La ley de Zipf y el punto de transición de Goffman en la indización automática

Rubén Urbizagástegui Alvarado *
Cristina Restrepo Arango **

Artículo recibido:
9 de marzo de 2011.
Artículo aceptado:
15 de junio de 2011.

Resumen

Con el fin de identificar las palabras con un alto valor semántico en el contenido temático de un artículo científico se explora el punto de transición de Goffman. Esta técnica se aplicó a 1,644 palabras diferentes identificadas en el texto. Las palabras fueron ordenadas en tablas de frecuencias en orden descendente para explorar cuatro posibilidades. En la primera exploración se tuvieron en cuenta tanto las palabras funcionales como las palabras con un alto valor semántico; en la segunda exploración se eliminaron las palabras funcionales; en la tercera exploración se lematizaron tanto las palabras funcionales como las palabras con un alto contenido semántico, en la cuarta exploración

* Universidad de California en Riverside, USA. ruben@ucr.edu
** Pontificia Universidad Javeriana, Colombia. crestrepoarango@yahoo.com

se eliminaros las palabras funcionales. Los resultados obtenidos en las cuatro exploraciones coinciden en la identificación de las palabras clave.

Palabras clave: Ley de Zipf; Punto de Transición de Goffman; Bibliometría; Cienciometría; Infometría

Abstract

Zipf's law and goffman's transition point in the automatic indexing

Rubén Urbizagástegui Alvarado and Cristina Restrepo Arango

To identify keywords with high semantic value in the thematic content of a scientific paper the Goffman's transition Point is applied. A total of 1,644 different words were identified in the text. These words were arranged in descending order of frequency to explore four possibilities. In the first examination both, functional words and words with high semantic value were taken into account. In the second examination, the functional words were removed; and in the third examination both functional words as well as words with high semantic content, were lemmatized. In the fourth examination only functional word were eliminated. The result shows the identification of specific keywords.

Keywords: Zipf's law; Goffman's Transition Point; Bibliometrics; Scientometrics; Infometrics

Introducción

La indización es un proceso de identificación del contenido de un documento y su descripción a través de términos verbales. De esa manera los conceptos identificados pueden ser representados por términos lingüísticos cuidadosamente seleccionados. Es una técnica de análisis sobre el contenido de un documento que busca expresar la información más significativa a través de la asignación de términos descriptores y crear así un lenguaje de mediación entre el usuario

y el documento. La estrategia de organización de esa información se basa así en *descriptores*, que son palabras clave cuyos conceptos representan el documento en el que están contenidos. Esos descriptores pueden ser escogidos según la capacidad del indizador; o bien siguiendo un conjunto de reglas de selección de las palabras clave de un vocabulario controlado. La indización se convierte de este modo en uno de los procesos básicos de la recuperación de la información, y tiene dos formas de expresión: indización manual que es realizada por una persona; e indización automática que es realizada a través de programas especiales ejecutadas por la computadora.

Uno de los problemas que enfrenta la indización manual es el tiempo disponible para su ejecución y el volumen de documentos esperando ser indizados. Ciertamente ambos problemas influyen en la calidad y adecuación del proceso, y otro de ellos se refiere al *dominio* del texto. La familiaridad y el conocimiento del indizador acerca de la terminología usada corrientemente en los *dominios científicos* son factores que inciden mucho en la calidad de la indización.

Irónicamente es en la indización manual donde su calidad se viene mostrando inadecuada, puesto que además de ser un proceso que conlleva jornadas extensas de trabajo y un elevado costo de ejecución, no puede minimizar la subjetividad del indizador (Bruzinga; Maculan & Lima, 2007). El conocimiento y familiaridad que el indizador tiene sobre el asunto determinará el grado de consistencia que alcance la indización. También el carácter variable de los campos de conocimiento exige una permanente actualización de parte del indizador. Igualmente es necesario tener en consideración las inconsistencias inter-indizadores (diferentes indizadores que les atribuyen diferentes términos a un mismo concepto/documento) y el intra-indizador (el mismo indizador que le atribuye diferentes términos a un mismo concepto/documento). Otro factor que perjudica la calidad de la indización es la posibilidad de que el indizador no domine el idioma en el cual está escrito el documento, por lo que la indización manual exigiría que el indizador, además de dominar los idiomas en que están escritos los textos también domine las terminologías usadas en cada campo científico.

Para superar los problemas señalados anteriormente se han impulsado las investigaciones en el campo de la indización automática. Este tipo de indización consiste en la mecanización del proceso de indización con el propósito de establecer prácticas que reduzcan la interferencia de la subjetividad del indizador, tanto en el análisis del documento como en la selección de los términos significativos para la indización (Mamfrim, 1991) lo que minimizaría al mismo tiempo los problemas impuestos por el idioma.

Uno de los mecanismos que se vienen explorando para identificar y seleccionar las palabras clave de un texto es la ley de Zipf, en especial el punto

de transición de Goffman. Hasta dónde saben los autores de éste artículo, en español existen pocos trabajos que se centren en el análisis e identificación de palabras clave para la indización, la construcción de tesauros y las listas de encabezamientos de asuntos que usen la ley de Zipf y el punto de transición de Goffman. Esas exploraciones están reducidas a los textos de Urbizagástegui Alvarado (1999), quien aplicó la ley de Zipf y el punto de transición de Goffman a un artículo escrito en inglés de Deanna B. Marcum. Encontró aquí un total de 1,025 palabras en el texto, pero sólo estaban presentes 395 palabras diferentes. Usando el punto de transición de Goffman logró identificar 4 palabras clave que describen adecuadamente el contenido del artículo. Concluyó que con esta ley se pueden identificar adecuadamente los términos de indización para un documento. El mismo autor exploró la aplicación de la ley de Zipf y el punto de transición de Goffman a la lírica textual de una grabación sonora de Martina Portocarrero (Urbizagástegui Alvarado, 2004). Encontró 1,999 palabras, de las cuales sólo 387 eran palabras diferentes. De estas palabras seleccionó 11 palabras clave que caracterizan la temática de la música ayacuchana cantada por de Martina Portocarrero. Sin embargo, esas palabras clave no son adecuadas para funcionar en la recuperación de la información. La aplicación de la ley de Zipf a la lírica textual en la que sus contenidos semánticos son metafóricos, parece aportar mejores resultados para la antropología y la etnología. También Jiménez-Salazar; Pinto y Rosso (2005) a través del punto de transición de Goffman estudiaron el problema de la indización en textos cortos. Utilizaron una colección de 48 resúmenes del campo de la lingüística computacional y del procesamiento de textos. El total de palabras encontradas fue de 956 con un promedio de 70.4 palabras por cada texto. Los resultados confirmaron que los términos con ocurrencia media obtenidos a través del punto de transición de Goffman representan muy bien a los textos analizados.

Como se observa más adelante en la revisión de la literatura, la mayoría de las exploraciones sobre este asunto están publicadas en portugués e inglés y sobre textos escritos en ambos idiomas. Por ello nuestro interés está dirigido a la identificación de palabras clave como indicadores del contenido semántico de un documento escrito en español, y que éstos sirvan para la recuperación del documento indizado en una base de datos bibliográficas en lenguaje español. Por eso el objetivo de este trabajo es explorar el punto de transición de Goffman, derivado de la segunda la ley de Zipf para identificar las palabras con un alto valor semántico en el contenido temático de un texto, y la posibilidad de usar estos términos como palabras clave para recuperar la información en bases de datos y catálogos de sistemas de información. Se pretende dar respuesta a la siguiente pregunta:

¿Es posible identificar palabras clave consistentes para la indización de documentos escritos en español aplicando el punto de transición sugerido por Goffman?

Para lograr este objetivo este texto está organizado en cuatro partes. La primera presenta la introducción y luego una revisión de la literatura de las aplicaciones de la ley de Zipf en la BCI y otros campos del conocimiento; la segunda describe la metodología empleada en él; la tercera presenta los resultados obtenidos con la aplicación de esta ley, las conclusiones, y la cuarta presenta finalmente la literatura revisada en la elaboración de este artículo.

Revisión de la literatura

La ley de Zipf ha sido estudiada en diversos campos del conocimiento; por ejemplo, ha sido aplicada para detectar las diferencias semánticas entre los ideogramas del idioma japonés y los fonogramas del idioma inglés (Nabeshima y Ukio-Pegio, 2004); en física ha sido usada para medir la transición de la fase líquida del gas (Ma, 2006); en la contabilidad como mecanismo para la detección de fraudes financieros (Huang, Yen,Yang, y Hua, 2008); en la economía para medir los ingresos (Wyt, 2005); en la demografía para medir la distribución de las poblaciones (Benguigui y Blumenfeld-Lieberthal, 2009; Cordoba, 2008; Black y Henderson, 2003) y en la biología para analizar los aspectos relevantes de la evolución (Bornbergbaue, 1997). La lista de aplicaciones en otros campos del conocimiento diferentes a la ciencia de la información es ilimitada y variada.

Uno de los primeros en explorar la frecuencia con que las palabras aparecían en un texto fue Jean-Baptiste Estoup, quien trabajó como taquígrafo para el parlamento francés. En el desempeño de sus funciones observó las regularidades en la longitud de las palabras en un texto, la aparición de nuevas palabras y la estructura cuantitativa del crecimiento de listas de palabras ordenadas de acuerdo con sus frecuencias de aparición (Estoup, 1908, 1916). Este modelo matemático que estudia la frecuencia de ocurrencias de las palabras en un documento fue también examinado por el físico Condon (1928) quien estudió los textos de L. P. Ayres y G. Dewey que contenían más de cien mil palabras en inglés, con éstas trazó el logaritmo de frecuencia de las palabras observadas frente al logaritmo de la frecuencia de apariciones en el texto, y observó una distribución negativa de las palabras sobre una línea recta. En la opinión de Condon (1928), la frecuencia de las palabras en un texto seguiría una ley cuantitativa de utilidad disminuyente muy similar a la ley de Weber-Fechner en psicología; por tanto, la frecuencia de uso de una

palabra mediría el efecto de su utilidad en la transmisión de ideas entre los individuos.

Posteriormente George Kingsley Zipf estudió también la modelación matemática de la frecuencia con que las palabras aparecen en un texto. Este lingüista se interesó por el estudio de los cambios fonéticos, la frecuencia de uso de los fonemas y su tendencia a cambiar en largos periodos de tiempo. En la década de los años 30 publicó un libro en el cual propuso el *principio de la frecuencia relativa* (Zipf, 1932) y otro donde aparece por primera vez el diagrama de Zipf sobre la frecuencia de ocurrencias de las palabras en los escritos en latín de Palutus (Zipf, 1935). A finales de la década del 40 publicó otro libro en el cual consideró que la principal razón del comportamiento humano es la teoría psicológica conocida como la ley del mínimo esfuerzo (Zipf, 1949), la cual establece que los seres humanos tendemos a minimizar el esfuerzo para obtener resultados exitosos. En este libro aplicó la ley del mínimo esfuerzo a las palabras que aparecen en el Ulises de James Joyce, organizó las palabras en forma descendente y concluyó que los individuos tienden a preferir las palabras más habituales sobre las poco utilizadas. Es decir, nos guiamos por el principio del mínimo esfuerzo que favorece lo común y dificulta lo desconocido. En general, la mayoría de las palabras frecuentes son también la más cortas y más fáciles de pronunciar (Wyllys, 1981; Bailón-Moreno, Jurado-Almeda, Ruiz-Baños y Courtial, 2005).

La propuesta de este lingüista (George Kingsley Zipf) se basa en contar el número de veces que se usa cada palabra en un texto más o menos extenso y ordenar las palabras de las más frecuentes a las menos frecuentes por rangos. Esta tendencia se explica porque siempre es más fácil escribir una palabra conocida que usar una menos conocida. Para autores como Erar (2002) la ley de Zipf es la probabilidad de que una palabra ocurra con cierta frecuencia en un documento, lo cual significa que hay mucho de azar en el uso de las palabras cuando transmitimos un mensaje. También refleja la actitud natural de los individuos a minimizar el esfuerzo en el uso del vocabulario y muestra la tendencia humana a inclinarse más por el mínimo esfuerzo en el uso de palabras conocidas que por la erudición, ya que con el tiempo hay disminución en la riqueza del vocabulario que usamos al escribir.

Esta ley es usada en el campo del procesamiento del lenguaje natural para desarrollar corpus lingüísticos, ontologías, taxonomías y otras aplicaciones, pues le ayuda a identificar el contenido temático de un documento o un conjunto de documentos. A pesar de esta ventaja la ley ha sido poco utilizada en el proceso de indización o elaboración de tesauros en bibliotecología y ciencia de la información (BCI). Su aplicación facilitaría el desarrollo de herramientas como los tesauros, más aún cuando los términos se extraen de la

propia literatura y las frecuencias de las palabras ayudarían a identificar los términos más usados en un área o especialidad del conocimiento, y por tanto identificarían el vocabulario corriente de un autor o autores presentes en un corpus de documentos. Se sabe que el lenguaje es fundamental para construir el conocimiento, puesto que el significado de una palabra representa una estrecha relación entre el pensamiento y el lenguaje que es difícil de discernir al ser un fenómeno del habla o un fenómeno del pensamiento (Vigotsky, 1998). Se entiende la construcción del conocimiento y su comunicación como un proceso dinámico e inseparable del lenguaje. Es a través de éste como el conocimiento pasa a existir y se consolida como tal. Las exploraciones de la aplicación de la informática y la estadística a la documentación en sus relaciones con la ley de Zipf son revisadas por Moreiro Gonzalez (2002).

En esa dirección Luhn (1957) fue uno de los primeros en proponer que la indización debería ser justificada por los propios documentos, en otras palabras, que la indización debería tener una "garantía literaria". Esta garantía literaria significa que el vocabulario de los documentos indexados debería ser aceptado como descriptor o término preferido en los tesauros, ya que éstos están garantizados por su uso real en los propios documentos. El método más simple de identificar los descriptores sería contar el número de veces que una palabra ocurre en un determinado documento, por esa razón los primeros experimentos volvieron los ojos hacia la ley de Zipf para crear una lista de palabras en orden descendiente de ocurrencias en un texto suficientemente extenso. El propio Luhn (1958) sugería que el vocabulario existente en un documento debería constituirse en la base para el análisis de su contenido, pues ésa sería la mejor manera de recuperarlo. La justificación para medir la significancia de las palabras por su frecuencia de uso se basaba en el hecho de que "un escritor normalmente repite ciertas palabras a medida que avanza o varía sus argumentos conforme profundiza en algún aspecto de su asunto. Este énfasis se toma como un indicador de su importancia. Cuanto más frecuentemente se encuentran las palabras en la compañía de otras en una oración, se le puede atribuir mayor significancia a cada una de esas palabras. A pesar de que ciertas palabras deben estar presentes para servir a la importante función de vinculación entre esas palabras, el tipo de significancia buscada en este caso no reside en esas palabras "[…] comunes que pueden ser separadas sustancialmente por métodos no-intelectuales y ser excluidas" (Luhn, 1958:160). Para el propio autor

> existe una probabilidad muy pequeña de que una determinada palabra sea usada para reflejar más de un concepto. También la probabilidad de que un autor use diferentes palabras para reflejar el mismo concepto es pequeña. Incluso si el

autor, por razones estilísticas, hace un esfuerzo razonable para seleccionar sinónimos, pronto se queda sin alternativas legítimas y cae en la repetición si la idea que buscaba está siendo cabalmente expresada (Luhn, 1958:160).

Para ser más claro en su propuesta, Luhn (1958:161) ofrece la siguiente *Figura 1.*

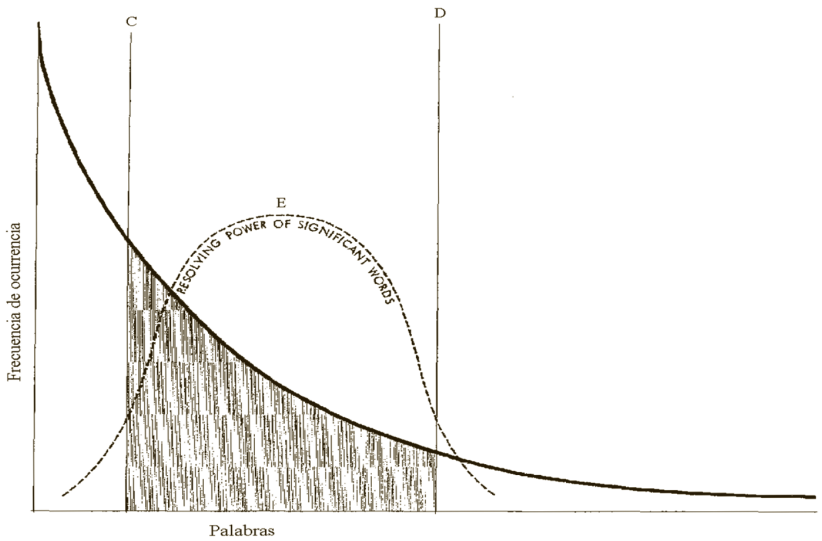

Fig. 1: Zona de ocurrencia de las palabras más significantes

En esta figura la presencia de muchas palabras en la región de más alta frecuencia (a la izquierda de C) tipificadas como "comunes" constituyen "ruido" en el sistema. Este ruido puede ser reducido por una técnica de eliminación a través de la cual las palabras "comunes" del texto se comparan con una lista de palabras sin significancia (stop-words). También por métodos estadísticos se podría establecer un punto de corte de las palabras de alta frecuencia para establecer un "límite de confianza". Si en la Figura 1 la línea C representa ese punto de corte, sólo las palabras del lado derecho tendrían significado. Como el criterio de corte es la frecuencia de ocurrencia de las palabras se podría establecer la línea D como límite inferior. Entonces las palabras agrupadas entre las líneas C y D (la parte sombreada) contendrían las palabras más significativas. En otras palabras, un punto medio de transición entre las palabras de alta y baja frecuencia de ocurrencia en un texto.

Siguiendo esa propuesta, Maia (1973) aplicó la ley de Zipf y el punto de transición de Goffman a tres artículos en el campo de la bibliografía de autores reconocidos publicados en portugués en revistas brasileras. Para el análisis seleccionó

las palabras compuestas, los nombres de instituciones y los establecimientos públicos, los títulos de publicaciones y los nombres de las conferencias. No incluyó en su análisis fechas e indicaciones numéricas de tiempo y denominó los artículos como Texto A, Texto B y Texto C. Para el artículo A encontró 6,395 palabras de las cuales 1,791 eran diferentes; para el artículo B encontró 2,210 y de ellas 839 eran diferentes; y para el artículo C encontró 1,290 palabras 535 las cuales eran diferentes. Los resultados que obtuvo con la aplicación del punto de transición de Goffman fueron: 2, 1.5 y 7 para los textos A, B y C respectivamente. Maia concluye afirmando que la ley de Zipf es válida para la lengua portuguesa, mientras que el punto de transición de Goffman para la lengua portuguesa tiene que sufrir una transformación y adecuación.

Ribeiro (1974) estudió una muestra de 10,093 unidades de editoriales extraídas del lenguaje periodístico que cubrían el campo de la política, las artes, los deportes, etc. del periódico *Jornal do Brasil* publicada en el periodo de 1959-1973 en portugués. Como se supuso que el discurso variaría con el correr del tiempo, se juzgó conveniente analizarlo en tres periodos más cortos: 1953-1963, 1964-1968, 1969-1973. Se encontró que las palabras clave variaban de acuerdo al el periodo examinado, lo que permitió la caracterización de acuerdo con los criterios en cuestión. En el primer periodo examinado (1959-1963) se obtuvieron palabras clave como *Política, Presidente, Brasil.* En el segundo periodo (1964-1968) la información más importante en comparación con el período anterior, fue la aparición de la palabra clave Nacional, que antes no figuraba entre las unidades lingüísticas más frecuentes. Las palabras clave fueron: *Gobierno, Inflación, Nacional* y a través de esa tríada se esbozó la situación brasileña de la época, sus preocupaciones y directrices de políticas: la discusión del problema inflacionario, el culto a lo nacional y la revalorización del gobierno del país. En el tercer periodo (1969-1973) se produjo la desaparición de la palabra *Nacional* y el aumento de frecuencia de la palabra *Inflación.* Estos resultados configuran un examen ideológico del discurso y las variaciones de este discurso durante un periodo de catorce años. Es dentro de ese marco ideológico como el periodismo brasileño analiza la inflación, y reafirma y reitera el nacionalismo brasileño.

Boyce (1975) analizó la efectividad de tres lenguajes de indización automática y sus procedimientos en el área de arritmias cardíacas en textos escritos en inglés. El procedimiento comenzó con el conteo de las palabras del texto completo de cada documento que ordenó en una lista por el rango de frecuencias de ocurrencia de las palabras en diecinueve documentos que trataban del tema de las arritmias cardíacas, y a éstas aplicó el punto de transición de Goffman. La relevancia de los índices la llevó a cabo un estudiante de medicina del cuarto año, quien leyó los diecinueve documentos. Después

de recuperar los documentos midió la sensitividad, especificidad y efectividad de ellos y los resultados indicaron que

> la indización automática basada en la frecuencia de ocurrencia de las palabras en los textos ofrecía una potencial utilidad alternativa para la asignación manual de términos de indización de vocabularios pre-construidos, en especial para la literatura de medicina (Boyce, 1975:384)

Pao (1977, 1978) partiendo de la ley de Zipf, analizó lo que se llama el *punto de transición* de Goffman para identificar las palabras clave en dos artículos en inglés. El primer artículo fue *On the Geometry of Libraries* de A. D. Booth, en el cual encontró 559 palabras diferentes y el número de palabras que ocurrieron una sola vez fue de 256 palabras. Usando el punto de transición de Goffman se logró identificar las palabras más significativas para indizar este texto. El segundo artículo fue *A Law of Ocurrences of Words of Low Frequency* de A. D. Booth y encontró 327 palabras diferentes 188 de las cuales aparecieron una sola vez. Nuevamente aplicando el punto de transición de Goffman se identificaron las palabras clave que representan este texto. La autora sugiere que la aplicación de esta ley permite extraer las palabras clave de un texto con la ayuda de la normalización, lo cual contribuiría a mejorar la recuperación de la información automatizada.

Basilio; Braga y Carvalho (1979) estudiaron dos textos en portugués, uno de lingüística y otro de procesos de automatización en la Biblioteca Nacional del Brasil. A las palabras ordenadas por frecuencias descendientes se les aplicó el punto de transición de Goffman y los resultados mostraron la plausibilidad del punto de transición de Goffman, y confirmaron que la constante 1 es válida para los textos en lengua portuguesa. Mamfrim (1991) buscando verificar la posibilidad de una indización derivativa a partir de textos integrales analizó, mediante el punto de transición de Goffman diez artículos sobre "bibliometría" publicados en lengua portuguesa en el Brasil. Observó que a pesar de variaciones en la longitud de los textos, el 60% del total de palabras en cada texto correspondía a palabras distintas en los textos. También el total de las palabras distintas correspondía al 30% del total de palabras existentes en los textos. En todos los textos analizados el punto de transición funcionó adecuadamente concentrando un conjunto de palabras claves que sugerían el contenido de los documentos. Goffman concluye considerando que

> a través de la frecuencia con la cual las palabras aparecen en los documentos, se puede vislumbrar el propio vocabulario de un asunto, lo que puede ser de gran

> ayuda en la construcción de vocabularios especializados, tesauros, identificación e incorporación de nuevos términos en una disciplina (Mamfrim, 1991:198).

Guedes (1994) partiendo de la premisa de que la frecuencia de ocurrencia de las palabras en un texto mide la significancia de las palabras en ese texto, realiza un estudio en el campo de mecánica de suelos. Seleccionó once notas técnicas, una literatura de discusión sobre el asunto y un artículo, haciendo un total de 13 documentos analizados y escritos en portugués. A la frecuencia ordenada de la ocurrencia de palabras le aplicó luego el punto de transición de Goffman verificando que la región identificada incluyera las frecuencias que correspondían a las palabras de mayor contenido semántico. Esas palabras identificadas fueron después comparadas con las palabras claves seleccionadas por un indizador especialista en mecánica de suelos, y se verificó la correspondencia entre ambos métodos de indización (manual y automatizado).

Quoniam, Balme, Giraud y Dou (1998) aplicaron la ley de Zipf para identificar las palabras clave de 4,703 documentos extraídos de la base de datos Pascal en el periodo de 1993 a julio de 1995 sobre la producción científica de Marseille (Francia). Posteriormente las palabras clave seleccionadas con la ley de Zipf fueron clasificadas en 8 zonas, con el fin de generar consultas que permitieran extraer los términos que representan el contenido de un documento.

Santos (2009) aplicó la ley de Zipf y el punto de transición de Goffman a los resúmenes de 100 cartas del archivo de Bertha María Julia Lutz (1894-1976) escritos en portugués, con el fin de identificar las palabras que tenían con un alto contenido semántico. Encontró 1,902 palabras, de las cuales 618 son palabras diferentes y 406 palabras que sólo aparecen una vez. La autora afirma que se alcanzó el objetivo principal de la investigación, ya que los resultados de la investigación apuntan a una zona de concentración de alto contenido semántico que puede ser utilizado en la indización temática de la correspondencia de la investigadora estudiada. Lapa y Corrêa (2010) aplicaron la ley de Zipf y el punto de transición de Goffman a los resúmenes de las tesis escritos en portugués en la Universidad Federal de Paraíba en Brasil, con el fin de seleccionar los términos relevantes para mejorar la recuperación en la biblioteca digital que almacena estas tesis. Para hacer esta selección usaron un software de fuente abierta que les permitió obtener el listado de palabras generadas a partir de los resúmenes de las tesis. A esta lista de palabras le aplicaron la ley de Zipf y el punto de transición de Goffman. Aunque no detallan los resultados obtenidos con las ecuaciones, concluyeron que la aplicación de estas fórmulas mostró un resultado satisfactorio en relación

con la precisión y la exhaustividad para mejorar la recuperación de la información.

Como se puede observar por la literatura revisada, la ley de Zipf y el punto de transición de Goffman parecen producir buenos resultados en la identificación de palabras con alto valor semántico en el contenido temático de un texto y posibilitan usar estos términos como palabras clave para la recuperación de la información en bases de datos y catálogos de sistemas de información.

Metodología

Como unidades de análisis se tomaron las palabras que aparecieron en el artículo *Análisis cariotípico de Capcicum pubescens (Solanaceae)* "rocoto" publicado por Guevara, Ciles y Bracamonte (2000). Para asegurar la homogeneidad en el conteo de las palabras se adoptaron las siguientes convenciones:

- Una palabra es considerada como una serie de caracteres tipográficos precedida y seguida de espacios en blanco en el texto;
- Las palabras unidas por guión fueron tratadas como una sola palabra;
- Las palabras que expresaron formas singulares o plurales fueron contadas como una sola;
- Las palabras fonéticamente diferentes fueron contadas como diferentes;
- Se omitieron los nombres de autores y nombres de instituciones, por ejemplo, Universidad Nacional Agraria La Molina (UNALM), Lima, Perú;
- Se retiraron del texto números, fórmulas y fechas, por ejemplo, 1964; 2,3 cm; 10 am;
- Algunas abreviaciones fueron completadas, por ejemplo, C. por Capsicum;
- Se retiraron abreviaciones como HCI; HCl 5N; DNA;
- Se retiró el nombre de los autores citados en el texto,
- Se eliminaron las referencias numéricas incluidas en la bibliografía citada, por ejemplo, (Darlington & La Cour, 1969).

Para identificar las palabras clave se usó el punto de transición de Goffman. La derivación y formulación matemática de esta ecuación puede ser consultada en el trabajo de Boyce (1975), Pao (1977, 1978) y Urbizagástegui Alvarado (1999). Aquí solamente se ofrece la ecuación ya derivada y expresada como:

$$n = \frac{\left(-1 \pm \sqrt{1 + 8I_n}\right)}{2}$$

Para el conteo de palabras se usó el software Hermetic Word Frequency Counter, el cual explora archivos en formato texto ANSI, XML y HTML, pero no reconoce los archivos con extensión .doc. Este software cuenta el número de ocurrencias de las palabras en un texto y las organiza alfabéticamente o por frecuencias. También cuenta frases ordenadas alfabéticamente o por frecuencias, de acuerdo con el rango y frecuencia con las que aparecen en el texto.

Resultados

Se encontró un total de 1,644 palabras en el texto, pero sólo estaban presentes 609 palabras diferentes. Esas palabras listadas de acuerdo con el orden de frecuencia descendente aparecen en la *Tabla Núm. 1* (Ver *Anexos*), donde se muestran las palabras ordenadas desde la primera palabra con 102 ocurrencias, hasta las últimas con sólo 1 ocurrencia al final de la lista, para hacer un total de 409 palabras únicas. Es evidente que las palabras funcionales (artículos, adjetivos, adverbios, preposiciones y conjunciones) son las más frecuentes en el texto y se situaron en el tope de la escala. También es obvio que algunas palabras de gran significado semántico para el texto están situadas entre o siguiendo a las palabras funcionales. Precisamente se afirma que la ley de Zipf está asociada a la identificación de grupos en los extremos de una lista de rangos de palabras. El primer grupo de las palabras de alta frecuencia y con poco significado semántico se agrupan en el tope de la lista, y el segundo grupo de palabras de uso menos frecuente se agrupa al final de la lista. Frente a esta situación Goffman introduce la idea de que las palabras más significativas de un texto se agruparían en una zona donde se encuentran las palabras de alta frecuencia con las de baja frecuencia; es decir, un punto intermedio de transición. Aplicando la ecuación del punto de transición de Goffman se obtuvo n = 28,11, lo que significa que el punto de transición se ubica entre las palabras que ocurrieron 28 veces, ésta es la palabra las. No obstante, al rededor de esta palabra ocurren otras palabras como *cromosoma* y *capsicum*, las cuales son más apropiadas para representar el contenido del texto, en otras palabras, *cromosoma* y *capsicum* serían las palabras clave identificadas en este caso.

Para explorar si la eliminación de las palabras funcionales mostradas en la *Tabla Núm. 1* (Ver *Anexos*) ayudan a identificar las palabras clave con mayor

precisión se elaboró la *Tabla Núm. 2* (Ver *Anexos*). Nuevamente el valor del punto de transición fue de 28,11, es decir, que las palabras de mayor significación semántica se ubican entre aquellas que ocurren 28 veces en esta tabla, en la cual se identificaron las palabras *cromosomas*, *capsicum*, *bandas* y *pubescens*. Estas palabras deberían ser elegidas como palabras clave.

Para explorar si ocurrirían cambios en la identificación de las palabras clave, en la *Tabla Núm. 3* (Ver *Anexos*) se lematizaron las palabras del texto; en otras palabras, se redujeron o agruparon las diferentes formas flexivas o variaciones morfológicas de una palabra a la forma canónica que se selecciona como lema o entrada o palabra clave. Aplicándose el punto de transición se obtuvo que n = 21,64, aproximadamente 22 palabras. Las palabras más significativas ocurren alrededor de aquellas palabras que tienen 22 ocurrencias en la *Tabla Núm. 3* (Ver *Anexos*). Éstas fueron las palabras *banda, bandas* y *bandeo.* Por lo tanto, las palabras más significativas deben ocurrir alrededor de estas tres palabras y estas fueron identificadas como las palabras *cromosoma, cromosomas, cromosómica, cromosómicas, cromosómico* y *capsicum*. De este resultado las palabras clave seleccionadas serían *cromosoma, banda* y *capsicum*.

Para experimentar el efecto que puede tener la eliminación de las palabras funcionales mostradas en la Tabla No. 3 (Ver anexos) se elaboró la Tabla Núm. 4 (Ver anexos). Como se puede observar en esa tabla, las 3 primeras palabras coinciden con los resultados obtenidos en los 3 experimentos anteriores. Por lo tanto las palabras que identifican el contenido del artículo y que podrían ser seleccionadas como palabras clave, son c*romosomas, capsicum*, *banda* y *pubescens.*

Conclusiones

La ley de Zipf a través de la aplicación del punto de transición de Goffman, de acuerdo con los resultados obtenidos en este trabajo facilita la identificación de las palabras clave de un documento o un conjunto de documentos. La eficiencia de este modelo se refleja en la lista de palabras lematizadas, de la cual se obtuvieron varias palabras que podrían ser utilizadas para hacer la indización del documento estudiado. Éstas son las palabras *cromosomas, capsicum, banda* y *pubescens* que facilitarían la recuperación de información en bases de datos bibliográficas en español. Asimismo, los datos que se obtuvieron en este trabajo corroboran la explicación de Zipf sobre las altas ocurrencias de palabras funcionales, ya que en este caso la palabra con el mayor número de ocurrencias fue el artículo "las". Esto confirma lo manifestado por la teoría

del mínimo esfuerzo, la cual sostiene que los seres humanos tendemos a minimizar el esfuerzo, en otras palabras, preferimos el uso de las palabras más conocidas y más fáciles de pronunciar sobre las menos conocidas.

En el campo de la ciencia de la información y la bibliotecología se deberían seguir explorando las posibilidades del uso de la ley de Zipf y el punto de transición de Goffman en la indización automática. De esta manera evitaríamos el llamado de atención hecho por Braga (1996) acerca de que la ley de Zipf es una de las más conocidas y, curiosamente, de menor aplicación práctica en sistemas de información. Ella afirma que lo correcto sería hablar de dos leyes de la ley de Zipf: la de alta frecuencia y la de baja frecuencia. Ambas leyes tal como son enunciadas, son meras "curiosidades" para los sistemas de recuperación de la información ya que es casi nula su aplicación práctica en los problemas que enfrentan los sistemas de recuperación de la información.

Bibliografía

Bailón-Moreno, R.; Jurado-Almeda, E.; Ruiz-Baños, R. y Courtial, J. P., "Bibliometric laws: empirical flaws of fit", en *Scientometrics*, 63(2):209-229, 2005.

Basilio, Margarida Maria de Paula; Braga, Lilian Maria; Pierotti y Maria de Lourdes Carvalho, "Estrutura de textos científicos em língua portuguesa: estudo bibliométrico-linguístico", en *Reunião Brasileira de Ciência da Informação* (2. : 1979 : Rio de Janeiro), [trabalhos apresentados], Río de Janeiro: IBICT, 1979.

Benguigui, L. y Blumenfeld-Lieberthal, E., *The temporal evolution of the city size distribution. Physica A: Statistical Mechanics and Its Applications*, 388(7):1187-1195, 2009.

Black, D. y Henderson, V., "Urban evolution in the USA", en *Journal of Economic Geography*, 3(4):343-372, 2003.

Bornbergbauer, E., "How are model protein structures distributed in sequence space?", en *Biophysical Journal*, 73(5):2393-2403, 1997.

Boyce, Bert, "Automatic and manual indexing performance in a small file of medical literature", en *Bulletin of the Medical Library Association*, 63(4):378-385, 1975.

Braga, Gilda Maria, "A representação da informação na descontrução do contexto", *Informare: Cadernos do Programa de Pós-Graduação em Ciencia da Informação*, 2(2):53-57, 1996.

Bruzinga, Graciane Silva; Maculan, Benildes Coura Moreira dos Santos y Lima, Gercina Ângela Borém de Oliveira. "Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações", en *VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação*, 28 a 31 de outubro de 2007, Salvador, Bahia, Brasil.

Condon, E. U. ,"Statistics of vocabulary", en *Science*, 68:1733, 1928.

Cordoba, J. C., "On the distribution of city sizes", *Journal of Urban Economics*, 63(1):177-197, 2008.

Erar, Aydin, "Bibliometrics or informetrics: displaying regularity in scientific patterns by using statistical distributions", en *Hacettepe Journal of Mathematics and Statistics*, 31:113-125, 2002.

Estoup, J. B., *Gammes sténographique*, París, Institut Sténographique, 1916.

_____, *Gammes sténographiques: recueil de textes choisis pour l'acquisition méthodique de la vitesse, précédé d'une introduction*, París: Institut Sténographique, 1908.

Guedes, Vânia Lisbõa da Silveira, "Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos", en *Ciencia da Informação*, 23(3):318-326, 1994.

Guevara, Misael; Ciles, María y Bracamonte, Olga, "Análisis cariotípico de Capsicum pubescens (Solanaceae) 'rocoto' ", en *Revista Peruana de Biología*, 7(2):1-10, 2000.

Huang, S. M.; Yen, D. C.; Yang, L. W.; Hua, J. S., "An Investigation of Zipf's law for fraud detection", en *Decision Support Systems*, 46(1):70-83, 2008.

Jiménez Salazar, Héctor; Pinto, David y Rosso, Paolo, "Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos", en *Procesamiento del Lenguaje Natural*, 35:383-390, 2005.

Lapa, Remi Correia y Corrêa, Renato Fernandes, "Seleção de descritores para a indexação automática de teses e dissertações da UFPE", en Encontro Nacional de Estudantes de Biblioteconomia, *Documentação, Gestão, e Ciência da Informação Os desafios do profissional da informação frente às tecnologias e suportes informacionais do século XXI: lugares de memória para a biblioteconomia 18 a 24 de julho de 2010*, Brasil: Universidade Federal da Paraíba.

Luhn, Hans Peter, "A statistical approach to mechanized encoding and searching of literary information", en *IBM Journal of Research and Development*, 1(4):309-317, 1957.

_____, "The automatic creation of literature abstracts", en *IBM Journal of Research and Development*, 2(2): 157-165, 1958.

Ma, Y. G. Moment analysis and Zipf law, *The European Physical Journal*, 30:227-242, 2006, disponible en: <http://cyclotron.tamu.edu/wci3/newer/chapV_5.pdf>. acceso el 30 de septiembre de 2010.

Maia, Elza Lima e Silva, "Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação", en *Ciência da Informação*, 2(2):99-138, 1973.

Mamfrim, Flavia Pereira Braga, "Representação de conteúdo via indexação automática em textos integrais de língua portuguesa", en *Ciência da Informação*, 20(2):191-203, 1991.

Moreiro González, José Antonio, "Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información", en *Anales de Documentación*, 5:273-286, 2002.

Nabeshima, Terutaka y Ukio-Pegio, Gungi, "Zipf's law in phonograms and Weibull distribution in ideograms: comparison of English with Japanese", en *BioSystems*, 73:131–139, 2004.

Pao, Miranda Lee, "Automatic indexing based on Goffman's transition of word occurrences", en *American Society for Information Science*, Meeting (40th : 1977 : Chicago, Ill.); *Information management in the 1980's : proceedings of the ASIS annual meeting 1977*, Volume 14 : 40th annual meeting, Chicago, Illinois, September 26-October 1, 1977; y Bernard M. Fry, compiler, Clayton A. White Plains, N.Y. *Knowledge Industry Publications for American Society for Information Science*, c1977, traducido por Rubén Urbizagástegui Alvarado.

_____, "Automatic text analysis based on Transition Phenomena of word occurrences", en *Journal of the American Society for Information Science*, 29(3):121-124, 1978.

Quoniam, L.; Balme, F.; Rostamg, H.; Giraud, E. y Dou, J. M., "Bibliometric law used for information retrieval", en *Scientometrics*, 41(1-2):83-91, 1998.

Ribeiro, Lais A., "Aplicação dos métodos estatísticos e da teoria da informação e da Comunicação na análise lingüística: estudo da linguagem jornalística", en *Ciência da informação*, 3(2):151-154, 1974.

Santos, María José V. C., Correspondência científica de Bertha Lutz: um estudo de aplicação da lei de Zipf e ponto de transição de Goffman em um arquivo pessoal, Ponto de Acesso, Salvador, 3(3):317-326, 2009.

Urbizagástegui Alvarado, Rubén, "Las posibilidades de la ley de zipf en la indización automática", en *B3: Ciencias de la Información*, [1999?], disponible en: http://b3.bibliotecologia.cl/ruben2.htm, acceso en: 20 de septiembre de 2010.

_____, "Martina Portocarrero: un acercamiento a través de la ley de Zipf", en *III Encuentro Internacional de Invierno*, ECI2004, [Lima], ECI, 2004, disponible en: <http://www.encuentrocientificointernacional.org/eci2004i/libroderesumenes.pdf>, acceso 3 de enero de 2011.

Vigotsky, L., *Pensamiento y Lenguaje*, Obras Escogidas, T II. Madrid: Visor, 1988.

Wyllys, Ronald E., "Empirical and theoretical bases of Zipf's law", en *Library Trends*, 30(1):53-64, 1981.

Wyt, Gerrit de, "Zipf's law in economics", *SCALES-Paper* N200503, 2005, disponible en: http://www.ondernemerschap.nl/pdf-ez/N200503.pdf, acceso en: 28 de septiembre de 2010.

Zipf, George Kingsley, *Human behaviour and the principle of least effort*, Cambridge, Mass, Adisson-Wesley Press, 1949.

_____, *The psycho-biology of language*, Boston: Houghton Company, 1935.
_____, *Selected studies of the principle of relative Frequency in language*, Cambridge, Mass, Harvard University Press, 1932.

ANEXOS

Tabla Num. 1: Palabras jerarquizadas

Rango R	Ocurrencias F	R x F C	Palabras
1	102	102	de
2	65	130	la
3	54	162	el
4	52	208	los
5	49	245	se
6	44	264	que
7	34	238	cromosomas
8	34	272	y
9	29	261	con
10	28	280	las
11	21	231	a
	21	231	es
12	20	240	como
13	19	247	del
14	17	238	capsicum
	17	238	una
15	16	240	bandas
	16	240	para
16	13	208	lo
17	11	187	no
18	10	180	pubescens
	10	180	son
19	8	152	al
	8	152	brazo
	8	152	constricción
	8	152	orceína
	8	152	pares
20	7	140	corto
	7	140	cromosómico
	7	140	ha
	7	140	heterocromatina
	7	140	presentan
	7	140	secundaria
	7	140	vegetales
21	6	126	(13 palabras diferentes)
22	5	110	(15 palabras diferentes)
23	4	92	(15 palabras diferentes)
24	3	72	(33 palabras diferentes)
25	2	50	(96 palabras diferentes)
26	1	26	(409 palabras diferentes)

Tabla Núm. 2: Palabras ranqueadas sin palabras funcionales

Rango R	Ocurrencias F	R x F C	Palabras
1	34	34	cromosomas
2	17	34	capsicum
3	16	48	bandas
4	10	40	pubescens
5	8	40	brazo
	8	40	constricción
	8	40	orceína
	8	40	pares
6	7	42	corto
	7	42	cromosómico
	7	42	heterocromatina
	7	42	presentan
	7	42	secundaria
	7	42	vegetales
7	6	42	(13 palabras diferentes)
8	5	40	(15 palabras diferentes)
9	4	36	(15 palabras diferentes)
10	3	30	(33 palabras diferentes)
11	2	22	(96 palabras diferentes)
12	1	12	(409 palabras diferentes)

Tabla Núm. 3 : Lematización de las palabras

Rango R	Frec F	R x F C	Palabras
1	213	213	la / las / le / lo / los / el
2	121	242	de / del
3	55	165	cromosoma/cromosomas/cromosómica/cromosómicas/cromosómico
4	49	196	se
5	44	220	que
6	34	204	y
7	29	203	a / al
	29	203	con
	29	203	es / son
8	25	200	un / una / uno
9	22	198	banda / bandas / bandeo
10	20	200	como
11	17	187	capsicum
12	16	192	esta /éstas / este / esto / estos
	16	192	para
	16	192	presentar / presentan / presente / presenta
13	15	195	par / pares / pareados
14	14	196	brazo / brazos

15	12	180	ha / han
16	11	176	análisis / analizadas / analizar / analizaron / analizarse
	11	176	color/coloración / colorante / colorea / coloreada /coloreadas / coloreados
	11	176	constricción / constricciones
	11	176	no
	11	176	pubescens
17	10	170	metafase / metafases / metafásicas / metafásicos
		10	170 permite / permitido / permitió
		10	170 teñida / teñidas / teñidos / tinción / tiñen
18	9	162	corta / corto / cortadas
	9	162	especie / especies
	9	162	heterocromatina / heterocromatínicas / heterocromatínicos
	9	162	identificar / identidad / identificación
	9	162	número / números / numerosa / numerosas
	9	162	región / regiones
	9	162	secundaria / secundarias
19	8	152	característica / características / característico / caracterización
	8	152	nuclear / núcleos / nucleolo / nucleolares
	8	152	orceína
	8	152	vegetal / vegetales
20	7	140	algunas / algunos
	7	140	debería / debido / debidamente
	7	140	centromérica / centroméricas / centromérico/ centrómero / centrómeros
	7	140	estudio / estudios
	7	140	fue /fueron
	7	140	longitud / longitudes
	7	140	obtener / obtiene / obtención / obtenidas / obtenidos
	7	140	pudo / puede / pueden
	7	140	telomérica / teloméricas / telómeros
21	6	126	(13 palabras diferentes)
22	5	110	(15 palabras diferentes)
23	4	92	(15 palabras diferentes)
24	3	72	(33 palabras diferentes)
25	2	50	(96 palabras diferentes)
26	1	26	(409 palabras diferentes)

Tabla Núm. 4: Lematización de palabras sin palabras funcionales

Rango R	Frec F	R x F C	Palabras
1	55	55	cromosoma/cromosomas/cromosómica/cromosómicas/cromosómico
2	22	44	banda / bandas / bandeo
3	17	51	capsicum
4	15	60	par / pares / pareados
5	14	70	brazo / brazos
6	11	66	color/coloración/colorante/colorea/coloreada/coloreadas/coloreados
	11	66	constricción / constricciones
	11	66	pubescens
7	10	70	metafase / metafases / metafásicas / metafásicos
8	9	72	especie / especies
	9	72	heterocromatina / heterocromatínicas / heterocromatínicos
9	8	72	nuclear / núcleos / nucleolo / nucleolares
	8	72	orceína
	8	72	vegetal / vegetales
10	7	70	telomérica / teloméricas / telómeros
11	6	66	(8 palabras diferentes)
12	5	60	(15 palabras diferentes)
13	4	52	(20 palabras diferentes)
14	3	42	(29 palabras diferentes)
15	2	30	(67 palabras diferentes)
16	1	16	(245 palabras diferentes)