Aplicación del algoritmo RAKE en la indización de documentos digitales

Marcial Contreras Barrera*

Artículo recibido:
14 de septiembre de 2015
Artículo aceptado:
10 de octubre de 2016

Resumen

En la actualidad existe una diversidad de documentos digitales y en particular de documentos textuales que, dada su naturaleza, permiten la aplicación de métodos automatizados de procesamiento, organización y análisis con el fin de obtener información de manera concisa y de forma eficiente.

Diversas áreas de estudio, como la informática, la bibliotecología, la lingüística computacional y la minería de texto, se encargan de desarrollar métodos para el procesamiento de documentos digitales con la meta de facilitar su representación, organización y recuperación, tanto en bibliotecas digitales como en bases de datos y catálogos. Estos métodos pueden ser de tipo

* Universidad Nacional Autónoma de México, México marcial@dgb.unam.mx

estadístico o lingüístico. En este artículo se estudia el método RAKE de tipo estadístico con la finalidad de identificar y extraer palabras clave multipalabra de los documentos digitales para su organización y recuperación, además de la aplicación del método en la indización automatizada de documentos.

Palabras clave: Método RAKE; Indización; Métodos Automatizados; Consistencia

RAKE algorithm application in digital document indexing

Marcial Contreras Barrera

Abstract

Currently there are a wide range of digital documents, particularly text documents that by their nature allow automated processing, organization and analysis methods for the purpose of retrieving information concisely and efficiently. Diverse areas of study such as computer science, library science, computational linguistics and text mining, among others, have developed digital document processing methods for the purpose of facilitating their representation, organization and retrieval in digital libraries, databases and catalogs. These methods are both statistical and linguistic in nature. In this paper, the RAKE statistical method is examined in order to identify and extract multiword keywords from digital documents to allow organization, retrieval and automated document indexing.

Keywords: RAKE Method; Indexing; Automated Methods; Consistency

Introducción

Las empresas, las bibliotecas —tradicionales o digitales— y cualquier organización tienen la necesidad de procesar, organizar, consultar y recuperar los diferentes tipos de documentos producidos para satisfacer las necesidades de información de sus usuarios. Una cantidad significativa de esa información se encuentra en formato digital, por lo que es importante realizar estudios encaminados a mejorar su procesamiento y disponibilidad.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

El estudio de la información digital, y particularmente el estudio de los documentos digitales, se realiza desde diferentes áreas del conocimiento, como el procesamiento del lenguaje natural —PLN—, la informática, la bibliotecología, los estudios de la información, la recuperación de información y la minería de texto, entre otras (*Figura 1*).

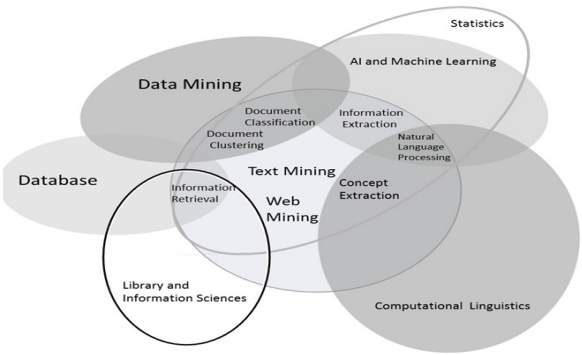

Figura 1. Áreas de estudio del documento digital textual
Fuente: Miner *et al*., 2012:31

La importancia del estudio de los documentos digitales radica en que permite identificar sus características y la posibilidad de extraer patrones con el objetivo de proponer métodos que ayuden a su procesamiento, organización, consulta, y recuperación.

Los métodos pueden tener diferentes finalidades, como la extracción de términos, la clasificación, la agrupación de documentos y la identificación de relaciones. Con el fin de cumplir con las tareas citadas, se pueden utilizar métodos basados en estadística, lingüística o la combinación de ambos.

Los métodos para el reconocimiento de términos tienen como objetivo la extracción de términos simples o compuestos a través de técnicas basadas en modelos matemáticos o lingüísticos, las cuales permiten la identificación de una manera eficiente de los términos. La extracción de términos es utilizada para identificar palabras clave a través de procedimientos manuales o automatizados. Los procesos automatizados tienen como finalidad la identificación de palabras de una manera rápida y eficiente (Urbizagástegui Alvarado y Restrepo Arango, 2011). La identificación y la extracción de términos desde la recuperación de la información tienen dos objetivos: principalmente sirven para crear una representación del contenido del documento y así

ser utilizados en los motores de búsqueda y bibliotecas digitales; en bibliotecología se usan para el proceso de indización automatizada y la creación de tesauros; en la minería de texto, para encontrar conceptos y establecer relaciones entre documentos.

Desde la terminología, la identificación de términos y sus definiciones permite la creación de diccionarios especializados, el desarrollo de ontologías utilizadas en la web semántica, la creación de glosarios e índices de manera automatizada y el agrupamiento de documentos. Además, la identificación de términos también permite la traducción automatizada, la indización de documentos, la recuperación y extracción de información, y la generación de texto (Barrón Cedeño, 2007).

Los métodos para la identificación y extracción de términos pueden variar, dependiendo del propósito y del área de estudio de los documentos. Un ejemplo de lo anterior es la organización de la información presentada en el navegador Yahoo, en el cual la información está organizada por temas. Esta organización se realiza mediante los métodos de categorización automatizados basados en la identificación de términos.

La indización es un proceso intelectual que determina los temas principales y secundarios contenidos en un documento. Para representar estos temas o materias, se utilizan palabras clave, descriptores, etcétera, de modo que la información contenida en los documentos se pueda representar mediante la combinación de estos términos representativos (Lévano, 2011).

En las bibliotecas tradicionales, la indización es efectuada por personal especializado, el cual realiza el proceso de forma manual e intelectual, dando como resultado palabras clave que son utilizadas como puntos de acceso en la búsqueda de la información.

Sin embargo, la identificación de términos suele ser una actividad que consume mucho tiempo, por lo que es necesario el uso de la tecnología para facilitar el proceso. En la actualidad se han desarrollado métodos automatizados que permiten agilizar la identificación y extracción manual con la meta de procesar el mayor número de documentos de manera rápida y precisa.

El método de la ley de Zipf y el punto de transición de Goffman son utilizados en la identificación de las palabras clave en los idiomas inglés, portugués y español como parte del proceso de indización automatizada. De acuerdo con Urbizagástegui Alvarado y Restrepo Arango (2011), se concluye que este método automatizado puede utilizarse adecuadamente para identificar palabras clave, aunque sólo se limita a las que están compuestas por una sola palabra; es decir, no es posible la identificación de palabras clave multipalabra.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

Para traspasar los límites de algunos de los métodos existentes, como el de Zipf, y debido a lo complejo del lenguaje natural, se han desarrollado métodos basados en la lingüística, dando origen al área del procesamiento del lenguaje natural —PLN— para el procesamiento de los documentos digitales. Los métodos desarrollados en esta área toman en consideración la estructura gramatical y partes de la oración para el análisis de los documentos, realizando un análisis léxico, morfológico, sintáctico y semántico (Hurwitz *et al.*, 2013).

Los análisis léxico y morfológico examinan las características de palabras como los prefijos, sufijos, las raíces de las palabras y partes de la oración —nombres, verbos, adjetivos, etcétera—; el etiquetado de partes de la oración asigna a cada palabra de un texto una etiqueta con la categoría gramatical a la que pertenece, como se muestra en la *Tabla 1*.

1. La	ART	el
2.bibliotecología	NC	bibliotecología
3. es	VSfin	ser
4. la	ART	el
5. ciencia	NC	ciencia
6. social	ADJ	social
7. que	CQUE	que
8. estudia	VLfin	estudiar
9. el	ART	el
10. manejo	NC	manejo
11. proceso	NC	proceso
12. transmisión	NC	transmisión
13. adquisición	NC	adquisición
14. producción	VLfin	producción
15. de	PREP	de
16. la	ART	el
17. información	NC	información

Tabla 1. Etiquetado de la oración

Los nombres de cada una de las etiquetas son los siguientes: SES: sujeto expreso simple; SEC: sujeto expreso compuesto; ST: sujeto tácito; SS: sujeto simple; PVC: predicado verbal compuesto; PVS: predicado verbal simple; N: núcleo; NC: núcleo compuesto; NS: núcleo simple; NV: núcleo verbal; OD: objeto directo; ART: artículo.

Otro de los métodos empleados en el procesamiento de los documentos digitales es el llamado algoritmo C-value/NCvalue (Barrón Cedeño, 2007), el cual es un método híbrido que utiliza la lingüística y la estadística para la extracción de términos, diseñado para la obtención de términos en inglés y posteriormente adaptado para términos en español. En la primera etapa del algoritmo se genera una lista de términos candidatos, basados en patrones sintácticos y en una lista de palabras que sirve para eliminar palabras candidatas que no pueden ser términos. En la segunda parte se realiza el cálculo de los términos candidatos, tomando en cuenta su longitud y frecuencia de aparición en el texto.

Finalmente, el método RAKE —Rapid Automatic Keyword Extraction—, de tipo estadístico, ha sido utilizado para identificar y extraer palabras clave compuestas para más de una palabra en documentos escritos en inglés. Por lo anterior, el objetivo de este trabajo es adaptar y aplicar el método RAKE para usarse en la identificación de términos formados por una o más palabras —multipalabra— en español, además de su aplicación en la indización automatizada. Al tener estos supuestos, ¿se puede utilizar el método RAKE para la identificación de palabras clave de forma eficiente en documentos escritos en español? ¿Se puede utilizar el método RAKE en el proceso de indización automatizada?

El método originalmente fue empleado en documentos en idioma inglés y, en este artículo, es adaptado para ser utilizado en el procesamiento de documentos en idioma español. Con las adaptaciones realizadas, se desarrolló un sistema de cómputo en el lenguaje de programación PHP y el manejador de bases de datos MySQL para evaluar su aplicación en la identificación de palabras clave en español y su aplicación en el proceso de indización automatizada. Con el sistema de cómputo se procedió a analizar y procesar documentos digitales textuales para identificar palabras simples y/o compuestas, y evaluar la aplicación del método RAKE en el proceso de indización.

De acuerdo con Zunde, citado por Gil, la consistencia es definida como "el grado de concordancia en la representación de la información esencial de un documento, por medio de un conjunto de términos de indización seleccionados por cada uno de los indizadores de un grupo" (Gil, 1999: 30). Por lo que la evaluación de la indización, manual o automatizada, se puede llevar a cabo tomando los criterios de consistencia o por el cálculo de exhaustividad y precisión en la recuperación de información. El mismo autor expone que la inconsistencia en la indización es inherente a ésta y no se debe ver como una anomalía.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

Para el cálculo de la consistencia manual y automatizada, Salton y McGill (1983: 100) propusieron una fórmula que también puede ser utilizada entre indizadores. La fórmula queda definida de la siguiente manera:

$$C=TA+B-T \qquad \text{Ecuación (1)}$$

Donde:

C = Consistencia entre indizadores o dos sistemas

T = Número de términos comunes asignados por los indizadores

A = Número de términos asignados por el indizador 1

B = Número de términos asignados por el indizador 2

De manera general, la identificación y extracción automática de términos en documentos es empleada en la traducción automatizada, la indización de documentos, la identificación de términos, la recuperación y extracción de información, la generación de texto y la generación de herramientas lexicográficas. En el siguiente apartado se describe el método RAKE y su aplicación en la identificación de palabras clave y en la indización de documentos.

Rapid Automatic Keyword Extraction —RAKE—

Rapid Automatic Keyword Extraction —RAKE— (Rose *et al*., 2010: 1-20) es un algoritmo utilizado para la extracción de palabras clave —*keywords*— compuestas por una o más palabras, basado en las estadísticas de las palabras y de las coocurrencias de las mismas; trabaja sobre documentos individuales para obtener palabras clave compuestas por una o más de una palabra, las cuales sirven de base para la descripción del contenido de los documentos, la indización de los mismos o en algún estudio de minería de texto.

Para el funcionamiento del algoritmo, es necesaria la definición de tres parámetros de entrada: lista de *stopwords* —*stoplist*—, lista de delimitadores de frases y lista de palabras delimitadoras. Las características de los tres parámetros de entrada son muy importantes para obtener la mayor precisión a la hora de identificar las palabras clave. Un ejemplo de las *stopwords* utilizadas en el algoritmo para el idioma español se muestra en la *Figura 2*. La lista presentada se obtuvo a partir de un análisis realizado a un conjunto de documentos del área de ingeniería y bibliotecología.

abandonar, abjurar, ablandar, ablandarse, abochornar, abofetear, abolir, abonarse, abordar, abortar, abovedar, abrasar, abrazar, abreviar, abrir, abrochar, abrumar, absorber, abuchear, abundar, aburrir, abusar, acabar, acallar, acampar, acaparar, acariciar, acceder, accionar, acechar, aceitar, acelerar, acentuar, acepillar, aceptar, acerar, acercar, achatar, acicalar, aclamar, aclarar, aclimatar, acoger, acolchar, acometer, acomodar, acompañar, aconsejar, acoplar, acordar, acorralar, acosar, acostar, acostarse, acotar, acreditar, acribillar, activar, actuar, acuartelar, acumular, acunar, quisiera, quiero, presenta, realizado, brinda, centra, debería, entendida, conocido, inserta, ofrecía, ocurra, escrito, presentan, usaron, verificado, aparición, aplicó, significa, representa, explora, ocurre, aplicó, partiendo, encontró, que, las, al, le, al, sus la, para, sobre, son, considerando, construcción, todos, tipos, dado, soluciones, pueden, ser, en, mezcla, usado, este, esta

Figura 2. Lista de stopwords en español

El análisis automatizado del documento comienza con la identificación de las palabras candidatas, tomando en consideración las palabras delimitadoras, los delimitadores de frases, la posición de *stopwords* y los signos de puntuación. Para poder identificar y extraer las palabras simples y compuestas, se aplica el método RAKE, el cual es descrito en la *Figura 3*. Una vez definidos los puntos anteriores, se aplican los siguientes pasos para la obtención:

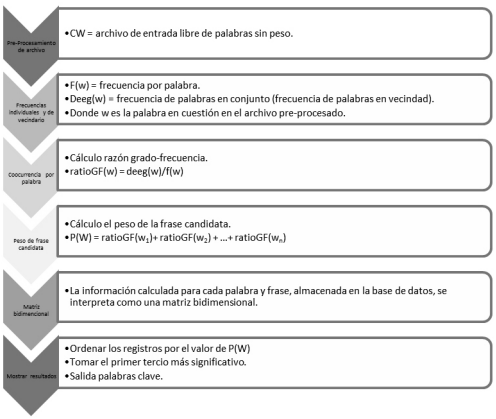

Figura 3. Método RAKE
Fuente: elaboración propia

El método RAKE fue implementado en un sistema de cómputo, tomando en consideración la secuencia de pasos descritos en la *Figura 3* y validado mediante los siguientes pasos:

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

1. Se seleccionaron varios artículos del área de ingeniería en cómputo y de bibliotecología, los cuales fueron convertidos de formato pdf a txt; posteriormente fueron analizados de manera manual con el fin de identificar la estructura del documento, identificar *stopwords*, signos de puntuación y cualquier elemento que sirviera de base para identificar las palabras clave simples y compuestas.
2. El documento en formato txt es procesado con el sistema de cómputo para obtener las palabras clave, además de realizar la comparación entre las palabras clave obtenidas y las palabras clave asignadas por el autor del artículo, con la finalidad de evaluar la precisión en la indización de documentos.

A continuación se describe la secuencia de etapas del método RAKE en el análisis del artículo de la *Revista Latina de Comunicación Social* en idioma español, con el título "Bibliotecas, centros de información y medios de comunicación en la sociedad de la información". La *Figura 4* muestra un segmento de texto del artículo para describir los pasos del método.

> La combinación de bibliotecas, centros de información y medios de comunicación como instrumento de difusión pueden ser elementos clave para reducir la brecha digital y el desnivel de conocimiento en los países con economías emergentes. La biblioteca ha sido tradicionalmente el establecimiento que sirve a la comunidad como fuente de acceso a información especializada, es también un elemento indispensable en los procesos de formación académica. Esta presta su servicio sobre la base de igualdad de acceso de todas las personas, independientemente de su edad, raza, sexo, religión, nacionalidad, idioma o condición social. Debe contar además con servicios específicos para quienes por una u otra razón no puede valerse de los servicios y materiales ordinarios. Es factible considerar el papel de los medios como instaladores de patrones de comportamiento y de formas de entender la realidad social, cultural, económica, política y de todo orden, por lo tanto se considera que la educomunicación puede ser un elemento fundamental para fomentar el uso de las bibliotecas y el desarrollo de habilidades informativas. (Asensio Baca y Cortés Montalvo, 2007)

Figura 3. Segmento de texto

El primer paso es la identificación de palabras para ser consideradas como términos, utilizando los signos de puntuación, las palabras delimitadoras y las frases delimitadoras, obteniendo como resultado la lista de frases candidatas mostradas en la *Tabla 2*.

centros de información ¬¬ medios de comunicación como instrumento de difusión ¬¬ ¬¬ ¬¬ elementos clave ¬¬ ¬¬ ¬¬ ¬¬ brecha digital ¬¬ ¬¬ ¬¬ desnivel de conocimiento ¬¬ los países con economías ¬¬ emergentes||| ||¬¬ biblioteca ¬¬ sido tradicionalmente ¬¬ establecimiento ¬¬sirve ¬¬ ¬¬ comunidad como fuente de acceso ¬¬ ¬¬ información especializada||| ||¬¬ también un elemento indispensable ¬¬ los procesos de formación académica||| ||¬¬ presta su ¬¬ servicio ¬¬ ¬¬ ¬¬ de igualdad de acceso de todas ¬¬

Tabla 2. Palabras candidatas

Una vez identificadas las palabras, se procede a crear una tabla que contiene la lista de palabras candidatas, como se muestra en la *Tabla 3*.

medios""comunicación
brecha""digital
Educación
Conocimiento
Investigación
desarrollo""habilidades""informativas
centros""documentación
educación""superior
Países
Bibliotecas""públicas

Tabla 3. Palabras candidatas

El segundo paso es el cálculo de las frecuencias F_w y del parámetro $deeg_w$ palabras vecinas con las que aparece w_i en cada frase y su frecuencia, como se muestra en la *Tabla 4*.

Palabra	deeg	F	ratioGF(w)
social	27	8	3.375
brecha	14	7	2
digital	14	6	2.3
desnivel	26	6	4.33
desarrollo	37	19	1.9473684210526
superior	21	11	1.9090909090909
habilidades	11	6	1.8333333333333
informativas	11	6	1.8333333333333

Tabla 4. Cálculo de frecuencias

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

Finalmente, se realiza el cálculo del peso de la frase candidata, definida como ratioGF(w), que es el resultado de dividir deeg/F; como ejemplo, se toma de base la palabra clave *brecha digital*, que tiene un ratioGF(digital) = 2.3 y ratioGF(brecha) = 2, dando como resultado *P(w)*=2.3+2=4.3.

Una vez realizado el cálculo del peso de la frase de todas las palabras, se obtiene una lista de 813 palabras, ordenada de mayor a menor. Según el algoritmo RAKE, se sugiere seleccionar una tercera parte del total de palabras (Rose *et al*., 2010), dando como resultado 272 términos, los cuales pueden estar compuestos por una o más palabras y, fijando un umbral de frecuencia de 4, se obtiene la lista de palabras de la *Tabla 5*.

Consecutivo	Compuesto	Veces	Valor calculado
1	Información	28	2.6326530612245
2	Bibliotecas	11	2.9230769230769
3	Comunicación	7	3.8947368421053
4	Sociedad	6	3.25
5	medios¨¨comunicación	5	8.0947368421053
6	brecha¨¨digital	5	4.3333333333333
7	Conocimiento	5	4
8	Educación	5	2.8823529411765
9	Investigación	5	1.1666666666667
10	Medios	4	4.2
11	Social	4	3.375
12	Biblioteca	4	1.8
13	Disposición	4	1.8
14	Unesco	4	1.4
15	Siglo	4	1

Tabla 5. Palabras clave calculadas

La tabla muestra las palabras que, por su frecuencia en el documento, son relevantes para describir el contenido del mismo, además de mostrar las multipalabras, tales como *medios de comunicación* y *brecha digital*. El método RAKE, al ser automatizado, reconoce cadenas de caracteres, las cuales ayudan a hacer la descripción del documento, pero por ser un método estadístico identifica términos genéricos y específicos, ya que no tiene la capacidad de reconocer conceptos. A partir de la lista, se debe determinar el número de palabras que finalmente serán consideradas como palabras clave del documento.

La evaluación de la indización manual o automatizada se lleva a cabo tomando en cuenta el criterio de consistencia en la indización o el criterio de exhaustividad y precisión en la recuperación. En este caso, se toma la consistencia para evaluar la indización automatizada realizada con el método RAKE. El proceso de evaluación se lleva a cabo de la siguiente manera: se identifican las palabras clave asignadas por el autor del artículo y las calculadas por el método RAKE. Las palabras clave asignadas por el autor del artículo son:

> *Bibliotecas públicas, información, medios de comunicación*, educomunicación, brecha digital, desnivel de conocimiento, sociedad de la información, gestión del conocimiento.

Las palabras identificadas por el algoritmo RAKE son:

> *Información, medios de comunicación, brecha digital*, educación, conocimiento, investigación, *bibliotecas públicas*, desarrollo de habilidades informativas.

Al realizar la comparación entre las palabras clave asignadas por el autor del artículo y las obtenidas por el método RAKE, podemos observar que existen palabras clave que el autor del artículo propone, pero nunca son mencionadas en el documento o sólo se mencionan una vez, como educomunicación, desnivel de conocimiento, sociedad de la información, y gestión del conocimiento. Estas palabras clave hacen referencia a temas en particular y su frecuencia es única. Por otra parte, las temáticas pueden quedar representadas por un conjunto de palabras clave que pueden pertenecer a dichas temáticas, como por ejemplo sociedad de la información, la cual contiene palabras como sociedad e información.

Al calcular la consistencia tomando como base la ecuación (1) se obtiene el resultado siguiente:

C = Consistencia entre indizadores, en este caso las palabras clave propuestas por el método automatizado y las palabras claves asignadas por el autor del artículo
T = 4; A = 8; B =8;
C = 4/((8+8)-4) = 0.33 de consistencia.

El resultado obtenido indica que existe 33 % de coincidencia entre las palabras asignadas por el autor y las identificadas por el método automatizado.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

Para tener más elementos de evaluación del método automatizado en la indización de documentos, se seleccionaron 10 artículos al azar del volumen 41, número 3 del año 2010 y del volumen 43, número 2 del año 2012 de la revista *Ciencias de la Información*, editada por el Instituto de Información Científica y Tecnológica (IDICT) en coordinación con la Sociedad Cubana de Ciencias de la Información (Socict). Los artículos se procesaron para obtener las palabras clave de cada uno de ellos por medio del método RAKE y, extrayendo las palabras clave asignadas por los autores, se calculó la consistencia como se explicó previamente y se obtuvieron los porcentajes mostrados en la *Tabla 6*.

Revista	Consistencia C=T/((A+B)-T)
1	28%
2	25%
3	25%
4	30%
5	16%
6	50%
7	28%
8	10%
9	28%
10	42%

Tabla 6. Consistencia

De acuerdo con Gil (1999: 30), "la tónica general es que la consistencia no se sitúe por debajo del 25 % ni por encima del 60 % [*sic*]"; es decir, que en el proceso de indización entre personas o entre sistemas, la identificación de palabras clave solamente concuerda entre 25 y 60 %, por lo que se puede comentar que el grado de precisión del método RAKE es adecuado en la identificación de palabras clave en idioma español. Por otra parte, el método también puede ser utilizado en la identificación de palabras, para la realización de minería de texto y alguna otra tarea en la que se requiera identificar palabras compuestas.

En este caso la consistencia es utilizada como una evaluación de una indización correcta, aunque determinar la evaluación no es fácil debido a diferentes factores que se presentan en la indización, como el número de palabras claves utilizadas en la indización, los errores que se pueden realizar en el análisis de los documentos o la falta de uso de vocabularios controlados en la indización, como es el caso del método RAKE, además de la subjetividad inherente a la indización.

Comentarios finales

El procesamiento y análisis de los documentos textuales digitales es una tarea compleja debido a la naturaleza del lenguaje natural, por lo que es necesaria la participación continua de diferentes áreas del conocimiento para su estudio; disciplinas como la lingüística, la estadística, la informática, la bibliotecología, la minería de texto, entre otras, deben de participar para desarrollar métodos y procedimientos, los cuales facilitan la organización, búsqueda, recuperación y análisis de los documentos de una manera eficiente.

Los métodos desarrollados por las diferentes áreas pueden ser de tipo lingüístico, estadístico o la combinación de ambos. El método RAKE de tipo estadístico puede ser utilizado en la identificación y extracción de palabras clave para documentos en idioma español y, por lo tanto, en el proceso de indización de documentos, teniendo como referencia la consistencia entre el proceso manual y automatizado, el cual se considera adecuado para el rango entre 25 y 60 %, de acuerdo con los cálculos realizados y mostrados en la *Tabla 6*. Con el desarrollo y aplicación de este tipo de tecnología se tienen los recursos que facilitan y agilizan el procesamiento de documentos digitales. Por lo tanto, se concluye que el método puede ser utilizado en el proceso de indización de documentos. Por otra parte, con el uso del sistema de cómputo se cuenta con la tecnología adecuada para realizar la organización de documentos de una manera eficiente en las bibliotecas.

Referencias

Asensio Baca, Gerardo y Jorge Cortés Montalvo. 2007. "Bibliotecas, centros de información y medios de comunicación en la sociedad de la información". *Revista Latina de Comunicación Socia*l 62 (Universidad de La Laguna-Laboratorio de Tecnologías de la Información y Nuevos Análisis de Comunicación Social). Fecha de consulta: 8 de abril de 2015. http://www.ull.es/publicaciones/latina/200710BacayCortes.htm

Barrón Cedeño, Luis Alberto. 2007. "Extracción automática de términos en contextos definitorios". Tesis de maestría, UNAM-Facultad de Ingeniería.

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

Gil Leiva, Isidoro. 1999. *La automatización de la indización de documentos* Gijón: Trea.

Hurwitz, Judith, Alan Nugent, Fern Halper y Marcia Kaufman. 2013. *Big data for dummies*. Hoboken: John Wiley & Sons.

Lévano, G. L. 2011. *Clasificación de colecciones*, 12 de junio. Fecha de consulta: 12 de agosto de 2013. http://www.ugel05.edu.pe/

Miner, Gary, Dursun Delen, John Elder, Andrew Fast, Thomas Hill y Robert A. Nisbet. 2012. *Practical text mining and statistical analysis for non-structured text data applications*. Waltham: Academic Press.

Parra, Sergio. 2010. "La ley de Zipf: la frecuencia con la que una palabra aparece en un texto". *Papel en blanco*, 25 de septiembre. Fecha de consulta: 1 de septiembre de 2014. http://www.papelenblanco.com/metacritica/la-ley-de-zipf-la-frecuencia-con-la-que-una-palabra-aparece-en-un-texto

Rose, S., D. Engel, N. Cramer y W. Cowley. 2010. "Automatic keyword extraction from individual documents", en *Text mining: applications and theory*, Michael W. Berry y Jacob Kogan (eds.). Hoboken: John Wiley & Sons.

Salton, Gerard. 1989. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston: Addison Wesley.

Salton, Gerard y Michael J. McGill. 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill. Computer science series XV.

Sierra, G., A. Barrón y E. Villaseñor. 2006. "C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español", conferencia presentada en el 7th Mexican International Conference on Computer Science (ENC 2006), San Luis Potosí, IEEE Computer Press.

Urbizagástegui Alvarado, Rubén y Cristina Restrepo Arango. 2011. "La ley de Zipf y el punto de transición de Goffman en la indización". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 25 (54) (México, UNAM-Instituto de Investigaciones Bibliotecológicas y de la Información): 71-92.

Zunde, Pranas. 1965. *Automatic indexing from machine readable abstracts of scientific document*. Washington D. C.: Unites States Air Force-Office of Aerospace Research.

Para citar este texto:

Contreras Barrera, Marcial. 2018. "Aplicación del algoritmo RAKE en la indización de documentos digitales". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32 (75): 109-123. http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2018.75.57951