos e datas inválidas;
4. filtro automático de nomes, tratando comentários, abreviaturas usuais, caracteres especiais e etc.;
5. fatores culturais, tais como a raridade da omissão do primeiro nome, a inserção de nome pós casamento, nomes não brasileiros incomuns na população local.

O medida gera um valor numérico entre 0 e 100 cujo limiar de separação entre nomes emparelháveis e não-emparelháveis precisou ser determinado empiricamente, conforme mostrado nos resultados. O limiar final é de 60.

Este método está descrito na Figura 1. Inicialmente, ambos os nomes em comparação são segmentados e sua "raridade" é verificada. A *raridade* de um nome é definida em relação à distribuição de ocorrência de nomes na população. Em nosso caso, verificamos empiricamente que ao fixar o percentil 7% como *limiar de raridade*, obtivemos os melhores resultados.

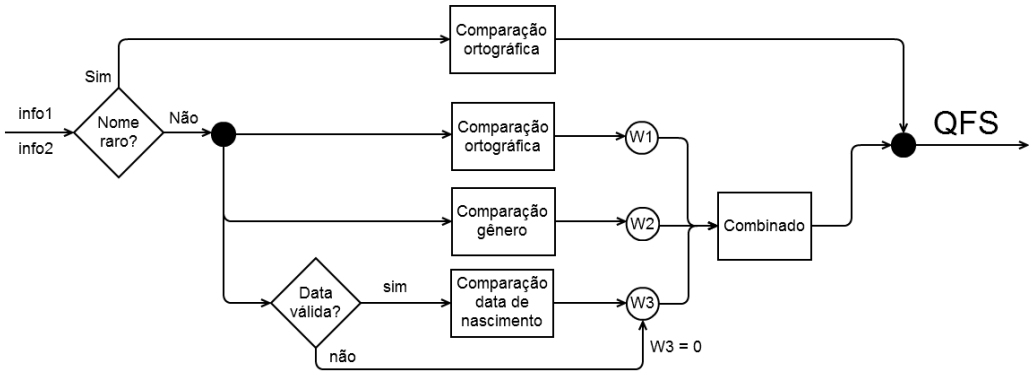

Figura 1. Diagrama de blocos do cálculo da medida QFS

2.2. Avaliação

Identificação de CPFHs

A avaliação da identificação CPFHs foi feita por amostragem. Trabalhamos com um número total de 1.475.353 registros nos quatro maiores bancos de dados da federação, contendo respectivamente 263.182, 830.853, 53.388 e 327.930 registros de pacientes. O número de pares com informações de identidade e CPFHs coincidentes foi de 123.197.

Fizemos uma amostragem aleatória de 10% dos registros emparelhados, a qual foi validada manualmente. Num universo de 11.993 pares identificados, encontramos uma taxa de acerto de **99,85%**. Este alto grau de confiabilidade na identificação de CPFHs fez com que este método fosse utilizado para avaliar o emparelhamento de nomes.

Identificação de Pacientes pela Medida QFS

O objetivo aqui era medir a qualidade da medida QFS. Inicialmente, trabalhamos sobre os pares identificados pelo mesmo CPFH. Nesta fase, o objetivo era identificar pacientes em vários bancos de dados, evitando o emparelhamento de pacientes com identificadores não equivalentes. Portanto, o objetivo era o de maximizar a *confiabilidade positiva*, P_{rel}, da identificação de pacientes, definida por $P_{rel} = \frac{T_p}{N_c}$, onde T_p são os verdadeiros positivos, F_p os falsos positivos e $N_c = T_p + F_p$ o número total de pares com mesmo CPFHs.

Utilizamos várias medidas. A medida *exata* considerou T_p como os pares com o mesmo CPFH e nomes idênticos na comparação de strings. Também utilizamos a medida Levenshtein normalizada (com limiar 60%), Lev_{60} [Bilenko et al. 2003], e o algoritmo soundex, $Soundex$ [Knuth 1981]. Comparamos com 4 variantes da medida QFS, as medidas QFS_x, com limiar $x \in \{20, 40, 60, 60\}$. Os resultados estão na Tabela 1, onde as medidas foram todas feitas com $N_C = 123.197$. Claramente a medida QFS_{20} predomina.

Medida	*Exata*	$Soundex$	Lev_{60}	QFS_{80}	QFS_{60}	QFS_{40}	QFS_{20}
T_p	115488	116517	117182	115423	117093	121893	**122487**
P_{rel}	0,9354	0,9458	0,9511	0,9369	0,9505	0,9894	**0,9966**

Tabela 1. Comparação da identificação do CPFH para diversas medidas

Notamos que os pares identificados pela medida QFS_{20} praticamente contém todos os outros pares identificados pelas outras medidas, mostrando grande eficácia. Similarmente, os pares identificados por Lev_{60} contém praticamente todas as demais, excetuando os identificados por $QFS_{20} e QFS_{40}$. Finalmente, os pares identificados pela medida exata (igualdade de cadeias de caracteres) estão contidos nos demais. Estes fatos nos dão a abrangência e a qualidade da medida QFS no contexto de identificação de CPFHs.

Avaliamos a medida QFS também no contexto livre, ou seja, sem levar em conta a identificação de CPFHs. Foi utilizado um córpus de 677 pares de nomes contendo erros artificialmente gerados. Para esta avaliação comparamos medidas de precisão, cobertura e medida-f, usando sua definição usual [Manning and Schütze 1999]:

Estas medidas foram computadas sobre o mesmo córpus para a identificação utilizando as seguintes medidas: a medida exata; a medida Levenshtein normalizada (com limiar 60%), Lev_{60} [Gusfield 1997]; o algoritmo soundex, $Soundex$ [Knuth 1981]; e as medidas $QFS_x, x \in \{20, 40, 60, 80\}$. Estes resultados estão ilustrados na Tabela 2.

Medida	*Precisão*	*Cobertura*	*Medida-F*
Exata	100,00%	45,91%	62,93%
Lev_{60}	79,81%	97,08%	87,60%
Soundex	88,53%	88,01%	88,27%
QFS_{20}	87,60%	92,71%	90,01%
QFS_{40}	93,90%	89,79%	91.80%
QFS_{60}	97,08%	87,72%	**92,17%**
QFS_{80}	97,07%	87,17%	91,85%

Tabela 2. Identificação de Nomes para Diversas Medidas

A Tabela 2 mostra que a medida QFS_{60} é a mais bem sucedida na identificação de nomes num contexto livre, obtendo o maior valor para Medida-F e a maior precisão das medidas não-exatas e possui confiabilidade de 95,05% (Tabela 1). Em geral, a Medida-F de todas as medidas QFS_x foram superiores às demais. A medida QFS_{60} foi, portanto, escolhida para aplicação na identificação de pacientes na federação de bancos de dados.

3. Conclusões e Trabalhos Futuros

Numa instituição com 50 anos de história, é praticamente impossível a qualquer base de dados permanecer incólume a erros e redundâncias no cadastro de informações. Em particular, o problema de cadastro de pessoas foi abordado, nesse tempo, com processos, premissas e sistemas de numeração diferentes.

A criação do algoritmo CPFH para a unificação de identificadores representou um grande avanço para a solução deste problema. No entanto, devido às diversas particularidades das bases de dados, fez-se necessária a elaboração de um algoritmo auxiliar. Este deveria ser capaz de unificar diferentes instâncias de um mesmo paciente, com base em atributos como nome, gênero e data de nascimento.

Resolvemos problemas como abreviação, mudança de nome, erros de digitação e omissões. Desenvolvemos um algoritmo cujos resultados encontrados foram muito superiores aos algoritmos usuais de comparação de texto. Tal método está sendo usado para integrar os registros antigos e atuais. Futuramente, trabalharemos na indexação baseada neste método, visando a aceleração de consultas por nomes similares.

Referências

Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., and Fienberg, S. (2003). Adaptive name matching in information integration. *Intelligent Systems, IEEE*, 18(5):16 – 23.

Filocomo, F., Finger, M., and Patrão, D. F. C. (2011). Resolução da heterogeneidade na identificação de pacientes. `http://www.lbhc.hcancer.org.br/wiki/RHIP`.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press.

Knuth, D. E. (1981). *The Art of Computer Programming, Volume III: Sorting and Searching*. The Art of Computer Programming. Addison-Wesley.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.