El Párrafo Virtual en la Generación de Extractos

[1]Claudia Bueno-Tecpanecatl, [2]David Pinto, [3]Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
14 sur y Av. San Claudio, Ciudad Universitaria, Edif. 135
Puebla, Pue., México, C.P. 72570
Tel. (+52-222)2295500 Ext. 7212, Fax (+52-222)2295672
[1]claudiabt@mail.cs.buap.mx, {[2]dpinto, [3]hjimenez}@cs.buap.mx

Resumen. En este artículo se presenta un método para la generación automática de extractos usando un concepto novedoso al que hemos denominado *párrafo virtual*; éste está compuesto de términos considerados representativos del contenido semántico del texto del cual se pretende obtener el extracto. Se usa como *corpus* un subconjunto de noticias provenientes del TREC–5 y se evaluan los resultados mediante un sistema de recuperación de información (SRI) basado en el modelo de espacio vectorial (MEV). La hipótesis de verificación consistió en evaluar un par de consultas supervisadas pertenecientes al mismo TREC–5 usando dos *corpora*: documentos completos y solamente extractos. Los resultados obtenidos son realmente alentadores, ya que posteriormente a la evaluación de las consultas, se observa que incluso en algunos casos, la precisión por niveles estándar es mejor cuando el SRI utiliza el conjunto de extractos como *corpus*.
Palabras clave: Extractos, Punto de Transición, Recuperación de Información.

1 Introducción

El crecimiento desmesurado de Internet ha generado grandes volumenes de información, difíciles de manipular aún para sistemas de búsqueda de información tan complejos como Google [10]. De esta manera, cuando se realizan consultas en estos sistemas, surge el problema de obtener información, que en su mayoría carece de significado para el usuario. Este fenómeno ha impulsado a investigar y desarrollar aplicaciones en el campo de tecnologías de recuperación y extracción de información que permitan obtener únicamente la información requerida [5].

El empleo del resumen de un documento es atractivo, pues reduce el espacio de almacenamiento y facilita el acceso a la información relevante. En este trabajo, usamos el concepto de extracto [3] para la elaboración de resúmenes automáticos.

El proceso de generación de extracto de un documento consiste en obtener oraciones pertenecientes al mismo documento consideradas como de mayor relevancia [3]. Para nuestro caso, determinamos la importancia de los párrafos de un documento a través de una función de similitud entre éstos y lo que hemos denominado como párrafo virtual (PV). El PV está compuesto de un conjunto de términos de alta importancia semántica para el texto en cuestión, los cuales se obtienen a través de la técnica del punto de transición [8].

Para la elaboración de este trabajo se ha tomado como referencia una recopilación de trabajos previos de extracción, como es el caso de [4] en donde

se presenta una propuesta para generar el extracto de un texto. Este trabajo principalmente se basa en el uso de una técnica denominada "expansión por *corpus*", que sirve para expandir cada término de un documento en base a su sentido. Las oraciones expandidas son procesadas, después se usa una función de similitud para la obtención de oraciones más significativas que constituyen el extracto del texto. En otro proyecto [7], se utilizan dos técnicas para la obtención del extracto de un texto y posteriormente se realiza una serie de consultas sobre los extractos para evaluar los resultados. La idea es innovadora, sin embargo, la cantidad de documentos utilizados es sumamente pequeña y adicionalmente, la evaluación de los juicios es realizada únicamente por tres jueces. Los resultados presentados en ese trabajo son interesantes, valdría la pena verificar el comportamiento de los algoritmos con un *corpus* mucho más grande. Una aportación más se encuentra en [6], donde se realiza una comparación de dos métodos que determinan automáticamente el extracto de un texto. El objetivo de ese trabajo es hacer patente la importancia del título de un texto. Se obtienen extractos de un texto usando funciones de similitud entre todas las oraciones (método uno) y usando una función de similitud entre el título del texto y las oraciones restantes (método dos). Los resultados muestran que ambos resultados son semejantes, sin embargo, el método dos se encuentra en un orden de complejidad lineal a diferencia del primer método que se encuentra en uno cuadrático.

Como se puede observar, existen diversos trabajos en generación automática de extractos y con resultados interesantes, que alientan a continuar investigando, con la finalidad de mejorar el desempeño.

Este artículo se inspira en los trabajos presentados en [6], con la finalidad de construir un sistema de complejidad lineal (con respecto al número de párrafos) para la obtención del extracto de un texto. El componente innovador del trabajo radica en la generación de un párrafo virtual que representa el significado de dicho documento. La calidad de los extractos obtenidos se evalúa mediante un SRI basado en el MEV. Dado un conjunto de consultas supervisadas, se asume que los extractos son de buena calidad si la precisión obtenida en el SRI es comparable a la precisión en el mismo SRI, usando como *corpus* al conjunto de documentos completos.

En las siguientes secciones se puntualiza dicha investigación. En la sección 2 se explica la descripción del método propuesto, en la sección 3 se presenta la descripción de los datos (*corpus* y consultas), la sección 4 muestra los resultados experimentales y por último las conclusiones y trabajos futuros son abordados.

2 Descripción del Método Propuesto

2.1 La Técnica del Punto de Transición (PT)

El PT surge a partir de las observaciones de George Kinsley Zipf, quién formuló la ley de frecuencias de palabras de un texto (Ley de Zipf), donde establece que las palabras con mayor frecuencia absoluta son las palabras cerradas, mientras que las menos frecuentes son aquellas que reflejan el estilo y riqueza del vocabulario

y por último las que aparecen en la zona media de la función de distribución de frecuencias son las que representan a los documentos [8]. El PT es la frecuencia de un término del texto que divide en dos a los términos de un vocabulario (en términos de alta y baja frecuencia). Esto significa que los términos más cercanos al PT, tanto de alta y baja frecuencia, pueden ser usados como palabras clave que identifiquen a un documento.

La fórmula usada para el Punto de Transición es la siguiente:

$$PT = \frac{\sqrt{1 + 8 * I_1} - 1}{2} \tag{1}$$

Donde I_1 representa el número de palabras que tienen frecuencia 1.

Booth [2] derivó la ley de términos de baja frecuencia de la cual proviene la ecuación 1, sin embargo, se presenta un inconveniente con respecto a los documentos que son demasiado pequeños, ya que para este tipo de textos, el valor obtenido para el PT regularmente se encuentra fuera de las frecuencias obtenidas en su vocabulario. A partir de este trabajo se observa que el PT puede ser obtenido por inspección, eligiendo dentro del vocabulario el primer término con la frecuencia más baja que no se repita. A partir de este punto se toma un porcentaje del 25% de términos de alta y 25% de baja frecuencia para obtener un rango de transición. Los experimentos realizados muestran que al tomar una banda de frecuencias desde un 15% hasta un 25% alrededor de PT, se obtienen los mejores resultados [1].

2.2 Método para la Generación de Extractos

El proceso de generación del extracto de un texto se presenta a continuación:

- Preprocesamiento. Se realiza el preproceso de cada uno de los documentos del *corpus* mediante la eliminación de las palabras cerradas (artículos, preposiciones, etc.), y el particionamiento del documento en párrafos.
- Obtención del vocabulario. Se calcula la frecuencia de ocurrencia de cada término no repetido que posea el documento preprocesado.
- Generación del párrafo virtual. Se aplica la modalidad para documentos pequeños del PT, tomando el término con la frecuencia más baja que no se repite de los términos del vocabulario de cada documento. Así, se obtienen los términos significativos o que se encuentren dentro de una vecindad del 25% alrededor del PT, los cuales constituyen lo que denominamos el párrafo virtual.
- Determinación de párrafos significativos. Se utiliza la función de similitud de Jaccard (ver ecuación 2) usada comunmente en el modelo booleano de representación de información [9], para determinar que tanto se parecen dos elementos. En este caso se compara cada párrafo del documento con el párrafo virtual que se obtuvo del PT. Los extractos de cada documento se conforman mediante los tres párrafos más relevantes.

$$sim(D, q) = \frac{\#(D \cap q)}{\#(D \cup q)} \tag{2}$$

3 Descripción de los Datos

3.1 *Corpus*

El TREC es una colección de documentos y de consultas supervisadas cuyo único propósito es apoyar a la investigación dentro de la comunidad del procesamiento del lenguaje natural, proporcionando la infraestructura necesaria para la evaluación de metodologías de la recuperación de información.

Existen diversos TREC que se han llevado a cabo año con año; de esa variedad se eligió el TREC–5. Este *corpus* consta de una recopilación de noticias (aproximadamente de 250 megabytes de información) que se obtuvieron del periódico mexicano "El Norte" de Guadalajara y 300 megabytes del periódico de "Agence France Presse". Aunque el TREC–5 consiste de 230,820 documentos, en nuestro caso obtuvimos un subconjunto para las pruebas. Nuestro *corpus* consta de 884 noticias, de las cuales el 33% consiste de las noticias relevantes de las consultas 26 y 28 del mismo TREC–5, y el 67% restante consiste de noticias no relevantes para las mismas consultas. Una descripción más detallada de estas consultas se presenta a continuación.

3.2 Consultas utilizadas

Una consulta es una oración en lenguaje natural que sirve para obtener información de interés. Para este artículo se trabaja con las siguientes dos consultas tomadas del TREC–5.

1. *Indicaciones de las relaciones económicas y comerciales de México con los paises europeos.* (Consulta 26)
2. *Indicaciones de las relaciones económicas y comerciales de México con los paises asiáticos, por ejemplo Japón, China y Corea.* (Consulta 28)

4 Experimentos

4.1 Evaluación de resultados

Es realmente complicado realizar una evaluación sobre la calidad de extractos obtenidos de un *corpus*. En [5] y [7], por ejemplo, se utilizan jueces humanos que generan su propio extracto, sin embargo, se observa un alto grado de desacuerdo entre ellos.

En este trabajo se evalua la calidad de los extractos usando un SRI construido sobre dos *corpora*(el primero constituido del *corpus* completo y el segundo de los extractos de este mismo *corpus*).

Este SRI calcula los pesos de representación de los documentos mediante el MEV propuesto por Salton [9] y evalua la relevancia de los mismos por medio de la fórmula de similitud del coseno del ángulo entre los vectores.

Nuestra hipótesis consiste en obtener gráficas de precisión y evocación similares. Si este es el caso, significará que la calidad del extracto es buena.

Tabla 1. Algunos ejemplos de párrafos virtuales

# de doc.	# de noticia	párrafo virtual
1	0000054	exportaciones estados unidos
2	0000182	presidente carta publicación económica económico inversión fiscal estados economistas unidos ritmo
3	0004750	empresarios europeos europea méxico
4	0008846	rendimiento incremento cartera 1992 emisión
.	.	.
.	.	.
.	.	.
884	0202886	china permitirá familiar dijo solidaridad sistema méxico

Después de haber obtenido los párrafos virtuales, se aplica la similitud de Jaccard dentro de cada documento para generar los extractos (mediante los tres párrafos más relevantes); véase el ejemplo siguiente de tres noticias con sus respectivos extractos.

1. La noticia SP94-0000054:
   - extracto 1. Indicó dirigente Camexa que los problemas para exportar Europa son conocimiento parcial los mexicanos las obligaciones para exportar las diferencias idioma transporte.
   - extracto 2. Entrevistado durante presentación Feria internacional maquinaria para envase embalaje empresario sealó que Alemania requiere productos textiles.
   - extracto 3. industrial mexicano debe voltear sus ojos sólo los pases hispanos especialmente Estados Unidos tiene que hacer esfuerzo para posesionarse mercado europeo insistió.
2. La noticia SP94-0000182:
   - extracto 1. creemos firmemente que plan política fiscal que contenga controles más rgidos sobre gasto federal una reducción del índice tributario fiscal IRA retiro individual) universalmente disponible incentivos eficientes efectivos para creación impuestos inversión empresaria influiría mucho para comenzar renovación Estados Unidos informó carta del AIV

- extracto 2. una carta Presidente grupo comercial Wall Street recomendó que política fiscal 1993 enfoque reducción del déficit implantación incentivos inversión ahorro promoción creación impuestos inversión empresaria

- extracto 3.sugieren reducción déficit Asociación Industria Valores recomendó consideración del Presidente Clinton sobre una serie medidas fiscales alentar crecimiento económico asegurar que Estados Unidos fortalezca papel como potencia económica líder

3. La noticia SP94-0004750:
   - extracto 1.Estados Unidos agregó depende poco más sus exportaciones Europa bloque asiático

   - extracto 2.Recordó que bloque estadounidense animal muy distinto bloque europeo donde las relaciones económicas equilibrio poder entre las naciones son más balanceadas las que Estados Unidos tiene que enfrentar Continente Americano

   - extracto 3.Unidos tuvo que formar propio bloque para competir pero podrá funcionar pronto como una sola nación porque tiene todavía que recorrer distancias económicas culturales que separan sus socios comerciales reiteró

4.2 Pruebas realizadas

Despues de evaluar las dos consultas supervisadas sobre el SRI, se observó que para el caso en que se utilizan los documentos completos como *corpus* se obtiene una baja precisión y una alta evocación global; este resultado es una consecuencia del alto número de documentos obtenidos (ver tabla 2), ya que de un total de 884 documentos se obtienen alrededor de 700.

Tabla 2. Evaluación sobre el corpus con documentos completos

	Consulta 26	Consulta 28
Documentos relevantes de acuerdo al TREC	154	155
Documentos relevantes de acuerdo al SRI	712	720
Presición global	0.1994	0.19444
Evocación global	0.92207	0.90322

En la tabla 3 se presenta la evaluación sobre el *corpus* que posee solamente los extractos; en este caso, la precisión global es mayor que en el caso anterior, y la cantidad de documentos obtenidos por el SRI es mucho menor que en el caso anterior, lo cual supone una mejora en el proceso de filtrado sobre el conjunto de documentos relevantes.

Las figuras 1 y 2 muestran el comportamiento del SRI sobre ambos *corpora*, para las consultas 26 y 28 del TREC–5.

Tabla 3. Evaluación sobre el corpus de extractos

	Consulta 26	Consulta 28
Documentos relevantes de acuerdo al TREC	154	155
Documentos relevantes de acuerdo al SRI	272	282
Precisión global	0.224265	0.223404
Evocación global	0.396104	0.406452

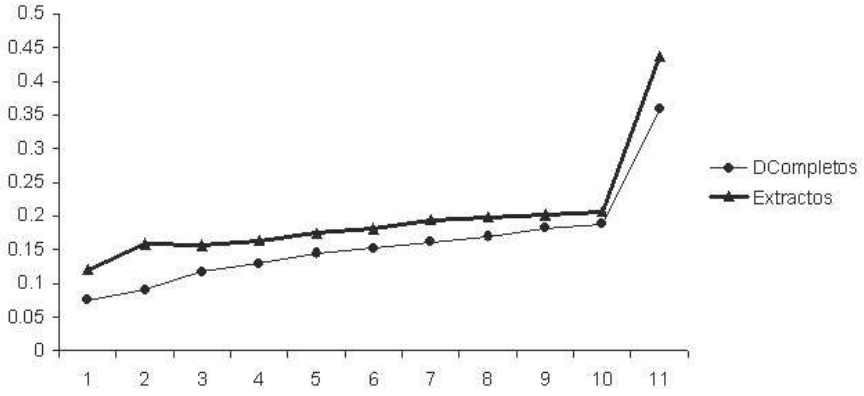

Fig. 1. Gráfica de precisión por niveles estándar para la consulta 26

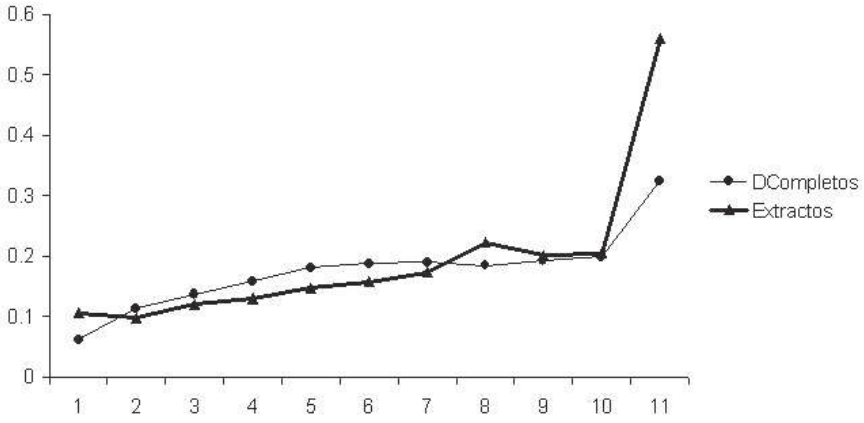

Fig. 2. Gráfica de precisión por niveles estándar para la consulta 28

5 Conclusiones

Se ha presentado un método novedoso para la generación de extractos de documentos. El método hace uso de la técnica del punto de transición para generar una oración virtual que es utilizada para obtener el conjunto de párrafos más representativos de un texto. Los tres párrafos más relevantes se consideran el extracto del texto.

Los resultados obtenidos muestran un comportamiento alentador, ya que a pesar de que se esperaba que el uso de extractos como *corpus* obtuviese una precisión por niveles estándar ligeramente por debajo de la obtenida por los documentos completos, se observó un incremento de la precisión. Esto presume que el método propuesto ha logrado eliminar términos de los textos que efectivamente no tienen un valor significativo de representación para el mismo documento. Se deberán validar estos resultados en todo el TREC–5, a fin de robustecer las conclusiones obtenidas.

A pesar de que el uso del SRI fue con la finalidad de medir la calidad de los extractos, se observó que el uso de extractos como mecanismo de representación de documentos puede ser una buena alternativa en la recuperación de información.

Referencias

1. Berenice, Reyes-Aguirre., Edgar, Moyotl-Hernández & Héctor, Jiménez-Salazar.: Reducción de Términos Indice Usando el Punto de Transición, *Facultad de Ciencias de Computación*, BUAP (2003).
2. Booth, Andrew.: A Law of Ocurrences for Words of Low Frequency, *Information and control* (1967).
3. Climent-Roca, S.: Sistemas de Resumen Automático de Documentos, *revista digital d'Humanitats*,3(1),(2001).
4. Héctor, Jiménez-Salazar., Hilario, Salazar-Martínez. & David, Pinto-Avendaño.: Text Extraction A Corpus-Based Approach, *XXX aniversario del programa educativo de computación*, BUAP, (2003).
5. Héctor, Jiménez-Salazar, David Eduardo, Pinto Avendaño: Recuperación de Información, *Notas de la Academia* (2003).
6. Hilario, Salazar-Martínez., David, Pinto-Avendaño., Héctor, Jiméz-Salazar.: Comparación de dos métodos que determinan automáticamente el extracto de un texto, *Taller de teconologías del Lenguaje Humano*,(ENC 2004), ISBN:970–692–170–2 (2004), 267–273.
7. Jiménez-Salazar, H., Pinto-Avendaõ, D., Salazar-Martínez, H.: Information Retrieval Based on Text Extraction, *1st. Indian International Conference on Artificial Intelligence*, (IICAI'03), ISBN:0–9727412–0–8,(2003).
8. Ruben, Urbizagástegui.: Las posibilidades de la Ley de Zipf en la indización automática, *Reporte de la Universidad de California Riverside*, (1999).
9. Salton, Gerard.: Advanced Information-Retrieval Models, *Automatic Text processing*, (1989).
10. Sergey, Brin., Lawrence, Page.: The anatomy of a large-scale hypertextual web search engine. *Computer Science Departament, Stanford University*, Stanford, CA 94305, USA, (1998).