Recuperación aproximada de direcciones postales

Approximate retrieval of postal addresses

Yaiza Temprado
Telefónica I+D
ytr@creativit.com

Resumen: En este artículo se presenta FuMaS (Fuzzy Matching System), un sistema que permite la recuperación eficiente de direcciones postales a partir de consultas con ruido. La recuperación difusa de esta información tiene innumerables aplicaciones, desde encontrar/limpiar duplicados en bases de datos (registros electorales, encontrar nidos de fraude postal, etc.) hasta corregir las entradas de los usuarios en sistemas tales como callejeros o cualquier tipo de formulario dónde haya que introducir una dirección postal. Los resultados de estos experimentos muestran que FuMaS es una herramienta muy útil para recuperar direcciones postales a partir de consultas con ruido, siendo capaz de resolver cerca del 85% de las direcciones con errores introducidas al sistema; una eficacia un 15% mayor que cualquier otro sistema similar probado.
Palabras clave: recuperación de información, record linkage, direcciones postales, minería de texto.

Abstract: In this article it is introduced FuMaS (Fuzzy Matching System), a system that allows efficient recovery of addresses through noisy queries. The fuzzy retrieval of this information has countless applications, from finding / clean duplicates in databases (voter registration, find nests of mail fraud, etc.) to correct the input from users on systems such as street directories or any type of form where an address has to be filled. The results of these experiments show that FuMaS is a very useful tool for addresses retrieval in noisy queries, being able to resolve about 85% of the addresses with errors which were introduced into the system; a 15% higher efficiency than any other similar system tested.
Keywords: information retrieval, record linkage, postal address, text mining.

1 Motivación

La integración de información es un área muy importante de investigación dentro de los campos de las bases de datos y de la minería de datos (Batini, Lenzerini y Navathe 1986).Integrar múltiples fuentes de información distintas, permite obtener una visión más completa y precisa del mundo, así como obtener conocimiento adicional del mismo. Uno de los problemas más usuales a la hora de integrar grandes bases de datos es el problema de detectar (para posteriormente limpiar o integrar) fragmentos de varios registros que tratan de las mismas entidades. Detectar varias registros tratando sobre las mismas entidades puede ser una tarea simple si la información que identifica a las identidades es la misma (y está completa) en todos los registros, pero esto no es algo común, ya que no siempre se tienen los mismos identificadores en todas las fuentes de datos a integrar (por ejemplo, lo que puede ser el dni en una base de datos, en otra puede ser el nombre). Otra fuente de dificultades se debe a la existencia de los mismos identificadores pero presentando algún tipo de ruido que hace que la coincidencia entre los mismos no sea perfecta.

La expresión *record linkage* (vinculamiento de regisros) se refiere precisamente al uso de técnicas algorítmicas para encontrar registros que, aunque no identifiquen exactamente de la misma forma a una entidad, sí que se refieren a la misma (Baeza-Yates y Navarro 1997). Dentro de la literatura, el proceso de linkado de registros se encuentra asociado a una gran variedad de nombres: hetereogeneidad de

identidad, identificación de identidades, identificación de instancias, mezclar/purgar, reconciliación de entidades, lavado de listas y limpieza de datos.

Las aplicaciones del vinculamiento de registros son innumerables, sobre todo en entornos administrativos: desde la gestión de relaciones con los clientes, detección del fraude, data warehousing, encontrar registros en diversas fuentes pertenecientes a un mismo paciente, etc.

El enfoque general de los algorimtos de vinculación de registros es determinar un coeficiente de similitud entre cada par de registros, lo cual es una tarea muy pesada, del orden de $O(n^2)$. Para poder determinar el coeficiente de similitud entre dos registros, usualmente se han utilizado distancias de edición (Navarro y Raffinot 2002) como puedan ser la distancia de Levenhstein o el algoritmo de Smith-Waterman. Estas distancias permiten calcular el número mínimo de operaciones sobre los caracteres (sustitución, inserción o eliminación) necesarias para convertir una cadena en otra. Existe una gran variedad de algoritmos de cálculo de distancias de dicción que pueden ser utilizados tanto dentro del proceso de vinculación de registros como en muchas otras aplicaciones tales como el procesamiento de señales con ruido, la recuperación de información con errores textuales y la biología computacional.

En este artículo se trata la vinculación de registros aplicada al problema del reconocimiento difuso de direcciones postales.

Las direcciones postales pueden considerarse como información estructurada, ya que contienen varios campos bien conocidos y delimitados (tipo de vía, nombre de la vía, código postal, municipio, etc.). Ahora bien, una dirección postal puede escribirse de múltiples maneras; de hecho la siguiente lista de direcciones correctas representa la misma dirección postal (calle Santa María Magdalena, número 5, 4º B, 28900, Madrid):

- C/ Santa Maria Magdalena, n. 5, 4B, 28900, Madrid
- C/ Sta Maria Magdalena, n. 5, 4B, 28900, Madrid
- C/ Santa M. Magdalena, n. 5, 4B, 28900, Madrid

Los 3 ejemplos anteriores muestran la misma dirección postal, pero escrita de 3 posibles maneras distintas. Esto es posible debido a la utilización de abreviaturas, sinónimos, contracciones y otras formas lingüísticas que permiten generar cierta ambigüedad o variedad de formas escritas para los mismos conceptos.

Además de estas formas correctas, también se pueden encontrar formas incorrectas que, pese a referirse a la misma dirección, muestran ciertos errores o datos faltantes. Como ejemplos:

- C/ Santa Maria Magdalena, Madrid
- C/ Santa Maria Madalena, n. 5, 4B, 28900, Madrid
- C/ Santa Marai Magdalena, n. 5, 4B, 28900, Madrid
- C/ Santa Maria Magdalena, n. 54, B, 28900, Madrid
- AV Santa Maria Magdalena, Madrid

No sólo se pueden encontrar varias formas correctas de escribir una dirección postal, ni siquiera se tiene que tener únicamente en cuenta el hecho de poder encontrar errores en las direcciones, sino que se pueden encontrar combinaciones de ambas, es decir, errores en direcciones que están ya utilizando formas alternativas:

- C/ Sta Marai Madalena, Madrid
- C/ S M Madalena, 5, 4B, 28900, Madrid

2 Descripción y Arquitectura del Sistema

La arquitectura del sistema de FuMaS es una arquitectura dividida en 3 capas, cada una de ellas representativa de un nivel de abstracción. FuMaS recibe como entrada una dirección postal y devuelve como salida un ranking de posibles direcciones postales, cada una con una medida de la calidad de la solución, ordenadas de mayor a menor relevancia.

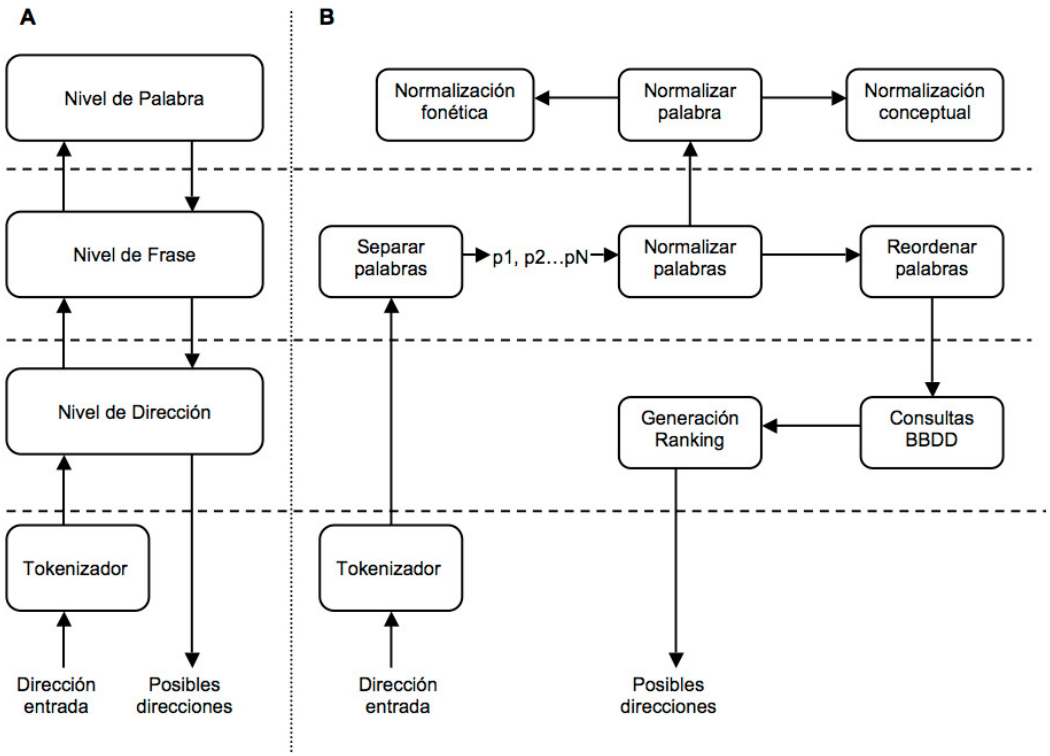
Como se puede ver en la subfigura 1.A, una dirección que se introduzca como entrada del sistema pasará por los siguientes niveles:

- Tokenización: Se puede considerar como un paso de preprocesamiento más que como un nivel de abstracción (por eso se contabilizan únicamente 3 capas). Este nivel se encarga de recoger una cadena de caracteres conteniendo

una dirección de entrada y la separa en los diferentes elementos constituyentes de la misma: tipo de vía, nombre de vía, número de portal, letra, escalera, código postal y municipio. Actualmente el tokenizador utilizado en FuMaS es un tokenizador *ad-hoc* que espera los campos en un determinado orden, aunque, gracias a la modularidad del sistema, se puede sustituir por un tokenizador más inteligente capaz de lidiar con diversos tipos de estructuras de dirección postal.

- Nivel de Dirección: Constituye la capa de mayor abstracción, ya que es la capa encargada de estudiar las coherencias entre todos los campos de una dirección postal. Ésto incluye las asociaciones entre códigos postales, municipios y nombres de calles, el número de portal asociado a un determinado domicilio, etc. Además, es el responsable de generar la salida del sistema: una lista de posibles direcciones postales, ordenadas de mayor a menor relevancia. Para poder generar el ranking de salidas del sistema se utiliza un algoritmo de cálculo de distancia de edición entre dos cadenas de caracteres (distancia de Levehnstein, Q-gram, etc.), aunque la elección del algoritmo concreto se ha dejado como un parámetro del sistema definible por el usuario.

- Nivel de Frase: Esta capa es la encargada de trabajar con las cadenas de palabras (frases). Principalmente se encarga de corregir los fallos debidos a la eliminación, inserción, sustitución o transposición de palabras dentro de una frase.
- Nivel de Palabra: Es la capa a más bajo nivel. Se encarga de corregir los fallos debidos a la eliminación, inserción, sustitución o transposición de letras dentro de una palabra, así como de normalizar las palabras, tanto a nivel fonético como conceptual. La normalización fonética se encarga de representar las palabras de una forma más cercana a como se pronuncian para corregir errores como la falta de una 'h' o la sustitución de una 'b' por una 'v'. La normalización conceptual se encarga de unificar distintas formas de escritura referentes a un mismo término (abreviaturas, sinónimos, etc.).



Esta arquitectura basada en niveles de abstracción permite un acercamiento más simplista a un problema bastante complejo de por si. Cada una de las capas puede modificarse, añadiendo mayor o menor complejidad en función de si lo que se busca son resultados más precisos o menor tiempo de ejecución. Esto ha permitido tener, en un periodo de tiempo relativamente corto, un prototipo del sistema plenamente funcional (aunque muy mejorable), para poder evaluar el rendimiento del mismo y, por ende, evaluar las posibilidades de éxito de la arquitectura propuesta.

3 Experimentos

Para poder evaluar el sistema, se ha construido una pequeña colección de 100 direcciones postales correctas y se han generado variantes incorrectas de las mismas, consiguiendo una colección de 300 direcciones postales en total. Las direcciones postales incorrectas se han generado teniendo en cuenta distintos posibles fallos a la hora de escribir las direcciones postales.

Con esta colección de direcciones, se ha comparado FuMaS con otro sistema similar (UNISERV), así como con 3 gestores de mapas online que incluyen algún tipo de corrección automática (Guía Campsa, Google Maps y LaNetro). Los resultados experimentales muestran que FuMaS es el sistema con mejor funcionamiento, logrando corregir cerca de un 85% de las direcciones postales introducidas en el sistema, seguido de UNISERV con un 70% de errores corregidos, la Guía Campsa con cerca del 60%, Google Maps con poco más del 40% y LaNetro que apenas es capaz de corregir errores básicos.

4 Conclusiones

En este artículo se ha presentado FuMaS, un sistema que permite la recuperación eficaz de direcciones postales con ruido. Los resultados experimentales muestran que FuMaS, a pesar de su pronto estado de gestación, es capaz de corregir un 85% de las direcciones erroneas introducidas al sistema, superando por más de 15 puntos a cualquier otro software similar evaluado.

FuMaS está lejos de ser un sistema acabado y cerrado; por ahora muestra una arquitectura capaz de lidiar con el sistema, y lo suficientemente modular como para permitir muchos grados de mejora. Por otra parte, los resultados presentados en este artículo representan la piedra de apoyo de una nueva línea de investigación que pretende abordar el problema de la recuperación aproximada de información estructurada, tanto en el dominio de las direcciones postales como en otros dominios donde la arquitectura de FuMaS pueda ser adaptada gracias a su gran flexibilidad.

En este sentido, cabe señalar que de los 3 niveles de abstracción de FuMaS, dos de ellos son muy generales (el de frase y el de palabra) y pueden ser reutilizados en su práctica totalidad en otros ámbitos.

Bibliografía

Baeza-Yates, R., y Navarro, G., A Practical Index for Text Retrieval Allowing Errors, CLEI, 1:273-282. 1997.

Batini, C., Lenzerini, M, y Navathe, S. B., A comparative analysis of methodologies for database schema integration. ACM Comput. Surv. 18(4):323-364, 1986.

Navarro, G., y Raffinot, M. Flexible Pattern Matching in Strings – Practical online search algorithms for texts and biological sequences. Cambridge University Press 2002.