Método para autocompletar consultas basado en cadenas de Markov y la ley de Zipf

Edgar Moyotl-Hernández, Mónica Macías-Pérez

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias Físico Matemáticas, Puebla,
México

{emoyotl, monica}@fcfm.buap.mx

Resumen. En este trabajo se presenta un algoritmo para autocompletar consultas, el cual genera semiautomáticamente términos que el usuario podría emplear para plantear adecuadamente una consulta y aumentar la efectividad de un Sistema de Recuperación de Información. Con el fin de determinar dichas palabras, se utilizan cadenas de Markov, n-gramas y el punto de transición de Goffman. Este método se aplicó a un corpus general construido con textos de Wikipedia y los resultados obtenidos en los experimentos sugirieron la inclusión de palabras importantes en la formulación de consultas, palabras consideradas relevantes de acuerdo con el modelo de espacio vectorial.

Palabras clave: Expansión de consultas, recuperación de información, cadenas de Markov, n-gramas, ley de Zipf, punto de transición de Goffman.

Method for Autocomplete Queries Based on Markov Chains and Zipf's Law

Abstract. This paper presents an algorithm to autocomplete queries, which semiautomatically generates terms that the user could utilize to properly write a query and increase thereby the effectiveness in an Information Retrieval System. In order to determine these words, Markov chains, n-grams and the Goffman's transition point are used. This method was applied to a general corpus elaborated with texts of Wikipedia and the results obtained in the experiments suggested the inclusion of important words in the query formulation, words considered relevant according to the vector space model.

Keywords: Query expansion, information retrieval, Markov chains, n-grams, Zipf's law, Goffman's transition point.

1. Introducción

La *recuperación de información (RI)* es el proceso por el cual se obtiene un conjunto de documentos cuyo contenido satisface la necesidad de información de

un usuario. Es decir, la recuperación de información intenta resolver el problema de encontrar y ordenar documentos relevantes que satisfagan la necesidad de información de un usuario [1,2]. El ejemplo más popular de esta tarea es el de los sistemas buscadores en Internet tales como Google,[1] Bing[2] o Yahoo[3].

Como se puede notar, el primer paso para llevar a cabo una recuperación consiste en que el usuario exprese su necesidad informativa como una *consulta*. Por lo tanto, el sistema parte de la consulta formulada por el usuario, no de la necesidad de información original, así que una formulación incorrecta o insuficiente (con palabras mal seleccionadas, mal escritas, con faltas de ortografía, etc.) no guiará adecuadamente al sistema durante el proceso de recuperación.

A este respecto, los mayores problemas que enfrenta el *sistema de recuperación de información (SRI)* son, por una parte, la dificultad del usuario para expresar claramente su necesidad en forma de consulta y, por otra, que cuando se describe un mismo concepto, las palabras (o términos) empleadas por el usuario y los autores de los documentos suelen diferir, impidiendo el establecimiento de correspondencias entre las consultas y los documentos [3]. Si bien los usuarios no tienen por qué conocer técnicas de RI, los resultados de su búsqueda mejorarían si se implementan técnicas de *expansión de consultas* (*QE*, por sus siglas en inglés, *Query Expansion*) para lograr que en la respuesta los documentos recuperados sean los documentos relevantes [1].

En este trabajo se presenta un método semiautomático para autocompletar consultas, el cual sugiere al usuario las palabras que podría emplear para especificar adecuadamente una consulta y mejorar su búsqueda, en un SRI que utilice el modelo espacio vectorial para la representación de documentos [4]. Para ello, se usa un procedimiento basado en la frecuencia con que aparecen juntas secuencias de n palabras contiguas, los *n-gramas*. A partir de estas propiedades estadísticas extraídas de un corpus, se construye una *cadena de Markov* para predecir términos de búsqueda que coincidan con el contenido de algún documento, con el fin de aumentar la probabilidad de que se obtengan resultados relevantes. Al mismo tiempo, se explora el uso del *punto de transición de Goffman*, derivado de la *ley de Zipf*, para identificar las palabras con un alto valor semántico en el contenido temático de un texto y la posibilidad de sugerir la inclusión de estos términos en las consultas.

Por otra parte, y a pesar de que existen ya varios trabajos relacionados con la temática, la gran mayoría de la investigación llevada a cabo se centra casi exclusivamente en textos escritos en inglés. Es por ello que los experimentos se realizaron sobre un corpus, no específico con respecto al género, de artículos de Wikipedia escritos en español.

El presente trabajo está organizado de la siguiente manera. En la sección 2, se introducen las características del modelado del lenguaje con cadenas de Markov, se describe el concepto de n-grama y se muestra su uso en la representación de frases. Posteriormente, en la sección 3 se revisan las aplicaciones del punto

[1] https://www.google.com
[2] https://www.bing.com
[3] https://www.yahoo.com

de transición de Goffman en la representación de textos. Luego, en la sección 4 se presenta un resumen de trabajos previos relacionados con el proceso de expansión de consultas y con los algoritmos propuestos, además estos últimos se describen a detalle. A continuación, en la sección 5 se describe el corpus, el proceso de preparación de datos y los experimentos, también se analizan los resultados. Finalmente, en la sección 6 se presentan las conclusiones y el trabajo futuro.

2. Modelado del lenguaje

Cualquier sistema informático que intente tratar el lenguaje natural necesita un modelo que le permita caracterizar y representar la lengua que trata. En los modelos de lenguaje probabilistas, el lenguaje es considerado como una fuente que genera una secuencia de palabras a partir de un conjunto finito de elementos, $V = \{w_i\}$, el vocabulario; y el objetivo es determinar la probabilidad de una secuencia de palabras específica [5], [6]. Esto es útil en diferentes aplicaciones, por ejemplo, corrección ortográfica y gramatical, traducción automática o asistida, reconocimiento de voz y de escritura, predicción de palabras en curso de captura, entre otras.

Como se mencionó anteriormente, el modelo probabilista se encarga de estimar la probabilidad de una frase, para ello, una frase s de longitud L se representa por una secuencia de palabras w_i, es decir, $s = w_1, \ldots, w_L$ o bien $s = w_{1,L}$. Así que, si se interpreta una frase como una sucesión de eventos dependientes, entonces, la probabilidad de cada palabra w_i depende de su historia o contexto $w_1, \ldots, w_{i-1}$, esto es:

$$\begin{aligned} P(s) &= P(w_{1,L}) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_L|w_{1,L-1}) \\ &= \prod_{i=1}^{L} P(w_i|w_{1,i-1}). \end{aligned} \tag{1}$$

Considere, por ejemplo, la frase $s =$ *"a b c d e"*. En este caso la probabilidad de s es:

$$\begin{aligned} P(s) &= P(a\ b\ c\ d\ e) \\ &= P(a)P(b|a)P(c|a, b)P(d|a, b, c)P(e|a, b, c, d). \end{aligned}$$

Nótese que el modelo puede predecir la siguiente palabra de una oración a partir de las palabras anteriores. La problemática consiste ahora en estimar la probabilidad $P(w_i|w_{1,i-1})$ para toda palabra del vocabulario y para todo contexto posible, es decir, tomada entre todas las secuencias posibles de palabras de V.

2.1. Cadenas de Markov

Se sabe que al generar una frase puede utilizarse cualquier palabra del conjunto previamente especificado, V, el léxico o vocabulario. Suponga que la frase evoluciona o cambia al agregar palabras a lo largo del tiempo, y sea w_t la última palabra de la frase en el tiempo t. Si se considera que las palabras futuras no están determinadas por las previas (la frase evoluciona de forma no determinista), entonces puede considerarse que w_t es una variable aleatoria para cada valor de t. Esta colección de variables aleatorias es la definición de *proceso estocástico*, y sirve como modelo para representar la evolución aleatoria de una frase a lo largo del tiempo.

Entonces, una determinada frase puede ser interpretada como la realización de un proceso estocástico de tiempo discreto. Si se considera que la próxima palabra (la evolución de este proceso) depende solamente de lo que ya fue escrito (su estado actual), se trata de un proceso markoviano. Por lo tanto, se puede reducir el contexto de la palabra w_i a la palabra más próxima y puede aproximarse de la siguiente forma:

$$P(w_i|w_{1,i-1}) \approx P(w_i|w_{i-1}). \tag{2}$$

Estos tipos de procesos son modelos en donde, suponiendo conocido el estado presente del sistema, los estados anteriores no tienen influencia en los estados futuros del sistema. Esta condición se llama *propiedad de Markov*.

En consecuencia, para aproximar la probabilidad de una oración, simplemente se necesita calcular el producto de las probabilidades de cambiar de un estado (palabra actual) a otro (siguiente palabra), lo que se conoce como *probabilidades de transición*. De modo que, la fórmula para el cálculo de $P(s)$ se obtiene sustituyendo la ecuación 2 en 1 y resulta lo siguiente:

$$P(s) = P(w_{1,L}) \approx \prod_{i=1}^{L} P(w_i|w_{i-1}). \tag{3}$$

Ahora, la probabilidad de la frase $s =$ *"a b c d e"* es:

$$\begin{aligned} P(s) &= P(a\ b\ c\ d\ e) \\ &= P(a)P(b|a)P(c|b)P(d|c)P(e|d)P(e). \end{aligned}$$

Como el conjunto de palabras distintas (el vocabulario) es finito y dado que en una oración la probabilidad de que ocurra una palabra depende de la palabra inmediata anterior, entonces, se puede formar una *cadena de Markov* donde las palabras representen los estados del proceso y la generación de una determinada frase represente la realización del proceso. Lo descrito anteriormente puede representarse gráficamente usando un diagrama como el de la Figura 1.

Los círculos se denominan nodos y representan los estados del proceso, las flechas son los arcos y representan las probabilidades de transición. Note que la suma de los elementos de cada orden es uno. En estas condiciones, es posible,

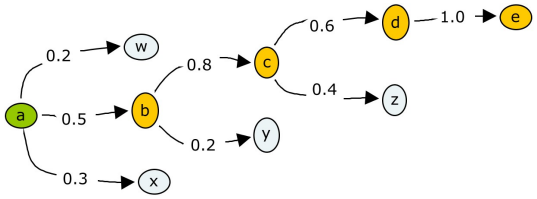

Fig. 1. Cadena de Markov con probabilidades entre palabras

conociendo las probabilidades de transiciones entre estados, calcular la probabilidad de toda cadena particular, en este caso, de una frase específica.

Por consiguiente, si se conocen las probabilidades de las transiciones entre estados, el modelo descrito puede funcionar como generador de consultas. Así por ejemplo, de la cadena representada por la Figura 1, se tiene que dada una primera palabra *"a"* se pueden presentar sugerencias para agregar *"w, b, x"* cada una con una probabilidad distinta y en el siguiente estado *"b"* (en caso de haber sido elegido) nuevamente se podrían presentar más sugerencias de búsqueda *"c, y"*, y así sucesivamente.

2.2. Modelo n-grama

El modelo n-grama es el más empleado y el que mejores resultados ha obtenido dentro del campo de la lingüística computacional. Los *n-gramas* de palabras son secuencias de n palabras consecutivas donde $2 \leq n \leq 7$ para la mayoría de las aplicaciones [5]. Así que, existen 2-gramas (bigramas), 3-gramas (trigramas), 4-gramas, 5-gramas, etc.

Aquí se presenta un ejemplo de n-gramas. De la frase: $s =$ *"a b c d e"* se obtienen los siguientes 2-gramas o bigramas: *a b, b c, c d, d e*. Y los siguientes 3-gramas o trigramas: *a b c, b c d, c d e*. Como se puede observar, el procedimiento es muy sencillo. De manera que, para determinar la probabilidad condicional de que ocurra la palabra w_i después de w_{i-1} se aproxima la probabilidad de un bigrama en particular, esto es:

$$P(w_i|w_{i-1}) \approx \frac{C(w_{i-1}w_i)}{C(w_{i-1})}, \tag{4}$$

en donde $C(w_{i-1}w_i)$ es el número de veces que ocurre el bigrama $w_{i-1}w_i$ y $C(w_{i-1})$ es el número de ocurrencias de la primera palabra del bigrama, w_{i-1}, aproximación presentada en [5], [6]. En suma, el modelo se calcula a partir de la frecuencia de todos los bigramas y de todos los unigramas (es decir, de n-gramas construidos de una sola palabra) en el corpus.

En general, con un modelo n-grama se aproxima la probabilidad de una palabra w_i dadas todas las palabras previas, por la probabilidad de w_i considerando sólo las $n-1$ palabras previas.

3. Ley de Zipf

En general, la mayoría de las palabras frecuentes son también las más cortas y más fáciles de recordar. Por lo que, es evidente que las *palabras funcionales* (también llamadas *palabras vacías* o *stop words*), tales como artículos, pronombres, preposiciones y conjunciones son las más frecuentes en el texto, mientras que las menos frecuentes son palabras que reflejan el estilo y riqueza del vocabulario del autor. Por lo tanto, las palabras que aparecen en la zona media de transición entre las de alta y baja frecuencia de ocurrencia son las que representan al documento [8].

El punto referente a la frecuencia, en torno al cual se encuentran estos términos significativos se llama *punto de transición de Goffman*, puesto que Goffman fue quien introdujo la idea de que las palabras más significativas de un texto se agruparán en una zona donde se encuentran las palabras de alta frecuencia con las de baja frecuencia, es decir, un punto intermedio de transición [9].

3.1. Punto de transición de Goffman

La ley de Zipf y en especial el punto de transición de Goffman han dado buenos resultados en la identificación de palabras clave para la indización y la construcción de tesauros, entre otras aplicaciones [9], [10]. Por ende, en este trabajo se utiliza el punto de transición de Goffman para identificar las palabras clave que representan los textos y se explora la posibilidad de incluir estos términos en las consultas para aumentar la efectividad del proceso de recuperación.

La fórmula para el cálculo del *punto de transición (pt)* es la siguiente:

$$pt = \sqrt{W}, \tag{5}$$

donde W es el número de total de palabras diferentes en el corpus, el tamaño del vocabulario V. La derivación y formulación matemática de esta ecuación puede ser consultada en el trabajo de [11] y [12].

Se puede observar que, cuando la frecuencia de una palabra es idéntica al pt su distancia a él es cero produciendo un valor de cercanía máximo. Por el contrario, si la palabra se encuentra alejada del pt, entonces su distancia aumentará. La medida que calcula esos valores para cada palabra a ambos lados del punto de transición de Goffman, la *distancia inversa al punto de transición (dipt)*, se define en el trabajo de [13] y su fórmula es:

$$dipt_i = \frac{1}{|pt - f_i| + 1}, \tag{6}$$

en donde pt se calcula de acuerdo con la ecuación 5 y f_i es la frecuencia de la palabra w_i en el corpus, es decir, el número de veces que la palabra ocurre en la colección dada. La posibilidad de dividir entre cero está prevista, por eso usa "+1". Con esta medida, los términos más cercanos al pt son aquellos que obtienen los valores más altos. Por supuesto, pt se puede calcular para un sólo texto y $dipt$ se puede calcular para distintas características de los textos incluyendo los n-gramas.

4. Expansión de consultas

En un SRI se suelen utilizar diferentes métodos para optimizar el proceso de búsqueda de información, uno de ellos es la expansión de las consultas ingresadas por el usuario. En general, se trata de un proceso por el que se toma la consulta original ingresada por el usuario y se le amplía con otros términos equivalentes o más adecuados para expresar un concepto [1], [14]. Existen numerosos trabajos sobre QE, en [3] se presenta una revisión de un gran número de métodos recientes que utilizan diversas fuentes de información[4] y emplean diferentes técnicas.

4.1. Trabajos relacionados

Según la literatura revisada, el problema en cualquier tipo de expansión de consultas es cómo definir cuáles términos están relacionados con los de la consulta. Algunos métodos utilizan cadenas de Markov para seleccionar dichos términos y las probabilidades de transición (es decir, las probabilidades de pasar de una palabra a otra) se estiman, en [15] mediante los tipos de relación semántica (polisemia, antonimia, sinonimia, etc.) entre las palabras, en [16] con la combinación de múltiples fuentes de conocimiento sobre sus relaciones y en [17] mediante las relaciones entre los significados de las palabras, todos obtienen excelentes resultados.

Por otra parte, la expansión de consultas puede ser desarrollada manual, automática o semiautomáticamente (interactivamente). En una expansión de consulta interactiva, el sistema sugiere términos y los presenta al usuario, así el usuario es quien toma la decisión final sobre la importancia relativa y la utilidad de un término [14]. Uno de los sistemas más conocidos de este tipo es la función autocompletar de Google,[5] la cual ofrece posibles términos de búsqueda para completar una consulta mientras el usuario está escribiéndola. Las sugerencias de la función autocompletar se generan automáticamente mediante un algoritmo que se basa en una serie de factores, incluida la frecuencia con la que otros usuarios buscaron una palabra y el contenido de las páginas web. Aunque el algoritmo está diseñado para reflejar la variedad de información disponible es posible que no muestre sugerencias para una palabra o un tema en particular.

En este trabajo se presenta un nuevo algoritmo y una variación del mismo para realizar expansión de consultas de forma semiautomática, requiriendo una interacción mínima del usuario. Estos algoritmos se basan en la probabilidad de que dos palabras aparezcan juntas y en la importancia relativa de esas palabras en una colección de documentos. A diferencia de los métodos descritos anteriormente toda la información se obtiene directamente del corpus, sin usar recursos lingüísticos externos.

[4] Los recursos lingüísticos más utilizados en la QE son diccionarios, tesauros y ontologías.

[5] https://support.google.com/websearch

4.2. Métodos propuestos

Las técnicas propuestas están basadas en la frecuencia de n-gramas en la colección de textos, esto garantiza que los usuarios utilicen las mismas palabras que aparecen en los documentos:

1. El método probabilístico combina la idea de la probabilidad condicionada, los n-gramas, con la noción de sucesos encadenados. La probabilidad de generar una determinada frase con este modelo es simplemente el producto de las probabilidades de los bigramas que la conforman. Esta primera propuesta establece un sistema de pesos en función de la probabilidad de cada n-grama en el corpus, como se muestra en la ecuación 3.
2. El método derivado de la ley de Zipf se basa en la idea de que las palabras más significativas de un texto se agruparán en una zona donde se encuentran las palabras de alta frecuencia con las de baja frecuencia, es decir, el punto de transición de Goffman. En este modelo, el peso de la frase se obtiene en función de la probabilidad de cada n-grama y de la distancia inversa de éste al punto de transición (ver ecuación 6).

En consecuencia, los métodos propuestos permiten ordenar las sugerencias con base en los valores de los pesos obtenidos.

4.3. Algoritmo

El algoritmo consiste en guiar al usuario para expandir el término inicial por uno más específico; permitir seleccionar y agregar términos relacionados con los de la consulta con el fin de precisar los documentos a recuperar; representar en forma adecuada un concepto de interés para el usuario. Dicho algoritmo se explica a continuación:

1. Generar y dividir en archivos cada uno de los tipos de n-gramas (unigramas, bigramas, trigramas, etc.) y contabilizar el número de veces que ocurren en el corpus. En los experimentos, el tamaño máximo de los n-gramas utilizados fue de 5.
2. Obtener el primer término de la consulta especificado por el usuario, el cual actúa como término inicial del n-grama.
3. Buscar en el archivo de $(n+1)$-gramas todas las combinaciones de palabras que comiencen con el n-grama y ofrecerlas al usuario como sugerencias. Para que el usuario pueda evaluar cuál es la mejor elección entre esas sugerencias se ordenan de acuerdo con los pesos de la métrica empleada.
4. Obtener el siguiente término de la consulta elegido por el usuario. Si el término está en la lista de sugerencias se agrega al n-grama, en caso contrario se convierte en término inicial de un nuevo n-grama.
5. Repetir los pasos 3 y 4 hasta que se obtengan consultas con los términos deseados o no se encuentren más sugerencias.

5. Experimentos

Ayudar al usuario a buscar es un desafío interesante que puede realizarse antes o durante el proceso de recuperación de información. Este trabajo presenta un estudio experimental sobre un corpus de artículos de Wikipedia escritos en español. El objetivo principal consiste en medir el funcionamiento de las técnicas diseñadas para sugerir términos de búsqueda con datos reales.

5.1. Medida de relevancia

En ambas técnicas se utilizó el valor promedio del parámetro $tfidf$ para evaluar los términos sugeridos, es decir, se obtuvo la suma de los pesos $tfidf_{ij}$ del término t_i en el documento d_j y se dividió por el número de documentos de la colección donde aparece dicho término, la frecuencia de documentos df_i (en inglés, *document frequency*). Los pesos de $tfidf_{ij}$ se normalizaron mediante la *normalización de coseno*, así los pesos de cada uno de los términos oscilan entre cero y uno, puesto que este proceso crea vectores unitarios. Intuitivamente, tf_{ij} mide la importancia relativa de un término en un documento, mientras que idf_i mide la importancia global de un término en todo el conjunto de documentos. El objetivo de tal esquema de pesado es mejorar la discriminación entre documentos y mejorar la efectividad en tareas de recuperación de información [4].

5.2. Wikicorpus

Para la realización de los experimentos se utilizó un corpus de textos de Wikipedia[6]. El Wikicorpus[7] es un corpus trilingüe (catalán, español e inglés) que contiene gran parte de Wikipedia del año 2006 y en su versión 1.0 contiene más de 750 millones de palabras [18]. De los tres corpora sólo se experimentó con el corpus en español; dicho corpus también integra El Corpus del Español Actual (*CEA*)[8].

Para la evaluación se aplicó la técnica de validación cruzada con 10 diferentes partes del corpus (10 *fold-cross validation*). Cada volumen de prueba tiene 25 904 documentos y un tamaño promedio de 45 MB. En la Tabla 1 se resumen estadísticas acerca de los n-gramas en estos volúmenes, el número de n-gramas promedio es obtenido después de eliminar palabras vacías.

El corpus en crudo no asigna categorías a los documentos, así que la experimentación se realizó sobre texto plano. La separación de palabras se llevó a cabo empleando los signos de puntuación, espacios en blanco o combinaciones de ellos como separadores. Una vez obtenidos los términos, se hizo la conversión de todos los caracteres a minúscula. Finalmente, se aplicó la operación de eliminación de palabras vacías.

[6] https://www.wikipedia.org
[7] http://www.cs.upc.edu/ nlp/wikicorpus
[8] http://spanishfn.org/tools/cea/spanish

Tabla 1. Estadísticas de los n-gramas en los volúmenes

	Únicos	Totales	Frec. máxima
1-gramas	263 616	4 111 204	12 426
2-gramas	1 590 716	2 273 378	4 338
3-gramas	1 156 542	1 242 148	425
4-gramas	668 754	687 443	336
5-gramas	373 407	379 219	231

5.3. Análisis de resultados

A continuación se presentan resultados de la aplicación de los métodos propuestos a palabras de distinto nivel de frecuencia en el corpus. En la Tabla 2 se muestran las 15 palabras utilizadas en los experimentos, palabras con frecuencia alta, mediana y baja. La primera columna es la palabra, la segunda es su peso promedio (ver Sección 5.1), la tercera es el valor de su frecuencia de ocurrencia y la cuarta es la frecuencia de documentos, todos son valores promedio en los 10 volúmenes.

Tabla 2. Palabras utilizadas en los experimentos

Palabra	$tfidf$	f_i	df_i	Palabra	$tfidf$	f_i	df_i	Palabra	$tfidf$	f_i	df_i
ciudad	0.039	12 426	4 671	tormentas	0.074	100	57	cosquillas	0.095	3	3
historia	0.025	8 747	5 190	nervios	0.077	100	66	microbio	0.122	3	2
guerra	0.042	6 739	2 643	glaciar	0.129	99	42	tesauro	0.135	3	2
mundo	0.034	5 145	2 825	robots	0.089	98	41	peculado	0.104	2	2
gobierno	0.046	4 892	1 978	guerrilla	0.077	97	60	estornudo	0.078	2	2

Puesto que la probabilidad de ocurrencia de una palabra se calcula dividiendo el número de veces que aparece una palabra entre el número total de palabras, se deduce que si la palabra es muy frecuente tiene mayor probabilidad de ocurrir y la cantidad de documentos a recuperar es enorme, mientras que, si la palabra no es frecuente se recuperarán pocos documentos. Por otro lado, la medida $tfidf$ indica qué tan relevante es la palabra dentro de un documento y en los documentos de la colección, por tanto, una palabra será importante si no es muy frecuente y aparece en pocos documentos (ver Tabla 2).

Para la evaluación de los métodos propuestos, el número de sugerencias se estableció a 10 o menos (ya que todas las sugerencias posibles pueden ser demasiadas). Como se explicó en la sección 4.2, para el enfoque basado en la ley de Zipf, primero se obtiene un cierto número de los n-gramas más probables (en los experimentos realizados se uso la cuarta parte del total de sugerencias

obtenidas) y luego, se presentan al usuario únicamente los más cercanos al punto de transición, en este caso al menos 10. Por ello, los términos sugeridos por ambos métodos podrían ser distintos.

Como ejemplo de la implementación del algoritmo (en uno de los volúmenes de prueba), en las Tablas 3, 4 y 5 se muestran las sugerencias obtenidas y el número de documentos a recuperar para las consultas *"guerra"* , *"nervios"* y *"cosquillas"*, cuando se elige la primera opción. Cabe mencionar que donde la sugerencia está vacía (es nula) es porque el algoritmo no encontró resultados o se alcanzó el número máximo de sugerencias, 5 en este caso.

Tabla 3. Sugerencias obtenidas para *"guerra"*

Método probabilístico		Método con la ley de Zipf	
Consulta	Sugerencia/Docs.	Consulta	Sugerencia/Docs.
guerra	mundial/804	guerra	independencia/189
	civil/558		civil/558
	independencia/189		vietnam/48
	estados/21		carlista/30
	santa/14		treinta/31
	muerte/5		marina/23
	ciudad/2		franco/31
	brasil/13		anglo/23
	alemania/16		malvinas/19
	paz/15		troya/19
guerra **mundial**	alemania/1	guerra **independencia**	estados/19
	grupo/1		venezuela/8
	ciudad/1		grecia/6
	movimiento/1		turca/4
	francia/2		alto/3
guerra **mundial**	**alemania** nazi/1	guerra **independencia**	**estados** unidos/19
guerra **mundial**	**alemania nazi**	guerra **independencia**	**estados unidos** aliada/1
		guerra **independencia**	**estados unidos aliada**

Nótese que aunque las frecuencias de los n-gramas son mucho más bajas que las frecuencias de las palabras simples, su uso aumenta la cantidad de documentos relevantes obtenidos para la consulta dada, por lo que disminuye la cantidad de documentos recuperados. Esta característica es útil en colecciones de documentos más grandes como la Web, en donde la consulta necesita ser más precisa para obtener más páginas relevantes.

Los resultados globales de los experimentos con las diferentes consultas de prueba se reportan en la Tabla 6. En ella se muestran las palabras iniciales de las consultas, la ponderación promedio de los términos sugeridos (de acuerdo con la función $tfidf$) y el promedio de documentos a recuperar con esas sugerencias. Estos resultados muestran que, en general, las dos técnicas sugieren términos

Tabla 4. Sugerencias obtenidas para *"nervios"*

Método probabilístico		Método con la ley de Zipf	
Consulta	Sugerencia/Docs.	Consulta	Sugerencia/Docs.
nervios	papel/1	nervios	craneales/5
	fuerza/1		longitudinales/1
	manos/1		laterales/2
	sonido/1		cruzados/2
	hojas/1		dibujan/2
	bastante/1		radiales/1
	capilla/1		bastante/1
	aparecen/1		encargados/1
	laterales/2		apoyan/1
	arco/1		perceptivos/1
nervios **papel**	operador/1	nervios **craneales**	opuesto/1
			contralateral/1
nervios **papel operador** arena/1		nervios **craneales opuesto**	
nervios **papel operador arena**			

Tabla 5. Sugerencias obtenidas para *"cosquillas"*

Método probabilístico		Método con la ley de Zipf	
Consulta	Sugerencia/Docs.	Consulta	Sugerencia/Docs.
cosquillas	lengua/1	cosquillas	lengua/1
	plantas/1		plantas/1
cosquillas **lengua**		cosquillas **lengua**	

relevantes a las consultas, si se considera que un término es relevante cuando recupera documentos que contienen los términos de la consulta.

Al comparar los datos mostrados en la Tabla 6 se observa que la técnica derivada de la ley de Zipf obtuvo mejores resultados al sugerir términos más importantes desde el punto de vista del modelo vectorial. A partir de este hecho se deduce que dicha técnica complementa a la técnica probabilística al agregar términos importantes y también comunes, términos de frecuencia media. Esto demuestra que el punto de transición de Goffman se puede utilizar para identificar palabras o n-gramas con un alto valor semántico en el contenido temático de textos.

Finalmente, otro punto interesante a destacar es que únicamente en consultas con palabras de frecuencia extremadamente baja, las dos técnicas van a sugerir las mismas palabras y recuperar los mismos documentos.

6. Conclusiones

Este trabajo explora una aplicación de las cadenas de Markov y la ley de Zipf para autocompletar consultas, es decir, generar secuencias de términos

Tabla 6. $tfidf$ promedio de los términos sugeridos/Documentos promedio a recuperar

	Método probabilístico			Método con la ley de Zipf		
Consulta	1er. Sug.	2da. Sug.	3ra. Sug.	1er. Sug.	2da. Sug.	3ra. Sug.
ciudad	0.050/63	0.056/6	0.083/1	0.059/80	0.072/1	0.072/1
historia	0.048/32	0.075/1	0.067/1	0.054/37	0.070/2	0.073/1
guerra	0.050/154	0.049/2	0.066/2	0.082/90	0.077/6	0.069/5
mundo	0.040/12	0.059/1	0.085/1	0.058/17	0.061/1	0.070/1
gobierno	0.047/38	0.072/7	0.067/1	0.055/41	0.073/2	0.047/4
tormentas	0.056/2	0.061/1	0.056/1	0.068/2	0.065/1	0.061/1
nervios	0.053/1	0.071/1	0.065/1	0.076/1	0.073/1	0.087/1
glaciar	0.049/1	0.069/1	0.061/1	0.123/1	0.083/1	0.057/1
robots	0.048/1	0.080/1	0.063/1	0.073/1	0.066/1	0.052/1
guerrilla	0.054/1	0.144/1	0.057/1	0.092/2	0.065/1	0.085/1
cosquillas	0.064/1	0.048/1	0.081/1	0.064/1	0.056/1	0.081/1
microbio	0.061/1	0.063/1	0.163/1	0.061/1	0.055/1	0.172/1
tesauro	0.069/1	0.066/1	0.074/1	0.069/1	0.053/1	0.131/1
peculado	0.060/1	0.091/1	0.103/1	0.060/1	0.086/1	0.086/1
estornudo	0.056/1	0.057/1	0.058/1	0.056/1	0.210/1	0.063/1

potencialmente útiles para obtener documentos relevantes a una necesidad de información. Las ventajas de emplear este enfoque para la expansión de consultas son, por un lado, su capacidad para aumentar la efectividad en la recuperación de información y, por otro, su sencillez y facilidad para implementarlo.

Los resultados obtenidos, sobre un corpus de textos de Wikipedia en español, demuestran que el método propuesto contribuye a resolver el problema que enfrentan los sistemas de recuperación de información, cuando el usuario parte de una formulación incorrecta de la consulta. La precisión de los sistemas de recuperación de información depende en gran medida de los términos que se encuentran en la consulta, es por ello que, intentar mejorar la consulta puede aumentar la cantidad y calidad de los documentos recuperados para satisfacer la necesidad de información dada por el usuario.

Como trabajo futuro, se planea combinar el método propuesto con los recursos lingüísticos adecuados y evaluar el algoritmo con otras colecciones de documentos y con otros idiomas.

Referencias

1. Baeza, R., Ribeiro, B.: Modern information retrieval, vol. 463. ACM Press, New York (1999)
2. Tolosa, G., Bordignon, F.: Introducción a la Recuperación de Información. Universidad Nacional de Luján, Argentina (2007)
3. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR), 44(1), article 1 (2012)
4. Salton, G., Wong, A., Yang, C.: A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18(11), pp. 613–620 (1975)

5. Moreno, A.: Lingüística Computacional. Editorial Sintesis, Madrid (1998)
6. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT press, Cambridge (1999)
7. Zipf, G. K.: Human Behavior and the Principle of Last-Effort. Addison-Wesley, Cambridge (1949)
8. Luhn, H. P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, vol. 1, pp. 309–317 (1957)
9. Urbizagástegui, R., Restrepo, C.: La ley de Zipf y el punto de transición de Goffman en la indización automática. Investigación Bibliotecológica, vol. 25(54), pp. 71–92 (2011)
10. Velasco, M., Díaz, I., Lloréns, J., Amescua, A., Martínez, V.: Algoritmo de filtrado multi-término para la obtención de relaciones jerárquicas en la construcción automática de un tesauro. Revista Española de Documentación Científica, vol. 22(1), pp. 34–49 (1999)
11. Booth, A.: A Law of Ocurrences for Words of Low Frequency. Information and Control, vol. 10(4), pp. 383–396 (1967)
12. Sun, Q., Shaw, D., Davis, C. H.: A model for estimating the occurrence of same-frequency word and the boundary between high-and low-frequency words in text. Journal of the Association for Information Science and Technology, vol. 50(3), pp. 280–286 (1999)
13. Moyotl, E., Jiménez, H.: Experiments in Text Categorization using Term Selection by Distance to Transition Point. Research on Computing Science, vol. 10, pp. 139–146 (2004)
14. Deco, C., Bender, C., Saer, J., Chiari, M.: Expansión de consultas utilizando recursos lingüísticos para mejorar la recuperación de información en la web. (2005)
15. Cao, G., Nie, J. Y., Bai, J.: Using Markov chains to exploit word relationships in information retrieval. In: Large Scale Semantic Access to Content (Text, Image, Video, and Sound). Le Centre de Hautes Etudes Internationales D'informatique Documentaire, pp. 388–402 (2007)
16. Collins-Thompson, K., Callan, J.: Query expansion using random walk models. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 704–711 (2005)
17. Gan, L., Wang, S., Wang, M., Xie, Z., Zhang, L., Shu, Z.: Query expansion based on concept clique for Markov network information retrieval model. In: Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 5, pp. 29–33. IEEE (2008)
18. Reese, S., Boleda, G. , Cuadros, M., Padró, L., Rigau, G.: Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In: 7th Language Resources and Evaluation Conference, LREC. La Valleta, Malta (2010)