Método de clústering de terminología quirúrgica para la corrección de nombres de procedimientos ingresados por texto libre.

Elián D. Bourdin, Hernán J. Navas,

Informática en salud, Sanatorio Finochietto

ebourdin@sanatoriofinochietto.com

Abstract. Los sistemas de información son una herramienta esencial para las instituciones de salud. La disponibilidad de la información, junto con su procesamiento eficiente, mejora la toma de decisiones en todos los niveles (asistenciales y gerenciales). Permiten entre otras cosas obtener estadísticas generales tanto de pacientes como de los procesos que en la institución se llevan a cabo. Sin embargo, una mala configuración en los parámetros puede llegar a dificultar significativamente los procesos de extracción de datos para el análisis estadístico. El Sanatorio Finochietto desde su apertura posee un HIS comercial que permite el manejo de la comunicación por vía electrónica entre los diferentes actores asistenciales y el personal administrativo. Todo lo que sucede en el Sanatorio queda registrado en el sistema. Luego del primer año de uso se hizo necesario extraer sistemáticamente la información para la toma de decisiones a alto nivel. La configuración del sistema permite el ingreso de texto libre dentro de la lista de procedimientos que se realizan en quirófano. Esta configuración trajo aparejada una distorsión muy grande en los cálculos estadísticos de los procedimientos llevados a cabo en la institución, por lo que se diseñó un proceso de corrección de los datos combinando múltiples técnicas informáticas.

1 Introducción

El procesamiento de la información es una parte importante en nuestras vidas, más aun en el campo de la salud. Es sabido que el manejo eficiente de los datos puede mejorar la toma de decisiones, la gestión administrativa y la educación del paciente [1]. Es por esto que la información es un requisito fundamental para la práctica médica y donde el acceso claro y sin impedimentos a los datos permite la resolución de muchos problemas.

El uso adecuado de sistemas informáticos depende de la capacitación en el área de los recursos humanos tanto personal administrativo, como al de enfermería y al médico. Para que el objetivo primordial que poseen los sistemas informáticos en salud, ligados

a mejorar la asistencia médica funcione correctamente y permita la extracción de datos de una manera clara, resulta crucial realizar una correcta configuración tanto de los parámetros, como de los permisos de accesos del sistema.

La documentación clínica electrónica puede llevarse a cabo a través de dos procesos. Por un lado, el de almacenamiento de datos a través de texto libre, lo que se traduce en una notable facilidad en la implementación del software, pero con importantes complicaciones en la utilidad del registro para el procesamiento estadístico. Por otra parte, tenemos el texto estructurado, que resulta el caso contrario al del texto libre, el cual posee mayores dificultades a la hora de la implementación informática, pero con un grado de utilidad mayor a la hora del procesamiento de los datos. Una correcta configuración del sistema, debe establecer el punto de cruce ideal entre las curvas que establecen la utilidad del texto libre y la facilidad de implementación de texto estructurado (Figura 1).

El Sanatorio Finochietto es una institución privada de salud del grupo ASE (Acción Social Empresaria), inaugurada en noviembre de 2013, con más de 133 camas de internación general para adultos, 24 de unidad de cuidados intensivos para adulto y 16 puestos de cuidados intensivos neonatales. Cuenta con 6 quirófanos generales inteligentes, 2 quirófanos ambulatorios, 2 salas de parto y 2 de preparto. Es el primer centro asistencial bio-eco-inteligente de Argentina ya que su estructura está diseñada para realizar un uso racional y responsable de los recursos naturales, como la reutilización de aguas grises y pluviales mediante terrazas verdes, o el sistema de inter-cambio geotérmico. Posee un sistema de gestión de edificios (Building Ma-nagement System - BMS) que permite la automatización integral de funciones claves para lograr dicho uso. Actualmente posee una certificación como miembro de la Red global de hospitales verdes y saludables.

El Sanatorio posee un HIS (Health Information System) comercial provisto por TIPS Salud que permite el manejo de la comunicación por vía electrónica entre los diferentes actores asistenciales y el personal administrativo. A su vez posee interfaces con todos los efectores de servicios complementarios (laboratorio, diagnóstico por imágenes, endoscopía, medicina nuclear, entre otros) y con el sistema de gestión de compras corporativo. El sistema se encuentra dividido por módulos para facilitar y ordenar la tarea de cada actor.

Los datos provenientes del sistema informático de la institución resultan de suma importancia para el procesamiento estadístico como soporte para la toma de decisiones médicas. En el este escenario descripto, se detectó un problema ligado a la carga de datos de procedimientos quirúrgicos en el módulo quirúrgico del HIS, donde ante la necesidad de cargar un nuevo término, se le brinda a un usuario genérico la posibilidad de hacerlo a través de texto libre, lo que puede derivar en una complicación importante y de difícil detección, ya que con un simple error de ortografía o el uso de diferentes abreviaturas o siglas, el sistema permite duplicar la información sin una verificación previa.

Esta situación produjo una gran cantidad de datos inútiles y redundantes, los cuales imposibilitaban la extracción de conclusiones.

El presente trabajo plantea una estrategia para la limpieza de estos datos, tomando ideas provenientes de los correctores ortográficos, los cuales en muchos casos utilizan

soluciones vinculadas a la codificación fonética como Metaphone, orientadas a cubrir errores de reconocimiento óptico basado en comparación de cadenas [2] y combinándolos con analogías en ciertos métodos informáticos para la limpieza de los datos.

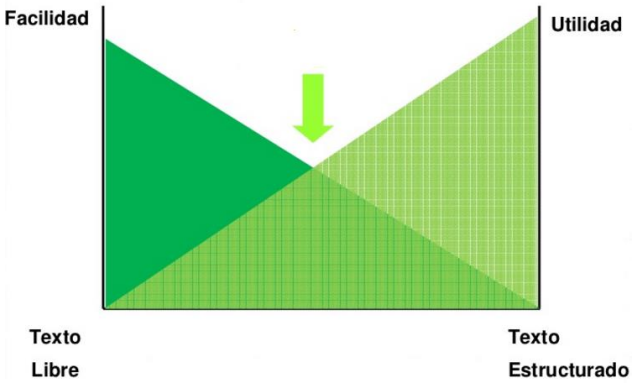

Fig. 1. Relación entre utilidad y facilidad de texto libre y texto estructurado. Podemos observar cómo se señala el punto de corte ideal entre las dos curvas [3].

2 Materiales y métodos

Se utilizó un set de datos confeccionado a través de la extracción por consultas de SQL en la base de datos principal del Sanatorio Finochietto, obteniendo de esta manera un total de 4458 procedimientos quirúrgicos únicos, llevados a cabo desde la apertura de la institución, hasta agosto de 2015. Estos se almacenaron como texto plano.

Los archivos del set de datos fueron transformados a formato FASTA. Este formato es muy común en bioinformática. Se basa en texto y soporta una secuencia o cadena de caracteres, nombres de secuencias y comentarios que preceden a la secuencia en sí. Es utilizado para representar secuencias de ácidos nucleicos o péptidos, empleando un código de una única letra para cada aminoácido o nucleótidos, es decir, una cadena de caracteres. Dado que los nombres de los procedimientos quirúrgicos son también representaciones de cadenas de caracteres, podemos utilizar esto como analogía con secuencias de nucleótidos o proteínas, para aplicar métodos de análisis de secuencias utilizados en bioinformática.

Para llevar a cabo esta tarea de formateo de datos, se utilizó PERL, un lenguaje de programación de propósito general originalmente desarrollado para la manipulación de texto y que ahora es utilizado para un amplio rango de tareas. Caracterizado por su simplicidad, además soporta tanto programación estructurada, como orientado a objetos y funcional. Incorpora un poderoso sistema de procesamiento de texto y una enorme colección de módulos disponibles.

Asimismo se utilizó el módulo de Metaphone de PERL, que contiene un algoritmo de codificación fonética con el objetivo de llevar a cabo una normalización de palabras [4], combinado con otro módulo, Levenshtein, el cual implementa la distancia de Levenshtein entre dos cadenas de texto [5]. Estos algoritmos se vinculan frecuentemente

en soluciones particulares de correctores ortográficos, asociados principalmente a cubrir errores de reconocimiento óptico, basados en la comparación de cadenas de texto (Bordignon et al., 2005).

El procedimiento de clústering de secuencias, proviene de técnicas bioinformáticas y se utiliza para el procesamiento de secuencias tanto nucleotídicas como proteicas. Esto fue llevado a cabo por el paquete de programas de CD-HIT, una de las herramientas más utilizadas para la eliminación de secuencias redundantes. Es un software de agrupamiento de secuencias que tiene como principal ventaja su velocidad. Es capaz de manejar bases de datos muy grandes a un costo computacional bajo. CD-HIT utiliza un algoritmo de agrupamiento incremental ordenando todas las secuencias (o palabras) en forma decreciente en función a su longitud. La más larga se convierte en representante de la primera agrupación. A continuación, cada secuencia restante se compara con los representantes de las agrupaciones existentes. Si la similitud con cualquier representante está por encima de un umbral determinado en la configuración de parámetros del programa, la secuencia es añadida al grupo, de lo contrario se define un nuevo grupo con la secuencia como representante [6].

3 Resultados

El set de datos analizado cuenta con 4458 procedimientos quirúrgicos únicos y se caracteriza por un alto grado de redundancia y procedimientos anotados con errores ortográficos.

Se utilizó PERL, con el módulo Metaphone, para llevar a cabo una normalización fonética sobre las palabras. Este método resulta muy robusto y se encuentra demostrado que presenta una buena performance cuando se aplica a nombres en lenguaje en inglés [7]. En este caso, el algoritmo se corrió sobre procedimientos con nombres propios de cirugías complejas, siendo necesaria una adaptación del método al español. Esto consistió en el reemplazo de acentos y caracteres propios del idioma tales como CH por X y Ñ por NY.

El próximo paso consistió en la definición arbitraria de subgrupos de secuencias, en función a la longitud de las mismas. Esto es necesario por las características de funcionamiento del algoritmo de clustering, el cual se basa en el agrupamiento incremental de cadenas de caracteres. En este punto resulta crucial tener en cuenta las características del set de datos, dado que ante una incorrecta configuración de parámetros, existen grandes posibilidades que secuencias de gran tamaño, engloben secuencias de menor tamaño de manera errónea.

- Grupo 1: longitud de secuencias de 3 a 10 caracteres.
- Grupo 2: longitud de secuencias de 11 a 20 caracteres.
- Grupo 3: longitud de secuencias de 21 a 30 caracteres.
- Grupo 4: longitud de secuencias de 31 a 40 caracteres.
- Grupo 5: longitud de secuencias de 41 a 55 caracteres.

A partir de la definición de los subgrupos de secuencias, se corrió CD-HIT con la siguiente configuración de parámetros en función a los grupos:

- Grupo 1: C = 0.97; n = 3; l = 3; s = 0,9.
- Grupo 2: C = 0.97; n = 5; s = 0.8.
- Grupo 3: C = 0.97; n = 5; l = 5; s=0.9.
- Grupo 4: C = 0.97; n = 5; l = 5; s=0.9.
- Grupo 5: C = 0.97; n = 5; l = 5; s=0.9.

Donde C hace referencia al porcentaje de identidad entre las secuencias, siendo el umbral de clustering. N es la longitud de la palabra. L es la longitud mínima de descarte de secuencias. S es el umbral de diferencia entre secuencias de menor longitud, agrupada dentro de otras de mayor longitud.

Este procedimiento nos permitió cumplir con dos tareas fundamentales, por una parte, eliminar casi en su totalidad la redundancia del conjunto de datos, ya que en una primera instancia la codificación fonética permite simplificar las cadenas de texto, sorteando de esta manera errores de ortografía más leves [8]. Por otra parte, la correcta configuración de los parámetros utilizados para la corrida de CD-HIT, posibilito agrupar los términos idénticos y muy similares.

Los procedimientos líderes de cada uno de los clusters de los diferentes grupos, fueron tomados como referencia, resultando 3552 secuencias con una tasa de redundancia sumamente baja. De esta manera se confeccionó un nuevo diccionario de términos.

Para una mayor explotación de la información brindada por CD-HIT, se analizaron los clusters. Posibilitando la detección de los procedimientos con mayor tasa de error en su escritura a la hora de cargar el dato por texto libre. Esta información es también muy importante, dado que permitió confeccionar una base de datos de “sinónimos”, que asistirá al usuario en el proceso de autocorrección de texto cuando necesite ingresar nuevamente procedimientos quirúrgicos en el sistema.

Este nuevo diccionario terminológico creado, no solo resulta de suma importancia para la asistencia futura en la incorporación de datos, sino para trabajar de manera retrospectiva, en la corrección de todo aquel término erróneo asociado a las fojas quirúrgicas ya existentes.

Se utilizó nuevamente el algoritmo Metaphone, pero en esta ocasión sobre las 10684 cirugías llevadas a cabo en la institución desde su apertura. Con los nombres normalizados tanto del nuevo diccionario de datos, como del set de cirugías totales es posible llevar a cabo un mapeo. Para esto se utilizó el módulo de distancia de Levenshtein de PERL, con el que se construyó una matriz de distancias, donde se ordenaron en las columnas las cirugías normalizadas y curadas y en sus líneas la totalidad de los procedimientos quirúrgicos del Sanatorio (Tabla 1).

Sobre la matriz de distancias conformada, se buscó el menor valor de cada uno de los vectores lineales, donde 0 indica la coincidencia plena de ambas cadenas normalizadas. Para casos donde la menor distancia es 1 se define como un alto grado de proba-

bilidad de que los procedimientos quirúrgicos sean coincidentes, mientras que al aumentar el valor de la distancia, disminuyen las probabilidades de asignación de una cirugía y resulta necesaria una supervisión por parte del usuario.

La matriz resultante permitió mapear con exactitud un 92% (9841, distancia 0) de procedimientos, mientras que un 5% (542, distancia 1 y 2) con alta probabilidad de éxito, dejando solamente un 3% (301, distancia mayor a 2) de los datos pendientes de revisión humana.

		T.R. 1	**T.R. 2**	**T.R. 3**	**…**	**T.R. N**
		N.F. (T.R.) 1	**N.F. (T.R.) 2**	**N.F. (T.R.) 3**	**…**	**N.F. (T.R.) M**
T.E. 1	**N.F (T.E.) 1**	D 1, 1	D 1,2	D 1, 3	…	D 1, M
T.E. 2	**N.F (T.E.) 2**	D 2, 1	D 2, 2	D 2, 3	…	D 2, M
T.E. 3	**N.F (T.E.) 3**	D 3, 1	D 3, 2	D 3, 3	…	D 3, M
…	**…**	…	…	…	…	…
T.E. N	**N.F (T.E.) N**	D N, 1	D N, 2	D N, 3	…	D N, M

Table 1. T.E. Término con posibilidad de error. N.F. normalización fonética. T.R. termino de referencia. D: distancia

4 Conclusiones

Establecimos una relación entre las secuencias proteicas o nucleotídicas, con los procedimientos quirúrgicos, dado que en definitiva ambos pueden tratarse de la misma manera.

Partiendo de un set de datos de baja confiabilidad, compuesto por 4458 términos referentes a procedimientos quirúrgicos, se utilizaron algoritmos de procesamientos de información provenientes de diferentes disciplinas para llevar a cabo de redundancia. Esto permitió eliminar 906 términos, en su mayoría incluidos originalmente en el diccionario del módulo quirúrgico por contener errores de ortografía y no coincidir idénticamente con las cadenas de caracteres ya existentes.

Por otra parte, se desarrolló un método capaz de relacionar de manera retrospectiva los procedimientos quirúrgicos originales, permitiendo de esta manera una corrección en la base de datos de referencia en la institución. Esto es de vital importancia para el correcto cálculo de las estadísticas relacionadas al módulo de quirófano.

Dentro de las fortalezas del algoritmo debemos mencionar la gran capacidad para procesar grandes cantidades de datos en muy poco tiempo.

Mientras que la mayor debilidad radica en la tasa de error que posee el método de clústering cuando la longitud de la palabra es pequeña [9]. Es aquí donde deben conocerse las características del set de datos que se está tratando y escoger la correcta conformación de los sub grupos de palabras y adaptar una correcta configuración de los parámetros necesarios para la corrida de CD-HIT. La perspectiva de trabajos futuros, apunta a mejorar esta deficiencia del algoritmo, probando distintos métodos de clustering.

5 Bibliografía

1. Margolis Álvaro: La informática en salud. Posibilidades y desafíos. Rev Med Urug. 12, 81–98 (1996).
2. Bordignon, F., Tolosa, G.H., Peri, J.A., Barrientos, D.: Método de corrección ortográfica basado en trigramas y distancia de edición. In: VII Workshop de Investigadores en Ciencias de la Computación (2005).
3. Middleton, Leavitt, M, Renner, K: Ambulatory practice clinical information management problems and prospects. Healthc. Inf. Manag. J. Healthc. Inf. Manag. Syst. Soc. Am. Hosp. Assoc. 11, 97 (1997).
4. Binstock A., Rex J.: Metaphone: A modern Soundex. Pract. Algorithms Program. Read. Mass Addit.-Wesley. 160–169, (1995).
5. Josh Goldberg, Neil Bowers: Modulo Levenshtein.
6. Li W., Godzik A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22, 1658–1659 (2006).
7. Snae, C.: A comparison and analysis of name matching algorithms. Int. J. Appl. Sci. Eng. Technol. 4, 252–257 (2007).
8. Gálvez, C.: Identificación de nombres personales por medio de sistemas de codificación fonética. Encontros Bibli Rev. Eletrônica Bibl. E Ciênc. Informação. 2, 105–116 (2006).
9. Li, W., Jaroszewski, L., Godzik, A.: Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics. 17, 282–283 (2001).