Una Revisión de la Generación Automática de Resúmenes Extractivos

A Review of the Extractive Text Summarization

MARTHA ELIANA MENDOZA BECERRA
Ingeniera de Sistemas, Magíster en Informática,
Dra. (c) en Ingeniería de Sistemas y Computación
Profesora Titular, Departamento de Sistemas, Facultad de Ingeniería Electrónica y Telecomunicaciones
Miembro del Grupo de I+D en Tecnologías de la Información
Universidad del Cauca
mmendoza@unicauca.edu.co
Popayán, Colombia

ELIZABETH LEON GUZMÁN
Ingeniera de Sistemas, Magister en Ingeniería de Sistemas,
Dra. en Ciencias de la computación e Ingeniería Informática
Profesora Asistente, Departamento de Ingeniería de Sistemas e Industrial, Facultad de Ingeniería
Directora del Grupo de Investigación en Minería de Datos
Universidad Nacional de Colombia
eleonguz@unal.edu.co
Bogotá, Colombia

Fecha recibido: 16/10/2012
Fecha de aprobación: 14/06/2013

RESUMEN

Las investigaciones en el área de generación automática de resúmenes de textos se han intensificado en los últimos años debido a la gran cantidad de información disponible en documentos electrónicos. Este artículo presenta los métodos más relevantes de generación automática de resúmenes extractivos que se han desarrollado tanto para un solo documento como para múltiples documentos, haciendo especial énfasis en los métodos basados en reducción algebraica, en agrupamiento y en modelos evolutivos, de los cuales existe gran cantidad de investigaciones en los últimos años, dado que son métodos independientes del lenguaje y no supervisados.

PALABRAS CLAVE: Generación automática de resúmenes de textos, reducción algebraica, agrupamiento, modelos evolutivos.

ABSTRACT

Research in the area of automatic text summarization has intensified in recent years due to the large amount of information available in electronic documents. This article present the most relevant methods for automatic text extractive summarization that have been developed both for a single document and multiple documents, with special emphasis on methods based on algebraic reduction, clustering and evolutionary models, of which there is great amount of research in recent years, since they are language-independent and unsupervised methods.

KEYWORDS: automatic text summarization, algebraic reduction, clustering, evolutionary models.

1. INTRODUCCIÓN

Actualmente se encuentra gran cantidad de información en documentos de texto digitales en internet y en las organizaciones, cuando un usuario está interesado en profundizar en una cierta temática, ésta puede estar contenida en gran cantidad de textos relacionados que difícilmente pueden ser leídos en su totalidad, teniendo que invertir mucho tiempo y esfuerzo para encontrar lo que está buscando; por esto contar con un resumen, en el cual se pueda identificar las principales temáticas contenidas en los documentos disponibles, es de gran ayuda. Un ejemplo de generación automática de resúmenes es la asignación de etiquetas a los grupos que se generan en el agrupamiento de documentos web, que actualmente se realiza usando términos estadísticamente sobresalientes u oraciones frecuentes que se extraen de snippets[1] de los documentos, los cuales en general son de mala calidad [1].

La generación automática de resúmenes de textos es una tarea del procesamiento de lenguaje natural [2], que tiene por objetivo resumir el contenido de un documento conservando la información importante en un resumen de tamaño corto, y se puede realizar sobre un único documento o múltiples documentos. La generación automática de resúmenes de textos se puede definir como la creación de una "breve pero exacta representación del contenido de un documento", los resúmenes deben conservar la información importante y deben ser cortos. Diferentes autores presentan la taxonomía de la generación automática de resúmenes de documentos desde varios puntos de vista [3, 4], a saber:

- La forma como el resumen puede ser extraído: Técnica extractiva vs. Técnica Abstractiva. Los primeros extraen secuencias de palabras (frases, oraciones o párrafos) del documento original y las copian en el resumen. Mientras que en la segunda son secuencias de palabras que no necesariamente están presentes en el documento original. La primera técnica aunque presenta problemas de consistencia y coherencia, es la más utilizada por su sencillez computacional. La segunda técnica presenta un reto en computación debido a que requieren de técnicas de generación de lenguaje avanzado [2].

- El nivel de procesamiento para crear el resumen: Enfoque Superficiales vs. Enfoque Profundos. Los primeros representan los documentos con características superficiales, por ejemplo: términos estadísticamente sobresalientes, términos posicionalmente más sobresalientes, términos de oraciones claves; al final extraen los términos para incluirlos en el resumen. Los segundos necesitan técnicas avanzadas de procesamiento de lenguaje natural y análisis semántico, por ejemplo: tesauros, relaciones sintácticas, relaciones de significado; y pueden producir tanto resúmenes extractivos como abstractivos.

- El propósito del resumen: Indicativos vs Informativos vs Críticos o Evaluativos. Los Indicativos dan información abreviada de los principales tópicos de los documentos, estos resúmenes tienen como objetivo, ayudar al usuario a decidir si es importante o no leer el documento y en promedio su tamaño oscila entre el 5 al 10% del texto original. Los Informativos buscan sustituir al documento original, retener información importante y detallada con un tamaño entre un 20 y 30% del tamaño original. Los Críticos capturan el punto de vista del autor del resumen, ejemplos de estos son las revisiones y las reseñas, pero este tipo de resúmenes están fuera del alcance de la generación automática de resúmenes de texto en la actualidad.

- La audiencia a la que va dirigido el resumen: Genéricos vs Basados en Consultas vs Enfocados en el usuario o en tópicos. En los genéricos se busca dar igual importancia a los tópicos principales del documento, debido a que van dirigidos a una comunidad amplia de usuarios. Los segundos pretenden generar un resumen de acuerdo a una pregunta específica. En los últimos se busca dar mayor importancia a las necesidades específicas de un usuario o se hace énfasis en unos tópicos en particular, está enmarcado en un paradigma más ajustado a recuperación de información (RI).

Otras características tienen en cuenta el número de documentos que se procesan, un documento vs múltiples [5]; el lenguaje en el que está escrito el documento, monolenguaje vs multilenguaje; y el tipo de documentos sobre el cual se hace el resumen, científicos, noticias, blogs, entre otros.

Esta revisión de literatura está enfocada en investigaciones que generan el resumen de forma extractiva, debido a su sencillez computacional y los resultados obtenidos en aplicaciones con gran cantidad

[1] Resúmenes de 20 palabras aproximadamente.

de documentos, lo cual hace que exista gran cantidad de trabajos que se enfocan en este tipo de resumen. También se presentan las medidas estándar utilizadas recientemente en la literatura para evaluar la calidad de los resúmenes extractivos generados automáticamente por un sistema.

Este artículo está organizado de la siguiente manera: en la sección 2 se presentan los trabajos más relevantes sobre generación automática de resúmenes de un solo documento, en la sección 3 se presentan los trabajos sobre generación automática de resúmenes para múltiples documentos, en la sección 4 se presentan las medidas utilizadas para evaluar los resúmenes generados automáticamente por un sistema, y finalmente en la sección 5 se presentan las conclusiones que se derivan de esta revisión bibliográfica.

2. GENERACIÓN AUTOMÁTICA DE RESÚMENES DE TEXTOS

Teniendo en cuenta el número de documentos procesados para la generación automática de resúmenes de textos, en la Figura 1 se muestra una taxonomía con los métodos utilizados para generar este tipo de resúmenes.

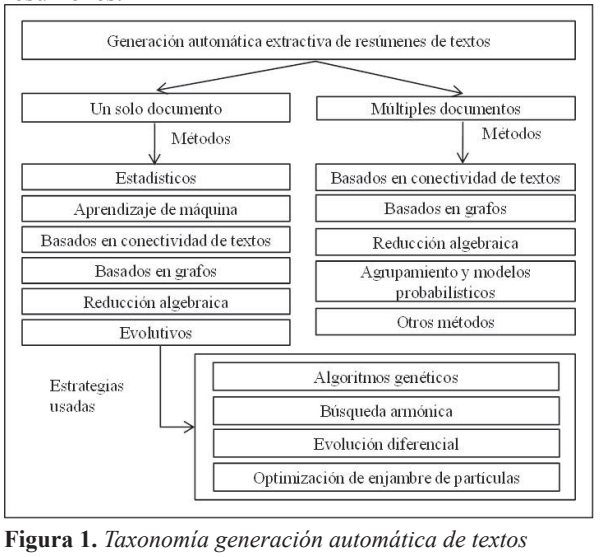

Figura 1. *Taxonomía generación automática de textos*

Basado en la Figura 1, en las siguientes secciones se realiza una revisión de las principales investigaciones en generación automática de resúmenes extractivos tanto para un solo documento como para múltiples documentos.

3. GENERACIÓN AUTOMÁTICA DE RESÚMENES DE UN SOLO DOCUMENTO

Existe gran cantidad de investigaciones que proponen métodos de generación automática de resúmenes de un solo documento, entre ellos están, los métodos estadísticos, basados en técnicas de aprendizaje de máquina, basados en conectividad de textos, basados en grafos, basados en técnicas de reducción algebraica y basados en modelos evolutivos. A continuación se presentan las investigaciones más representativas de estos métodos.

3.1 Métodos estadísticos

Entre los primeros trabajos de generación automática de resúmenes se encuentran los desarrollados por los investigadores de IBM, Luhn [6] y Baxendale [7] en 1958, el primero bajo la premisa de que un escritor repite ciertas palabras cuando está escribiendo sobre un tópico específico, se basa en características estadísticas como frecuencia de la palabra y de la oración, para extraer las oraciones más importantes del texto. Por su parte Baxendale, asume que las oraciones más importantes del documento se encuentran en las primeras y últimas oraciones de cada párrafo. Unos años después Edmundson [8] en 1969, propuso un sistema de extracción de oraciones que utilizó la frecuencia de las palabras y la posición de la oración, junto con otras dos características: presencia de palabras de referencia (como "importante" o "relevante") y presencia de palabras del título del documento.

Luego la tradicional formula de recuperación de información en un espacio vectorial, denominada *tf-idf* fue usada para la creación de resúmenes por Salton en 1988 [9], en este caso, *tf* es la frecuencia del término *i* en la oración, *idf* es la frecuencia invertida de las oraciones que contienen este término. El cálculo de la relevancia de las oraciones se realiza sumando los pesos obtenidos por cada término contenido en la oración, las oraciones con puntaje superior a un valor se seleccionan para ser parte del resumen.

Más adelante Lin y Hovy [10] en 1997 estudiaron la importancia de una única característica (la posición de la oración), basados en que cada género o dominio específico presenta regularidades en la estructura del discurso, y que las oraciones del tópico central tienden a ocurrir en algunos lugares específicos (por ejemplo, títulos, resúmenes, etc.), entonces, adaptaron el método

de posición hacia el óptimo de un género, llamándolo la política de posición óptima (OPP).

El proceso general de estos métodos estadísticos se muestra en la Figura 2: primero se realiza la extracción de las palabras clave o de las oraciones, después se realiza el cálculo del puntaje de cada oración de acuerdo a las características estadísticas seleccionadas (por ejemplo: frecuencia de palabra significativa, posición de la oración en el documento, etc.), luego se ordenan y se seleccionan las oraciones de acuerdo al puntaje obtenido por cada una de estas, y por último se obtiene el resumen con las oraciones que obtuvieron los puntajes más altos.

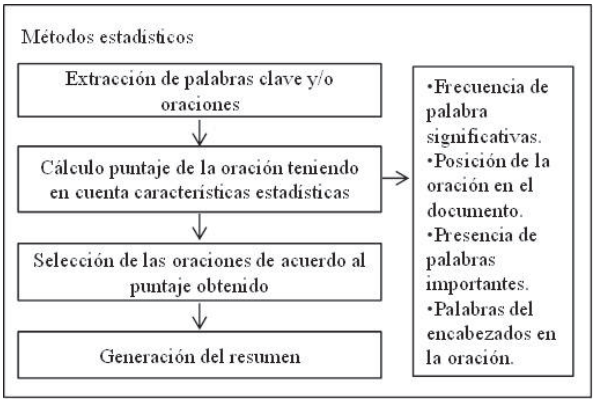

Figura 2. *Proceso general de métodos estadísticos*

3.2 Métodos basados en técnicas de aprendizaje de máquina

Entre los métodos de aprendizaje de máquina se encuentran los que usan clasificadores bayesianos para definir la probabilidad de que una oración sea incluida en un resumen, asumiendo independencia de las características y con un conjunto de datos de entrenamiento. Kupiec [11] en 1995, trató la selección de extractos como un problema de clasificación estadístico, ajustó las características propuestas por Edmundson e incluyó la longitud de la oración y la presencia de palabras en mayúsculas; a cada oración se le asignó una puntuación de acuerdo a la fórmula bayesiana y sólo las oraciones con puntajes más altos fueron extraídas. Aone et al. [12] en 1999, utilizaron el clasificador Naive Bayes, pero con más características, durante la evaluación con un conjunto de datos encontraron mejores resultados con las palabras significativas y la información posicional, pero con otra fuente de datos los resultados fueron diferentes, sugiriendo que los resultados de la generación automática de resúmenes de textos depende del tipo de documentos que se procesa. Luego Lin [13] en 1999, no asume que las características son independientes entre sí y modela el problema de extracción de las oraciones usando árboles de decisión, examinando diversas características y su efecto sobre la extracción de oraciones. Este sistema extrae las oraciones de los documentos de acuerdo a una consulta. Algunas de las nuevas características que incluyeron fueron: el query signature (número de palabras de la consulta que contiene la oración), IR signature (las m palabras más destacados en el conjunto de documentos), datos numéricos, nombre propio, pronombre o adjetivo, día de la semana o mes y cita; encontrando que la construcción del resumen es sensible al tópico del conjunto de documentos y a los términos de las consultas. Osborne [14] en el 2002, tampoco asume independencia de características y usa modelos log-lineales, adicionando una probabilidad previa (prior).

Svore et al. [15] en 2007 propusieron un algoritmo basado en redes neuronales, el modelo se entrenó a partir de etiquetas que identifican las mejores oraciones y teniendo los valores de las características para cada oración del documento, de esta forma el sistema aprende del conjunto de entrenamiento la distribución de las características de las mejores oraciones y se obtiene una lista ordenada de las oraciones de un documento. El ordenamiento se realizó con el algoritmo RankNet, un algoritmo de redes neuronales que es usado para ordenar un conjunto de entradas, en este caso, el conjunto de oraciones de un documento. Este trabajo incluye nuevas características como la similitud de una oración con el título y características que se producen de información desde registros de consultas de motores de búsqueda de noticias de Microsoft y entradas de Wikipedia.

Shen et. al [16] en 2007, proponen usar los campos aleatorios condicionales (CRF) para generación automática de resúmenes extractiva. CRF combina el modelo oculto de markov (HMM) y la regresión logística (LR); CRF es una versión discriminativa de HMM (los modelos discriminativos son buenos para problemas de clasificación). Los autores tratan el problema de generación de resúmenes como un problema de etiquetamiento de secuencias, el objetivo es producir una etiqueta de secuencia que denote si la oración se debe incluir en el resumen, maximizando toda la secuencia de oraciones (maximizando la probabilidad de la etiqueta de secuencia global y la consistencia entre las diferentes etiquetas en la secuencia).

Además Wong K-F et. al [17] en 2008, plantean la generación de resúmenes extractivos usando aprendizaje de máquina supervisado por medio de máquinas de vectores de soporte probabilísticos (PSVM), y aprendizaje semi-supervisado por medio

del co-entrenamiento de PSVM y el clasificador de naive bayes (NBC); además proponen la combinación de características (superficiales, contenido, relevancia, evento), cuyos pesos se obtienen por medio de aprendizaje supervisado.

Estos métodos de aprendizaje de máquina son supervisados, y el proceso general se muestra en la Figura 3: primero se realiza un entrenamiento con un conjunto de datos para encontrar las probabilidades o los pesos de las características. Luego se realiza el proceso de generación del resumen sobre un conjunto de datos de prueba, para lo cual: primero se extraen las palabras clave o las oraciones, después se realiza el cálculo de las características (estadísticas, contenido, relevancia, evento), por último se ordenan y se seleccionan las oraciones de acuerdo al puntaje obtenido por cada una de estas, para de esta forma obtener el resumen.

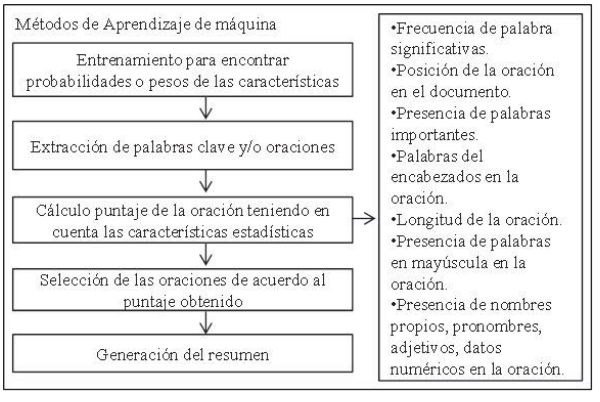

Figura 3. *Proceso general de métodos de aprendizaje de máquina*

3.3 Métodos basados en conectividad de textos

Barzilay y Elhadad [18] en 1997, proponen realizar la tarea de generación automática de resúmenes utilizando análisis lingüístico, por medio del concepto de cadena léxica, esto es, una secuencia de palabras relacionadas en el texto, que abarca distancias cortas (palabras u oraciones adyacentes) o largas (todo el texto). El método se realiza en los siguientes pasos: segmentación del texto, identificación de cadenas léxicas, y uso de las cadenas léxicas fuertes para identificar las oraciones que serán extraídas. Para encontrar las cadenas léxicas usaron WordNet, y las relaciones entre los miembros de las cadenas fueron medidas en términos de distancia WordNet. Estas cadenas fueron calificadas por su longitud y homogeneidad, seleccionando las cadenas léxicas más fuertes, luego de cada una de estas cadenas se seleccionan oraciones para crear el resumen.

Los métodos basados en conectividad de textos con cadenas léxicas, son no supervisados y dependientes del lenguaje. El proceso general se muestra en la Figura 4: primero se realiza la extracción de palabras candidatas del documento que existan en WordNet, luego se revisa si esa palabra se puede incluir en una cadena léxica existente, de lo contrario se crea una nueva cadena léxica, por último se seleccionan las cadenas léxicas más fuertes, y de estas cadenas se seleccionan las oraciones para crear el resumen.

Como parte de métodos basados en conectividad de textos que usan estructuras retóricas, se encuentra Ono et al. [19] que en 1994 presentan un modelo para extraer la estructura retórica del discurso, por medio de un árbol binario que representa las relaciones entre las oraciones. Los pasos de extracción de la estructura retórica son: análisis de la oración, extracción de las relaciones retóricas, segmentación de expresiones retóricas, generación de todas las posibles estructuras retóricas y selección de la estructura candidata con el puntaje de penalidad más bajo. Luego el sistema calcula la importancia de cada oración basado en la importancia relativa de las relaciones retóricas, y de forma iterativa se podan nodos del árbol de acuerdo a las penalidades. La lista de los nodos de la estructura final produce el resumen. Los resultados para los artículos técnicos fueron buenos, debido a que estos contienen expresiones retóricas y claves lingüísticas que permiten que el sistema pueda extraer la estructura retórica.

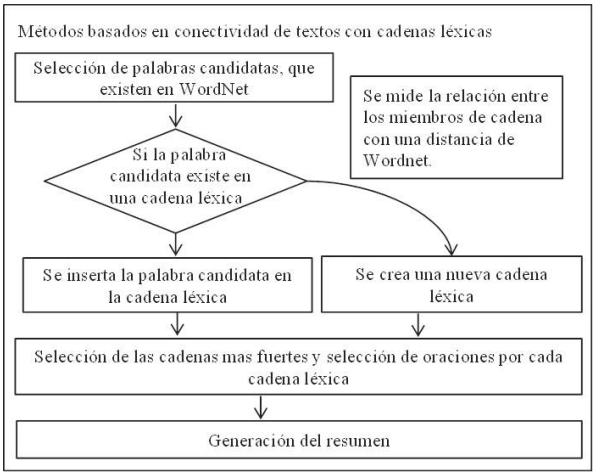

Figura 4. *Proceso general de métodos basados en conectividad de textos con cadenas léxicas*

Luego Marcu en 1998 [20, 21] utiliza la Teoría de la Estructura Retórica (RST) que se da entre dos piezas de texto que no se superponen: el núcleo y el satélite; el núcleo de una relación retórica se puede comprender independiente del satélite, pero no a la viceversa.

Define para un texto varias estructuras retóricas, a las cuales se les calcula un puntaje de acuerdo a una fórmula establecida (una combinación lineal de pesos de ciertas métricas o la aplicación de ciertas funciones de similitud sobre el árbol con la estructura retórica), la estructura retórica con el mayor puntaje es seleccionada para generar el resumen. Uno de sus trabajos se enfocó en encontrar el valor de los pesos de siete métricas para calcular el puntaje de cada estructura del discurso.

Estos métodos basados en conectividad de textos con manejo de estructuras retóricas, al igual que con cadenas léxicas también son no supervisados y dependientes del lenguaje. El proceso general se muestra en la Figura 5: primero se realiza la extracción de segmentos retóricos del documento original, luego por medio de un algoritmo se generan las diferentes estructuras retóricas, se procede a establecer el puntaje de cada estructura retórica de acuerdo a las métricas que defina el algoritmo, y se selecciona la estructura que obtenga el mayor puntaje, después se ordenan los segmentos retóricos de la estructura retórica seleccionada para generar el resumen del documento de acuerdo con este orden.

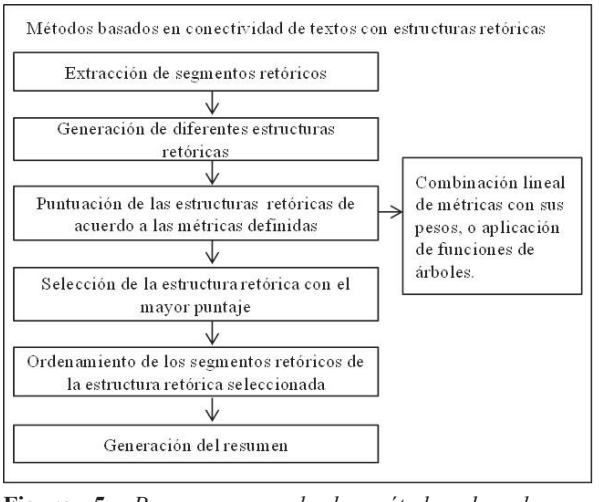

Figura 5. *Proceso general de métodos basados en conectividad de textos con estructuras retóricas*

3.4 Métodos basados en grafos

En los algoritmos de ordenamiento de grafos, la importancia de un vértice dentro del grafo es iterativamente calculada desde el grafo completo. Mihalcea [22] en el 2004 utiliza el modelo basado en grafos, por medio de un algoritmo llamado TextRank, para la extracción de palabras clave y la extracción de oraciones. En el primer caso, un vértice es una unidad (secuencia de una o más unidades léxicas extraídas del texto), y los bordes definen las relaciones entre las unidades; un borde es adicionado entre dos unidades léxicas que co-ocurren dentro de una ventana de máximo N palabras. Luego se ejecuta iterativamente el algoritmo de ordenamiento hasta que converja. Por último se ordena el grafo de acuerdo a los puntajes de forma descendente y se seleccionan los vértices con el puntaje más alto. Para la segunda tarea de extracción: un vértice es una oración del texto y los bordes definen las conexiones entre las oraciones. Estas conexiones son definidas usando una relación de similitud (el solapamiento de dos oraciones puede ser determinado como el número de palabras comunes entre las representaciones léxicas de dos oraciones). Luego el algoritmo es aplicado sobre las oraciones del grafo y se ordena de acuerdo a los puntajes de los vértices, de igual forma las oraciones que se ubican en los vértices de la parte de arriba del grafo se incluyen en el resumen.

Estos métodos basados en grafos son no supervisados e independientes del lenguaje. El proceso general se muestra en la Figura 6: primero se realiza la extracción de palabras claves o de las oraciones, luego se crea un grafo, en el cual los vértices son las oraciones y los bordes la relación entre las oraciones (que se mide por medio de una función de solapamiento del contenido de ambos vértices), se itera el grafo hasta que converja y por último se ordena el grafo de acuerdo al puntaje obtenido en cada vértice, para de esta forma obtener el resumen.

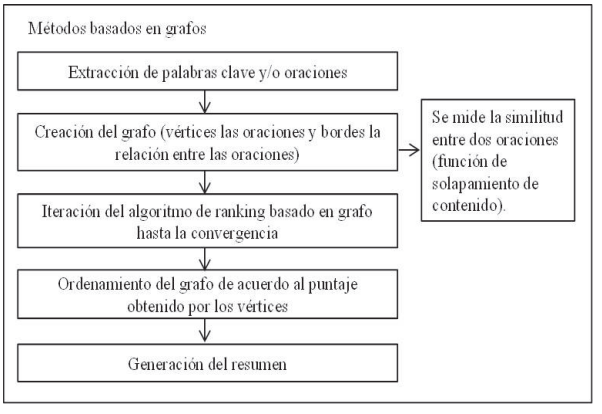

Figura 6. *Proceso general de métodos basados en grafos*

3.5 Métodos basados en reducción algebraica

El análisis Semántico Latente (LSA por sus siglas en inglés, Latent Semantic Analysis) es una técnica matemática para extraer e inferir relaciones contextuales entre palabras escritas en un discurso.

Gong y Liu [23] en 2002 propusieron usar LSA para la generación automática de resúmenes genéricos, aplicando la descomposición de valores singulares (SVD por sus siglas en inglés, Singular Value Decomposition). El proceso de análisis semántico está compuesto por dos pasos. El primero es la creación de una matriz de términos por oración $A = [A_1, A_2, \ldots, A_n]$, donde cada columna A_i representa un vector con el peso de la frecuencia de términos de la oración *i* en el documento. Si hay *m* términos y *n* oraciones en el documento, entonces la matriz A para el documento será de $m \times n$ (donde $m >= n$). La matriz A es dispersa porque cada término no aparece normalmente en cada oración. El siguiente paso consiste en aplicar SVD a la matriz A, así, $A = U\Sigma V^T$, donde, $U = [u_{ij}]$ es una matriz de columnas ortonormales de $m \times n$ cuyas columnas son llamadas vectores singulares de izquierda, $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ es una matriz diagonal de $n \times n$, cuyos elementos diagonales son valores singulares no negativos en orden descendente ($\sigma_1 >= \sigma_2 .. >= \sigma_r > \sigma_{r+1} = .. = \sigma_n = 0$) y $V = [v_{ij}]$ es una matriz ortonormal de $n \times n$, cuyas columnas se denominan vectores singulares derechos. La dimensionalidad de las matrices es reducida a las *r* dimensiones más importantes y, por tanto, U' es $m \times r$, Σ' es $r \times r$ y V^T es una matriz de $r \times n$.

Desde el punto de vista matemático, SVD produce un mapeo entre el espacio *m*-dimensional especificado por los vectores de frecuencia de términos y el espacio vectorial singular *r*-dimensional. Desde la perspectiva del procesamiento de lenguaje natural (NLP por sus siglas en inglés, Natural Language Processing), SVD deriva la estructura semántica latente del documento representado por la matriz A, es decir, un desglose del documento original en *r* vectores base linealmente independientes los cuales expresan los principales "tópicos" del documento. SVD capta las interrelaciones entre los términos, y permite que los términos y las oraciones puedan ser agrupados sobre la base "semántica". Si un patrón de combinación de palabra es destacado y recurrente en un documento, este patrón es capturado y representado por uno de los vectores singulares, la magnitud de este vector indica el grado de importancia de este patrón dentro del documento. Las oraciones que contengan este patrón de combinación de palabras serán proyectadas en este vector singular, y la oración que mejor represente este patrón tendrá el valor del índice más grande dentro del vector. Partiendo de que cada patrón de combinación de palabras describe un tópico en el documento, cada vector singular representa cada tópico y la magnitud de su valor singular representa el grado de importancia de este tópico. Para el resumen, este método selecciona las oraciones cuya representación vectorial tengan la longitud más grande, escogiendo la oración con el ponderado más grande a través de todos los tópicos.

El método de generación automática de resúmenes propuesto en [23] usa la representación de un documento para escoger las oraciones que van en el resumen basándose en la importancia relativa de los "tópicos" descritos por la matriz V^T. El algoritmo de generación automática de resúmenes escoge para cada "tópico" la oración más importante para ese tópico: por ejemplo, la oración K_{TH} elegida es la que tiene el valor del índice más grande en el K_{TH} vector singular derecho en la matriz V^T. El inconveniente principal de este método es que cuando las *l* oraciones son extraídas de los *l* tópicos más altos, se tratan como igualmente importantes, por lo tanto, un resumen puede incluir oraciones acerca de "Tópicos" que no son particularmente importantes. Por esto Steinberger y Ježek en el 2004 [24], cambiaron el criterio selección para incluir en el resumen las oraciones cuya representación vectorial en la matriz $\Sigma^2 \bullet V$ tengan la "longitud" más grande, en lugar de las oraciones que contiene el mayor valor del índice para cada "Tópico". Esto se basa en la idea de que elegir las oraciones con el mayor peso combinado a través de todos los tópicos importantes, posiblemente incluye más de una oración relacionada con un tópico importante, en lugar de una oración para cada tópico. Más formalmente: después de computar la SVD de la matriz de término por oraciones, se calcula la longitud de cada vector oración en $\Sigma^2 \bullet V^T$, el cual representa su puntaje.

Otro método de generación automática de resúmenes que usa LSA fue propuesto por Yeh et al.[25] en el 2005, después de realizar SVD sobre la matriz de términos por oración y reducir la dimensionalidad del espacio latente, reconstruyen la correspondiente matriz $A' = U'\Sigma'V'^T$. Cada columna de A' denota la representación semántica de la oración, que es usada en lugar del vector de frecuencia basado en palabras clave, para la creación de un grafo de relaciones del texto que representa la estructura de un documento, luego un algoritmo de ordenamiento es aplicado al grafo resultante.

Luego Lee et. al en el 2009 [26], proponen usar la factorización de matriz-no-negativa (NMF por sus siglas en inglés, Non-negative matrix factorization) para seleccionar las oraciones a incluir en el resumen. NMF representa objetos individuales como una combinación lineal no negativa de información extraída desde un volumen grande de objetos. NMF puede trabajar con un volumen de información grande de manera eficiente, debido a que la matriz original no

negativa se descompone en dos matrices no negativas dispersas y distribuidas. NMF descompone una matriz A no negativa de $m \times n$, en la multiplicación de una matriz de característica no negativa (NSFM) de $m \times r$ (W) y una matriz de semántica variable no negativa (NSVM) de $r \times n$ (H), donde r es más pequeño que m o n, de forma que los tamaños de W y H son más pequeñas que la matriz A. Se usa una función objetivo para satisfacer la aproximación de A=WH, luego W y H son actualizados de forma iterativa hasta que convergen al umbral o exceden el número de repeticiones definidos. Inicialmente se parte de una matriz A con los ponderados de la frecuencia del término en la oración, se aplican NMF para obtener la matriz W y H; y se define la relevancia genérica de la oración (GRS) que hace referencia a cuánto la oración refleja los tópicos principales (esto es representado por las características semánticas) y por último selecciona las k oraciones con los valores más altos de relevancia genérica. NMF logro mejores resultados que LSA, seleccionando oraciones más significativas que los métodos relacionados con LSA, además NMF puede usar características semánticas que son más fáciles de interpretar y entender la estructura innata de los documentos.

Los métodos basados en reducción algebraica son no supervisados e independientes del lenguaje. El proceso general se muestra en la Figura 7: primero se realiza la extracción de palabras claves o de las oraciones, luego se crea la matriz de términos por oración y se realiza la descomposición matricial, por último se seleccionan las oraciones con los valores más altos, para de esta forma obtener el resumen.

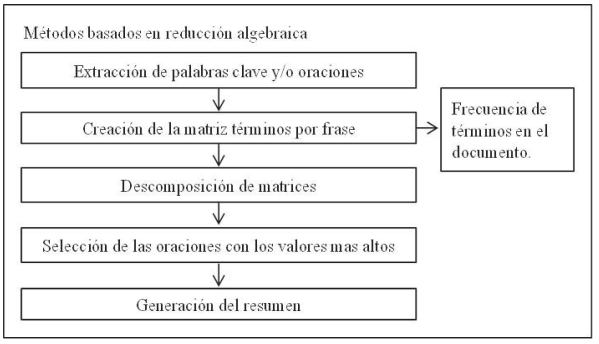

Figura 7. *Proceso general de métodos basados en reducción algebraica*

3.6 Métodos basados en modelos evolutivos

La generación automática de resúmenes se puede considerar como un problema de optimización global o de búsqueda, por lo tanto existen investigaciones que están abordando este problema basados en algoritmos evolutivos, algunas para optimizar los pesos de las características de las fórmulas que permiten hacer el cálculo del puntaje de las oraciones y otras investigaciones para generar los resúmenes. A continuación se presentan las más relevantes:

3.6.1 Métodos evolutivos para la optimización de pesos de las características

Kiani y Akbarzadeh [27] en el 2006, proponen la generación de resúmenes usando una combinación de algoritmos genéticos (GA por sus siglas en inglés, Genetic algorithms) y programación genética (GP por sus siglas en inglés, Genetic Programming) para un sistema difuso. Los GA fueron usados para optimizar las funciones de membresía de un sistema difuso y la GP para optimizar el conjunto de reglas de este sistema. Esta optimización es una entrada para el sistema difuso y la decisión de acuerdo a las reglas IF-THEN; estas reglas tienen en cuenta el valor de seis características estadísticas medidas para cada oración del documento.

Luego Fattah y Ren en el 2009 [28], proponen varios modelos para generación de resúmenes: basados en GA, Modelo de regresión matemática (MR por sus siglas en inglés, mathematical regression), Red neuronal hacia adelante (FFNN por sus siglas en inglés, feed forward neural network), Red neuronal probabilística (PNN por sus siglas en inglés, probabilistic neural network) y Modelo de mezcla gaussiano (GMM por sus siglas en inglés, gaussian mixture model). El modelo tiene dos modos de operación: modo de entrenamiento, en el cual los cinco modelos se entrenan para obtener los pesos de las diez características utilizadas; y modo de prueba, en el cual se calculan las características (estadísticas y de similitud) para cada oración teniendo en cuenta los pesos calculados en el modo de entrenamiento. Las oraciones con el más alto puntaje son seleccionadas para hacer parte del resumen. En el AG un cromosoma es representado como la combinación de todos los pesos de las diez características, para cada generación se produjeron miles de soluciones, se evaluó la función objetivo de cada solución y se conservaron los diez más aptos para cruzarse con los nuevos en la siguiente generación, se evaluaron cien generaciones para obtener los pesos. En MR, el vector de salida es igual a la matriz de entrada (cuyas filas son las oraciones y las columnas las características) multiplicado por el modelo estadístico lineal del sistema (pesos de las características). Una FFNN permite clasificar una oración como parte o no del resumen basado en sus características, usa diez neuronas de entrada (las diez características), 20 neuronas ocultas y una neurona de salida (las oraciones que tengan un valor de salida alto se consideran parte

del resumen y las que tengan un puntaje bajo no). PNN es una herramienta eficiente para clasificar datos de alta dimensionalidad, los pesos y funciones son respaldados por la probabilidad bayesiana. Cada neurona recibe un vector X (los valores de las diez características de una oración) y la salida representa la probabilidad de que el vector X pertenezca a la clase. En GMM, la función de probabilidad de densidad para una cierta clase (categoría) de vector de características X es la suma o mezcla de ponderados de las *k* distribuciones gausianas de clase condicional, dado un conjunto de vectores de entrenamiento de una cierta clase, un conjunto inicial de medias es estimado usando agrupamiento k-*means*, luego la mezcla de pesos, medias y covarianzas son entrenadas iterativamente usando el algoritmo de maximización de las expectativas (EM); de esta forma, se crea un modelo de clase dependiente para cada categoría, luego un vector de características X puede ser clasificado en una de las dos categorías (resumen o no resumen).

En [29], proponen un modelo de generación de resúmenes basado en programación genética (GP), que utiliza cadenas de caracteres de longitud fija (llamados cromosomas) los cuales se expresan como árboles de expresión (ET). GP inicia con una población de soluciones candidatas aleatoria en forma de cromosomas, los cuales se mapean a ET, luego se evalúan y seleccionan basados en la función de aptitud para reproducirse con operaciones de modificación genética. La función de aptitud está conformada por ocho características y un peso para una de ellas. Para cada generación se producen miles de soluciones, se evalúa la aptitud de cada solución y se conservan los ocho más aptas para cruzarse en la siguiente generación, se evalúan miles de generaciones hasta obtener la combinación adecuada de pesos de las características. Luego se aplica el GP con esta combinación de pesos y todas las oraciones del documento son organizadas en forma descendente de acuerdo al puntaje obtenido y las oraciones con puntajes más altos son seleccionadas como el resumen basado en la tasa de compresión (tamaño del resumen).

En [30] en el 2009, proponen un modelo para generación de resúmenes que integra la lógica difusa y la inteligencia de enjambre. Con el modelo de enjambre [31] se obtienen los pesos que son usados para ajustar los puntajes de las características (usaron cinco), luego estos puntajes son la entrada del sistema de inferencia difuso para producir el puntaje final de la oración, estas se ordenan de acuerdo al puntaje y las primeras *n* oraciones son seleccionadas para el resumen.

En [32] en el 2010 proponen un modelo de optimización evolutivo difuso llamado FEOM para hacer agrupamiento de documentos y generación de resúmenes. El método de generación automática de resúmenes se basa en el concepto de agrupamiento de oraciones del documento, luego se seleccionan las oraciones más importantes de cada grupo para obtener el resumen. FEOM utiliza algoritmos genéticos, genera una población aleatoria como el grupo inicial de soluciones para agrupamiento, cada individuo de la población es una cadena de números reales, los tres operadores evolutivos (selección, cruce y mutación) son empleados para producir nuevos hijos hasta el criterio de terminación. Se aplican tres parámetros de control (coeficiente de distribución, distancia relativa, efecto de evolución) para regular la probabilidad de cruce y de mutación de cada solución.

Nuevamente en [33] en el 2010, proponen un modelo diversidad híbrido difuso-enjambre, que combina tres métodos basados en: diversidad, enjambre y difuso-enjambre. El método basado en diversidad, hace grupos de oraciones que son ordenadas en un árbol binario de acuerdo a los puntajes, luego aplica Maximal Marginal Importance (MMI) para seleccionar las oraciones a incluir en el resumen. El método basado en particle swarm optimization (PSO) binario es usado para optimizar el peso correspondiente a cada característica de la función objetivo; la posición de la partícula es una cadena de bits, donde uno significa que la característica correspondiente es seleccionada y un cero lo contrario; después de tener los pesos se calcula el puntaje para cada oración y se escogen las oraciones con el mayor puntaje para ser incluidas en el resumen. En el método basado en enjambres y lógica difusa, el algoritmo difuso calcula el puntaje de la oración a través de un sistema de inferencia, que parte de los pesos encontrados con PSO, luego convierte el resultado del proceso de inferencia (puntajes finales de las oraciones), y al final se ordenan las oraciones de acuerdo al puntaje obtenido y se obtiene el resumen. Al final utiliza otro procedimiento para seleccionar las oraciones de los resúmenes obtenidos por cada uno de los tres métodos anteriores.

En [34] en 2010 proponen un enfoque independiente del lenguaje para generación automática de resúmenes extractivos basado en la optimización lineal de diferentes medidas usando un algoritmo genético (AG) denominado MUSE. El AG es usado para optimizar la combinación de pesos lineales de treinta y una características que son usadas para puntuar las oraciones. En el entrenamiento, la función de aptitud mide la calidad del vector representado por medio de

la medida de recuerdo de ROUGE-1. En la etapa de inicialización, la población inicial se genera con una distribución normal estándar y normalizada, si un valor es negativo se considera como una "penalidad". En la selección una proporción de la población es seleccionada para generar nuevos hijos de acuerdo a los más aptos. En la reproducción los nuevos genes se introducen en la población, y las nuevas soluciones se generan desde los genes seleccionados por medio de los operadores de apareamiento, un par de soluciones padres se seleccionan aleatoriamente y la nueva solución es creada usando cruce y mutación. Por último el puntaje de cada oración es calculado con los valores de las características y los pesos encontrados con el AG.

3.6.2 Métodos evolutivos para la generación automática de resúmenes

En el 2008 [35] proponen un método de generación automática de textos basado en un algoritmo genético para la extracción de oraciones. Definen un buen resumen como la colección de oraciones legibles que están relacionadas y discuten acerca del título del documento. Por lo tanto contemplan tres factores a tener en cuenta en la función objetivo: (i) Factor relación con el tópico: Medida de similitud de las oraciones del resumen hacia el título del documento. (ii) Factor Cohesión: Similitud entre las oraciones del resumen. (iii) Factor de Legibilidad: Similitud de una oración del resumen con la siguiente. Se construyen algunos vectores aleatorios, esta población se usa para generar los hijos, un hijo lo generan dos padres. Después todos los padres y los hijos se colocan juntos, y se usa la función de aptitud para escoger la mitad de la nueva población, que serán los padres de la nueva generación. El cruce se hace con dos padres para crear dos hijos, la mutación que usan es de un bit.

En el 2008 [36] proponen un método de generación de resúmenes basado en un algoritmo de búsqueda armónica. Este trabajo se basan en los mismos tres factores utilizados en [35], se define un buen resumen como la colección de oraciones legibles que están relacionadas y discuten acerca del título del documento y por lo tanto contemplan los mismos tres factores a tener en cuenta en la función objetivo. El vector solución o armonía es de longitud *n* (cantidad de oraciones en el documento), una entrada en el vector de cero significa que esa oración no pertenece al resumen y un uno que si pertenece. Se ejecuta el algoritmo de búsqueda armónica hasta que se cumpla el criterio de terminación, entonces el armónico que se encuentra en la memoria armónica con el mejor valor de la función objetivo se selecciona y se obtiene el resumen.

Luego en [37] en el 2009 proponen un modelo de generación de resúmenes para un documento basado en un algoritmo de evolución diferencial discreto, cuya función objetivo utiliza agrupamiento de oraciones y como medida de similitud entre oraciones la medida de google.

Estos métodos basados en modelos evolutivos son no supervisados e independientes del lenguaje. El proceso general es como se muestra en la Figura 8: primero se realiza la extracción de palabras claves o de las oraciones, luego se define la función objetivo (utilizando características estadísticas y de similitud) y se ejecuta el algoritmo evolutivo, por último se seleccionan las oraciones del vector solución con el mejor valor de aptitud, para de esta forma obtener el resumen.

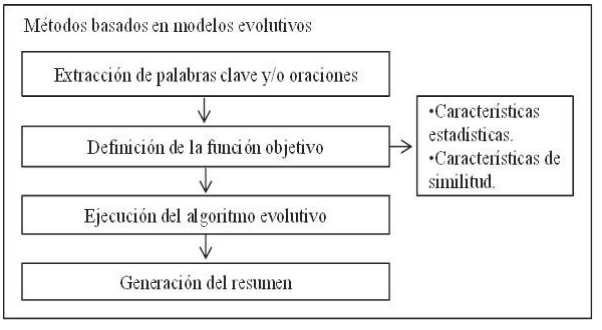

Figura 8. *Proceso general de métodos basados en modelos evolutivos*

4. GENERACIÓN AUTOMÁTICA DE RESÚMENES DE MÚLTIPLES DOCUMENTOS

También existe gran cantidad de investigaciones que proponen métodos de generación automática de resúmenes de múltiples documentos, entre ellos, métodos basados en conectividad de textos, basados en grafos, basados en técnicas de reducción algebraica, basados en agrupamiento, modelos probabilísticos y otros métodos.

4.1 Métodos basados en conectividad de textos

En el 2005 [38] proponen la generación de resúmenes de múltiples textos basado en cadenas léxicas, aplicando el método de generación de resúmenes de un solo documento [18] revisado anteriormente para cada documento, generando de esta forma un resumen de

resúmenes. Para el manejo de redundancia de oraciones en los documentos, toma la oración con puntaje más alto, luego la siguiente y se calcula la similitud entre estas dos oraciones, si esta similitud es menor a un umbral se deja en el resumen; y se repite nuevamente el procedimiento. En la Figura 9, se muestra el proceso que se lleva a cabo para este método (no supervisado y dependiente del lenguaje), se puede observar que el proceso es igual al mostrado en la Figura 4 para un solo documento, pero agregando un proceso de eliminación de redundancia, debido a que se pueden presentar oraciones similares en varios documentos.

4.2 Métodos basados en grafos

En el 2005 [39] , propone un método que utiliza los mismos principios del ordenamiento basado en grafos que fueron aplicados en generación automática de resúmenes de un solo documento [22]. Primero se realiza el resumen de cada documento y luego se resumen los resúmenes de los documentos individuales utilizando el mismo método de grafos. Con respecto a la similitud entre las oraciones, manejan un umbral máximo de similitud entre oraciones.

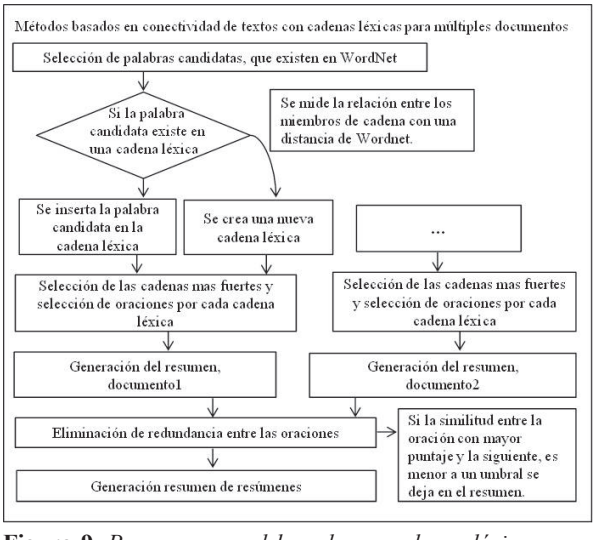

Figura 9. *Proceso general basado en cadenas léxicas para múltiples documentos*

También en el 2010 [40] propone un algoritmo para uno y múltiples documentos basado en grafos usando el concepto de prominencia local para indicar la importancia de una oración en un documento en particular y prominencia global para indicar la importancia de una oración en el conjunto de documentos; además tienen en cuenta la posición de la oración en el documento y en el conjunto de documentos. En este algoritmo primero se construyen los grafos de afinidad (cada grafo es representado por una matriz de afinidad) para reflejar las diferentes clases de relaciones entre las oraciones, luego de forma iterativa se calculan los puntajes de prominencia local y global de las oraciones basado en los grafos, cuando el algoritmo converge se obtienen los puntajes de la oraciones, por último las oraciones con puntaje local alto de un documento se escogen como resumen de éste, y las oraciones con puntaje global alto se escogen como resumen del conjunto de documentos.

En la Figura 10, se muestra el proceso que se lleva a cabo para este método (no supervisados e independiente del lenguaje), se puede observar que este proceso también es igual al mostrado anteriormente en la Figura 6 para un solo documento, pero agregando un proceso de eliminación de redundancia, debido a que los documentos tratan la misma temática y se pueden presentar oraciones similares en el resumen.

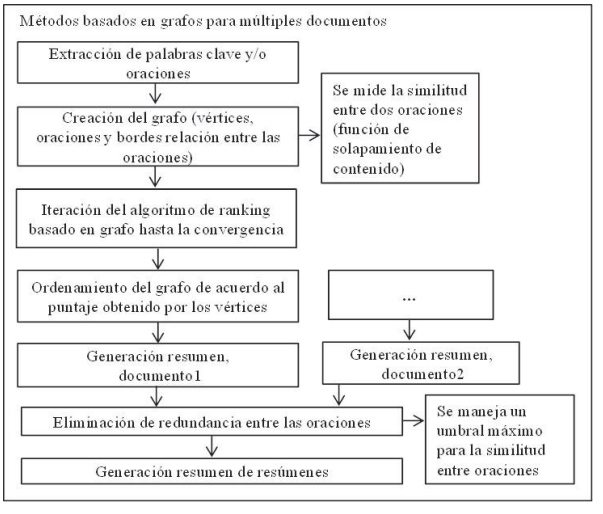

Figura 10. *Proceso general basado en grafos para múltiples documentos*

4.3 Métodos basados en reducción algebráica

En 2005 [41], proponen un enfoque de generación automática de resúmenes de múltiples documentos orientado a consulta basado en LSA y en relevancia máxima marginal (MMR), llamado EMBRA. Antes de aplicar MMR, se aplica LSA y se tienen en cuenta las siguientes restricciones: Tiempo (preservar el orden temporal de los eventos), Secuencia (preservar el orden original de las oraciones), Grupos (grupos con oraciones similares), Contexto (recrear el contexto precedente original). El algoritmo es determinístico y optimiza localmente de la bolsa de oraciones extraídas, determina la oración con mayor puntaje y la mueve al final del resumen destino, y repite el proceso hasta que todas las oraciones de la bolsa son insertadas.

En el 2007 [42] proponen una extensión de LSA, primero se crea una matriz de términos por oración que incluye todas las oraciones del conjunto de documentos, el puntaje se calcula de la misma forma que se hace para un único documento, y se seleccionan las oraciones con mayor puntuación para el resumen. Para evitar la redundancia, antes de incluir una oración en el resumen, se revisa si ya existe una oración similar, que debe estar cerca a la consulta del usuario. Este método favorece las oraciones largas, porque estas probablemente contendrán más términos importantes que una corta, por esto, se divide la puntuación de la oración por el *número de términos*^{lk}, donde *lk* es el coeficiente de longitud.

En la Figura 11, se muestra el proceso que se lleva a cabo para estos métodos (no supervisados e independientes del lenguaje), se puede observar que este proceso también es igual al mostrado anteriormente para un solo documento, agregando un proceso de eliminación de redundancia, debido a que se pueden presentar oraciones similares en varios documentos.

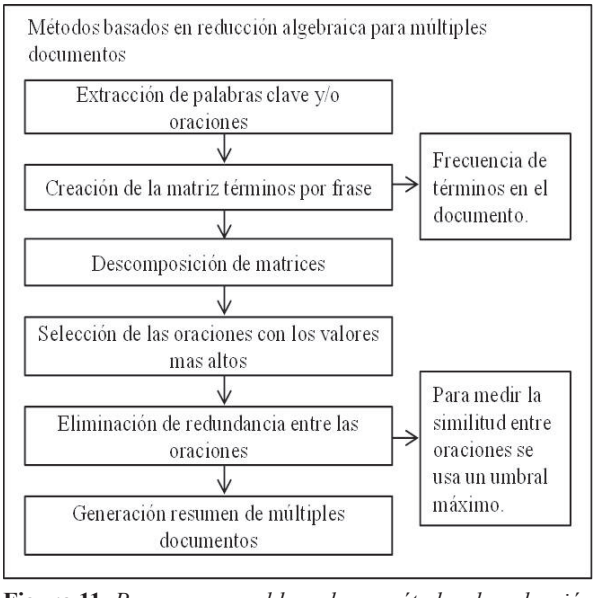

Figura 11. *Proceso general basado en métodos de reducción algebraica para múltiples documentos*

4.4 Métodos basados en agrupamiento y modelos probabilísticos

En el 2004 [43], proponen un generador de resúmenes llamado MEAD, que usa los centroides de grupo producidos por un sistema de detección y seguimiento de tópicos. Antes de generar el resumen, identifican los artículos sobre un evento, este proceso es llamado Detección y seguimiento de tópico (TDT). MEAD utiliza MMR para eliminar la redundancia en el resumen y recibe como entrada *n* oraciones de un grupo de documentos y la tasa de compresión *r*; y como salida *n*r* oraciones del grupo con los puntajes más altos. El puntaje de cada oración tiene en cuenta características como: valor del centroide, valor posicional y solapamiento con la primera oración. A este valor se le resta una penalidad por redundancia. Cada documento de cada grupo se califica y se ordena la oración de cada documento de acuerdo al puntaje.

En el 2008 [44] proponen un modelo de lenguaje, factorización con bases dadas (FGB por sus siglas en inglés factorization with given bases), que simultáneamente hace agrupamiento y generación de resúmenes, por medio de la matriz de términos por documento y la matriz de términos por oración. Los autores buscan minimizar la divergencia entre los documentos dados y el modelo de reconstrucción de términos, este proceso de minimización obtiene dos matrices que representan las probabilidades de los documentos y de las oraciones dado los grupos, lo que permite asignar cada documento al tópico con la más alta probabilidad y el resumen se forma por las oraciones con la más alta probabilidad en el tópico. Al inicio del proceso, después de tener las dos matrices, el sistema ejecuta factorización no negativa sobre la matriz de términos por documento usando como base la matriz de términos por oración, dando lugar a la matriz de documento por tópico y oración por tópico.

En el 2009 [45] proponen dos técnicas para generación automática de resúmenes de documentos, la primera técnica consiste en adicionar la características de similitud con la primer oración en el método MEAD[2] (CPSL), y la segunda técnica es combinar CPSL con LEAD (se escoge la primera y la última oración del párrafo, este método es bueno para los artículos de noticias). La técnica CPSL se basa en MEAD, con la variación que calculan la similitud entre cada oración y la primera oración (por medio de la ley de cosenos). La técnica LESM, aplica por aparte los métodos LEAD y CPSL para obtener el resumen, luego se extraen solo las oraciones comunes a estos dos resúmenes para obtener el resumen, pero para completar el porcentaje de las oraciones del resumen, aplican nuevamente LEAD sobre las oraciones que no concuerdan con el resumen, las ordena y toma la primera oración que no concuerda del primer documento, luego chequea si ya se cumple con el porcentaje deseado como criterio de terminación del algoritmo, si no se logra el porcentaje, se toma la primera oración que no concuerda del

[2] Una versión anterior de MEAD, la versión más nueva se explicó como primer referencia de esta sección.

segundo documento y así sucesivamente, cuando ya se han seleccionado las primeras oraciones de todos los documentos, se sigue el proceso pero con la última oración de cada documento.

En el 2009 [46] proponen un algoritmo de agrupamiento orientado a consulta, para ello se trata la consulta como si perteneciera al conjunto de documentos, los grupos se mezclan en un solo grupo y utilizan la Relevancia marginal máxima (MMR) modificada para extraer las oraciones del resumen. El algoritmo de agrupamiento realiza las siguientes tareas: construir la matriz de oraciones (se incluye la oración de la consulta) por términos, construir la matriz de similitud entre las oraciones (ley de cosenos), se generan los centroides iniciales de grupo (aleatoriamente una oración es el primer centroide y luego la oración menos similar a ésta es el segundo, los otros se escogen de forma que sean los más lejanos a los actuales), para determinar si una oración debe estar en un centroide del grupo se hace por medio de la similitud entre la oración y el centroide (en cada paso iterativo el valor de similitud de las oraciones del grupo es re-calculada), si el valor es mayor o igual al umbral la oración se coloca en el centroide, si el valor es menor al umbral entonces se remueve del centroide del grupo.

En el 2009 [47] propone un método de generación de resúmenes orientado a consulta basado en PLSA, el cual permite representar las oraciones y la consultas como distribuciones de probabilidad sobre tópicos latentes. PLSA permite modelar los documentos como una mezcla de tópicos. El resumen se produce en tres pasos: (1) Crear la matriz de términos por oración y entrenar el modelo PLSA sobre esta matriz; (2) Calcular las diferentes características a nivel de oración basado en la similitud de las distribuciones de las oraciones y de la consulta sobre los tópicos latentes; (3) Calcular el puntaje de la oración como la combinación lineal de los puntajes de las características y ordenar las oraciones de acuerdo al puntaje, luego utilizar MMR para seleccionar las oraciones y penalizar las oraciones candidatas basado en su similitud con el resumen parcial.

En el 2009 [48] proponen un modelo de tópicos bayesiano basado en las oraciones (BSTM por sus siglas en inglés, Bayesian Sentence-based Topic Models), usando la matriz de términos por documento y la matriz de términos por oración. Este es un modelo probabilístico generativo, el algoritmo recibe como entradas la matriz de términos por documento, términos por oración y el número de tópicos latentes; y como salida se obtiene una matriz de oraciones por tópico y una matriz auxiliar de documentos por tópico. El artículo presenta las distribuciones de probabilidad para seleccionar las oraciones dados los tópicos, además usa un algoritmo bayesiano variacional para estimar los parámetros del modelo. El BSTM se diferencia del FGB (de los mismos autores [44]), en que la ubicación de documento-tópico es marginalizada (en el algoritmo esto se refleja en el ajuste de Dirichlet), esto incrementa la estabilidad en la estimación de los parámetros de oración-tópico.

En el 2011 [49] proponen un método que simultáneamente agrupa y ordena las oraciones, a diferencia de otros métodos que hacen estas dos tareas pero de forma secuencial. Este método utiliza una estructura que representa el grupo de oraciones, que permite identificar las oraciones que comparten el mismo contenido como aquellas que son proyectadas sobre esta estructura y como las oraciones más importantes dentro del grupo aquellas que poseen longitudes de proyección más grandes. Para generar el resumen, extraen las oraciones más sobresalientes desde cada grupo hasta completar el tamaño del resumen.

4.5 Otros métodos

En el 2004 [50] , proponen un método de generación automática de textos usando la entropía como medida para ordenar las oraciones de acuerdo a la relevancia (basado en el conocimiento pasado en un dominio en particular). Los documentos disponibles en un dominio forman el conjunto de datos de entrenamiento. Los valores de entropía calculados son aplicados a cada una de las oraciones en el conjunto de documentos y se ordenan las oraciones. Antes de aplicar la fórmula de entropía, para detectar y remover la redundancia utilizaron la representación de un grafo dirigido, cada nodo es una oración y se establece un enlace entre oraciones si más de tres palabas no vacías son comunes a ellas.

En el 2008 [51] proponen un método de generación automática de resúmenes de múltiples documentos basado en el descubrimiento de la característica de la palabra (WFS), creando siete características. Para extraer la palabra del tópico y del evento: si la palabra aparece en diferentes párrafos del mismo documento es una palabra de evento; si la palabra aparece distribuida en todos los documentos es una palabra de tópico. Lo primero que hacen es calcular los valores de las características para cada palabra (utilizando el modelo de regresión), luego construyen el modelo espacio vectorial, para calcular el valor de la oración suman

el valor de cada palabra que contiene ésta, luego seleccionan las oraciones del resumen y calculan el valor de similitud de las oraciones extraídas (eliminando las oraciones redundantes).

En el 2010 [52] proponen un método de consenso ponderado que combina los resultados de diferentes métodos de generación automática de resúmenes de un solo documento (WCS), para ello el método resuelve un problema de optimización que busca encontrar la mínima distancia ponderada entre una lista de ordenamiento de las oraciones y las listas de ordenamiento de las oraciones de cada método. El algoritmo recibe como entrada la lista de oraciones ordenadas por cada generador de resúmenes, luego de forma iterativa: calcula una lista que es un consenso de estas listas con un ponderado fijo para cada generador de resúmenes y luego calcula la distancia euclidiana entre estas listas y la lista consensuada para obtener un nuevo valor ponderado para cada generador de resúmenes. Esto se repite hasta que el algoritmo converge y se ordenan la lista de forma ascendente.

En el 2010 [53] proponen un enfoque de abajo hacia arriba para ordenar las oraciones extraídas en la generación automática de resúmenes de múltiples documentos. El algoritmo propuesto iterativamente concatena segmentos de texto (una secuencia de oraciones) hasta que un segmento es ordenado con todas las oraciones. Para poder definir el orden y la asociación de dos segmentos de texto definen cuatro criterios: Cronología, orden cronológico de las oraciones por fecha de publicación o por orden en el documento; Cercanía de temas (si se tienen dos segmentos A y B), se refiere a la asociación de dos segmentos basado en su similitud temática (usan ley de cosenos), se asigna el valor más alto cuando el tópico tratado por el segmento A es igual al del segmento B; Precedencia, una oración seleccionada puede presuponer información de otras oraciones que no fueron seleccionadas (usan ley de cosenos), se calcula la similitud para cada oración anterior de la oración b (del segmento B) en el documento original con cada oración del segmento A; Sucesión, evalúa la cobertura de la información posterior al segmento A colocando el segmento B después del A (usan ley de cosenos), se calcula la similitud para cada oración posterior de la oración a (del segmento A) en el documento original con cada oración del segmento B. Luego utilizan resúmenes humanos como datos de entrenamiento para encontrar la combinación óptima de los criterios propuestos, de esta forma integran los cuatro criterios y definen la función que representa la asociación de dirección y fuerza de dos segmentos. Esta tarea la abordan como un problema de clasificación binaria y emplean máquinas de soporte vectorial (SVM) para modelar la función.

En el 2010 [54] proponen un modelo híbrido para generar resúmenes de múltiples documentos, compuesto de dos pasos: un modelo generativo para descubrir patrones y un modelo de regresión para hacer inferencia. Los autores usan un modelo de tópicos jerárquico para obtener las características latentes de las oraciones y calcular el puntaje de éstas en el grupo de documentos. Luego entrenan un modelo de regresión basado en las características léxicas y estructurales de las oraciones. Por último se usa el modelo para puntuar oraciones de nuevos conjuntos de documentos para generar el resumen.

En el 2010 [55] propone un método de generación de resúmenes para múltiples documentos basado en técnicas de agrupamiento y en un algoritmo de optimización de enjambres de partículas discretas (PSO), la función objetivo tiene en cuenta el agrupamiento de las oraciones y la similitud entre la oraciones se realiza por medio de la medida de similitud de google. Después en 2011 [56] proponen un modelo basado en máxima cobertura y mínima redundancia, modelando la generación de textos como un problema de programación lineal entera. Para obtener un buen resumen optimizan las siguientes tres propiedades: Relevancia (unidades de información que son relevantes al usuario), Redundancia (no contener unidades de texto que transmitan la misma información), Longitud (se limita la longitud del resumen). La función objetivo es (maximizar) la similitud entre las oraciones del resumen y las oraciones de todo(s) los documento(s), pero eliminando la redundancia (similitud entre las oraciones del resumen); con una restricción por la longitud del resumen. Entonces la función objetivo y la restricción, se plantea como un problema de programación lineal entera; la función objetivo final es una combinación lineal de ésta función basada en la similitud de coseno y basada en la medida de similitud de google. Como algoritmos para resolver el problema de encontrar la solución óptima utilizaron: ramificación y poda, que han sido estudiados de forma eficiente con estos algoritmos (si el tamaño del problema es limitado, algunas veces se puede obtener una solución exacta dentro de un tiempo aceptable por medio del algoritmo B&B); y el algoritmo PSO binario con un tamaño del enjambre de 30 partículas, 500 iteraciones y con una posición mínima y máxima permitida de 50 cada una. Luego en el 2012 [57] propone un modelo orientado a restricción, un enfoque orientado a la cobertura y otro

orientado a la diversidad; el problema es formulado como programación entera cuadrática y se resuelve el problema con PSO.

También en el 2011[58] proponen un algoritmo de generación automática de resúmenes de múltiples documentos basado en evolución diferencial con parámetros adaptativos para el cruce y la mutación, cuya función objetivo es la división entre cobertura y redundancia; luego en el 2012 [59] basado también en evolución diferencial para definir la función objetivo utilizan el problema de la p-mediana y tienen en cuenta los siguientes factores: relevancia (oraciones relevantes al contenido principal de los documentos), cobertura del contenido (oraciones que traten los subtópicos del documento), diversidad (evitar oraciones redundantes que tratan la misma información) y longitud del resumen (definido por el tamaño definido para el resumen); en el mismo 2012 en [60] proponen este mismo algoritmo con una variación en la función objetivo propuesta en [58].

5. ASPECTOS RELEVANTES DE LOS MÉTODOS

Los métodos de generación de resúmenes extractivos revisados en este artículo presenta varios aspectos importantes que son: la definición de las características que serán tenidas en cuenta para darle un puntaje a cada una de las oraciones de los documentos y el método utilizado para obtener el valor estas características o de los pesos asociados a las mismas.

En cuanto a la definición de las características, por tratarse de resúmenes extractivos, las más utilizadas como se observó en este documento son, estadísticas (posición y longitud de la oración, frecuencia de términos y de palabras clave, presencia de nombres propios, pronombres, etc.) y medidas de similitud entre cada par de oraciones de los documentos o entre cada oración y todas las oraciones del documento (o el título para el caso de resumen de un solo documento). En el caso de los métodos basados en modelos evolutivos se define una función objetivo, la cual también contempla estas características estadísticas y de similitud.

Con respecto a la obtención del valor de estas características o de los pesos asociados a las mismas, se encuentra gran variedad de métodos. Algunos factores importantes de estos son, la necesidad de tener en cuenta el lenguaje al aplicar el método, de datos de entrenamiento y el manejo múltiples tópicos; la importancia de estos factores se debe a:

- Independencia del lenguaje: Permite que la generación de resúmenes, se pueda realizar sin importar el lenguaje en el que esté escrito el documento, de esta forma, el método puede ser usado en diferentes lenguajes al usado en el método propuesto.

- Supervisado/No Supervisado: En los esquemas supervisados es necesario contar con suficiente conjuntos de documentos para poder realizar tanto el entrenamiento como la prueba; por su parte los esquemas no supervisados solo necesitan de los conjuntos de documentos para la prueba.

- Manejo de múltiples tópicos: Permite que el resumen generado contemple oraciones de los diferentes tópicos tratados en el documento o conjunto de documentos. Este aspecto no se tiene en cuenta en algunas de las investigaciones de múltiples documentos y en la mayoría de un solo documento, debido a que se asume que éste contiene información solo de un tópico, como sucede con los artículos de noticias recolectados por la Conferencia de entendimiento de documentos (DUC, por sus siglas en inglés Document Understanding Conference). Sin embargo, aún en el caso de las noticias, se pueden presentan información de sub-tópicos. Además existen otros tipos de documentos que contienen información de varios tópicos o sub-tópicos y el resumen debería contemplar oraciones de cada tópico.

La presencia de estos tres factores en los métodos de generación automática de resúmenes de textos (uno o múltiples documentos) se aprecia en la Tabla 1.

Como se aprecia en la Tabla 1, los métodos que son independientes del lenguaje, no supervisados y que contemplan el manejo de múltiples tópicos son los basados en las técnicas de reducción algebraica, evolutivos y de agrupamiento. Estos dos últimos recientemente están generando mucho interés por parte de los investigadores, debido a los buenos resultados obtenidos por estos.

Tabla 1. *Cuadro comparativo métodos*

Método	Indep. lenguaje	Super./ No Super.	Múltiples tópicos
Estadísticos	SI	No supervisado	NO
Aprendizaje de máquina	SI	Supervisado	NO
Conectividad de textos	NO	No supervisado	NO
Grafos	SI	No supervisado	NO
Reducción algebraica	SI	No supervisado	SI
Evolutivos	SI	No supervisado	SI
Agrupamiento, PLSA	SI	No supervisado	SI

6. EVALUACIÓN DE RESÚMENES

Después de tener un panorama de los métodos más relevantes para la generación automática de resúmenes de textos, es importante revisar que evaluación se pueden aplicar a esta tarea y qué medidas son las más aceptadas para realizar la evaluación de la calidad de los resúmenes generados.

6.1 Conceptos de evaluación

En evaluación de la generación automática de resúmenes, existen dos tipos de evaluación importantes [61]: intrínseca que busca medir la calidad del resumen creado y extrínseca que mide que tan bien se ejecuta el resumen dentro de una tarea en particular. La evaluación intrínseca parte de que existe un estándar para comparar, que puede ser un conjunto de datos (texto/resumen) existente o con un resumen "ideal" creado por un humano. Este tipo de evaluación presenta el problema de variabilidad del ser humano al realizar esta tarea, por lo que se requiere de un diseño y análisis experimental. De otro lado la evaluación extrínseca requiere seleccionar una tarea que puede utilizar la generación automática de resúmenes y medir el efecto de utilizar resúmenes automáticos en lugar del texto original. En esta evaluación los problemas son la selección de la tarea y los indicadores para la medición.

6.2 Evaluación automática de resúmenes

En 1998 [62], analizaron tanto la evaluación intrínseca como la extrínseca. Para reducir la subjetividad en el primer caso, ya que se utiliza un resumen "ideal", utilizan el modelo estadístico para partir de varios resúmenes hechos por humanos y de estos construyen el resumen "ideal", los autores demuestran que la longitud del resumen afecta el resultado en este caso, ya que en la experimentación encontraron que medidas como la precisión y el recuerdo son sensibles a la longitud del resumen, cuando el resumen es corto los resultados son más fiables que cuando son más largos, también indican que la precisión y el recuerdo no son las mejores medidas para calcular la calidad del resumen (un pequeño cambio en la salida del resumen, por ejemplo, reemplazar una oración por otra también buena) puede afectar significativamente el puntaje del sistema. En el caso de la evaluación extrínseca (basada en una tarea) escogieron una tarea de recuperación de información basada en la consulta del usuario, sin embargo, como no disponían de generadores de resúmenes basados en consultas, utilizaron generadores genéricos involucrando la consulta del usuario dentro del tópico principal del documento. Los criterios de evaluación fueron: tiempo requerido, precisión de la decisión y confidencia de la decisión. La experimentación mostró que no hay co-relación entre la longitud y el mejoramiento de la tarea.

En el 2002 Harman y Over [61] presentan un reporte de evaluaciones de generación automática de resúmenes del 2001 en DUC, el tipo de resúmenes a evaluar fueron genéricos y el tipo de evaluación fue intrínseca. Los documentos que se escogieron fueron artículos de noticias, con al menos 10 oraciones cada documento. La implementación contemplo: 60 conjuntos de documentos con aproximadamente 10 documentos, analistas de información retirados del Instituto Nacional de estándares y tecnología (NIST, por sus siglas en inglés National Institute of Standards and Technology) definieron los conjuntos de documentos y crearon resúmenes manuales de 100 palabras para cada documento y resúmenes de cada conjunto de documentos con cuatro longitudes (50, 100, 200, 400), los sistemas generadores de resúmenes también crearon resúmenes genéricos de 100 palabras para cada documento, y resúmenes de cada conjunto de documentos con las mismas cuatro longitudes. Treinta conjunto de documentos (con los resúmenes manuales) se destinaron a entrenamiento y los otros treinta conjuntos (sin los resúmenes) se utilizaron para la prueba. Luego NIST presento los resultados de esta evaluación, que realizó el mismo personal que creo los resúmenes manuales y se basó en una comparación manual de los resúmenes realizados por los generadores de resúmenes con los resúmenes construidos por los humanos. Esta comparación manual

se realizó con el apoyo de una herramienta llamada SEE, para evaluar dos áreas: la calidad del resumen (gramaticalidad, cohesión y organización/coherencia), y la cobertura del resumen con respecto al contenido del documento. Esta evaluación fue muy importante ya que fue la primera evaluación intrínseca a gran escala de generación automática de resúmenes para uno y múltiples documentos.

En el 2004 [63] introduce una herramienta llamada ROUGE (Recall-Oriented Understudy for Gisting Evaluation) que incluye medidas para automáticamente determinar la calidad de un resumen generado con un resumen ideal, midiendo la similitud entre resúmenes por medio de N-gramas. Además presenta variaciones de ROUGE entre las que están: ROUGE-N (Estadísticas de co-ocurrencia de N-gramas) que es un N-grama de recuerdo entre un resumen candidato y uno de referencia, donde N es la longitud del N-grama; ROUGE-L (Subsecuencia común más larga) que toma el resumen como una secuencia de palabras e incluye la secuencia común más larga de n-gramas; ROUGE-W (Subsecuencia común más larga ponderada) que otorga un peso mayor a la secuencia más larga de N-gramas pero donde estos sean secuenciales en el texto. Para evaluar la eficacia de ROUGE utilizaron los juicios humanos hechos en DUC en los años 2002 y 2003, sobre resúmenes de un solo documento (100 palabras), resúmenes de un solo documento muy cortos (10 palabras) y resúmenes de múltiples documentos (10, 50, 100, 200, 400 palabras). Como resultado encontraron que ROUGE-2, ROUGE-L, ROUGE-W y ROUGE-S; lograron muy buena correlación con los juicios humanos de los datos de DUC 2001 y 2002, trabajando bien para tareas de generación automática de resúmenes de un solo documento. ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4 y ROUGE-SU9 obtuvieron los mejores resultados para los resúmenes muy cortos. Y por último ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-S9, ROUGE-SU4 y ROUGE SU9, funcionaron razonablemente bien para tareas de generación automática de resúmenes de múltiples documentos.

Más recientemente en [64] (2008) proponen una variación de ROUGE, llamada ROUGE-C, que permite evaluar resúmenes sin tener resúmenes humanos de referencia, reemplazando estos resúmenes de referencia con el documento fuente y aplicando ROUGE-C-N, para los puntajes de similitud. Para la evaluación tomaron los datos de DUC de 2001 a 2005, algunas de los aspectos encontrados fue que para la generación automática de resúmenes de un solo documento ROUGE-C se ejecutó mejor para resúmenes de 10 palabras, y para múltiples documentos la co-relación aumento cuando el tamaño del resumen era más grande. Encontraron que con restricciones de las condiciones apropiadas y una aceptable disminución de la eficiencia ROUGE-C se co-relaciona bien con métodos que dependen de los resúmenes de referencia (incluyendo los juicios humanos).

7. CONCLUSIONES

Para la generación automática de resúmenes extractivos de un solo documento, las investigaciones más recientes proponen métodos basados en reducción algebraica y en modelos evolutivos. Estos dos métodos presentas dos ventajas: (i) son métodos independientes del lenguaje, de esta forma, la generación de resúmenes de un solo documento, se pueden realizar sin importar el lenguaje en el que esté escrito el documento; (ii) son métodos no supervisados, por lo cual no es necesario contar con mayor cantidad de conjuntos de documentos, para poder realizar tanto el entrenamiento como la prueba, lo cual en muchos casos es difícil de encontrar.

En cuanto a la generación automática de resúmenes extractivos de múltiples documentos, algunas investigaciones primero propusieron un método de generación de resúmenes de un solo documento y luego un método de generación de resúmenes de múltiples documentos como una extensión del método de un solo documento, es decir, primero generan el resumen por cada documento del grupo de documentos y luego eliminan redundancia; éste es el caso de los métodos basados en conectividad de textos, basados en grafos y reducción algebraica. Otras investigaciones proponen el método para múltiples documentos sin que exista un trabajo previo del método para un solo documento (de los mismos u otros autores), este es el caso de métodos de agrupamiento y modelos evolutivos.

Un aspecto importante a tener en cuenta en los métodos de generación de resúmenes de múltiples documentos, es la eliminación de redundancia, esto debido a que los documentos están relacionados con una misma temática, de esta forma, se evita que en el resumen se encuentren oraciones muy similares.

Las investigaciones más recientes de generación automática de resúmenes extractivos de múltiples documentos proponen métodos basados en reducción algebraica, agrupamiento y en modelos evolutivos. Estos métodos también presentan dos ventajas: son métodos independientes del lenguaje y son métodos

no supervisados. Aunque algunas de las investigaciones asumen que los documentos a resumir pertenecen a un mismo tópico (artículos de noticias recolectados por DUC) y no contemplan el manejo de múltiples tópicos; es importante que el método contemple este aspecto, debido a que las noticias manejan sub-tópicos y además existen otros tipos de documentos que contienen información de varios tópicos o sub-tópicos, lo que hace necesario que el resumen contemple oraciones de cada tópico o sub-tópico.

Aunque en el caso de la generación automática de resúmenes de un solo documento se han obtenido muy buenos resultados con métodos basados en reducción algebraica y modelos evolutivos; y en el caso de múltiples documentos con métodos basados en reducción algebraica, agrupamiento y modelos evolutivos; investigaciones recientes basadas en estos últimos, tanto para un documento como para múltiples documentos han mostrado en algunos casos mejores resultados, haciendo que la investigación en esta área sea prometedora, y dejando la posibilidad de explorar en la aplicación de otros modelos evolutivos no utilizados en la actualidad.

En cuanto a la evaluación de los resúmenes generados automáticamente por sistemas propuestos, las medidas más usadas hasta hace unos años han sido medidas estándar de recuperación de la información como: precisión, recuerdo y medida F. Sin embargo, en la última década se han planteado nuevas medidas diseñadas especialmente para evaluar la calidad de los resúmenes generados; estas medidas son propuestas por ROUGE (basadas en N-gramas), que permiten medir el solapamiento por medio de n-gramas y no de oraciones. Sin embargo, dada la complejidad de evaluar la calidad de un resumen generado por un sistema con el generado por un humano, esta área sigue abierta a nuevas propuestas que permitan medir de forma más adecuada la calidad de estos resúmenes.

8. AGRADECIMIENTOS

El trabajo realizado en este artículo fue soportado por la Universidad del Cauca por medio del proyecto VRI-3029 y de la Universidad Nacional de Colombia sede Bogotá.

9. REFERENCIAS

[1] S. Osiński and D. Weiss, "A concept-driven algorithm for clustering search results," *Intelligent Systems, IEEE,* vol. 20, pp. 48-54, 2005.

[2] D. Das and A. F. T. Martins, "A Survey on Automatic Text Summarization," ed, 2007.

[3] K. Ježek and J. Steinberger, "Automatic Text Summarization (The state of the art 2007 and new challenges)," in *Znalosti 2008*, Bratislava, Slovakia, 2008, pp. 1-12.

[4] T. Simone and M. Marc, "Summarizing scientific articles: experiments with relevance and rhetorical status," *Computational Linguistics,* vol. 28, pp. 409-445, 2002.

[5] Z. Jiaming, L. Han Tong, L. Ying, and S. Aixin, "Automatic text summarization in engineering information management," in *Proceedings of the 10th International Conference on Asian digital libraries: looking back 10 years and forging new frontiers*, Hanoi, Vietnam, 2007.

[6] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development,* pp. 159-165, 1958.

[7] P. Baxendale, "Machine-made index for technical literature - an experiment.," *Journal of Research Development,* vol. 2, pp. 354-361, 1958.

[8] H. P. Edmundson, "New Methods in Automatic Extracting," *Journal of the ACM (JACM),* vol. 16, pp. 264-285, 1969.

[9] G. Salton, "Automatic Text Processing," *Addison-Wesley Publishing Company.,* 1988.

[10] C.-Y. Lin and E. Hovy, " Identifying topics by position," *In Proceedings of the Fifth conference on Applied natural language processing. San Francisco, CA, USA.,* pp. 283-290, 1997.

[11] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and development in information retrieval*, Seattle, Washington, United States, 1995, pp. 68-73.

[12] C. Aone, M. E. Okurowski, J. Gorlinsky, and B. s. Larsen, " A trainable summarizer with knowledge acquired from robust nlp techniques.," *Advances in Automatic Text Summarization,* vol. Mani, I. and Maybury, M. T., pp. 71-80, 1999.

[13] C.-Y. Lin, "Training a selection function for

extraction. ," *In Proceedings of CIKM '99. New York, NY, USA,* pp. 55-62, 1999.

[14] M. Osborne, "Using maximum entropy for sentence extraction," in *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Phildadelphia, Pennsylvania, 2002.

[15] K. Svore, Vanderwende, L., and Burges, C., "Enhancing single-document summarization by combining RankNet and third-party sources," *In Proceedings of the EMNLP-CoNLL,* pp. 448-457, 2007.

[16] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *Proceedings of the 20th International Joint Conference on Artifical intelligence*, Hyderabad, India, 2007, pp. 2862-2867.

[17] K.-F. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester, United Kingdom, 2008.

[18] R. Barzilay, Elhadad, M, "Using Lexical Chains for Text Summarization," *In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain.,* pp. 10–17, 1997.

[19] K. Ono, Sumita, K., and Miike, S., "Abstract generation based on rhetorical structure extraction.," *In Proceedings of Coling '94. Morristown, NJ, USA,* pp. 344-348, 1994.

[20] D. Marcu, "Improving summarization through rhetorical parsing tuning," *Proceedings of The Sixth Workshop on Very Large Corpora. Montreal, Canada,* pp. 206-215, 1998.

[21] D. C. T. Marcu, "The rhetorical parsing, summarization, and generation of natural language texts," PhD thesis, University of Toronto. Adviser-Graeme Hirst., 1998.

[22] R. Mihalcea, Tarau, P. , " Text-rank - bringing order into texts," *In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain.,* 2004.

[23] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latentsemantic analysis," in *Proceedings of ACM SIGIR*, New Orleans, USA, 2001.

[24] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *In Proceedings ISIM '04* 2004.

[25] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management,* vol. 41, pp. 75–95, 2005.

[26] J.-H. Lee, S. Park, C.-M. Ahn, and D. Kim, "Automatic generic document summarization based on non-negative matrix factorization," *Information Processing & Management,* vol. 45, pp. 20-34, 2009.

[27] A. Kiani and M. R. Akbarzadeh, "Automatic Text Summarization Using Hybrid Fuzzy GA-GP," in *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2006, pp. 977-983.

[28] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Computer Speech & Language,* vol. 23, pp. 126-144, 2009.

[29] P.-K. Dehkordi, F. Kumarci, and H. Khosravi, "Text Summarization Based on Genetic Programming," in *In Proceedings of the International Journal of Computing and ICT Research*, 2009, pp. 57-64.

[30] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm based text summarization," *Journal Computer Sciences,* vol. 5, pp. 338–346, 2009.

[31] M. S. Binwahlan, N. Salim, and L. Suanmali, "Swarm Based Text Summarization," in *In Proceedings of the International Association of Computer Science and Information Technology - Spring Conference. IACSITSC '09*, 2009, pp. 145-150.

[32] W. Song, L. Cheon Choi, S. Cheol Park, and X. Feng Ding, "Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization," *Expert Systems with Applications,* vol. 38, pp. 9112-9121, 201*1*.

[33] M. S. Binwahlan, N. Salim, and L. Suanmali, "Fuzzy swarm diversity hybrid model for text summarization," *Information Processing and Management,* vol. 46, pp. 571-588, 2010.

[34] M. Litvak, M. Last, and M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 2010, pp. 927-936.

[35] V. Qazvinian, L. Sharif, and R. Halavati, "Summarising text with a genetic algorithm-based sentence extraction," *International Journal of Knowledge Management Studies (IJKMS),* vol. 2, pp. 426-444, 2008.

[36] E. Shareghi and L. S. Hassanabadi, "Text summarization with harmony search algorithm-

based sentence extraction," in *Proceedings of the 5th International Conference on Soft computing as transdisciplinary science and technology* Cergy-Pontoise, France, 2008.

[37] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization," *Expert Systems with Applications,* vol. 36, pp. 7764-7772, 2009.

[38] Y.-M. Chen, X.-L. Wang, and B.-Q. Liu, "Multi-document summarization based on lexical chains," in *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, Vol. 3, 2005, pp. 1937-1942.

[39] R. Mihalcea, Tarau, P., " An Algorithm for Language Independent Single and Multiple Document Summarization.," *In Proceedings of the International Joint Konference on Natural Language Processing, Korea.,* 2005.

[40] X. Wan, "Towards a Unified Approach to Simultaneous Single-Document and Multi-Document Summarizations," in *In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, 2010, pp. 1137–1145.

[41] B. Hachey, G. Murray, and D. Reitter, "The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space," in *Proceedings of the Document Understanding Conference (DUC)*, Vancouver, Canada, 2005.

[42] J. Steinberger and M. Křišťan, "LSA-Based Multi-Document Summarization," in *Proceedings of 8th International PhD Workshop on Systems and Control*, Balatonfured, Hungary, 2007.

[43] D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management,* vol. 40, 2004, pp. 919-938.

[44] D. Wang, S. Zhu, T. Li, Y. Chi, and Y. Gong, "Integrating clustering and multi-document summarization to improve document understanding," in *Proceedings of the 17th ACM conference on Information and knowledge management*, Napa Valley, California, USA, 2008, vol. 5, 2011, pp. 1-26.

[45] M. Ali, M. K. Ghosh, and A. Al-Mamun, "Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation," in *International Conference on Future Computer and Communication, 2009. ICFCC 2009.* , 2009, pp. 93-96.

[46] M. Xiao-Chen, Y. Gui-Bin, and M. Liang, "Multi-Document Summarization Using Clustering Algorithm," in *Proceedings of the International Workshop on Intelligent Systems and Applications*, 2009, pp. 1-4.

[47] L. Hennig, "Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis," in *International Conference RANLP*, Borovets, Bulgaria, 2009, pp. 144–149.

[48] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-Document Summarization using Sentence-based Topic Model," in *Proceedings of the ACL-IJCNLP*, Suntec, Singapore, 2009, pp. 297–300.

[49] X. Cai and W. Li, "A spectral analysis approach to document summarization: Clustering and ranking sentences simultaneously," *Information Sciences,* vol. 181, 2011, pp. 3816-3827.

[50] G. Ravindra, N. Balakrishnan, and K. R. Ramakrishnan, "Multi-document Automatic Text Summarization Using Entropy Estimates," in *SOFSEM 2004: Theory and Practice of Computer Science*, ed, 2004, pp. 73-82.

[51] W. Meng, W. Xiaorong, L. Chungui, and Z. Zengfang, "Multi-document Summarization Based on Word Feature Mining," in *Proceedings of the 2008 International Conference on Computer Science and Software Engineering*, 2008, pp. 743-746.

[52] D. Wang and T. Li, "Many are better than one: improving multi-document summarization via weighted consensus," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and development in information retrieval*, Geneva, Switzerland, 2010, pp.

[53] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," *Information Processing and Management,* vol. 46, 2010, pp. 89-109.

[54] A. Celikyilmaz and D. Hakkani-Tur, "A Hybrid Hierarchical Model for Multi-Document Summarization," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010., 2010, pp. 815–824.

[55] R. M. Aliguliyev, "Clustering techniques and discrete particle swarm Optimization algorithm for multi-document," *An international journal Computational Intelligence,,* vol. 26, 2010, pp. 420-448.

[56] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova, and C. A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model," *Expert Systems with Applications,* vol. 38, 2011, pp. 14514-14522.

[57] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "CDDS: Constraint-driven document summarization models," *Expert Systems with Applications,* vol. 40, 2013, pp. 458-465.

[58] R. M. Alguliev, R. M. Aliguliyev, and C. A. Mehdiyev, "Sentence selection for generic document summarization using an adaptive differential evolution algorithm," *Swarm and Evolutionary Computation,* vol. 1, 2011, pp. 213-222.

[59] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization," *Knowledge-Based Systems,* vol. 36, pp. 21-38.

[60] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, "Multiple documents summarization based on evolutionary optimization algorithm," *Expert Systems with Applications,* vol. 40, 2013, pp. 1675-1689.

[61] D. Harman and P. Over, "The DUC summarization evaluations," in *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, California, 2002, pp. 44-51.

[62] H. Jing, R. Barzilay, K. Mckeown, and M. Elhadad, "Summarization Evaluation Methods: Experiments and Analysis," in *AAAI Symposium on Intelligent Summarization* 1998, pp. 51-60.

[63] C. Lin, "Rouge: a package for automatic evaluation of summaries," in *In Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 25-26.

[64] H. Tingting, C. Jinguang, M. Liang, G. Zhuoming, L. Fang, S. Wei, and W. Qian, "ROUGE-C: A fully automated evaluation method for multi-document summarization," in *Proceedings of the IEEE International Conference on Granular Computing*, 2008, pp. 269-274.