Análisis de una representación de textos mediante su extracto*

Ernesto Miñon Romero, David Pinto & Héctor Jiménez-Salazar
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
ernesto_minon@yahoo.com.mx dpinto@cs.buap.mx hjimenez@aleteya.cs.buap.mx

Resumen

En este trabajo usamos un método para obtener el extracto de un texto, que no utiliza recurso lingüístico alguno. El método se basa en la llamada "oración virtual", un conjunto de términos presumiblemente de alto contenido semántico. Para obtener la oración virtual es necesario especificar un umbral que determine la extensión de dicha oración. Los experimentos llevados a cabo contemplan diferentes umbrales y varias subcolecciones de TREC-5, tomando como marco de evaluación la Recuperación de Información. Los resultados obtenidos muestran que existe una correspondencia inversamente proporcional entre los tamaños de las consultas y los umbrales establecidos.

1. Introducción

En la actualidad, con el uso de internet, la información es cada vez más abundante. Esto es provocado por el crecimiento exponencial de los sitios web y el número de usuarios potenciales que contribuyen con información a la web. Es por ello muy importante la creación de herramientas que permitan obtener información relevante y útil para el usuario de forma eficaz y rápida. La generación de extractos es una opción que permitiría, por ejemplo, aligerar la representación de los textos mediante su extracto, siempre que la aplicación mantenga su efectividad.

La generación del extracto de un texto tiene como finalidad obtener las oraciones más importantes del texto original. Hay muchos trabajos que proponen métodos para extraer las oraciones más importantes. Una propuesta muy sencilla y además efectiva es la que hizo D. Marcu [1], la cual se basa en determinar la oración de un texto que más aporta a la similitud entre un conjunto de oraciones y el texto original. Inspirado en el trabajo anterior está [2], donde se usa un método basado en la expansión de las oraciones (reemplazando cada palabra por sus contextos tomados de un *corpus*) para calcular la similitud entre una oración expandida y su complemento en el texto; *i.e.* el texto sin la oración. En [12], se propone un método basado en una gráfica dirigida con el fin de asignar una puntuación a las oraciones que establecen relación con las demás del texto original. Ha habido, además, algunas aplicaciones del extracto en la Recuperación de Información (RI) [3].

En este trabajo usamos un método, muy eficiente en tiempo, que se basa en la llamada "oración virtual" [13]. De manera sucinta, el método realiza el cálculo de la similitud entre cada una de las oraciones del texto con la oración virtual, y aquellas que tienen mayor puntuación son elegidas como extracto del texto. Para obtener la oración virtual es necesario especificar un umbral que determine la extensión de dicha oración. Aquí presentamos el resultado de experimentos llevados a cabo con diferentes umbrales, y en varias colecciones de textos, tomando como marco de evaluación la RI.

Así, el extracto puede reemplazar al documento en un sistema de RI y entonces, a través de la medición del desempeño, evaluar la pertinencia del extracto, o bien comparar el grado de representatividad de diferentes extractos. Utilizamos el modelo booleano de RI como una primera prueba para conocer el desempeño del extractor de textos.

En las secciones que restan, se presenta el método seguido para generar un extracto a partir de un texto, los resultados experimentales obtenidos en varias subcolecciones del TREC5, y las conclusiones.

2. Método

Como se ha dicho, la metodología de evaluación se basa en el desempeño de un sistema de RI usando los extractos en lugar de los textos. Por ello empezaremos por precisar el modelo empleado para la RI. Inmediatamente después especificamos cómo se obtiene la oración virtual, y, por último, la puntuación de las oraciones que permite hacer el extracto de un texto.

*Este trabajo fue parcialmente apoyado por BUAP-VIEP 3/G/ING/05.

2.1. Modelo Booleano

El modelo booleano para RI representa y recupera documentos que empatan con los términos de una consulta. Más específicamente, sea $C = \{T_1, ..., T_k\}$ una colección de textos y $C' = \{T'_1, \ldots, T'_k\}$ la colección preprocesada (eliminación de palabras cerradas; conjunciones pronombres, artículos, etc.). Sea $V = \cup_i T'_i$ el vocabulario de la colección, y $V_0 = [v_i]_{i \leq n}$, el vocabulario ordenado lexicográficamente. La representación de un texto T' es el vector $\overrightarrow{T'}_j = [t_i]_{i \leq n}$, donde

$$t_i = \begin{cases} 1 & \text{si } t_i \in T'_j, \\ 0 & \text{si } t_i \notin T'_j. \end{cases} \quad . \tag{1}$$

Los documentos recuperados bajo el modelo booleano son T_j tales que $r_j = \overrightarrow{T_j} \cdot \overrightarrow{q} \neq 0$. El valor r_j ayuda a establecer el *ranking* de los documentos encontrados. Usaremos, en lugar del producto interno entre vectores la función de similitud de Jaccard cuyos valores están normalizados en $[0, 1]$.

2.2. Oración virtual

Con base en que el punto de transición es útil para la identificación de palabras clave de un texto [5], se han realizado pruebas con el fin de representar en forma reducida un texto; por ejemplo la selección de términos. Nuestro caso es muy semejante, ya que requerimos esencialmente elegir las oraciones "más representativas" de un texto.

El punto de transición (PT) es una frecuencia que divide en dos al vocabulario de un texto: los de alta y baja frecuencia. El PT empezó a usarse en la indización de textos [5], y posteriormente en otras aplicaciones como Recuperación de Información [13] [14], y en Categorización de Textos [7] [8] [9]. Esta diversidad de aplicaciones proviene del hecho de que el PT da una pauta para extraer términos con alto contenido semántico [11].

Los términos determinados por este método son aquellos que tienen una frecuencia media; esto es, términos alrededor del PT. Una fórmula para obtener el PT de un texto T es: $PT_T = (\sqrt{1 + 8 \cdot I_1} - 1)/2$, siendo I_1 es el número de términos con frecuencia 1 en T.

Llamamos oración virtual a un conjunto de términos alrededor del PT. Es necesario, entonces, definir un umbral que permita determinar dicho conjunto. Para el texto T, y la frecuencia $fr(\cdot)$ de cada palabra el vocabulario de T es $V_T = \{(x, y) | x \in T, y = fr(x)\}$. Definimos la oración virtual de T con umbral p como:

$$Ov_T = \{x | (x, y) \in V_T, PT_T \cdot (1-p) \leq y \leq PT_T \cdot (1+p)\} \tag{2}$$

Es decir, Ov_T es una vecindad de términos cuya frecuencia está alrededor del PT.

2.3. Obtención del extracto

Para obtener el extracto de cada uno de los textos, T, que forman la colección de prueba, se aplicó el siguiente procedimiento:

Preproceso

Dividir T en oraciones: $T = [O_i]_i$.

Eliminar las palabras cerradas (*stopwords*) de T.

Representación

Calcular el punto de transición: PT_T

Calcular la oración virtual: Ov_i

Obtener la similitud entre la oración virtual y cada una de las oraciones del texto, O_i. Para este experimento utilizamos el índice de Jaccard:

$$sim_i = \frac{\#(Ov_T \cap O_i)}{\#(Ov_T \cup O_i)}$$

Elección de oraciones

Ordenar descendentemente las oraciones $O_i \in T$ según sim_i, y tomar las tres primeras oraciones.

En el apéndice se muestra un ejemplo de la aplicación del algoritmo.

3. Experimento

El experimento consistió en evaluar una serie de consultas supervisadas de la colección TREC-5. Se calculó el desempeño, en cada una de 3 subcolecciones de TREC-5, sobre los documentos representados por su extracto, y sin él. Enseguida explicaremos las características de cada colección, detalles sobre el procedimiento seguido en cada prueba, y los resultados.

3.1. Composición de las colecciones

TREC-5 es una colección de textos periodísticos controlados, para la que se tiene 50 consultas con sus textos relevantes. Esto nos permite evaluar la efectividad de un SRI.

Debido a que deseamos conocer el desempeño de nuestro sistema de RI booleano, hemos tomado en una etapa inicial, varias subcolecciones de TREC-5 correspondientes a las consultas. Esto es, seleccionamos 2 consultas al azar y formamos una colección con los textos relevantes a ambas consultas y, además agregamos el doble de noticias no relevantes para ambas consultas. La tabla 1 muestra, por columna: el identificador de la colección, la consulta, y la colección dividida en el número de textos relevantes y el número de textos no relevantes para dicha consulta.

Id	Consulta	Textos Relev.	Textos no Relev.
5	Maquiladoras en la economía mexicana	257	709
10	México es importante país de tránsito en la guerra anti-narcótica	206	622
11	Derechos a las aguas de los ríos en la región fronteriza entre Mexico y los Estados Unidos	105	622
24	Prevención de SIDA en México	131	981
25	Programa de Privatización de Empresas Públicas Mexicanas	359	981

Tabla 1. Descripción de las subcolecciones de prueba.

3.2. Pruebas

Cada prueba consideró la obtención de la oración virtual por documento. El experimento se llevó a cabo utilizando cuatro diferentes umbrales: 0.1, 0.2, 0.3 y 0.4. Esto permitió conocer la efectividad de varias bandas de frecuencias alrededor del punto de transición. Obtenidas las oraciones virtuales correspondientes a cada documento de la colección, se calculó el índice de Jaccard entre las consultas de cada colección y la oración virtual de cada texto. Esta prueba se aplicó a los 3 *corpora* generados, utilizando el modelo booleano clásico.

3.3. Evaluación de resultados

Después de realizar las pruebas anteriormente descritas, se procedió a evaluar los resultados obtenidos. Para ello usamos medidas estándares; *i.e.* precisión, P, evocación, R y F_1 [6]:

$$P = \frac{\#\text{documentos relevantes que obtuvo el sistema}}{\#\text{documentos que obtuvo el sistema}}, \quad (3)$$

$$R = \frac{\#\text{documentos relevantes que obtuvo el sistema}}{\#\text{documentos relevantes}}, \quad (4)$$

$$F_1 = (2 \cdot P \cdot R)/(P + R). \quad (5)$$

Con base en las anteriores fórmulas se calcula la precisión por niveles de evocación [10]. Los resultados, para cada subcolección (consulta) y variando el umbral p de la oración virtual se muestran en las figuras 1, 2, 3, 4 y 5. Notemos que para las consultas 24 y 5, su precisión es mayor que la de las demás. Justamente, estas consultas son cortas, y la representación de los textos por vectores más dispersos (debidos al extracto) tienen dificultad en incluir índices diversos que se presentan con mayor regularidad en consultas largas.

La tabla 2 exhibe la cantidad de términos del vocabulario de las colecciones para diferentes umbrales y el tamaño del vocabulario original. La reducción mínima en nuestras pruebas (que correspondería al umbral 0.4) es de 28% del total de términos.

La figura 6 muestra, para cada umbral, los valores de precisión promediados sobre las colecciones. Dicha gráfica resume el comportamiento del extracto con base en el umbral. Por último, podemos ver globalmente la influencia de los umbrales elegidos en la tabla 3. Esta tabla resume aún más la gráfica previamente aludida, y muestra que para consultas largas tiende a ser mejor el umbral $p = 0.1$.

Id	#Tér. Extracto (umbral 0.4)	#Tot Térm. (colección)	% Reduc.
5	13,787	47,536	29
10 y 11	12,182	41,940	29
24 y 25	14,657	54,247	27

Tabla 2. Número de términos en las representaciones.

Consulta	umbral				
	0.05	0.1	0.2	0.3	0.4
5	0.184	0.185	0.209	0.229	**0.242**
10	**0.329**	0.323	0.300	0.292	0.286
11	0.144	**0.193**	0.189	0.160	0.166
24	0.154	0.156	0.163	**0.172**	0.167
25	0.192	**0.193**	0.170	0.185	0.197

Tabla 3. Valores promedio F_1 para diferentes umbrales.

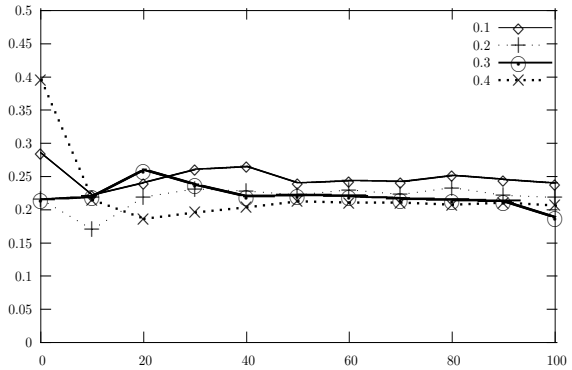

Figura 1. Precisión por niveles de evocación para la consulta 10.

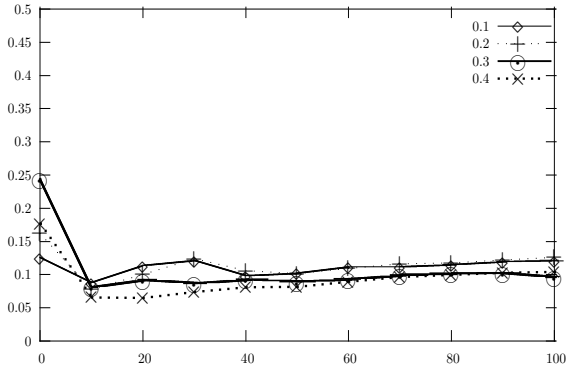

Figura 2. Precisión por niveles de evocación para la consulta 11.

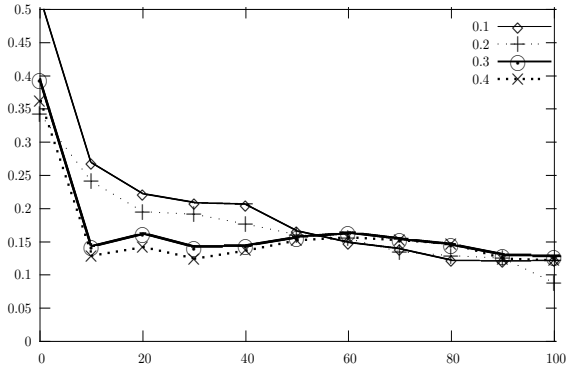

Figura 3. Precisión por niveles de evocación para la consulta 24.

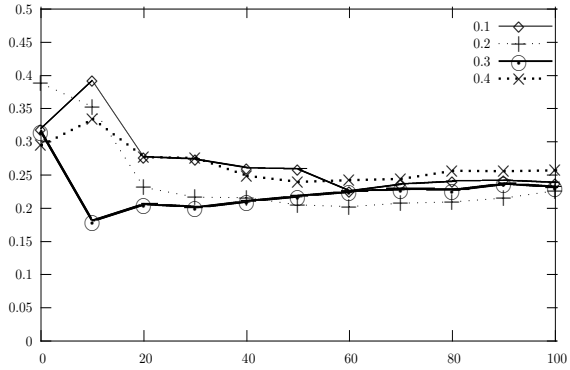

Figura 4. Precisión por niveles de evocación para la consulta 25.

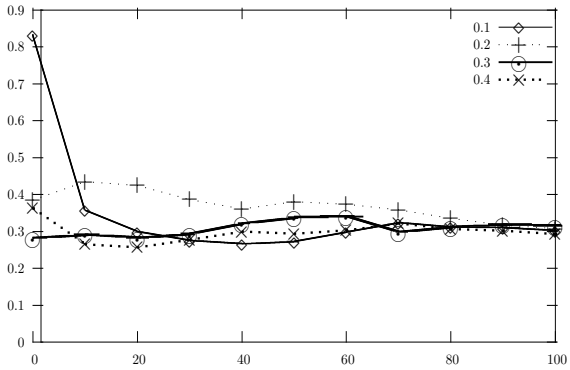

Figura 5. Precisión por niveles de evocación para la consulta 5.

4. Conclusiones

Hemos presentado los resultados de probar diferentes umbrales para ser utilizados en la obtención del extracto de un texto. La efectividad del extracto fue medida dentro de un sistema de RI, utilizando subcolecciones de TREC-5. Aunque el desempeño del reemplazo del documento por su extracto en un sistema de RI no es comparable con el documento completo, sí podemos concluir lo siguiente:

- La reducción de la dimensionalidad del extracto obtenido con el método propuesto es significativamente alta.
- Se dilucida una correspondencia entre umbrales grandes y consultas cortas, y viceversa; *i.e.* a mayor tamaño de consulta se prefiere menor umbral.

No es posible, para los extractos, identificar una regularidad con respecto al tamaño de las consultas independientemente del umbral. La dependencia muy probablemente se deba a que la raleza de los vectores booleanos es mayor cuando el umbral es bajo, y por tanto, las consultas largas tendrán más probabilidad de empatar con los índices elegidos.

Lo anterior se enmarca en la selección de términos para diferentes propósitos que toman como base la representación de textos. Asimismo, es importante considerar que resultados más definitivos tendrán que apoyarse en colecciones de texto variadas y un análisis más concienzudo; por ejemplo, determinar cuáles términos alrededor del PT vician la representación.

Referencias

[1] Daniel Marcu: The Automatic Construction of Large-scale Corpora for Summarization Research, *ACM-SINGIR*, pp. 137-144, 1999.

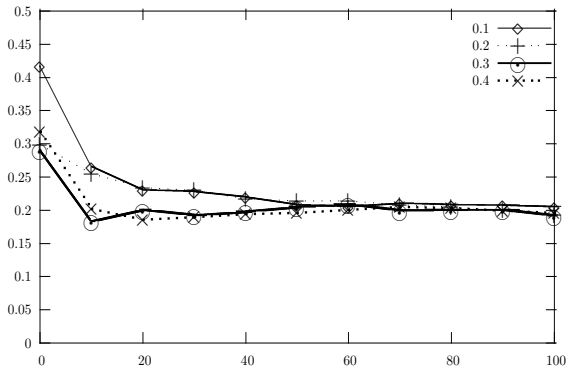

Figura 6. Precisión promedio por niveles de evocación de las 5 consultas.

[2] Salazar, H.; Jiménez, H.& Pinto, D.: Text Extraction: a Corpus-based Approach, en *Memorias del 30 Aniversario del PE en Computación de la BUAP*, ISBN-9688637114, pp 92-94, 2003.

[3] Jiménez, H,; Pinto, D. & Salazar, H.: Information Retrieval based on Text Extraction, en *Proc. of 1st Indian Int. Conf. on AI*, 2003.

[4] Salton, G., Wong, A. & Yang, C.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(11) pp 613-620, 1975.

[5] Urbizagástegui, A.R.: Las Posibilidades de la Ley de Zipf en la Indización Automática, `http://www.geocities.com/ResearchTriangle/2851/RUBEN2.htm`, 1999.

[6] van Rijsbergen, C.J.: *Information Retrieval*. London, Butterworths, 1999.

[7] Moyotl, E. & Jiménez, H.: An Analysis on Frecuency of Terms for Text Categorization, *Proc. of SEPLN-04, XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, pp 141-146, 2004.

[8] Moyotl, E. & Jiménez, H.: Experiments in Text Categorization Using Term Selection by Distance to Transition Point, *Proc. of CIC-04, XIII Congreso Internacional de Computación*, pp 139-145, 2004.

[9] Moyotl, E. & jiménez, H.: Enhancement of DPT Feature Selection Method for Text Categorization, *Proc. of CICLing-2005, Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pp 706-709, 2005.

[10] Baeza-Yates, R.: *Modern Information Retrieval*, Addison Wesley, 1999.

[11] Luhn, H.P.: The Automatic Creation of Literature Abstracts, *1RE National Convention*, IBM Journal, 1958.

[12] Mihalcea, Rada: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, p. 170-173, 2004.

[13] Claudia Bueno-Tecpanécatl, David Pinto & Héctor Jiménez-Salazar: El párrafo virtual en la generación de extractos, en *Advances in Computer Science in México*, A. Gelbukh & H. Calvo (Eds.), p. 83-90, 2005

[14] Rubí J. Cabrera, David Pinto, Darnes Vilariño & Héctor Jiménez-Salazar: Una nueva ponderación para el modelo de espacio vectorial para recuperación de información, en *Advances in Computer Science in México*, A. Gelbukh & H. Calvo (Eds.), p. 75-82, 2005

Apéndice

A continuación se muestra el fragmento de uno de los textos (subcolección 10), la oración virtual obtenida, y las oraciones de este texto ordenadas según su similitud con la oración virtual.

No.	Oración
0	la paz, abril 13.
1	las autoridades locales consignaron a la justicia federal a ram pereda gastelum como presunto responsable de la posesi de 4 toneladas 694 kilogramos de cocaína, informla procuraduría general de la república.
2	el subdelegado de la pgr, ismael gonzález contreras, dijo que se intensificará la vigilancia por el decomiso del mayor cargamento de droga asegurado en este a, el pasado fin de semana en esta ciudad.
3	expuso que pereda gastelum fue consignado al juzgado de distrito por la posesi del estupefaciente y una arma prohibida, un rifle ak-47 ”cuerno de chivo”.
4	expuso que se trata del decomiso más grande en lo que va del a en el país, en un hecho sorpresivo por haberse encontrado en esta ciudad, donde los cargamentos que se aseguran son menores y por la vía del transbordador.
5	el procurador estatal de justicia, genaro canett yee, dijo que el hallazgo del cargamento de droga en la zona industrial resulta sorpresivo, porque no se conoce a la entidad como el paso de grandes cantidades de droga hacia estados unidos.
6	pero, admiti que la vigilancia tendrá que reforzarse en el marco de los convenios de cooperaci con las autoridades federales del ramo, toda vez que es el segundo cargamento importante en un a, localizado en la zona urbana de la ciudad.
7	los judiciales federales encontraron la droga en una de las naves del parque industrial de la paz, dividida en 2 mil paquetes.
8	se estima que el valor del cargamento en el mercado del narcotráfico podría alcanzar 450 mil millones de dares.
9	decomisan cocaina la procuraduría general de la república informde un decomiso de 425 kilogramos de cocaína en un operativo conjunto con las autoridades en la zona fronteriza con guatemala, donde dos mexicanos fueron aprehendidos por el ejcito de ese país.

Para este texto la oración virtual resultante fue: OV =`[autoridades kilogramos cargamento país guatemala avioneta]`, y las 5 oraciones con mayor similitud a OV:

sim	No.	Oración
0.1818	12	la dependencia dijo que el sistema hemisférico de información proporcionó datos a los gobiernos de méxico y guatemala de que una avioneta proveniente de costa rica, vía océano pacífico, con dirección al norte ingresaría al país.
0.1818	9	decomisan cocaina la procuraduría general de la república informó de un decomiso de 425 kilogramos de cocaína en un operativo conjunto con las autoridades en la zona fronteriza con guatemala, donde dos mexicanos fueron aprehendidos por el ejército de ese país.
0.1578	23	en diferentes acciones realizadas en ocho estados del país, la pgr incautó dos toneladas 64 kilogramos de mariguana y capturó a 24 presuntos narcotraficantes
0.1304	1	las autoridades locales consignaron a la justicia federal a ramón pereda gastelum como presunto responsable de la posesión de 4 toneladas 694 kilogramos de cocaína, informó la procuraduría general de la república.
0.1250	16	en esa área, según reportes de la pgr, también el pasado 16 de febrero un avión cessna arrojó 117 kilogramos de cocaína sobre los márgenes del usumacinta, como en otras tres acciones cuantiosas detectadas en los municipios de balancan y huimanguillo.