Revista de Gestión Empresarial y Sustentabilidad

http://rges.umich.mx

Hacia Resúmenes de Texto Inducidos por Consultas Semánticamente Aumentadas

J Guadalupe Ramos Díaz[1]

Ramón Guztavo Ramos Díaz[2]

Héctor Oceguera Soto[3]

[1]Tecnológico Nacional de México / Instituto Tecnológico de La Piedad, jgramos@pricemining.com

[2]Universidad Michoacana de San Nicolás de Hidalgo, rg.omega77@gmail.com

[3]Instituto Tecnológico de La Piedad, hos6509@hotmail.com

Revista de Gestión Empresarial y Sustentabilidad, Vol. 4, Núm. 1, Enero – Diciembre 2018, ISSN: 2448-6043, pp. 13 – 29, Fecha de Recepción 31 de Julio de 2018, Fecha de Aceptación 16 de Octubre de 2018

Hacia Resúmenes de Texto Inducidos por Consultas Semánticamente Aumentadas

Resumen

Gente platicando con dispositivos inteligentes será, tal vez, el estilo más usado de interfaces de computadoras en el futuro cercano. Ahora, este fenómeno ocurre cuando una persona interactúa, por ejemplo, con un agente de software y se le hace una consulta, de hecho, esto es común con los teléfonos inteligentes y las tabletas. Normalmente, el sistema operativo de los dispositivos provee un cierto tipo de asistente donde el usuario pregunta y el software responde.Un gran reto de los asistentes es la reducción de la información para poderla presentar al usuario.

En este trabajo se presenta una técnica algorítmica que construye automáticamente un reporte de texto a partir de un grupo de documentos y a partir de una consulta, en particular, se extraen los fragmentos más similares a la consulta. El cálculo estándar de similitud de coseno se enriquece a partir de un componente semántico que se obtiene mediante la aplicación de la técnica de Análisis Semántico Latente. En este sentido recuperamos no sólo el texto con similitud estándar sino aquellos que comparten un nivel de semejanza de significado. Así pues, preparamos un reporte de texto (resumen) a partir de múltiples documentos con similitud semántica con respecto a una consulta de usuario.

Palabras Clave: *Web Semántica, resumen, extracción de información*

Abstract

People talking with intelligent devices, may be, will be the most used style in the following days for human computer interfaces. Now, this phenomenon occurs when a person interacts, for instance, with a software agent asking or querying for something, indeed it is common with smartphones and tablets. Regularly smart devices' system operative offers a certain typeof assistant, the user states a query and the software answer her.

A big problem for that type of software assistants is the taskof information reduction in order to prepare and to present asummarized report from many information sources.

In this work we present an algorithmic technique that automaticallyconstructs a text report induced by a query from a setof documents. The method extracts text excerpts from documentsby considering the most similar text w.r.t. the query as a retrievalcriterion. The well known cosine method for similarity calculationis enriched by a semantic flavor incorporated by means of theresults of the application of the now standard technique for semanticdiscovering, i.e., Latent Semantic Analysis. In this setting,we recover not only text with literal similarity but also thosethat share a level of meaning likeness. In this way we prepare asummary from source documents with semantic similarity w.r.t.a user query by exploiting latent semantic analysis.

Keywords: *Semantic Web, summary, information extraction*

Introducción

Cada día es más común la interacción entre personas y dispositivos computacionales por medio de lenguaje hablado. Las razones asociadas a este fenómeno son muchas, por un lado las pantallas de lectura escritura de los dispositivos de mayor uso, i.e., de los teléfonos inteligentes son muy pequeñas, es decir, ofrecen una comunicación anti-natural eincómoda y por otro el volumen de información disponible en la Web es cada vez mayor. Así pues, podemos percibir que el uso de asistentes de software capaces de atender peticiones verbales tiene un pronóstico de uso creciente. De hecho, el advenimiento del Internet de las Cosas, que implica millones de dispositivos interconectados y con ello la posibilidad de emitir instrucciones de control, sin teclados, empuja aún más la necesidad de interfaces de órdenes mediante lenguaje hablado.

Una de las tareas que se espera que un agente de software pueda llevar a cabo es la de preparar reportes o resúmenes de información, esto es, un usuario lanza una consulta, el agente busca, encuentra y discrimina información y entrega finalmente un extracto de los textos consultados, esto es $A(Q, d_1,\ldots,d_n)\rightarrow S$ donde A es una función que recibe como argumentos una consulta de usuario Q, una lista de documentos $d_1 \ldots d_n$, denominada la *colección* y entrega como resultado un resumen de texto S.

En el presente documento se introduce un método para abordar la solución a este problema, primero se presenta el modelo formal en la sección "Modelo Vectorial", después en la Sección "Agregación de Componente Semántico en los Resúmenes**"** se analiza la técnica de Análisis de Semántica Latente y como se puede usar para construir resúmenes. En la "Creación de Resúmenes Semánticamente Aumentados a Partir de Consultas de Usuario" se bosqueja el método para crear resúmenes, para, enseguida en la sección "Prototipo" presentar evidencias del desarrollo y finalmente en la sección "Conclusiones" se discute trabajo relacionado y se concluye.

El Modelo Vectorial

En esta sección introducimos la representación formal estándar para documentos de texto escritos en lenguaje natural, dado que se trata de un marco formal ampliamente usado para el procesamiento de lenguaje natural.

El modelo de espacio de vectores para indexado automático de texto fue formulado por Salton et al. 1975, y se considera una técnica de representación estándar en el ámbito de recuperación de información en el que las entidades de análisis, i.e., documentos de texto se comparan unos con otros.

Dado un documento de texto *d*, un diccionario de términos es un conjunto cuyos elementos son las diferentes palabras presentes en el documento *d* (término y palabra se emplean indistintamente). *V*(*d*) denota el vector asociado al documento *d*, cuyos componentes son los pesos de cada elemento en el diccionario. Se asume que los pesos de los elementos (valor numérico asociado a cada palabra) se computan mediante el esquema *tf* (*term frequency*), esto es, el valor de un componente particular se da de acuerdo al número de veces que aparece la palabra en el documento *d*.

El conjunto de documentos en una colección puede abstraerse como un conjunto de vectores en un espacio vectorial. Al representar un documento mediante un vector, existe un eje para cada término. Esta representación pierde el orden relativo de los términos que tenían en el documento origen (Manning et al., 2008). A esta visión de documento se le suele llamar *modelo de bolsa de palabras*, donde el orden de los términos en el documento se ignora, pero el número de ocurrencias en dicho documento sí se considera. No obstante,

es intuitivo pensar que dos documentos con similar representación de bolsa depalabras son similares en contenido.

Ejemplo 1: Dado *texto* = "sitios Web, servicios Web o aplicaciones Web-basadas", el diccionario (sin distinción de mayúsculas y minúsculas) se compone de {web, sitios, servicios, basadas, aplicaciones}, entonces *V*(*texto*) es <3,1,1,1,1>. Aquí el término Web aparece 3 veces, sitios 1,etc.

En el Ejemplo 1 aparecen símbolos tipográficos como ","o "-", regularmente este tipo de elementos de texto se ignoran en la representación vectorial. De igual manera hay palabras que no agregan significado y que no se toman en cuenta, por ejemplo, la palabra "o". Estas palabras aparecen en textos de cualquier tema, son extremadamente comunes, y por tanto no ayudan a distinguir el significado de un texto, se les llama de manera estándar *stop words*.

Las etapas para analizar documentos de texto en procesamiento de lenguaje natural son:

- Filtrado: Se trata de eliminar símbolos tipográficos: ",.:;?"
- Remoción de *stop words*
- Construir bolsa de palabras (diccionario del documento)
- Indexar, esto es construir la representación vectorial del documento.

El objetivo de nuestra técnica es descubrir y extraer aquellos fragmentos de texto provenientes de los documentos que presenten la mejor similitud con respecto a una consulta de usuario. Una consulta de usuario es también un fragmento de texto escrito en lenguaje natural, por ejemplo "*técnicas aplicadas a sitios Web, servicios Web, o aplicaciones Web basadas"*. Este tipo de consulta es lanzada comúnmente a un buscador Web o bien a asistentes o sistemas de ayuda. La similitud es una medida típica entre fragmentos detexto en lenguaje natural. La forma estándar de cuantificar la similitud entre dos textos *t*1 y *t*2 es computar la similitud de coseno de sus representaciones vectoriales *V(t1)* y *V(t2)* (Manning et al., 2008).

Definición 2 (similitud): La similitud entre fragmentosde texto *t1*, *t2* se define por *sim(t1, t2) =V(t1)V(t2) / |V(t1)| |V(t2)|* donde el numerador representa el producto punto de los vectores *V(t1)V(t2)*, y el denominador es el productode sus longitudes euclidianas.

El producto punto de los dos vectores *v,w* es$\sum_{i=1}^{n} v_i\, w_i$ mientras que la longitud euclidiana de *t1* es$\sqrt{\sum_{i=1}^{n} V_i^2(t1)n}$ es el número máximo de palabras distintas entre *t1* y *t2*. Una similitud total computa un valor de 1.

Ejemplo 3: Si *t1* =”World Wide Web” y *t2* =”sitios Web, servicios Web o aplicaciones Web-basadas” entonces el diccionario para *t*1 y *t*2 considerados como el total del contexto es “world, wide, web, sitios, servicios, basadas, aplicaciones”. La representación vectorial es: *V(t1)* = <1,1,1,0,0,0,0>, *V(t2)* = <0,0,3,1,1,1,1>. Por lo tanto *sim*(*t*1,*t*2) es $3/\sqrt{3}\sqrt{13}$, i.e., 0.48.

A grandes rasgos, nuestro método tomará una consulta de usuario y entonces computará muchas pruebas de similitud entre la consulta y el texto de los documentos buscando textos similares a la consulta con la finalidad de producir un reporte automático. En este trabajo incluimos los resultados de un análisis de semántica latente (Furnas et al., 2017) sobre los documentos para agregar un componente semántico a los resúmenes que se pueden confeccionar con nuestro método.

Agregación de Componente Semántico en los Resúmenes

En la presente sección se analiza la incorporación de semántica a los resúmenes de texto. Para ello, primero se introduce la técnica de análisis de semántica latente y su mecanismo matemático denominado descomposición de valores singulares, enseguida, se hace un análisis de las capacidades aportadas por la técnica formal y la necesidad de nuevos métodos con el fin de fijar la base de la propuesta de este trabajo.

A. Análisis de semántica latente

El análisis de la semántica latente: LSA (del inglés *Latent Semantic Analysis*) es un modelo computacional dentro de la Inteligencia Artificial que busca replicar la construcción de significados (conceptos) de los humanos.

LSA es una teoría y método para la extracción y representación del significado a partir del uso contextual de palabras mediante cálculos estadísticos aplicados a un corpus extenso de texto. La idea subyacente es que el agregado de todos los contextos de palabras en las cuales aparece o no aparece una palabra específica provee un conjunto de restricciones

mutuas que en el volumen de texto determina la cercanía o similitud de significado de palabras y conjuntos de palabras unas con respecto a otras.

A grandes rasgos una palabra tiene un significado a partir de la ocurrencia en común con otro conjunto de palabras, por ejemplo, en contextos donde se habla de una computadora será recurrente encontrar también los términos, computador uordenador. Algo similar podría pasar con la palabra ratón, es muy posible que en el mismo espacio de términos aparezca la palabra roedor. Justamente ese tipo de relaciones son las que descubre LSA.

La primera etapa de LSA consiste en transformar el texto en una matriz M en la cual cada renglón i se emplea para representar una única palabra y cada columna para representar un documento j. Cada celda $M_{i,j}$ contiene la frecuencia con la cual la palabra i aparece en el documento j. Los valores de frecuencia pueden ser normalizados con algún otro esquema para representar la aportación de la palabra al documento, e.g., *tf.idf* (Manning et al., 2008).

Enseguida, en LSA se aplica la técnica de descomposición de valores singulares: SVD (del inglés *Singular Value Decomposition*) a la matriz M. SVD es una técnica estándar que se aplica en algebra lineal sobre matrices, es una forma específica de análisis factorial. En la matriz original M los términos y documentos son mutuamente dependientes entre ellos.

En SVD una matriz rectangular se descompone en elproducto de otras tres matrices, i.e., $M=USV^T$, las cuales tendrán vectores singulares o valores singulares. La matriz resultante U contendrá la representación vectorial delas palabras, las cuales tendrán independencia lineal de la relación con los documentos, mientras que V, contendrá la representación vectorial de los documentos cuyos componentes serán linealmente independientes de la relación con palabras en M. Finalmente S es una matriz diagonal en losque se encuentran en orden descendente los valores singulares que representan las relaciones que mantienen ambas matrices de vectores singulares. Esta técnica está ampliamente cubierta por diferentes librerías en el ámbito matemático o en distintos *frameworks* de ciencia de datos, en este trabajo se emplea únicamente como una herramienta, nuestro objetivo es desarrollar productos de tecnología avanzada con bases científicas, sin embargo para una revisión abundante el lector puede consultar (Furnas et al., 2017).

La matriz original puede reconstruirse multiplicando las tres matrices resultantes. Cuando se realiza la reconstrucción sepueden elegir los primeros k elementos: $M' = U_k S_k V_k^T$, con

ello, se obtiene una nueva matriz M' en la que el ruido introducido por los elementos más insignificantes se elimina.

Así, los nuevos valores $M'_{i,j}$ descubren relaciones latentes entre las palabras y los documentos.

Ejemplo 4: Consideremos las siguientes cuatro frases:

1) *d1* =La computadora traía software de marca

2) *d2* =Un ordenador es útil solo con software de marca

3) *d3* =El hardware de ordenadores (o computadora) puedeser genérico

4) *d4* =Software de marca y hardware genérico va biencon mi computadora

De aquí el diccionario de la colección es: {*computadora,software, marca, ordenador, hardware, genérico}*. De acuerdo a lo dicho arriba, el primer renglón es para abstraer el término"computadora" y la columna 1 ser 'a la del primer documento. Por ello tenemos 6 renglones, uno para cada elemento del diccionario y 4 columnas correspondientes a 4 documentos.

Entonces $M_{1,4}$ se refiere al número de veces que aparece el término computadora en el documento 4 y así sucesivamente. Al aplicar la técnica SVD obtenemos $M = USV^T$

$$M = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} U = \begin{bmatrix} 0.49 & 0.21 & 0.35 & 0.77 \\ 0.47 & -0.50 & 0.02 & -0.17 \\ 0.47 & -0.50 & 0.02 & -0.17 \\ 0.26 & 0.14 & -0.93 & 0.22 \\ 0.35 & 0.47 & 0.08 & -0.39 \\ 0.35 & 0.47 & 0.08 & -0.39 \end{bmatrix}$$

$$S = \begin{bmatrix} 3.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.74 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.19 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.6 \end{bmatrix} V^T = \begin{bmatrix} 0.45 & 0.38 & 0.46 & 0.67 \\ -0.46 & -0.50 & 0.73 & 0.08 \\ 0.32 & -0.75 & -0.36 & 0.45 \\ 0.70 & -0.22 & 0.35 & -0.58 \end{bmatrix}$$

Al reconstruir la matriz M considerando k=2 obtenemos

$$M' = \begin{bmatrix} 0.542 & 0.413 & 0.986 & 1.085 \\ 1.067 & 0.993 & 0.044 & 0.929 \\ 1.067 & 0.993 & 0.044 & 0.929 \\ 0.265 & 0.195 & 0.557 & 0.577 \\ 0.133 & 0.019 & 1.115 & 0.821 \\ 0.133 & 0.019 & 1.115 & 0.821 \end{bmatrix}$$

Si computamos la similitud entre los vectores conformadospor el renglón 1 <1,0,1,1> y 4 <0,1,1,0> de M, esto es comparar la similitud entre los términos ordenador y computadora, obtendremos como resultados 0.4, mientras que hacer la misma operación con los vectores del renglón 1 y 4 en M' nos daremos cuenta que hay un resultado de 0.99. Esto resulta revelador ya que en la colección se usa indistintamente ordenador y computadora, aunque

coinciden ambas en sólo undocumento, las correlaciones con el resto de términos permiten descubrir un parecido latente aún mayor. Estas evidencias de relaciones serán aprovechadas más adelante por nuestro método.

La técnica permite descubrir evidencias de relaciones no visibles en la matriz original, si bien esto es importante, la realidad es que en muchas aplicaciones de procesamiento de lenguaje hace falta tratar con mayor granularidad el texto, enese sentido hacen falta métodos que exploten los resultadosde SVD, en la sección IV abundamos.

B. Construcción de significado

En este apartado abordamos la discusión de construcción de significados y la granularidad de aplicación de técnicas a métodos de procesamiento de lenguaje natural.

En la literatura se hace énfasis en presentar la técnica LSA como un proceso de construcción de significados a partir de la coocurrencia de términos en un conjunto de documentos, es verdad que la aparición reiterada de los términos en contextos de discurso de un mismo tema implica que las palabras que ahí aparecen aportan cierta semántica a la propia definición del tema y su significado.

A partir de la matriz reconstruida por SVD se pueden hacer nuevas medidas de similitud entre dos términos y descubrir relaciones de cercanía de significado que originalmente no eran observables. Sin embargo, no puede garantizarse con absoluta certeza que dos términos que ahora tienen más similitud son en realidad sinónimos, al menos no como lo podría garantizar un tesauro construido por humanos. Así pues, requerimos incorporar tales medidas de aproximación de similitud en nuevos métodos que exploten esa evidencia de codependencia de términos.

Por otro lado, muchos métodos de manipulación de texto no solamente requieren tratar al documento como una unidad textual, que de hecho es algo inmediato al reconstruir la matriz M', a veces es necesario analizar fragmentos para construir resúmenes, en este sentido, nuevamente, es necesario el diseñode métodos que consideren la evidencia de las relaciones recuperadas a partir de LSA pero que puedan aplicarse a gránulos pequeños de texto, como hacemos enseguida.

Creación de Resúmenes Semánticamente Aumentados a Partir de Consultas de Usuario

En la presente sección abordamos la construcción de resúmenes a partir de una consulta de usuario. Para la confección del resumen se tomaría en cuenta la evidencia de codependencias de términos arrojadas por LSA y materializadas en la matriz reconstruida a partir de SVD. Definamos formalmente consulta de usuario y gránulo de texto, conceptos que se requerirán más adelante:

Definición 5 (Consulta de usuario): Sea Q la consulta de un usuario, esto es, una cadena de texto tal que $dict(Q) + dict(d1) + \ldots + dict(d_n)$ constituyen el diccionario del contexto de análisis, *dict* es una función que devuelve el diccionario del texto representado por su argumento, + es la operación de concatenación de cadenas, y $d_1 \ldots d_n$ son el conjunto de documentos en el contexto de análisis a considerarpor el método.

Definición 6 (Gránulo de texto): Sea g un fragmento cualquiera de texto de un documento d tal que $dict(g)$ es subconjunto contenido de $dict(d)$ y $|dict(g)| <= |dict(d)|$

Para efectos del método se pretende que se puedan computar similitudes entre Q y g a efecto de entregar un resumen compuesto por fragmentos g de los documentos d, y sólo aquellos que presentan mayor similitud con respecto a Q. En la siguiente sección se detallará la forma de incorporar el resultado de LSA como componente semántico de los resúmenes.

A. Consideraciones de diseño del método

LSA permite a partir de la reconstrucción de la matriz de codependencias de términos y documentos encontrar nuevos valores numéricos que evidencian las relaciones justamente entre términos y documentos. Sin embargo, en la práctica, para construir herramientas de procesamiento de texto en lenguaje natural es necesario acceder a fragmentos de texto más pequeños al documento, por ejemplo, para construir resúmenes es necesario poder manipular párrafos o líneas,es decir la granularidad (variedad de tamaño de gránulos) es más fina y por cierto podría ser variable en función del método. Para determinar codependencia entre gránulos y términos habría que aplicarse cada vez el cálculo LSA y

eso sería muy poco práctico porque implicaría reconstruir en cada ocasión matrices en función de la unidad de tamaño de texto seleccionada.

En este sentido, los valores arrojados por LSA en la matriz reconstruida deben servir como referencia para poder manipular los trozos de texto sujetos a análisis. Nuestro método, debe recoger los resultados de LSA y aplicarlos en el tratamiento del texto independientemente del tamaño del gránulo.

Por otro lado, a partir de LSA no se puede garantizar la igualdad semántica absoluta, es decir no podemos determinar que dos términos son realmente sinónimos, esto se debe a que la similitud se estima con base al conjunto de texto analizado, mientras que en el lenguaje natural dos términos son sinónimos porque las personas así lo han determinado. En contraste la igualdad semántica, sí se podría enunciar con un tesauro.

Así pues, se determinan dos requerimientos de diseño:

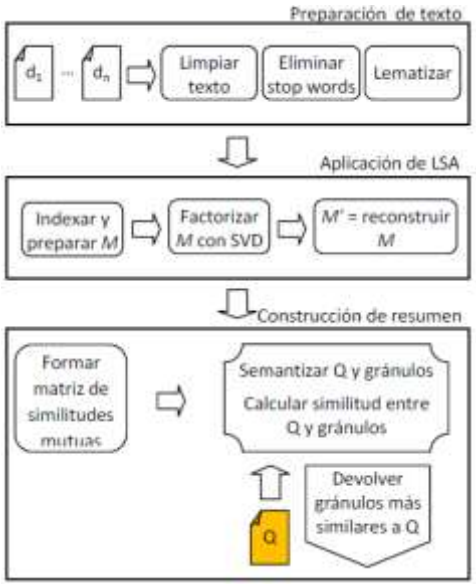
- Tamaño de gránulo independiente del método
- Aprovechamiento de los valores que permiten determinar similitud en la matriz reconstruida por LSA

Los cuales recogemos para la formulación de nuestro método para crear resúmenes.

B. Procesos del generador de resúmenes

En la Figura 1, se introduce la arquitectura de la herramienta que opera a partir de un conjunto de documentos y una consulta de usuario

Figura 1. Arquitectura de la herramienta.



Fuente: Elaboración Propia

A grandes rasgos, en la figura 1 se distinguen tres procesos del método:

a) Preparación de texto: En esta fase se leen los documentos de texto que constituyen el contexto de análisis. El texto viene escrito de la manera en que lo haya redactado el autor así que es necesario un proceso de limpieza y preparación para la aplicación del método.

b) Aplicación de LSA: El texto se representa en una matriz de manera tal que pueda ser aplicado el análisis de semántica latente.

c) Construcción de resumen: Finalmente, para construirel resumen se elabora una estructura de datos que va a contener las similitudes mutuas de todos los términos del contexto de análisis. El objetivo será aprovechar esas similitudes mutuas(las máximas) para que el método sea capaz de recoger más información.

C. Pasos del método a detalle

En este apartado se explica a detalle cada uno de los pasos que conforman los procesos del método para crear resúmenes con componente semántico.

1) Limpiar texto: En este paso eliminamos signos de puntuación, números, signos de interrogación y exclamación, el objetivo es dejar únicamente palabras que pertenezcan al lenguaje natural.

2) Eliminar stop words: En el lenguaje existe un conjunto de palabras que no aportan significado puesto que se usan como complemento o conectores de otros términos, tal es el caso de los artículos y los adverbios. Algunos ejemplos de *stop Word* son: la, el, los, y, para, que, etc. En este punto se eliminan, de tal manera que nos quedamos con palabras clave que tienen significado en algún contexto temático.

3) Lematizar y aplicar stemming *a los términos:* Existen palabras que aunque se escriben diferente comparten una misma raíz y por tanto significado, por ejemplo, computador y computación. Algunas varían sólo en la terminación en función del género y número, por ejemplo niño, niña, niñez; para estas palabras no es conveniente que se consideren de manera separada, sino más bien como una sola unidad. Si bien en la literatura existen modelos estándar para atender este fenómeno en nuestra versión de laboratorio no atendemos este proceso, esto se debe a que estamos trabajando enuna herramienta con base a tokens formales (en el contexto de expresiones regulares y lenguajes formales) con lo que esperamos cumplir con esta fase.

4) Indexar y preparar la matriz de codependencias M: En este punto tenemos un conjunto de palabras de todos los documentos, el paso que sigue es colocar en cada celda $M_{i,j}$ la frecuencia de aparición del término i en el documento j. Este hecho constituye la indexación de los documentos en la matriz de codependencias de términos y documentos.

5) Factorizar M con SVD: La factorización de la matriz M se puede llevar a cabo con cualquier herramienta que implemente la técnica SVD, por ejemplo Mathematica, o bien alguna de los populares *framewoks* de ciencia de datos que son usuales hoy en día. En nuestra herramienta procedimosde esta manera.

6) Reconstruir M para obtener M': A partir de las matrices obtenidas por la factorización de M se computa la multiplicación entre ellas, pero considerando sólo los primeros k términos para obtener M'. En este sentido, la literatura sugiere que si tenemos más de 300 términos ajustemos a 300. Al reducir la dimensionalidad se elimina ruido de términos

insignificantes, semánticamente hablando, y se revelan relaciones no evidentes en el contexto de términos.

7) Formar matriz de similitudes: En esta parte secomputa la similitud de un término con respecto al resto, y eso se hace para todos. El objetivo es saber para cada término,con cuáles se parece más, esta información es crucial para el sentido de nuestro método.

8) Semantizar Q y g gránulos de cada documento d:
Inicialmente deberemos saber la definición del gránulo que usaremos es decir, un gránulo será un párrafo o una línea, o un bloque. En esta parte enriqueceremos los vectores de Q y g a partir de la información de la matriz de similitudes. A los vectores enriquecidos les llamaremos $V\Delta(Q)$ y $V\Delta(g)$. En la versión de journal se detalla paso a paso la semantización de los vectores.

9) Calcular similitud entre vectores enriquecidos: Conel anterior procedimiento tendremos vectores enriquecidos semánticamente. Así pues tomamos los $V\Delta(Q)$ y $V\Delta(g)$ para todos los gránulos de cada documento y calculamos su similitud.

10) Devolver gránulosmás similares a Q: De cada documento tendremos que hay gránulos que arrojan mayor similitud con Q, pues esos serán los que tomaremos para construir el resumen. Al final tendremos un resumen formado por losgránulos que más se parecieron a Q habiendo considerado un enriquecimiento semántico.

Prototipo

Como parte del presente trabajo se ha desarrollado un prototipo en lenguaje Java que lleva a cabo la creación de resúmenes a partir de un conjunto de documentos. En lasección IV-C se describió cada uno de los pasos implicados en el proceso, en la Figura 1 se muestran esquemáticamente.

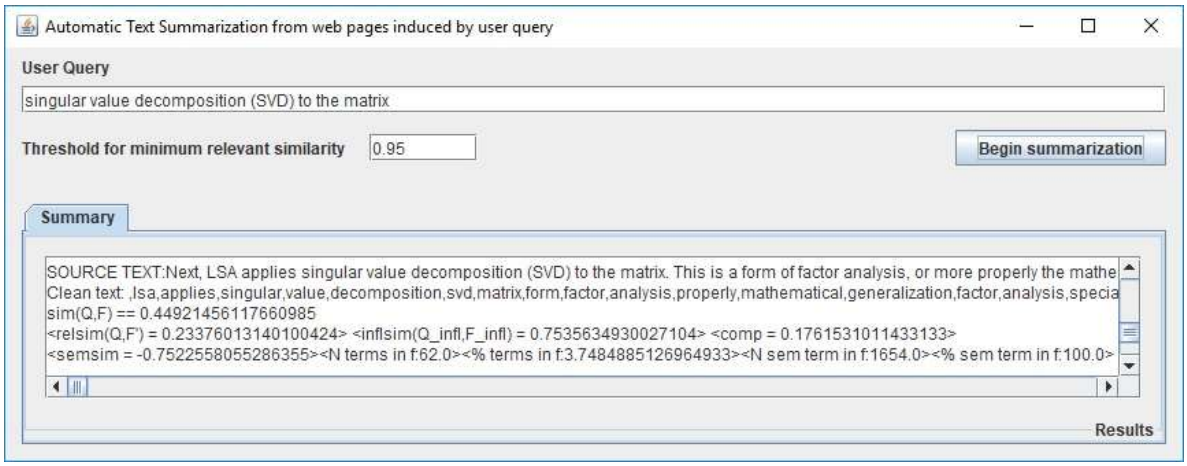
La fase de preparación de texto e indexación de la matriz M se resuelve mediante una librería de desarrollo propio, mientras que la aplicación de LSA, i.e., la descomposición de M mediante SVD y su reconstrucción se ha probado mediante librerías de terceros, no se

ha hecho una evaluación exhaustiva, como parte de este trabajo, para determinar la librería más adecuada para ser incluida en un producto industrial; en todo caso, consideramos el proceso SVD como una herramienta central, sin embargo, el objeto de este documento es ilustrar una técnica nueva construida sobre un cálculo matemático estándar.

Finalmente, la determinación de fragmentos de texto importantes, a partir de su similitud con una consultade usuario se desarrolló como un módulo de código nuevo para el método introducido y se acopló a nuestra librería de desarrollo propio en el área de sistemas del Instituto Tecnológico de La Piedad.

Figura 2, se muestra una salida típica de la herramienta para generar resúmenes, se trata de una pantalla que indica una serie de mediciones sobre el texto, que están presentadas en la versión journal del trabajo.

Figura 2. Resumen generado.



Fuente: Elaboración Propia

Conclusiones

Este trabajo se puede considerar una continuación de (Ramos et al., 2013) en el cual se hacían mediciones de similitud sólo entre palabras literales ignorando las relaciones semánticas, en este sentido se obtiene una mejora. En (Castillo et al., 2012) se hace un filtrado de páginasweb (fragmentos de página web relevantes) en donde se considera la distancia sintáctica entre términos, no se explotan las coocurrencias entre términos. En

(Gong et al., 2011) se aplica LSA para crear resúmenes sin tener en cuenta una pregunta de usuario. Consideramos que la generación de resúmenes de manera automática aprovechando las relaciones ocultas entre las palabras y que son reveladas por LSA constituye un tema de importancia para la generación de herramientas de software capaces de discriminar información. Si bien LSA supone la creación de una matriz que revela relaciones entre términos, por sí solo no es suficiente para crear herramientas que manipulen texto, una de las razones es que la manipulación detexto exige procesar fragmentos de texto de tamaño variable, en ese sentido hemos introducido aquí un mecanismo que puede explotar la semántica latente a partir de gránulos detexto en lugar de documentos. Los resultados revelan que en general se gana en semejanza entre documentos si se explotan las relaciones semánticas. Así pues la posibilidad de manipulación de fragmentos de texto, garantiza no sólo la posibilidad de determinar resúmenes sino también de avanzar en otros procesos de manipulación de texto. Los resultados de la herramienta dan cuenta de la utilidad del método introducido.

Referencias

Salton G., A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," Commun. ACM, vol. 18, no. 11, pp. 613–620,1975.

Manning C., P. Raghavan, and H. Schütze, An Introduction to Information Retrieval. Cambridge University Press, 2008.

Furnas G.W., S. C. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure," SIGIR Forum, vol. 51, no. 2, pp. 90–105, 2017.

Ramos J.G , L. Campoy, S. Ruiz, N. Jasso, and J. C. Solorio, "Preparing text reports from web pages employing similarity tests," in Mexican International

Conference on Computer Science, ENC 2013, October 30 - Nov. 1, 2013, 2013, pp. 13–19.

Castillo C., H. Valero, J.G. Ramos, and J. Silva, "Information Extraction from Webpages Based on DOM Distances," in CICLing (2), 2012, pp.181–193.

Gong Y. and Liu X., "Generic text summarization using relevance measure and latent semantic analysis," in Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development Information Retrieval, SIGIR'01. ACM, 2001, pp. 19–25.