Universidad Central de Las Villas

Hospital Universitario “Arnaldo Milian Castro”

ESTANDARIZACIÓN DE CADENAS DE TEXTO EN APLICACIONES MÉDICAS

TEXT STRING STANDARDIZATION IN MEDICAL APPLICATIONS

Autores:

MSc. Beatriz López Porrero

Profesora Auxiliar Departamento de Ciencia de la Computación, UCLV.

Carretera a Camajuaní # 199 km 2 ½ Santa Clara, Villa Clara, Teléfono (42)281212

e-mail: blopez@uclv.edu.cu

Dr.CT Ramiro A. Pérez Vázquez

Profesor Titular Departamento de Ciencia de la Computación, UCLV

Carretera a Camajuaní # 199 km 2 ½ , Santa Clara, Villa Clara, Teléfono (42)281212

e-mail: rperez@uclv.edu.cu

Dr. Francisco Parrilla Arias

Especialista Segundo Grado en Anatomía Patológica. Hospital “Arnaldo Milián Castro” Santa Clara. Villa Clara.

Dirección: Santa Clara, Villa Clara.

Dr. Daniel Artiles Martínez

Especialista Primer Grado en Anatomía Patológica. Hospital “Arnaldo Milián Castro” Santa Clara. Villa Clara.

Dirección: Calle Lubián No. 10 e/ Marta A. Y Padre Chao. Santa Clara. Villa Clara.

Resumen

En este artículo se expone un marco de trabajo para la estandarización de atributos tipo texto, donde el elemento más importante lo constituye el agrupamiento de dichas cadenas usando una modificación a la distancia de edición propuesta por Levenshtein, donde se trata de acercar palabras que puedan resultar unas de otras por errores tipográficos, de tal manera que el costo de sustitución utilizado en la distancia depende de la posición de los caracteres en el teclado. Se muestra los resultados obtenidos en el intento por estandarizar las cadenas que representan las causas de muertes determinadas en las necropsias realizadas en el departamento de Anatomía Patológica del Hospital Universitario “Arnaldo Milián Castro” de la ciudad de Santa Clara.

Palabras claves: estandarización de cadenas, distancia de edición, limpieza de datos.

Abstract

This paper presents a framework for the standardization of string attributes. The most importan step is the clustering of the string. A modification of Levenshtein distance is used in order to create the clusters. The used modification produces clusters in which the strings are close to each other in the typographical sense. In the proposed distance, the cost of the substitution operation depends of the distance between characters in the keyboard. The paper shows the results of the standardization of strings that represent death causes. The experiment was made in the Pathological Anatomy Department in University Hospital “Arnaldo Milián Castro” in Santa Clara city.

Key Words: string standardization, edition distance, data cleaning.

Introducción

Uno de los problemas comunes que afectan la calidad de los datos es la no estandarización de los mismos [1] este se presenta frecuentemente en atributos de tipo cadenas, en que un mismo objeto u entidad es descrito utilizando cadenas de texto diferentes.

Estos atributos están presentes en cualquier base de datos, en particular aparecen frecuentemente en aplicaciones médicas para indicar nombres de personas, causas de diferentes fenómenos, etc.

En los sistemas actuales de bases de datos se trata de evitar la entrada libre de texto con el uso de diferentes controles que puedan restringir de alguna manera las cadenas a entrar. Se utilizan habitualmente cuadros de listas, cuadros combinados, botones radiales, árboles, etc. Tratando siempre que el usuario escoja uno de los establecidos, pero esto no siempre es posible. En ocasiones porque las listas serían interminables, en otros porque no se pueden establecer a priori los elementos que deben contener.

Esto hace que en el atributo en la base de datos queden valores diferentes que refieren a la misma entidad. Por ejemplo en la base de datos de Necropsias del Hospital “Arnaldo Milián Castro” existe un atributo que es la “Causa de Muerte”. Este tiene, entre otros, los siguientes valores:

1. shock mixto (septico e hipovolémico)
2. shock mixto (séptico***/***hipovolémico)
3. shock mixto(septico e hipovolemico)
4. muerte subita cardiovascular en un corazon estructuralemente sano
5. muerte subita cardiovascular en un corazon estructuralemte sano
6. muerte subita cardiovascular en un corazon estructuralmente sano

Aquí se observa que los tres primeros (1-3) hacen referencia a la misma causa de muerte, se diferencian sólo en la forma en que se han escrito (uso de paréntesis, conjunción o /). Los otros tres valores (4 – 6) también se refieren a la misma causa, en este caso la diferencia la hace la palabra “estructuralmente” que solo en el caso 6 está bien escrita. Estas diferencias, hacen que si se hiciera algún cálculo estadístico con este atributo aparezcan 6 causas de muerte diferentes, cuando realmente son sólo 2. Se ha tomado como ejemplo una pequeñísima porción de la tabla que tiene una información parcial del 2009.

Este atributo, por su variabilidad, tiene una entrada “libre” en el sistema de Necropsia, lo que posibilita que se introduzcan errores tipográficos, además como el reporte puede ser realizado por cualquier patólogo, cada uno puede usar un término diferente para referirse a la misma causa de muerte y se den situaciones como las descritas anteriormente.

El proceso de homogenizar los datos de un determinado atributo se conoce como estandarización, que en muchos casos se hace como parte del análisis y diseño del sistema, y en otros hay que hacerlo necesariamente como un proceso de limpieza de los datos para que los mismos ganen en calidad.

En el presente artículo se presenta un marco de trabajo apoyado por una herramienta computacional (DBStandardS) que ayuda a estandarizar atributos de tipo cadena.

Desarrollo

El marco de trabajo consiste en 4 pasos esenciales:

- Elección de la tabla y el campo a unificar.
- Realización de sustituciones previas.
- Agrupamiento.

- Reemplazo de valores.

El primer paso es elemental. Se debe elegir sobre qué tabla trabajar, y dentro de ella el atributo particular que se estandarizará. En la herramienta de software se utiliza una conexión ActiveX Data Object (ADO) de tal manera que el gestor de datos puede ser cualquiera.

El paso dos consiste en realizar algunas sustituciones que puede ser a partir de abreviaturas, apócopes, siglas, etc., de uso común y que pueden influir en el proceso. Por ejemplo, puede que se conozca en el campo aparezca con frecuencia “insuf” como apócope de la palabra “insuficiencia”, o “Glez” de “González”, etc. Entonces se pueden realizar estas sustituciones, lo que hace que los valores ganen en uniformidad. En el software se organizan estas sustituciones en catálogos y hay varios definidos previamente para apellidos, direcciones, etc. Pero se pueden definir todos los que sean necesarios. Estas sustituciones se almacenan en ficheros XML con la estructura siguiente:

```
<sustitucion>
    <quien>insuf</quien>
    <pque>insuficiencia</pque>
</sustitucion>
```

El tercer paso de agrupamiento pretende concentrar las cadenas similares en el mismo grupo facilitando el proceso de sustitución, de esta forma las sustituciones se harán en el grupo y se evitará tener que revisar el conjunto de datos completo. En este paso se puede utilizar cualquier técnica de agrupamiento, y como distancia entre elementos se propone utilizar una extensión de la distancia de edición de Levenshtein [2]. La distancia de edición entre dos cadenas se define como el mínimo número de operaciones de inserción, eliminación y sustitución que hay que hacer en

una cadena para convertirla en la otra. Estas operaciones se llaman operaciones de edición. En la propuesta inicial de Levenshtein todas las operaciones tienen igual costo y el mismo es unitario. Luego han aparecido otras extensiones en que se asignan costos diferentes a las operaciones.

La extensión que se propone, se justifica a partir del hecho que la escritura incorrecta de las cadenas se puede producir por las siguientes razones:

- errores ortográficos (en español, por ejemplo se cambia frecuentemente la “b” por “v”, “s” por “c” o viceversa, se omite la letra “h”, etc.);
- errores cuando son tecleados los datos, por oprimir teclas incorrectamente (por ejemplo si es necesario teclear la letra “a” puede ser oprimida la letra “q” porque ambas están cercas en el teclado), y
- errores de sonido (alguien dice “maría” y la persona que está entrando lo datos oye “manía”), también es frecuente que por sonido se omitan sonidos como “s” y “r” al final de las palabras.

La idea que sustenta la extensión que se propone es que el costo de los reemplazos, inserciones y eliminaciones, no es igual en todos los casos. Este planteamiento aparece en la literatura [3] [4] [5] [6], pero no de la forma en que se aborda en el presente trabajo.

En una primera propuesta solamente se trabajará con la operación de reemplazo, orientada en la siguiente dirección: el costo de la operación de sustitución de dos caracteres está dado por la distancia que existe entre las teclas correspondientes en un teclado. Por lo tanto en el caso de caracteres que son adyacentes, el costo de la operación es uno, en otros casos el costo es mayor que uno. Esta proposición hace que dos cadenas estén más cercas si la sustitución ocurre entre caracteres cercanos en el teclado.

En la herramienta el agrupamiento se realiza utilizando el método PAM propuesto en [8], y el teclado que se utiliza para buscar la distancia entre caracteres es el teclado “QWERTY”, que es el más usado en Cuba. Se ha incluido además la operación de transposición [7], que consiste en permutar dos letras.

Experimentación

El software construido tiene una aplicación general, pero se ha decidido utilizar en aplicaciones médicas donde aparecen cadenas de caracteres cuya estandarización es importante para poder hacer cualquier análisis estadístico de los datos.

Se utilizó la base de datos Necropsia del Hospital Arnaldo Milián Castro de Santa Clara, donde se almacenan los resultados de las necropsias realizadas en dicho hospital. De la base de datos se tomó un fragmento que contiene 263 necropsias realizadas en el 2009. Este caso resulta interesante para el estudio pues cada patólogo escribe el informe de sus necropsias, lo que hace que puedan existir algunas distorsiones en los datos de entrada. En dicha base de datos existe una tabla denominada “Protocolos”, donde hay varios campos que tienen las características mencionadas, pero solamente se presentan algunos resultados obtenidos con el campo CDM (Causa Directa de Muerte). Este campo tiene 57 cadenas diferentes.

En este caso no se aplicó el paso de sustituciones previas, porque no es habitual el uso de abreviaturas.

Al correr el programa con 10 cluster se obtuvieron agrupamientos que realmente muestran la misma causa escrita de diferentes formas. Ejemplos de algunos clusters se muestran en la Tabla 1.

1	shock mixto (septico e hipovolémico) shock mixto (séptico/hipovolémico) shock mixto(septico e hipovolemico)
2	arritmia cardiaca arritmia cardíaca arritmia cardiaca irreversible arritmia cardíaca irreversible arritmia irreversible arritmia ventricular irreversible
3	hipertension endocraneana hipertensión endocraneana hipertensión endocraneana(hemorragia protuberancial de duret) hipertension endocraneana. hemorragia de duret hipertensión endocraneana. hemorragia de duret del puente hipertensión endocraneana. hemorragia de duret izquierda. hipertensión endocraneana. muerte encefalica
4	disfunción encefálica (muerte encefálica) disfunción ventricular aguda disfuncion ventricular izquierda aguda disfunción ventricular izquierda aguda distres respiratorio del adulto insuficiencia cardio-respiratoria aguda insuficiencia cardio-respiratoria insuficiencia cardio-respiratoria aguda insuficiencia cardiorrespiratoria aguda insuficiencia respiratoria aguda insuficiencia ventricular aguda

Tabla 1: Ejemplos de clusters obtenidos con el DBStandardS

Como se puede observar, se han agrupado causas de muertes similares pero que han sido escritas con diferencias que hacen que sean tomadas como diferentes. Después de realizar las sustituciones, quedaron sólo 37 causas de muerte diferentes. En la tabla Protocolos existen otros campos que indican otras causas que pudieron provocar la muerte que presentan el mismo problema.

Conclusiones

La estandarización de atributos textuales en bases de datos médicas es una necesidad para lograr la calidad de los datos.

El marco de trabajo que se propone ayuda a resolver el problema de la estandarización, pero no constituye una herramienta automática, sino que exigen la participación de los especialistas en la toma de decisiones en cuanto a las cadenas a sustituir y se mostró efectivo en la estandarización de las causas de muerte en la base de datos de Necropsias del Hospital Arnaldo Milián Castro.

La modificación propuesta de distancia ayuda a acercar cadenas que puedan representar el mismo objeto y por lo tanto hace que queden en el mismo cluster.

Bibliografía:

1. Loshin D. Data Standardization. Embarcadero Technologies; 2002 . [Citado: 05/04/2007]. Disponible en http://www.embarcadero.com/resources/technical_papers.html

2. Levenstein V.I. Binary codes capable of correcting deletions, insertions and reversals. Sov Phys. Dokl. 1966; 10(8):707-710.

3. Xiaomin W. Liang N. Feature Weighting for Edit Distance. Thesis for the degree of Master of IA of KULeuven, 2003.

4. Ukkonen E. On approximate string matching. Proceedings of the 1983 International FCT Conference. Lecture Notes In Computer Science. 1983; 158: 487-495.

5. Cormode G. Muthukrishnan S. The String Edit Distance. Matching Problems with moves. ACM Transactions on algorithms. 2007; 3(1).

6. Cohen W. Ravikumar P. A Comparison of String Distance Metrics for Name-Matching Tasks. Proceeding of the IJCAI, 2003.

7. Lowrance R. Wagner R.A. An extension of the string-to-string correction problems. J ACM. 1975; 22(2):177-183.

8. Kaufman L. Rousseeeuw PJ. Finding Groups in Data. New York: John Wiley & Sons Inc; 1990.