RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

TextLec: Método para la segmentación por tópicos en textos científico-técnicos

MSc. Laritza Hernández Rojas,
Dr. C. José E. Medina Pagola

RT_007 **Noviembre 2009**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2143
ISSN 2072-6260
Versión Digital

SERIE GRIS

REPORTE TÉCNICO
Minería de Datos

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

TextLec: Método para la segmentación por tópicos en textos científico-técnicos

MSc. Laritza Hernández Rojas,
Dr. C. José E. Medina Pagola

RT_007 **Noviembre 2009**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

TextLec: Método para la segmentación por tópicos en textos científico-técnicos

MSc. Laritza Hernández Rojas, Dr. C. José E. Medina Pagola

Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
lhernandez@cenatav.co.cu

RT_007 CENATAV
Fecha del camera ready: 24 de marzo de 2009

Resumen: El presente reporte de investigación se realizó en el departamento de Minería de Datos del CENATAV, responsable del procesamiento y la extracción de información en documentos digitales en esta institución. De ahí que su propósito sea la exposición de la elaboración de un método para segmentar automáticamente textos por tópicos sobre colecciones de documentos científico-técnicos, donde se intenta lograr una cohesión léxica considerable de los segmentos que se obtienen, así como evitar la innecesaria interrupción de los mismos, con similar o superior eficacia a otros métodos existentes. Para ello es necesario presentar el Marco Teórico de la investigación, mostrando el estudio y análisis crítico del estado actual de los métodos de segmentación por tópicos. Luego se aborda el diseño de un método de segmentación por tópicos, nombrado TextLec, que se sustenta en el uso de la cohesión léxica como señal de cambio de tópico, del modelo de espacio vectorial como forma de representación de las unidades textuales, de la medida del coseno para determinar la similitud entre dos unidades textuales, de la teoría computacional de Skorochod'ko sobre la estructura lineal del discurso y en el uso de una ventana de párrafos inferiores (por debajo) a cada párrafo, con vista a localizar el párrafo cohesionado más lejano a cada párrafo y evitar la interrupción de los tópicos. Finalmente, se muestra la validación del método propuesto a partir de *corpus* textuales representativos del universo investigado y su comparación con algunos de los métodos encontrados, resultando más adecuado que las anteriores propuestas.

Palabras claves: segmentación por tópicos, segmentación del discurso, cohesión léxica.

Abstract: This technical report was carried out at CENATAV, particularly at the Data Mining department which is the one in charge of processing and extracting information from digital documents. Thus the objective is to expose the development of a method to automatically segment texts by topics for scientific and technical collections that try to achieve a strong lexical cohesion of the obtained segments and to avoid the unnecessary interruption with a similar or higher accuracy to other existing methods. For this aim, it is necessary to preset the Theoretical Framework of the research, showing the study and critical analysis of the related works on thematic of segmentation by topic. Later, one method of segmentation by topic called TextLec is presented. This is supported by the use of lexical cohesion as a cue of topic change of the Vector Space Model as a way to represent text units, the cosine measure to determine the similarity between two textual units, the Skorochod 'ko computational theory about the linear structure of discourse and the use, for each paragraph, of a paragraphs lower window (paragraph below) to find the farthest cohesive paragraph inside the window and to avoid topic interruptions. Hence, the validation method is showed using text from the universe studied and we compared it with some of the methods we found, aiming at outperforming other proposals.

Keyword: Topic Segmentation, Discourse Segmentation, Lexical Cohesion.

1 Introducción

Actualmente, el cúmulo de documentos electrónicos de naturalezas disímiles es enorme y aumenta día a día. Las computadoras permiten almacenar y controlar dicha información, evitando su deterioro con el paso del tiempo, a diferencia de lo que ocurre con los documentos impresos. Pero, además de las computadoras para controlar y almacenar grandes volúmenes de información, cada vez se hace más necesaria la existencia de herramientas de procesamiento de textos eficaces y eficientes que faciliten la comprensión de dicha información; por ejemplo, la recuperación de información (IR, por sus siglas en inglés), la confección automática de resúmenes, la detección y seguimiento de tópicos, entre otras. Pero, durante el estudio y desarrollo de las soluciones de tales tareas, se ha detectado que existen problemas más específicos que han requerido de su estudio y del desarrollo de herramientas automáticas para resolverlos.

Una de las necesidades identificadas la constituyen las herramientas automáticas que permitan segmentar un texto por tópicos. Por ejemplo, en la recuperación de información, más específicamente en la recuperación de pasajes, se necesitan los métodos de segmentación por tópicos para devolver los segmentos o pasajes más relacionados con la consulta que realizaría un usuario, en lugar del documento completo [22]. La confección automática de resúmenes de textos también sería más robusta si se tuviese conocimiento de todos los subtópicos que forman un documento, porque estos subtópicos se podrían utilizar como guía para una selección balanceada de las ideas principales que conformarían el resumen de todo el documento [2]. Como último ejemplo, se tiene que en la detección y seguimiento de tópicos, la segmentación se necesita para la división en noticias individuales de un flujo de transmisión continua, teniendo en cuenta los cambios de tópicos de una a otra [46].

Aunque se han encontrado algunas aproximaciones para resolver el problema de la segmentación por tópicos, los resultados que éstas logran no siempre son de alta calidad, debido a la perdida de cohesión y coherencia en algunos de los segmentos que se obtienen. Por ejemplo, una de las deficiencias observadas es la incorrecta interrupción de segmentos, dejando fuera oraciones o párrafos, o sea, dejando inconclusa la información o el mensaje contenido en un segmento. Cuando esto sucede, además, pueden obtenerse segmentos espurios con esas oraciones o párrafos que no fueron incluidos dentro del segmento correspondiente; también pueden obtenerse segmentos de baja cohesión al incluirse en éstos esas unidades textuales que quedaron excluidas de su segmento.

En este trabajo se expone la propuesta de un método para segmentar automáticamente un texto por tópicos, logrando una cohesión léxica considerable de los segmentos que se obtengan y evitando la innecesaria interrupción de los mismos, con similar o superior eficacia a otros métodos existentes. Inicialmente, este método se concibió para documentos científico- técnicos, dada la necesidad reportada por el departamento de Minería de Datos del CENATAV. No obstante, puede aplicarse a otros tipos de documentos con características similares a las que se asumen en este trabajo.

Este reporte de investigación está estructurado como sigue a continuación. En la introducción se ubicó y esbozó el problema de investigación, las necesidades y conveniencias de resolverlo, así como los objetivos de investigación. En la Sección 2 se dan las nociones necesarias de algunos términos y fenómenos propios asociados a los cambios de tópicos así como a su identificación, además, se describen las principales etapas de preprocesamiento del

texto que son necesarias para obtener un buen desempeño en la segmentación. En la Sección 3 se exponen algunos de los distintos enfoques y métodos asociados a la segmentación por tópicos y se comentan las principales deficiencias de estos. En la Sección 4 se propone un método de segmentación por tópico que intenta mejorar la eficacia de los métodos criticados en la Sección 1 que resultaron más adecuados a las intenciones de esta investigación. En la Sección 4 se evalúa y valida el método propuesto, empleando los métodos de evaluación de la segmentación por tópico más apropiados y *corpus* textuales representativos del universo investigado. Posteriormente, en la Sección 5 se exponen las conclusiones y las tareas futuras a emprender con la investigación. Por último, se relacionan las fuentes bibliográficas consultadas.

2 Segmentación por tópicos y trabajos relacionados

A través del estudio y análisis de la literatura sobre el procesamiento de texto se han encontrado varios métodos que se basan en distintas interpretaciones de la segmentación de textos. Una de las más aceptadas estructura el problema de la segmentación en dos clases: la segmentación por unidad textual (por ejemplo, párrafos [6], [9], [10]), y la segmentación por unidades de tópicos. Este trabajo enfocará su atención sólo en el segundo problema; o sea, en la segmentación por tópicos.

Alrededor de la problemática de la segmentación por tópico existen muchos aspectos que deben ser considerados tanto en el terreno computacional como en el lingüístico; por ejemplo, la distinción entre tópicos y subtópicos, la organización de los subtópicos en el texto, las señales lingüísticas que permiten identificar computacionalmente los cambios de tópicos, así como también es necesario considerar los aspectos inherentes al preprocesamiento computacional del texto original; es decir, es necesario conocer cómo llevar el texto a un estado que permita su procesamiento eficaz y eficiente desde un punto de vista computacional. A continuación se abordarán dichas cuestiones.

2.1 Segmentación por tópicos

Cuando se comienza a leer, se empieza a elaborar una primera idea de cuál es el tópico de lo que se lee. En este trabajo se entenderá por tópico aquello de lo que se habla, o a lo que se refiere el texto o discurso[1][5], [48- 50].

El tópico se desarrolla de forma secuencial y se va confirmando a medida que avanza la lectura. Pero, mientras avanza la lectura es muy frecuente que se desarrollen nuevos contenidos que determinen un cambio parcial de dicho tópico; o sea, existe un tópico global para todo el discurso mientras se tienen tópicos parciales, relacionados y desarrollados secuencialmente para cada parte o segmento discursivo[2]. Estos tópicos parciales más comúnmente se les conoce como subtópicos [43], [44], [47].

[1] El término discurso sirve para referir conceptos diferentes en distintos contextos, además de ser empleado en ocasiones incorrectamente, prestándose a confusión. Con dicho término, en este trabajo, se estará haciendo alusión a una forma escrita (texto) de comunicación más extensa que una oración.

[2] Los estudios sobre el Análisis del Discurso establecen el acuerdo de que existe una estructura discursiva que determina en gran medida la coherencia y evolución del discurso, dividiéndose este último en unidades llamadas segmentos discursivos. Éstos, a su vez, pueden estar relacionados de diferentes modos. Pero, aún los especialistas

Teniendo en cuenta lo anterior, la tarea de segmentación por tópico se puede definir más formalmente como: el proceso automático que identifica en un texto los cambios de tópicos, ya sean estos parciales o globales.

2.2 Segmentación del discurso

La estructura de los subtópicos en el discurso se puede asumir de varias formas. Según los estudios sobre la teoría computacional de la estructura del discurso, existen dos tipos de estructuras de discurso: lineal y jerárquica. Estas estructuras se diferencian fundamentalmente por el nivel de granularidad con el que se segmenta el discurso.

Una de estas teorías, propuesta por Skorochod'ko en 1972 y vigente en la actualidad, plantea que el texto tiene una estructura lineal que puede representarse por una red semántica determinada por la presencia de las relaciones semánticas entre las unidades textuales dentro de un documento (en este caso, con unidades textuales el autor se refiere a oraciones o párrafos) [42]. Skorochod'ko determina el grado de relación semántica entre dos unidades textuales, en base a las palabras idénticas o relacionadas semánticamente entre dichas unidades textuales. Por ejemplo, un grafo completamente conexo pudiera indicar una discusión densa de un tópico, mientras que una cadena larga de conectividades podría indicar una discusión secuencial de un tópico. A continuación, en la Fig. 2.1 se muestran los cuatro tipos de texto más importantes propuestos por Skorochod'ko, teniendo en cuenta la estructura de los mismos.

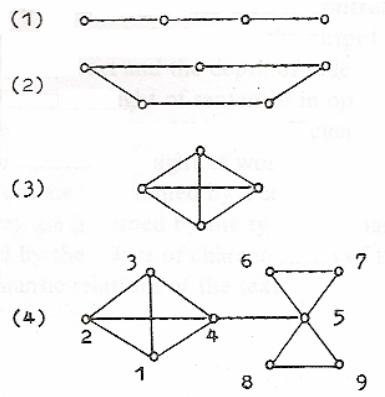


Fig. 2.1. Tipos de estructura textos propuestos por Skorochod'ko[3]

La definición de cada texto según su estructura es la siguiente:
(1) *Cadena*: cuando solamente las unidades vecinas están relacionadas.
(2) *Anillo*: cuando sólo las unidades vecinas están relacionadas, pero también existe relación entre la primera y la última.

discrepan en cuanto a los tipos de relaciones que definen la estructura discursiva y los aspectos del discurso que determinan dichas relaciones.

[3] Esta figura se extrajo del artículo original de Skorochod'ko referido en [42].

(3) *Monolítico*: cuando todas las unidades del texto se relacionan.
(4) *Monolítico por partes*: cuando hay porciones de texto que son monolíticas, pero hay pocas conexiones entre estas porciones.

Basado en lo planteado por Skorochod'ko, Reinar plantea que los documentos que tienen una estructura de cadena resultan ser los más fáciles de segmentar, separando las unidades textuales vecinas. Lo anterior puede apreciarse mejor con ayuda de la figura 2.1. Los que tienen una estructura de anillo también pueden ser segmentados sin muchos problemas si estos se transforman previamente en una cadena; o sea, si la relación entre la primera unidad y la última se ignora, entonces la estructura del documento se convertiría en una cadena que ya se conoce cómo segmentar. Los documentos monolíticos por partes también pueden ser segmentados, descomponiéndolos por sus partes monolíticas, las cuales se asume que representan tópicos individuales. Los textos monolíticos son los que presentan el mayor problema para ser segmentados, porque todas sus unidades textuales están relacionadas [36].

En 1986, Grosz y Sidner, a diferencia de Skorochod'ko, plantearon que el discurso presenta una estructura jerárquica, estando compuesto por tres estructuras interrelacionadas: estructura lingüística, estructura intencional y un estado atencional [11].

La estructura lingüística captura las relaciones entre unidades textuales consecutivas (o adyacentes) y divide el texto en segmentos discursivos. Éstos pueden incluir a otros segmentos o estar incluidos en algún segmento, conformándose de esta forma una jerarquía. La estructura intencional se refleja en la estructura lingüística, la misma modela los objetivos generales y los objetivos específicos de la discusión, comprende los propósitos (o intenciones) asociados con los segmentos discursivos y las relaciones entre dichos propósitos. Estas relaciones son identificadas por la estructura lingüística a través de indicadores lingüísticos como los sintagmas de entrada, los cuales se ampliarán más adelante. El estado atencional, por su parte, refleja el foco de atención de los participantes del discurso en la medida en que éste avanza, siendo la estructura lingüística la que restringe los cambios en el estado atencional.

Estas diferencias entre las estructuras que pueden adquirir el discurso evidencian una de las complejidades de la tarea de segmentación por tópicos, dada por la naturaleza subjetiva de decidir los límites adecuados de los subtópicos, es decir, lo que para algunos puede resultar correcto o suficiente, para otros no.

2.3 Señales que indican cambios o continuidad de tópicos

Todos los métodos de segmentación por tópicos requieren del uso de indicadores o señales lingüísticas para identificar los cambios de tópicos entre las unidades textuales. Tales unidades textuales pueden ser lo mismo palabras, oraciones, párrafos o bloques de textos formados por combinaciones de éstos. Es necesario que se note que para seleccionar adecuadamente una señal u otra, es imprescindible que se tenga en cuenta el tipo de texto que se va a segmentar; por ejemplo, documento científico, narrativo, informativo u otro. A continuación se expondrán algunas de dichas señales, distinguiendo, cuando sea necesario, en qué situaciones son más convenientes.

2.3.1 Cohesión léxica

Los resultados de varias investigaciones en el área de la segmentación por tópicos han mostrado que la cohesión léxica es un elemento muy útil para detectar los cambios de tópicos, asumiendo

que las unidades textuales que están fuertemente relacionadas por cohesión léxica, usualmente constituyen un segmento que abarca un tópico simple, o lo contrario si no están relacionadas.

Cohesión léxica es un término lingüístico que fue definido por Halliday y Hasan en 1976 como una de las relaciones de significado que existen entre las unidades textuales en un texto [13], [14]. Según Halliday y Hasan, la cohesión léxica abarca dos aspectos diferentes, la reiteración y la colocación[4]. En este trabajo sólo se hará referencia al primero, por ser éste el que más se utiliza en la detección de cambios de tópico. Estos autores definieron la reiteración como una forma de cohesión léxica que se refleja a través de la repetición de un elemento léxico o a través del uso de su sinónimo, casi sinónimo, hiperónimo o hipónimo. De éstos, los más utilizados han sido la repetición y la sinonimia, los cuales se explicarán brevemente a continuación.

2.3.1.1 Repetición

La repetición, o mera reiteración léxica, ocurre cuando se repite un elemento léxico en su identidad material y semántica [24]. Esto puede verse en el ejemplo siguiente.

Ejemplo:

María esperó por el *ómnibus* hasta las tres de la tarde. El *ómnibus* la llevará junto a sus padres antes de caer la noche.

Cabe destacar que un elemento léxico no necesita estar en la miasma categoría gramatical para reconocerse como repetido; como ocurre con coma, comiendo y comida. La ocurrencia de uno significa la repetición de cualquiera de los otros [14].

La repetición excesiva de palabras en el texto en ocasiones es mal empleada, porque suele entorpecer la lectura del mismo, esto es considerado como un problema de estilo en la redacción. Por el contrario, un buen uso de este recurso puede indicar continuidad del tópico y del sentido. Además, en determinados tipos de texto, la repetición no sólo es considerada como cuestión de estilo, sino que es necesaria y se exige. Esto ocurre, por ejemplo, en los textos científico-técnicos, que son el objetivo fundamental de la segmentación que se propondrá en este trabajo.

La repetición de términos no sólo es muy usada por los métodos de segmentación por tópicos, existe un número considerable de tareas de procesamiento que también la emplean. Este hecho está fuertemente ligado a que su cálculo requiere de poco costo computacional.

2.3.1.2 Sinonimia

[4] La colocación es referida como la tendencia que tienen algunas palabras a co-ocurrir en un idioma determinado. Tomando el mismo ejemplo que utilizan estos autores, se puede ver que existe una relación de colocación entre "smoking" y "pipe", lo cual hace que la ocurrencia de "pipe" en la cuarta línea sea cohesiva [14].

> "A little man of Bombay/ Was smoking one very hot day/ But a bird called a snipe/ Flew away with his pipe/ Which vexed the fat man of Bombay"

La sinonimia o igualdad de semas[5] ocurre cuando se repite el significado de un elemento léxico mediante otro elemento léxico o una frase. A continuación se muestran dos ejemplos de sinonimia.

Ejemplo 1: *cuaderno* y *libreta*

Ejemplo 2: *Ocaso* y *Caída del sol*

Algunos autores plantean que no existe la sinonimia absoluta entre dos palabras o frases, y que el uso de una u otra se determina por el contexto. Por ejemplo, dos palabras como “planta” y “fábrica”, que al parecer resultan muy similares, no siempre pueden utilizarse como sustituta una de la otra; o sea, decir “¡Cómo ha crecido la fábrica en los últimos meses!” no es lo mismo que decir ““¡Cómo ha crecido la planta en los últimos meses!”. Lo anterior muestra que la sinonimia, pese a ser un buen indicador de la continuidad de los tópicos en el texto, hace que los métodos de segmentación que lo utilizan sean dependientes del dominio.

2.3.2 Sintagmas de entrada

Los sintagmas de entrada son indicadores lingüísticos de la estructura discursiva; en ocasiones, se denominan palabras indicio. Éstos se encargan de guiar las inferencias que se realizan en la comunicación. Los sintagmas de entrada son expresiones tales como: pues bien, en primer lugar, por su parte, dicho sea de paso y otras.

Algunos autores emplean sintagmas de entrada para identificar cambios de tópicos que son relativamente independientes del dominio [25]. Otros, en cambio, utilizan expresiones indicios que son específicos del dominio. Esto implica que deba crearse una lista nueva de sintagmas de entrada para cada fuente distinta de procedencia de los documentos que serán segmentados. Realizar este trabajo manualmente es muy costoso, mientras que automatizarlo también demanda un notable trabajo manual, requiriendo la creación de *corpus* anotados. No obstante, los sintagmas específicos del dominio se consideran más fiables para indicar los cambios de tópicos; por ejemplo, en el dominio de los flujos de transmisiones continuas de noticias es posible encontrar sintagmas de entrada muy específicos como, por ejemplo, saludos (buenos días o buenas tardes), los cuales ocurren casi siempre al inicio de un segmento de transmisión. Reinar, por ejemplo, utilizó una lista de sintagmas de entradas que denominó *domain cues* [36], [37]. Esta lista fue construida manualmente y separada en varias categorías (nueva persona, saludos, comienzos introductorios, presentación a próximas historias o de otros locutores); él hace una división en categorías porque considera que no todos los sintagmas de entrada señalan un cambio de tópico con la misma fuerza.

2.3.3 Entidades nombradas

Las entidades son objetos en el mundo; por ejemplo, lugares o personas. El nombre de una entidad es una frase que se refiere de manera única al objeto correspondiente, ya sea por su nombre propio, acrónimo, apodo o abreviación. Algunos ejemplos de nombres de identidad son: Rosa María, Castillo del Morro, CENATAV, etc. La anotación o identificación de entidades

[5] Sema es la unidad mínima de significado lexical o gramatical (tomado de la RAE). El conjunto de todos los semas de una palabra es el significado o semema (tomado de Belaaichi, A.: CAMPO SEMANTICO. Material docente.King Saud University).

nombradas siempre se hace de acuerdo al significado que éstas tienen en su contexto; o sea, la anotación de las entidades depende de cómo se usan. Los distintos estudios sobre este fenómeno han llegado al acuerdo de que existen cuatro tipos de entidades nombradas [36]:

- *Persona*: las entidades de persona están limitadas a humanos identificados por un nombre, apodo o alias.
- *Título/Rol*: títulos personales o roles. Están limitados a títulos que se encuentran cerca del nombre de la persona a la que describen.
- *Organización*: las entidades de organización están limitadas a corporaciones, instituciones, agencias de gobierno y otros grupos de gente definidos por una estructura organizacional establecida.
- *Lugar:* las entidades de lugar incluyen nombres de lugares definidos política o geográficamente (ciudades, provincias, países, regiones internacionales, conjuntos de agua, montañas). Los lugares incluyen también estructuras hechas por el hombre como aeropuertos, autopistas, calles, fábricas y monumentos.

En el dominio de los flujos de transmisiones continuas de noticias, las entidades nombradas son muy útiles para la detección de los cambios de tópicos por su poder discriminante. La reiteración de un mismo nombre de persona, organización o lugar, es poco probable que se produzca en noticias sobre distintos tópicos, convirtiéndose en un indicio confiable de que dos piezas de textos están dentro del mismo tópico.

2.3.4 Primer uso de la palabra

Como se vio anteriormente en las secciones sobre la cohesión léxica y las entidades nombradas, los cambios del uso de un conjunto de términos léxicos (vocabulario) permiten detectar, con un buen grado de confianza, los cambios de tópicos. Esta sección, también, refiere al primer uso de un conjunto de términos léxicos para indicar la entrada a nuevos tópicos. El número de palabras usadas por primera vez en un documento, lógicamente disminuye a medida que avanza el discurso porque el vocabulario del autor es finito; además, también puede observarse que las ocurrencias de nuevos grupos de palabras en un documento suelen coincidir con los cambios de subtópicos [21].

2.4 Preprocesamiento

Como se mencionó en la sección introductoria de esta sección, existe una etapa prácticamente inviolable antes de comenzar cualquier tarea de procesamiento textual, conocida comúnmente como preprocesamiento. La necesidad del preprocesamiento se debe fundamentalmente al aumento considerable de la diversidad o heterogeneidad del formato de los documentos digitales y tiene el objetivo de mejorar el desempeño de las tareas propias de procesamiento. En esta etapa se genera como resultado un conjunto de palabras o términos más pequeño y de mayor calidad que el original.

En la etapa de preprocesamiento el texto sufre una serie de transformaciones necesaria para facilitar la extracción de información y por tanto, facilita la detección de unidades textuales relacionadas con un mismo tópico. Dentro de éstas se destacan la confección del texto plano, la

eliminación de las palabras vacías y la extracción de raíces o lemas, las que serán más detalladas en las secciones que continúan.

Es indispensable destacar que no todos los autores consideran necesario realizar cada una de las etapas mencionadas, e incluso existen algunos que incluyen otras más específicas.

2.4.1 Conversión de documento a texto plano

Muchos de los formatos de textos digitales son propietarios como, por ejemplo, los "*.doc*". Esto hace que sea necesario llevarlos a un formato que sea abierto como los "*.txt*" [6]. Además, para aumentar el rendimiento de los sistemas de procesamiento de textos es necesario hacer más ligeros los documentos; o sea, minimizar el espacio que ocupan en disco. Por otra parte, para la mayoría de las tareas de procesamiento de texto, lo que suele ser importante es el contenido del mismo y no su formato. Una de las formas más populares de resolver estas problemáticas es la conversión de los documentos a archivos de texto plano.

El texto plano, texto llano, o texto simple, como también se le conoce, son sólo caracteres, sólo texto sin formatear; es decir, sin códigos de tipos de letras, negritas, cursivas, formatos de párrafos, etc.

En la etapa de creación del texto plano, en ocasiones se reducen las mayúsculas a minúsculas, y generalmente se identifican y eliminan los signos de puntuación y los acentos, estos últimos son frecuentes en algunos idiomas como, por ejemplo, el español.

Después de este tratamiento queda eliminada toda la información que puede resultar superflua en los documentos y se obtiene un fichero que está listo para ser leído y procesado por cualquier sistema.

2.4.2 Reducción de palabras vacías

Las palabras vacías o *stop words* son palabras que se consideran carentes de utilidad; o sea, que están carentes de todo significado para alguna tarea o intención; por ejemplo, en la segmentación, los artículos no son adecuados a la hora de determinar la similitud por repetición de términos entre unidades textuales, debido a que es muy probable que aparezcan en casi todas.

Entonces, previamente a la segmentación, se crea una lista de términos vacíos y se verifica la presencia de cada palabra en la misma. Esta lista se forma por las preposiciones, conjunciones, artículos, pronombres, así como todas aquellas palabras que suelen ser poco discriminantes por su elevada frecuencia de aparición en el texto.

2.4.3 Extracción de raíces o lemas

Las tareas de extracción de raíces y la extracción de lemas pertenecen al nivel morfológico del procesamiento del lenguaje natural, pero los términos lema y raíz léxica, en ocasiones, tienden a ser confundidos, por lo que se creyó necesario definir cada uno para una mejor comprensión de las diferencias entre ambas tareas. El objetivo principal de dichas tareas es obtener, en el mínimo número de caracteres posibles, el máximo de información del término.

[6] Los formatos propietarios están protegidos por una patente o derechos de autor. Los abiertos, en cambio, están públicos y son patrocinados, habitualmente, por una organización de estándares abiertos, y libre de restricciones legales de uso.

La raíz léxica o lexema es la unidad léxica primaria de una palabra, que lleva los aspectos más significativos del contenido semántico y que no se puede reducir en componentes más pequeños. Por ejemplo, los términos *hablan* y *hablando* se reducirían a la raíz *habl*.

El lema es cada una de las entradas de un diccionario o enciclopedia. El lema define un conjunto de palabras con la misma raíz léxica, y que pertenece a la misma categoría gramatical (verbo, adjetivo, etc.). La lematización pretende normalizar los términos pertenecientes a una misma familia y por tanto próximos en significado, reduciéndolos a una forma común o lema, que no coincide necesariamente con la raíz. Por ejemplo, los términos *hablan* y *hablando* se reducirían al lema *hablar*.

2.5 Trabajos relacionados

En esta sección se expondrán las cuestiones principales de algunos de los métodos más relevantes dirigidos a la identificación de unidades de tópicos en el texto, haciendo énfasis en los que resultaron estar más relacionados con el problema a resolver. Dichas cuestiones son básicamente: el tipo de texto sobre el que se trabaja, las señales lingüísticas utilizadas para determinar la similitud entre las unidades textuales y los recursos utilizados, los criterios empleados para identificar los cambios de tópicos, así como las deficiencias fundamentales de estos métodos.

2.5.1 Segmentación por tópicos globales de Ponte y Croft

Ponte y Croft, en 1997, propusieron un método de segmentación que tiene como objetivo el seguimiento de tópicos de un programa de transmisión de noticias y la identificación de tópicos en una base de datos documental [34]. Este trabajo se enfoca en textos que tienen oraciones relativamente pequeñas, en los que las oraciones dentro de los segmentos de tópicos tienen relativamente pocas palabras en común; estas características realmente tornan más complejo el problema de la segmentación. Ponte y Croft consideraron que las oraciones son las unidades mínimas de segmentación; o sea, no se identifican segmentos menores a una oración. Estos autores hacen uso de una técnica de expansión de consultas, mediante la cual intentan encontrar rasgos comunes entre las oraciones para facilitar la identificación de aquellas que corresponden a un mismo tópico. Su método está divido en cuatro etapas fundamentales.

Como primer paso, se intenta encontrar las palabras o frases semánticamente relacionadas entre un par de oraciones, usando el método de Análisis de Contexto Local (LCA, por sus siglas en inglés) propuesto por Xu y Croft en [27]. Con el LCA cada oración original es vista como una consulta a la base de datos del LCA y, a partir de cada consulta, se retornan las 100 palabras o frases más asociadas con ella. Estas palabras o frases conforman conceptos que serán utilizados en lugar de la oración original.

Luego, se define una ventana[7] de oraciones de tamaño fijo con la cual se recorre todo el texto para determinar la similitud de los conceptos asociados a las oraciones de cada ventana. La similitud entre los conceptos de dos oraciones se determina según la cantidad de conceptos que

[7] *Ventana* es un término común en la segmentación por tópicos, este se refiere a una pieza o bloque de texto formado por un cierto número de unidades textuales, según el interés de cada autor.

ellas tienen en común. Los autores utilizan una ventana para eliminar cálculos de similitud innecesarios entre oraciones que están muy distantes en el texto.

El tercer paso consiste en asignar a cada ventana una puntuación como posible segmento. Dicha puntuación se define como la suma de la similitud interna de la ventana más las sumas de las dos similitudes externas, derecha (o por debajo) e izquierda (o por arriba). Tales similitudes fueron definidas de la siguiente forma:

- La similitud interna: es la suma de todos los valores de similitud entre las oraciones de la ventana.
- La similitud externa derecha: es la suma de los valores de similitud entre cada oración de la ventana con cada oración de la ventana adyacente derecha.
- La similitud externa izquierda: es la suma de los valores de similitud entre cada oración de la ventana con cada oración de la ventana adyacente izquierda.

Finalmente, se pasa a determinar los segmentos, para lo que se considera cada posible segmentación; o sea, cada posible secuencia de tópicos en los que pueda ser dividido el texto. El objetivo es encontrar la mejor segmentación en base a la puntuación de cada ventana con el menor costo computacional posible, para lo que se usa un método de programación dinámica.

Este método constituye una propuesta interesante para resolver el problema de la segmentación en textos muy pequeños donde la cantidad de palabras en común que tienen las unidades textuales es muy poca e incluso puede ser nula. Pero, por lo general, los métodos que realizan expansión de consultas tienen la dificultad de que se restringen sólo a los textos que coinciden con el idioma de la base de conocimiento utilizada, la cual comúnmente es un diccionario electrónico o un tesauro[8]. Por otra parte, usualmente el procesamiento de dichas bases de conocimiento requiere de un alto esfuerzo ingenieril. Además, generalmente se corre el riesgo de que en el proceso de expansión se obtengan muchos términos espurios, causando un solapamiento entre los conceptos de las oraciones, lo que provocaría imprecisiones en la segmentación.

2.5.2 Segmentación por tópicos globales de Stokes, Carthy y Smeaton

Stokes, Carthy y Smeaton en 2004 propusieron un sistema llamado SeLeCT, con similar objetivo a la propuesta de Ponte y Croft [45], [46]. Este sistema toma un fichero que contiene un flujo de transmisión continua de noticias con la intención de retornar segmentos del fichero que contengan noticias individuales. Para ello, los autores se enfocan en la identificación de secuencias léxicas en el fichero, definidas como un cluster de palabras semánticamente similares, bajo el supuesto de que estos clusters de palabras pueden coincidir con noticias individuales.

En el proceso de segmentación de SeLeCT se distinguen tres etapas. En la primera se seleccionan y preprocesan los términos que los autores consideraron claves; algunos de éstos fueron sustantivos comunes, sustantivos compuestos y adjetivos.

En la segunda etapa se crean las secuencias léxicas, buscando las relaciones entre los términos resultantes de la primera etapa, utilizando el tesauro WordNet y algunas reglas de estadísticas de co-ocurrencia (por ejemplo, Osama Bin Laden y the World Trade Center) para determinar los términos semánticamente similares. El procedimiento se basa en un algoritmo de

[8] Un tesauro es una herramienta que permite encontrar las palabras que mejor expresan un concepto, a diferencia de los diccionarios que explican el significado de las palabras.

clustering single-pass, donde la primera secuencia léxica se obtiene a partir del primer término, que se toma como semilla. Luego, cada término subsiguiente se añade a una secuencia existente si éste está relacionado con al menos otro término en dicha secuencia. Además de la similitud, el método exige que el término se añada a la última secuencia actualizada más similar a él. Por otra parte, exige que la distancia entre dos términos relacionados sea menor que un número máximo de términos que estará en correspondencia con la fuerza de la similitud entre los términos, o sea, mientras mayor sea la similitud mayor será la distancia permitida entre los dos términos. En resumen, el procedimiento básicamente consiste en añadir un término a la secuencia si éste se considera "aceptable" bajo las condiciones anteriormente mencionadas, de lo contrario, este término será la semilla de una nueva secuencia, así hasta que todos sean ubicados.

Por último, en la tercera etapa se identifican los límites de los segmentos. Para ello, primeramente se determina la fuerza del límite entre cada par de oraciones consecutivas del texto, $w(i, i+1)$, donde la fuerza del límite entre las oraciones i y $i+1$ será la suma del número de secuencias léxicas que terminan en la oración i más el número de las que comienzan en la oración $i+1$. Cuando se ha calculado la fuerza del límite entre todas las oraciones consecutivas, se obtiene la media de estos valores con aquellos que son distintos de cero. Esta media se considera como la fuerza mínima permisible para que un punto entre dos oraciones consecutivas sea considerado cómo límite de segmento.

Luego, estos posibles límites son filtrados, identificando aquellos que distan menos de una cota máxima de oraciones y dejando solamente el que tenga el valor de fuerza más alto. Cuando los valores de fuerza coinciden, se escoge el más lejano en el texto. Estos autores consideraron que la cota debe ser un valor tan pequeño que no sea una longitud razonable para una noticia.

SeLeCT tiene una característica relevante que consiste en asumir que las palabras que tienden a co-ocurrir juntas en el contexto de las noticias suelen estar relacionadas semánticamente, lo que permite identificar con mayor exactitud las noticias individuales en un flujo de transmisión continua. Con este fin también utiliza el tesauro WordNet, pero esto trae como consecuencia la presencia de algunas de las principales dificultades de la propuesta de Ponte y Croft, en cuanto a la imposibilidad de segmentar otros textos escritos en un idioma que no corresponda con el del tesauro.

2.5.3 Segmentación jerárquica del discurso de Morris y Hirst

El trabajo de Morris y Hirst en 1991 se destaca dentro de los muy pocos métodos de segmentación dirigidos a determinar la estructura jerárquica del discurso. En esencia, estos autores, al igual que Stokes, Carthy y Smeaton, identifican secuencias léxicas en un texto pero, en este caso, para reconocer la estructura jerárquica del discurso propuesta por Grosz y Sidner [31]. Ellos consideran que dichas secuencias son un buen indicador de la estructura del texto, bajo el supuesto de que éstas son un resultado directo de unidades textuales que tratan sobre una misma cosa; o sea, que las secuencias léxicas determinan segmentos de textos con una fuerte unidad de significado. Tales secuencias se definieron como: secuencias o cadenas de texto formadas por palabras cercanas relacionadas mediante cohesión léxica.

Morris y Hirst comienzan por seleccionar las palabras candidatas a formar parte de la secuencia léxica. Ellos no consideraron los pronombres, las preposiciones ni otras palabras de alta frecuencia en el texto.

Luego, para determinar las secuencias léxicas, se apoyaron en la cuarta edición del *Roget's International Thesaurus*, desarrollada en 1977. Este tesauro no estaba en un formato legible para ser leído por la computadora, debido a esto, los autores se vieron obligados a construir manualmente las secuencias léxicas.

El tesauro agrupa las palabras por categorías básicas, y cuenta con un índice que indica en cuál categoría está cada palabra. Por otra parte, hay tres niveles anteriores al nivel de las categorías básicas; el nivel superior está formado por ocho clases (*abstract relations, space, physics, matter, sensation, intellect, volition, and affections*). Cada clase se dividió en subclases, y estas en sub-subclases, adquiriendo el tesauro una estructura jerárquica. Las categorías están separadas en pares que tienen como etiquetas palabras que son antónimos; por ejemplo *life* y *death*. Cada categoría contiene una serie de párrafos para agrupar las palabras más relacionadas. Dentro de cada párrafo, los grupos de palabras mucho más relacionadas se separan por punto y coma, y estos grupos pueden tener referencias cruzadas o punteros a otras categorías o párrafos relacionados.

Para construir las secuencias, se consideró que dos palabras pertenecían a una misma secuencia si se encontraban una de otra a una distancia de pocas oraciones y si estaban relacionadas mediante cohesión léxica, cumpliendo alguna de las cinco condiciones siguientes:

1. Tienen una categoría común en sus índices de entradas.
2. Una tiene una categoría en su índice de entrada que contiene un puntero a una de las categorías de la otra.
3. Una es etiqueta de una de las categorías de la otra.
4. Tienen categorías que están en clases o subclases relacionadas semánticamente.
5. Tienen una categoría que tiene un puntero a la misma categoría.

Los autores también identifican cuáles secuencias están a continuación de otras y cuáles están incluidas dentro de otras.

Este método logra segmentar los textos a un nivel fino de detalle; pero, como puede observarse, el mismo es muy dependiente de la estructura del tesauro que se utilice y del idioma del mismo. Además, debe notarse que acceder y procesar automáticamente dicho tesauro para identificar las secuencias léxicas, teniendo en cuenta las cinco condiciones especificadas por los autores, requiere de un gran esfuerzo ingenieril. Por otra parte, seleccionar las palabras que están relacionadas por cohesión léxica mediante este proceso, donde no se cuantifica la fuerza de dicha relación, restringe la selección a los criterios utilizados para construir el tesauro.

2.5.4 Segmentación jerárquica del discurso de Gruenstein, Niekrasz y Purver

Gruenstein, Niekrasz y Purver en 2005 y 2006 crearon una arquitectura de un asistente automático de oficina para representar, anotar y analizar el discurso desarrollado durante una reunión [12], [32]. Su objetivo es crear componentes que permitan comprender y resumir una reunión, así como ayudar a la confección colaborativa de documentos durante el curso de la misma. Como parte de esta herramienta, ellos propusieron un esquema de anotación de reuniones, enfocándose en dos tipos de estructuras de anotación. Una para marcar aquellas partes más relevantes a la reunión o aquellas partes de la reunión donde se toman acuerdos que deben ser cumplidos por algunos participantes luego de concluida la reunión. Y otra para la segmentación secuencial y jerárquica de la reunión en diferentes tópicos.

Para la segmentación por tópicos, estos autores propusieron un esquema de anotación jerárquico de dos niveles. En el nivel superior, la reunión se segmenta completa y

secuencialmente. Los límites de segmentos se ponen en los puntos del discurso donde ocurren interrupciones muy notables, además, en aquellos puntos a partir de los cuales el tópico del discurso cambia considerablemente. En el nivel inferior del esquema, los segmentos mayores son opcionalmente subsegmentados sin que exista solapamiento entre los nuevos segmentos que se obtienen. Los segmentos menores significan una digresión temporal o una discusión más enfocada del tópico del segmento mayor.

Se crearon cuatro nombres de tópicos reservados para aquellas partes que suelen ser estándares a todas las reuniones:

- Agenda: parte de la reunión en que la agenda se presenta y se discute.
- Introducción: discurso que da inicio oficial a la reunión.
- Final: discurso que concluye oficialmente la reunión.
- Dificultades técnicas: un período de la reunión en que hay dificultades técnicas con el equipo magnetofónico.

Dígitos: porción dedicada a tareas de lectura de dígitos que se encuentran en el corpus de reuniones ICSI[9].

Salvo la Agenda, los autores consideraron que el resto de los nombres reservados tienen el propósito de resaltar porciones de la reunión que no se consideran partes propias de ésta. Además de estos nombres, se les da la posibilidad a los anotadores (personas que realizan las anotaciones) de poner nombres descriptivos a los tópicos que identificaban sin ninguna restricción de formato. Por otra parte, los anotadores están libres de asignar límites de segmentos en cualquier parte del discurso; por ejemplo, un cambio de locutor no tenía que coincidir necesariamente con un cambio de tópico, en cambio un tópico podía comenzar y terminar durante la intervención de un locutor.

Estos autores comentaron que se encontraban desarrollando un segmentador automático, entrenando un clasificador con las anotaciones que ellos obtuvieron de 65 reuniones de los corpus ICSI y ISL [8], [26]. No obstante, este esquema de segmentación representa un primer paso muy positivo, debido a que son muy pocos los trabajos de segmentación jerárquica que se han realizado hasta el momento. Además, estos autores experimentan en un dominio que tampoco ha sido muy explorado, discursos en los que intervienen más de un locutor.

2.5.5 Segmentación lineal del discurso de Kozima

Dentro de esta problemática se destaca como uno de los primeros trabajos el de Kozima en 1993; éste tiene como foco de atención los textos narrativos[10] [29]. Este autor propuso un indicador de la estructura del texto, al cual llamó *Lexical Cohesion Profile,* (LCP). Con el LCP se registra la cohesión léxica mutua entre todas las palabras de una cadena de texto y, en base a estos valores, se identifican los límites de los segmentos bajo el supuesto de que un alto valor de cohesión refleja en buena medida la unidad semántica del segmento.

[9] El corpus ICSI tiene una porción dedicada a la tarea de lectura de dígitos, en la que los participantes de las reuniones leen en voz alta largas cadenas de dígitos. Esta tarea fue designada para proveer un conjunto de entrenamiento de vocabulario restringido para desarrolladores del reconocimiento del habla.

[10] Un texto narrativo responde a “qué pasa”. En estos textos se cuentan hechos reales o de ficción que le suceden a los personajes que participan. Tales hechos conducen al lector de una situación final a una inicial.

El LCP de un texto *T* de n palabras, $T = \{w_1, w_2, \ldots, w_n\}$, se definió como una secuencia de cohesiones léxicas $LCP = \{c(S_1), c(S_2), \ldots, c(S_n)\}$, donde S_i es una cadena de texto que se forma mediante una ventana de palabras de tamaño fijo, que tiene su centro sobre la i-ésima palabra de *T*.

El primer paso del método de Kozima es determinar la cohesión léxica $c(S_i)$ de las cadenas de textos, calculando la similitud entre las palabras de la cadena con ayuda de una red semántica que se va construyendo sistemáticamente desde el diccionario inglés *Longman Dictionary of Contemporary English*, LDOCE, [28]; esta red fue nombrada *Paradigme*. Cada palabra *w* del texto se asocia a un nodo en la red *Paradigme*. Para cada palabra se determina un valor de importancia $s(w) \in [0, 1]$. Este valor es la relación que existe entre la frecuencia de la palabra w en un corpus determinado y la cantidad de palabras de dicho corpus[11]. Además, para cada secuencia de texto S_i existe un patrón de activación $P(S_i)$ que se produce activando el nodo de cada $w \in S_i$ con una fuerza de $s(w)^2 / \sum s(w_i)$. Entonces $c(S_i)$ se determina mediante la siguiente expresión:

$$c(S_i) = \sum_{w \in S_i} s(w) a(P(S_i), w), \qquad \textbf{(1)}$$

donde $a(P_i(S_i), w)$ es el valor de actividad del nodo asociado con *w* en el patrón $P(S_i)$. El autor intenta representar en $c(S_i)$ la homogeneidad semántica de S_i.

Luego, considerando los valores registrados por el LCP, se especifica un límite de segmento bajo las siguientes suposiciones:

1. Si la secuencia S_i está dentro de un segmento, entonces S_i tiende a ser cohesiva y el valor de $c(S_i)$ tiende a ser alto.
2. Si la secuencia S_i atraviesa un límite de segmento, entonces S_i tiende a variar semánticamente y el valor de $c(S_i)$ tiende a ser bajo.

Estos autores emplean un método muy interesante para calcular la similitud entre las palabras, que permite determinar la cohesión léxica de una pieza de texto en la cual existen pocas palabras repetidas, como es el caso de los textos narrativos, brindando información sobre la homogeneidad semántica de dicha pieza. No se conoce que el método de segmentación que estos autores proponen haya sido probado para otros géneros. Una de las causas es que requiere una red semántica para calcular las puntuaciones del LCP, la cual no está públicamente disponible [36]. Por otra parte, para aplicar este método sobre otro idioma distinto al inglés sería necesario contar con diccionarios del idioma de interés que permitan un tratamiento computacional.

[11] Para estimar la importancia de una palabra se utilizó el *West's corpus* (1953), según los autores este corpus contaba con 5,487,056. Por ejemplo, la importancia de la palabra *red* y la palabra *and,* que aparecieron con una frecuencia de 2,308 y 106,064 respectivamente, se determinó de la siguiente forma:

$$s(red) = \frac{-\log(2308/5487056)}{-\log(1/5487056)} = 0{,}500955,$$

$$s(and) = \frac{-\log(106064/5487056)}{-\log(1/5487056)} = 0{,}254294.$$

2.5.6 Segmentación lineal del discurso de Hearst

Otra de las propuestas de segmentación lineal es la de Hearst de 1993 a 1997, que constituye uno de los estudios más interesantes y completos sobre la identificación de estructuras de subtópicos [15]-[22]. Hearst propuso un método que intenta dividir textos explicativos[12] en unidades de discurso de múltiples párrafos, al que denominó TextTiling. El autor asumió que este tipo de texto tiene una estructura monolítica por partes y que la misma puede ser reconocida utilizando más de una señal lingüística como, por ejemplo, la cohesión léxica o el primer uso de la palabra, suponiendo que: si un grupo de términos léxicos o vocabulario se usa durante el curso de la discusión de un subtópico y este subtópico cambia, entonces una porción significativa del vocabulario cambia también.

TextTiling se inicia con una fase de preprocesamiento. En dicho preprocesamiento se eliminan los *stopwords* y se extraen las raíces léxicas de los términos; además, los documentos se dividen en secuencias, o seudo-oraciones, de un tamaño predefinido de los términos resultantes sin considerar los signos de puntuación.

Luego, se procede a determinar una puntuación léxica para los espacios entre grupos de seudo-oraciones. TextTiling propone un método para calcular dicha puntuación que se basa en la repetición de términos como mecanismo de cohesión léxica[13]. Dicho método de puntuación compara bloques adyacentes de seudo-oraciones y asigna una puntuación de similitud entre estos bloques, teniendo en cuenta la cantidad de palabras que ellos tienen en común. Los bloques se forman por una cantidad especificada de seudo-oraciones, se representan mediante el modelo de espacio vectorial y la similitud entre ellos se calcula usando la medida del coseno.

Sean dos bloques de texto b_1 y b_2, cada uno con k seudo-oraciones, donde $b_1 = \{s_{i-k}, \ldots, s_i\}$ y $b_2 = \{s_{i+1,} \ldots, s_{i+k+1}\}$, la puntuación léxica del espacio i entre estos bloques, corresponde a cuán similares son las seudo-oraciones desde la i-1 a la i con las seudo-oraciones desde la $i + 1$ a la $i + k + 1$.

Finalmente se pasa a la identificación del límite. Teniendo en cuenta la puntuación léxica, se asigna una puntuación de profundidad a cada espacio entre oraciones donde ocurra un valle. Un valle, según el autor, son los puntos donde baja dicha puntuación léxica; el autor para determinar esto utiliza un valor umbral que se basa en el promedio de las puntuaciones léxicas.

La puntuación de profundidad del valle corresponde a cuán fuertemente cambiaron las señales para un subtópico a ambos lados del valle, basándose en la distancia desde el valle a los dos picos que lo forman. En otras palabras, si una baja puntuación léxica es precedida y sucedida por una alta puntuación léxica, esto se asume como indicador de un cambio en el vocabulario, que corresponderá, según lo supuesto, con un cambio de subtópico. Lo anterior se puede ilustrar mediante un ejemplo hipotético, utilizando de apoyo la Fig. 2.2, similar a la que usó el autor en su trabajo, teniendo en cuenta que el eje x representa los espacios entre los bloques y el eje y las

[12] Un texto explicativo se desarrolla en base “al por qué y al cómo”; o sea, no se limita a informar, sino que se define por su intención de hacer comprender a su destinatario por qué un fenómeno o un acontecimiento actúa de un modo determinado.

[13] Hearst propone el uso de más de tres métodos para calcular las puntuaciones léxicas, repetición de términos en un bloque de texto, primer uso de la palabra, y confección de secuencias léxicas de términos relacionados, pero este ultimo método, los autores decidieron no incluirlo en TextTiling. En este trabajo sólo se hace alusión al primero porque tiene varios aspectos coincidentes con el método que se propone en este trabajo de tesis.

puntuaciones léxicas de estos espacios. Según este ejemplo, la puntuación de profundidad para el punto *i* será $(y_{i1} - y_{i2}) + (y_{i3} - y_{i2})$.

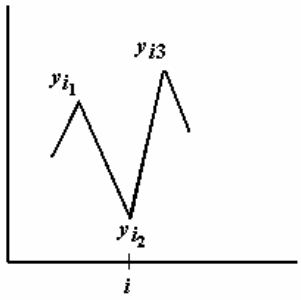


Fig. 2.2. Curva de puntuación léxica de los espacios entre los bloques

Luego, las puntuaciones de profundidad se ordenan y se usa este orden para determinar los límites de los segmentos, siendo las posiciones con puntuaciones más altas las de mayor probabilidad para que ocurran los límites.

Este algoritmo puede resultar adecuado si se aplica sobre documentos científico-técnicos, ya que éstos, generalmente, están formados por múltiples párrafos que explícitamente explican o enseñan sobre un tópico, y en que suelen repetirse frecuentemente las palabras mas relacionadas con dicho tópico. Además, no hace uso de tesauros o diccionarios electrónicos para identificar las unidades textuales relacionadas por cohesión léxica, sino que emplea el modelo de espacio vectorial para representar dichas unidades y la medida del coseno para calcular la similitud entre éstas. Pero presenta una dificultad que provoca que segmentos que contienen subtópicos simples sean interrumpidos; esta dificultad provoca, además, que la cantidad de segmentos obtenidos sobrepase considerablemente la cantidad que es considerada como válida. Esto ocurre cuando existe un párrafo corto u otro como, por ejemplo, citas textuales o párrafos que ejemplifiquen una determinada situación, que interrumpe una cadena de párrafos cohesionados. TextTiling no detecta esto; en cambio, cuando la puntuación léxica decrece notablemente en una zona del texto, se reconoce como un valle y es muy probable que se asigne un límite de segmento.

2.5.7 Segmentación lineal del discurso de Heinonen

Heinonen en 1998, similar a Hearst, propuso un método para la segmentación lineal de textos de múltiples párrafos. Pero, a diferencia de Hearst, su método emplea una ventana que recorre todo el texto y determina, para cada párrafo, su párrafo más similar dentro de ella, con vista a disminuir el efecto que tienen sobre la segmentación algunos párrafos como los mencionados en el caso de Hearst [23]. Este método de segmentación es esencialmente útil cuando se necesita controlar la longitud de los segmentos. Heinonen usa un método de programación dinámica para garantizar como resultado una segmentación óptima, teniendo en cuenta la longitud y la cohesión léxica de los segmentos que se obtengan.

Primeramente, al igual que en TextTiling, el autor propone que el texto pase por una etapa de preprocesamiento para eliminar los *stopwords* y reducir la palabras a su raíz léxica. Luego, los párrafos se representan usando el modelo de espacio vectorial y la similitud entre los párrafos se

calcula, basada en la repetición de términos, mediante la medida del coseno, también similar a como se hizo en TextTiling.

Posteriormente, se construye un vector de cohesión *(Cohe$_1$…Cohe$_n$)* con todos los párrafos del documento, donde a cada párrafo se le asocia el valor de similitud más alto que se obtuvo dentro de su ventana, que está formada por varios párrafos a su alrededor, párrafos por encima y párrafos por debajo.

Luego, se procede a determinar los límites de segmento utilizando un método de programación dinámica. El método considera todas las segmentaciones posibles y determina la de mínimo costo. El algoritmo calcula de forma secuencial, del primero al último, el mínimo costo de segmentación por párrafo considerando la siguiente expresión:

$$Cost_i = min(CostS(S_i^i)...CostS(S_i^1)), \quad \textbf{(2)}$$

$$Cost_0 = 0 \ , \ Cohe_0 = 0, \quad \textbf{(3)}$$

$$CostS(S_i^k) = Flon(S_{ik}) + Cohe_{k-1} + Cost_{k-1}, \quad \textbf{(4)}$$

$$Flon(S_{ik}) = clen("cantidad\ de\ palabras\ en\ S_{ik}", p, h), \quad \textbf{(5)}$$

donde Cost$_i$ es el costo de segmentar el texto que se forma desde el párrafo *i* hasta el primer párrafo, como si esta porción fuera un texto independiente; o sea, en cada iteración se divide el problema de segmentar el texto completo en el subproblema de segmentar sólo la porción de texto que se forma desde cada párrafo hasta el primero.

Cada porción de texto puede segmentarse de varias formas $S_i^i, S_i^{i-1}, ..., S_i^1$, según la cantidad de párrafos que éste tenga, donde S_i^i corresponderá a la segmentación que separa al párrafo *i* del resto, S_i^{i-1} a la segmentación que incluya a *i* en un segmento con *i*-1 pero separados del resto, así sucesivamente. Esto hace que se tenga un costo $CostS(S_i^k)$ por cada forma de segmentación, que se determina en relación a la longitud $Flon(S_{ik})$ del segmento S_{ik}, a la cohesión léxica del párrafo *k*-1 con su entorno, *Cohe*$_{k-1}$, y el costo de la solución anterior, *Cost*$_{k-1}$; o sea, el costo de la segmentación de la porción de texto del párrafo *k*-1 al primero.

Como puede verse, el algoritmo considera la longitud de los segmentos. Para esto utiliza una función de costo de longitud, *clen(x, p, h)*, que determina la correspondencia entre la longitud de un segmento y la longitud deseada para éste, donde *x* es la longitud real del segmento, *p* la longitud deseada, y *h* un parámetro de escala para ajustar el peso de las longitudes.

Entonces, como el objetivo es obtener la segmentación de mínimo costo, *Cost*$_i$ será igual al menor de todos los costos de segmentar la porción de texto correspondiente a *i*. Además, por cada párrafo se determina su límite de segmentación, que será el último párrafo del segmento anterior al segmento que lo contiene. Este límite queda determinado por la expresión:

$$LimP_i = k - 1 \ \ donde \ \ Cost_i = CostS(S_i^k). \quad \textbf{(6)}$$

Lo anterior se puede ilustrar mejor mediante el siguiente ejemplo hipotético: Si se tiene un texto de tres párrafos, se comienza por determinar el costo de segmentación hasta el párrafo 1, el cual depende solamente de su longitud, porque sólo hay una única forma de segmentarlo. Luego, en el segundo paso, se comienza a complicar el proceso, porque el costo de la segmentación hasta el párrafo 2 ($Cost_2$) ya depende de la primera solución, $Cost_2 = min(CostS(S_2^2), CostS(S_2^1))$. Ya en el tercer paso, la segmentación se complica aún más, quedando $Cost_3 = min(CostS(S_3^3), CostS(S_3^2), CostS_3^1)$, donde el costo de S_3^3 dependerá de la solución tomada en el segundo paso, o sea, $CostS_3^3 = Flon(S_{33}) + Cohe_2 + Cost_2$.

Heinonen logra determinar una correspondencia óptima entre la longitud de los segmentos que se obtienen, la longitud deseada para éstos y el valor de similitud asociado con cada párrafo, y logra disminuir el efecto de los párrafos que pueden interrumpir un segmento. Sin embargo, su método también tiene un inconveniente. El vector de cohesión del documento asocia cada párrafo con el valor de similitud más alto en su ventana, pero este valor puede pertenecer tanto a la similitud con un párrafo que esté por encima como a un párrafo que esté por debajo del párrafo en cuestión. El algoritmo – teniendo en cuenta tal valor y no distinguiendo esta situación – puede decidir la inclusión de un párrafo en un segmento que está por debajo de él. Como puede observarse, permitir que la alta similitud se observe con párrafos por encima, para decidir la inclusión de un párrafo en un segmento por debajo de él, es incorrecto. Esto debilita uno de los presupuestos del método, posibilitando que obtengan segmentos de baja cohesión léxica en los que existan párrafos que no sean similares al resto de los párrafos que lo forman y que, en cambio, lo sean a otros que se encuentran en un segmento contiguo. Además, este método tiene otra dificultad, requiere de la especificación de la longitud aproximada de los subtópicos, que realmente es un valor impredecible y que no suele ser el mismo para todos los subtópicos de un texto. Otra dificultad relacionada con la especificación de la longitud se produce cuando se intenta establecer una correspondencia entre ésta y la longitud real del segmento que se forma, porque esto provoca la interrupción de un subtópico cuando el algoritmo determina que se produce esta correspondencia.

3 Método TextLec

Dadas las limitaciones detectadas en los métodos de segmentación evaluados, se hizo necesario concebir una nueva propuesta que supere dichas limitaciones, con vista a obtener un producto que satisfaga las necesidades expuestas. La nueva propuesta redunda en un método de segmentación nombrado TextLec.

TextLec inicialmente se concibe para la segmentación de textos científico-técnicos, basándose en algunas de las características presentes en estos textos y que son de utilidad al proceso de segmentación. No obstante, puede aplicarse a otros tipos de textos con características similares a las que se asumen en este trabajo.

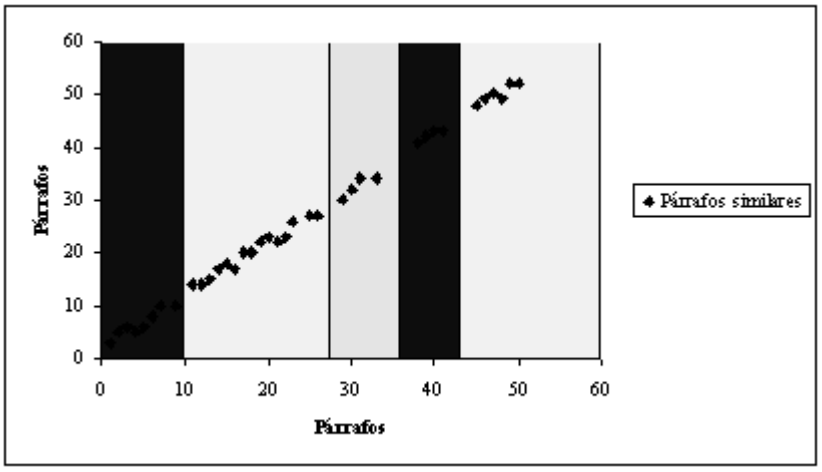
3.1 Características de los textos científicos

Los textos de este tipo usualmente son textos de múltiples párrafos que explícitamente explican o enseñan sobre un tópico. El universo de las palabras utilizadas, para expresar dicho tópico, se sitúa en cualquier ámbito de la ciencia y la tecnología y las más significativas, en relación a dicho tópico, usualmente se repinten con más frecuencia que el resto. Esta última característica se extiende a los subtópicos; es decir, en cada subtópico, las palabras más frecuentes son las más relevantes a él. Esto hace válido suponer que, en este tipo de texto, la repetición de términos es un elemento confiable para identificar aquellas unidades textuales que están relacionadas con un mismo subtópico. Otra característica presente en los textos científico-técnicos, relevante a la segmentación, radica en que los mismos suelen estar divididos por secciones interrelacionadas de unidades de texto que discuten densamente algún subtópico.

En la figura que se muestra a continuación, en las regiones sombreadas de las gráficas, pueden distinguirse ejemplos de distintas porciones de párrafos considerados similares porque comparten varios términos en común. Se consideró que dos párrafos estaban relacionados si su similitud, calculada a través de la medida del coseno, no era menor a 0.18. En la Figura 3.1 a) y b), cada punto indica el párrafo similar más alejado a cada párrafo, el cual se determina en un margen de 3 y 10 párrafos respectivamente.

Este epígrafe tiene aproximadamente 55 párrafos y fue utilizado por Heinonen en la evaluación de su método.

Este texto se formó con el primer epígrafe del capítulo 2 del libro Mars, nombrado "*Evidence of it*" y escrito por Percival Lowell[14].



(a

[14] Lowell P.: Mars. S.n: s.e; 1895. Disponible en: http://www.wanderer.org/references/lowell/Mars/ [Consultado: 7 de marzo del 2007].

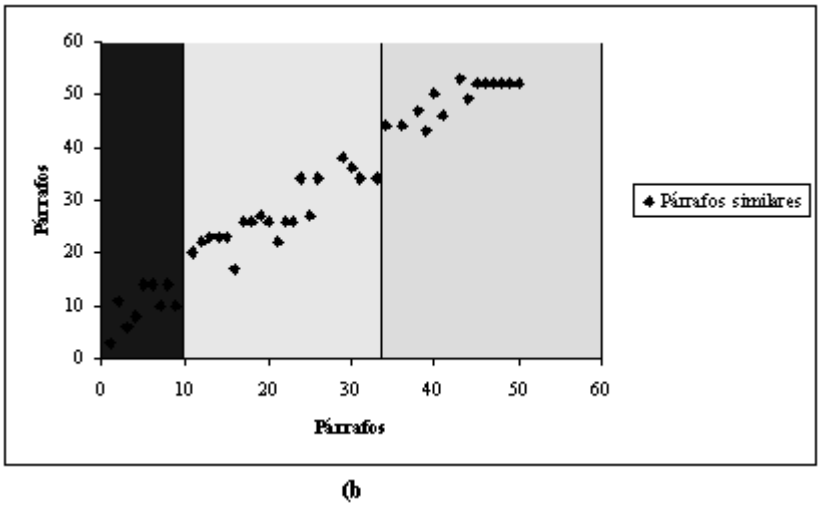


Fig. 3.1. Ejemplo que ilustra la estructura de un texto científico-técnico en secciones de párrafos similares

Lo anterior se corresponde con la estructura monolítica por partes definida por Skorochod'ko como un tipo de estructura lineal, la que se comentó en la sección anterior. Según este autor, los documentos que presentan dicha estructura no son los más difíciles de segmentar descomponiéndolos por sus partes monolíticas, asumiendo que cada una contiene un subtópico. Por otra parte, para satisfacer los objetivos de este trabajo se considera que es suficiente una segmentación lineal, primero por que no se pretende hacer una segmentación a un nivel fino de granularidad y porque la identificación de una estructura lineal no precisa del alto costo de implementación.

3.2 Método TextLec

TexteLec es una nueva propuesta de método de segmentación por tópicos que intenta identificar los cambios de subtópicos en textos con un contenido científico-técnico, utilizando la repetición de términos y asumiendo que los subtópicos tienen una distribución lineal en dichos textos.

Este método reconoce a los párrafos como las unidades textuales, asumiendo que todas las oraciones que pertenecen a un mismo párrafo tratan el mismo tópico. La representación de los párrafos se basa en el Modelo de Espacio Vectorial, al igual que las propuestas de Hearst y Heinonen.

Por otra parte, TextLec se basa en que los cambios de vocabulario en un documento coinciden con los cambios de subtópicos, similar a lo supuesto por Hearst. Esto, por su parte, permitió asumir que los párrafos cercanos que mantienen una cohesión léxica significativa entre sí, en cuanto a los términos léxicos que usan, están asociados al mismo tópico; es decir, que los párrafos que tienen un número significativo de términos en común deben pertenecer al mismo segmento.

Antes de dar comienzo al proceso de segmentación, los textos pasan por una etapa de preprocesamiento, en la que se consideró eliminar los *stopwords*, y lematizar. Las cuestiones particulares a esta etapa se comentarán en la próxima sección.

Esencialmente, en esta nueva propuesta se distinguen dos etapas básicas: control de los párrafos cohesionados más lejanos y detección de los cambios de tópicos. Cada una de éstas se expone a continuación.

3.2.1 Control de los párrafos cohesionados más lejanos

Cuando se concluye la etapa de preprocesamiento se determina la cohesión léxica entre los párrafos cercanos y se decide si ésta es suficiente para que los párrafos pertenezcan al mismo segmento. En este trabajo se considera que dos párrafos mantienen una cohesión léxica significativa o suficiente para pertenecer al mismo segmento si, después del cálculo de la misma, ésta es mayor que un umbral determinado. A partir de este momento, el término cohesionados se utilizará para referirse a párrafos tales que su cohesión léxica no esté por debajo de dicho umbral[15].

Por otra parte, se define una ventana inferior para cada párrafo; ésta se forma solamente con algunos párrafos por debajo del párrafo en cuestión, a diferencia de Heinonen que toma párrafos por encima y por debajo. El uso de esta ventana permite disminuir el efecto de párrafos cortos u otros que interrumpen una cadena de texto cohesiva, ya que se calcula la cohesión léxica de cada párrafo con todos los párrafos que están dentro de su correspondiente ventana. Además, el uso de una ventana permite reducir cálculos innecesarios. Una expresión más formal de la ventana inferior V_i para un párrafo *i* es la siguiente:

$$V_i = \{p_{i+1}, \ldots, p_r\}, \tag{7}$$

$$r = \begin{cases} i + \Delta \quad si \quad i + \Delta \leq n \\ n \qquad otro\ caso \end{cases}, \tag{8}$$

donde Δ es la cantidad de párrafos que forman la ventana, la cual puede variar con el tamaño del documento o con el objetivo de la segmentación[16].

Para elegir el párrafo que representa la cohesión dentro de la ventana, se ha decidido – a diferencia de Heinonen, que busca el valor de cohesión más alto – considerar el párrafo cohesionado más lejano. Esto se hace con el objetivo de controlar mejor el posible fin (límite inferior) de segmento para cada párrafo, suponiendo que dicho límite no se encuentra antes del párrafo cohesionado más lejano al párrafo en cuestión. Para controlar el párrafo cohesionado más lejano se ha propuesto el uso del vector *Parf*. El valor de la componente i-ésima de *Parf* será el número de párrafo cohesionado con *i* que esté más lejano a *i* dentro de su ventana. Es

[15] Dado que no todos los autores utilizan el mismo estilo de redacción (unos pueden utilizar el recurso de la repetición de términos más que otros) se dificulta la selección de un umbral común para todos los textos. En la siguiente sección, correspondiente a las evaluaciones experimentales del método, se exploran diferentes umbrales con vista a determinar el más adecuado para la identificación de los límites de segmento.

[16] Es bueno notar que aumentando el tamaño de la ventana es posible obtener segmentos más largos, porque se incrementa la posibilidad de encontrar un párrafo cohesionado más lejano, aunque esto puede disminuir la cohesión del segmento.

posible que un párrafo no tenga algún párrafo cohesionado dentro de la ventana; en este caso se considera que el párrafo cohesionado más lejano es él mismo. Más formalmente, la expresión de *Parf* puede definirse de la siguiente forma.

Sea $T = \{p_1, \ldots, p_n\}$ un texto de n párrafos y sea ξ un umbral de cohesión léxica entre párrafos, el vector de los párrafos cohesionados más lejanos se define como:

$$Parf = (coh_1, \ldots, coh_n) \text{ donde } coh_i \in T, \quad \textbf{(9)}$$

$$coh_i = \begin{cases} p_k \in V_i \quad si \;\; siml(p_k, p_i) \geq \xi, \; y \\ \qquad\qquad \neg\exists\, p_j, p_j \in V_i : siml(p_j, p_i) \geq \xi \;\; y \; j > k, \\ p_i \qquad otro\ caso \end{cases} \quad \textbf{(10)}$$

donde *siml* es una función mediante la cual se determina numéricamente la cohesión léxica entre dos párrafos.

A continuación se explica cómo se representan los párrafos mediante el modelo de espacio vectorial, y cómo se calcula la cohesión léxica entre ellos empleando la medida del coseno.

3.2.2 Obtención de la representación y la similitud entre los párrafos

Una de las formas de representación de las unidades textuales más utilizada es el modelo de espacio vectorial, VSM, por sus siglas en inglés; ésta se usa en la segmentación de forma similar a como se hace con los documentos en la IR [37]. El VSM permite calcular con bastante eficacia y eficiencia la similitud entre dos párrafos según la cantidad de términos que coinciden entre ellos.

Mediante el VSM, los párrafos se transforman en vectores dentro de un espacio multidimensional, donde las componentes son los términos diferentes resultantes del preprocesamiento. Dicho de otro modo, suponiendo que se tiene un documento de n párrafos, $P = \{p_1, p_2 \ldots p_n\}$, formado por un conjunto de k términos únicos $T = \{t_1, t_2 \ldots, t_k\}$, el párrafo i podrá modelarse como un vector de la siguiente forma:

$$p_i \rightarrow \overrightarrow{p_i} = (w(t_1, p_i), \ldots, w(t_k, p_i)), \quad \textbf{(11)}$$

donde $w(t_j, p_i)$ es el peso del término t_j en el párrafo p_i.

El uso de los pesos en la IR tiene el objetivo de normalizar la frecuencia de los términos en documentos muy extensos para evitar, por ejemplo, que sus tamaños influyan demasiado en el cálculo de sus relevancias en la recuperación. En la segmentación también es de utilidad este esquema de pesado.

Existen varias fórmulas para determinar los pesos; por ejemplo, una de las más utilizadas es la siguiente:

$$w_j = \frac{TF(t_j, p_i)}{cant(p_i)}, \quad \textbf{(12)}$$

donde w_j es el peso del término t_j del párrafo p_i, el término $TF(t_j,p_i)$ representa la frecuencia con la que aparece el t_j en p_i y $cant(p_i)$ corresponde a la cantidad de términos de *pi*.

Por otra parte, en el VSM se asume que no existe ninguna relación o dependencia entre los términos y, a partir de ello, se asume que las dimensiones son ortogonales. Por ejemplo, considerando que se tiene un documento con tres términos, *modelo*, *espacio*, *vectorial*, entonces se tendría un espacio *E* tridimensional, como se muestra en la Figura. 3.2.

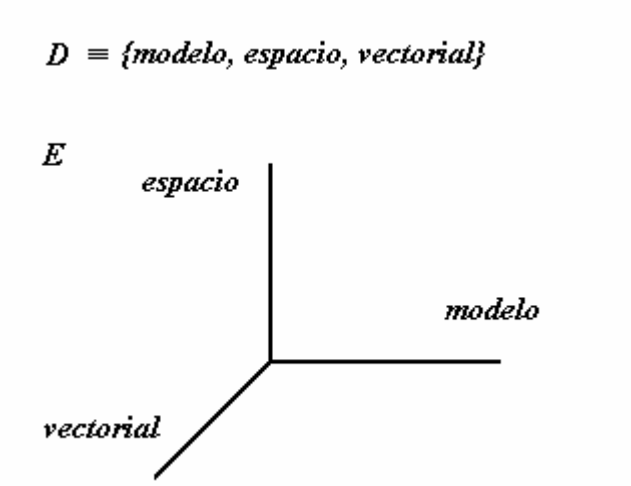


Fig. 3.2. Espacio tridimensional definido por los términos: modelo, espacio, vectorial

Después de representar dos párrafos se puede determinar su cohesión léxica. Esto se hace calculando la similitud entre los vectores que los representan; a mayor cercanía, mayor similitud. La cercanía entre los vectores se puede calcular utilizando la medida del coseno, la cual se expresa mediante la siguiente expresión:

$$simil(p_i, p_j) = \frac{\overrightarrow{p_i} \bullet \overrightarrow{p_j}}{\left|\overrightarrow{p_i}\right| \times \left|\overrightarrow{p_j}\right|} = \frac{\sum_{r=1}^{k} w_{ri} \times w_{rj}}{\sqrt{\sum_{r=1}^{k} w_{ri}^2} \times \sqrt{\sum_{r=1}^{k} w_{rj}^2}}, \quad \textbf{(13)}$$

donde w_{ri} es la componente *r* del vector de p_i.

Geométricamente, la interpretación de esta expresión se corresponde con el coseno del ángulo que se forma entre los vectores, (ver Figura. 3.3). El valor de similitud será un valor entre 0 y 1. Cuando los párrafos son iguales la similitud alcanza el valor de 1, y 0 cuando ellos son completamente diferentes; o sea, que no comparten ningún término.

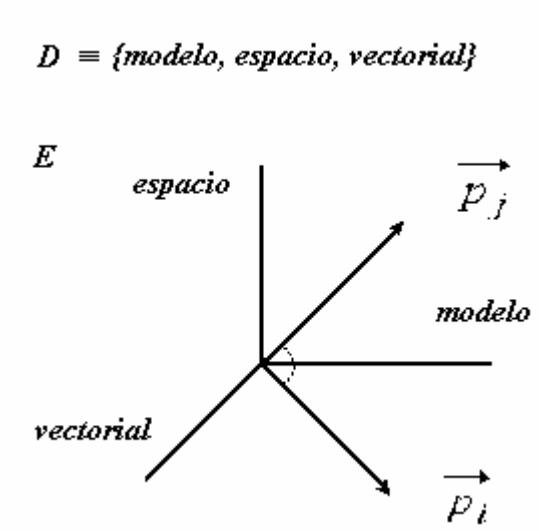


Fig. 3.3. Representación de la interpretación geométrica de la similitud entre dos párrafos según la medida del coseno

3.2.3 Detección de cambios de tópicos

El proceso de segmentación consiste en incluir secuencialmente párrafos en un segmento hasta que se incluya el último párrafo que esté cohesionado con alguno de los párrafos del segmento, considerándose este párrafo como el límite inferior de dicho segmento, bajo el supuesto de que, si todos los párrafos que tienen una cohesión léxica suficiente para pertenecer al mismo tópico están incluidos dentro de un segmento, entonces el límite inferior de éste coincide con un cambio de dicho tópico.

El proceso comienza creando un segmento, que está formado por el primer párrafo del texto; luego se controla en una variable, nombrada *MaxInf*, el párrafo cohesionado más lejano del segmento creado, que en este caso coincide con el valor de $Parf_1$, porque hasta el momento dicho segmento sólo incluye al primer párrafo. Luego, se analiza dónde incluir al segundo párrafo. Éste se incluye en el primer segmento si no se encuentra después del párrafo cohesionado más lejano al segmento, es decir, si no es mayor que *MaxInf*. En caso contrario se crea un nuevo segmento que se inicia con el segundo párrafo. Posteriormente, se actualiza el valor del párrafo cohesionado más lejano al último segmento creado, teniendo en cuenta ahora el valor de $Parf_2$. Este proceso continúa hasta que el último párrafo se incluye en un segmento.

Este proceso de identificación de los cambios de tópicos se formalizó en un algoritmo de segmentación nombrado TextLec, cuyo seudo-código se muestra en la Figura 3.4. Durante el proceso se usa la variable *MaxInf* que permite conocer el párrafo cohesionado más lejano del segmento en proceso; por tanto, el párrafo controlado por ella será posiblemente el que cierre este segmento. Además, se emplea un vector que controla para cada segmento su posible límite inferior, este vector recibe el nombre de *Lim*. El elemento Lim_k contiene el último párrafo del segmento *k*; por ejemplo, si $Lim_k=5$ entonces se tiene un segmento *k* que termina en el párrafo 5.

```
Algorithm: TextLec
 Input: TxP – Matriz de términos por párrafos
        N - total de párrafos
Output: Lim – límites de segmentos
1)      /* Determinar Parf según las expresiones 2.3 y 2.4 */
2)      Parf = DeterminaParf(TxP, ξ, •);
3)      /* Determinar los posibles límites de segmentos */
4)      MaxInf = 0;
5)      k = 0;
6)      for i = 1 to N do begin
7)        if MaxInf = i-1 then begin
8)          Limk = MaxInf;
9)          k = k + 1;
10)       End
11)       MaxInf = max( Parfi, MaxInf );
12)     End
13)     /* Remover los segmentos de longitud menor a • */
14)       j = 1;
15)       for i = 2 to k do
16)         if Limj - Limi < • then
17)           if simlinfi <  simlsupi
18)             then Limj = Limi;
19)         else begin
20)           j = j + 1;
21)           Limj = Limi;
22)         End
23)     k = j;
24)     Limk = N;
```

Fig. 3.4. Seudo-código del proceso de segmentación del método TextLec

El método recibe una matriz de términos por párrafos, donde los términos de cada párrafo son los que se obtienen como resultado del preprocesamiento del texto original.

La primera operación del método consiste en crear el vector de los párrafos cohesionados más lejanos *Parf,* según como se explicó en la sección anterior. Luego, para generalizar el método, se asigna $Lim_0 = 0$.

Durante la ejecución se analiza el resto de los párrafos del documento, determinando si se incluyen o no dentro del último segmento que se está procesando. A continuación se explican las tres situaciones en la que puede encontrarse el párrafo *i*, cuando se analiza su inclusión en el segmento que se procesa en ese instante:

- No existe un párrafo dentro del segmento que sea cohesionado con *i* o cohesionado con un párrafo después de *i*, (*MaxInf* = *i*-1). En este caso se toma el párrafo *i*-1 como límite inferior del segmento ($Lim_k = MaxInf$), y se incluye a *i* en un nuevo segmento.
- *i* es el párrafo cohesionado más lejano del segmento (*MaxInf* = *i*). En este caso se incluye a *i* dentro del segmento.

- El párrafo *i* pudiera estar o no cohesionado con algún párrafo del segmento, pero existe al menos un párrafo del segmento cohesionado con un párrafo que está después de *i*, ($MaxInf > i$). En este caso no se interrumpe el segmento y se incluye a *i* en el segmento.

Después de la inclusión del párrafo *i* en un segmento, ya sea el mismo que se procesaba o uno nuevo, se debe actualizar el párrafo cohesionado más lejano al último segmento creado, verificándose si el párrafo cohesionado más lejano de *i* ($Parf_i$) está más lejos que *MaxInf*.

Cuando esta parte del proceso termina se tiene una primera aproximación de los segmentos de textos, encontrándose en *Lim* los posibles límites inferiores de dichos segmentos.

En el proceso recién explicado es posible que se obtengan segmentos muy cortos; es decir, de muy pocos párrafos. Estos segmentos usualmente son llamados segmentos espurios. Los segmentos espurios pueden no resultar convenientes para determinadas aplicaciones; por ejemplo, aquellas que intentan segmentar textos de múltiples párrafos. Esto ocurre debido a que los párrafos que forman este tipo de segmento tienen una cohesión léxica por debajo del umbral definido con los párrafos de ambos segmentos adyacentes.

Por tal motivo se decide eliminar los segmentos espurios como sigue. Se establece que la longitud mínima de un segmento válido es el tamaño escogido para definir la ventana de párrafos, considerándose como espurios los segmentos que no cumplan con esta longitud. Los párrafos que forman los segmentos espurios se incluyen en el segmento adyacente (superior o inferior) más similar, como se muestra en las instrucciones 14-22 del seudo-código. Para determinar si un segmento espurio es más similar al segmento adyacente superior o al segmento adyacente inferior se consideran los valores de *simlsup* y *simlinf* respectivamente. Estos valores se corresponden con la máxima similitud que existe entre algún párrafo del segmento espurio con algún párrafo del segmento adyacente correspondiente. Más formalmente:

$$similsup_i = max(Max^i_{Lim_{i-1}+1}, \ldots, Max^i_{Lim_i}), \quad \textbf{(14)}$$

$$Max^i_j = max(siml(p^i_k, p^i_j), \ldots, siml(p^i_{Lim_{i-1}}, p^i_j)), \quad \textbf{(15)}$$

$$k_i = \begin{cases} j-\Delta \;\; si \quad j-\Delta > Lim_{i-2} \\ Lim_{i-2}+1 \;\; otro\ caso \end{cases}. \quad \textbf{(16)}$$

De forma similar se determina *simlder*, variando solamente la expresión:

$$k_i = \begin{cases} j+\Delta \;\; si \quad j+\Delta \leq Lim_{i+1} \\ Lim_{i+1} \quad otro\ caso \end{cases}. \quad \textbf{(17)}$$

Como puede notarse en las expresiones anteriores, no se consideran todos los párrafos de los segmentos adyacentes para determinar los valores de similitud. Esto se debe a que se decidió mantener los criterios expuestos en la sección 2.2.1 cuando se definió la ventana de párrafos empleada para calcular la similitud de un párrafo con cada párrafo, dentro de su correspondiente ventana.

Debe notarse que, después de aplicarse esta segunda parte del proceso de segmentación, existe la posibilidad de que se obtengan segmentos formados únicamente por segmentos

espurios; es decir, que varios segmentos espurios pueden unirse y formar un segmento válido por su longitud. Se considera que deben analizarse en trabajos futuros las consecuencias de este efecto en la efectividad del método propuesto.

Finalmente, el proceso de segmentación termina dejando *k* segmentos con límites inferiores en el vector *Lim*.

4 Evaluación

Esta sección está dedicada a mostrar el desempeño del método propuesto mediante resultados experimentales. En la Sección 4.1 se comentan las dificultades de evaluar los resultados de la segmentación por tópicos y las soluciones encontradas. En la Sección 4.2 se presentan los resultados de los experimentos utilizando algunas de estas soluciones.

4.1 Métodos de evaluación empleados

Evaluar los resultados de los algoritmos de segmentación por tópico tiene dos dificultades fundamentales. La primera está dada por la naturaleza subjetiva de detectar los límites físicos adecuados de los subtópicos, en la que pueden incluso estar en desacuerdo varias personas que decidan efectuar esta tarea; esto hace difícil seleccionar un corpus de referencia para realizar las comparaciones de los resultados que se obtienen [21], [33], [46].

Usualmente, esa dificultad se resuelve comparando el resultado de los métodos de segmentación contra las marcas, encabezados o subtítulos, que en ocasiones especifica el autor de un documento para identificar los subtópicos; pero estas marcas no siempre se precisan o no suelen precisarse bajo los mismos criterios. Algunos comparan sus resultados en términos de cuán bien el método de segmentación distingue un documento de otro en un fichero de documentos concatenados y dónde se distingan diferentes tópicos. Mientras tanto, otros comparan sus resultados con el resultado de una segmentación manual basada en el juicio de varias personas.

La segunda dificultad es que la importancia de los tipos de errores depende de las aplicaciones en donde se necesitan las técnicas de segmentación. Por ejemplo, en la recuperación de información (IR, por su siglas en inglés) se pueden aceptar límites de segmento que difieran en unas pocas oraciones del límite real del segmento. En cambio, para la segmentación de un flujo de transmisión continua de noticias es muy importante la exactitud de la ubicación de los límites.

Por otra parte, encontrar una métrica de evaluación adecuada para determinar la exactitud de un algoritmo de segmentación es un tema que ha generado mucha polémica. Dos de las medidas de evaluación que han sido utilizadas por muchos autores son *Precision* y *Recall*, las cuales son medidas estándares en las experimentaciones con sistemas de IR. En la estimación de la exactitud de la segmentación las métricas *Precision* y *Recall* suelen definirse de la siguiente forma.

Precision: El porciento o índice que representan los límites de segmento correctamente detectados por el algoritmo del total de límites detectados por el algoritmo.

Recall: El porciento o índice que representan los límites de segmento correctamente detectados por el algoritmo del total de límites reales detectados en la segmentación de referencia.

Estas medidas de evaluación resultan muy convenientes en aplicaciones donde la exactitud de la localización de los límites de los segmentos es muy importante. Pero no es así en aquellas aplicaciones que no lo requieren, porque penalizan muy fuerte al algoritmo cuando encuentran límites que no coinciden exactamente con los límites de la segmentación de referencia, y no tienen en cuenta si existe proximidad entre ellos. Otra dificultad es que hay una compensación inherente entre ellas; o sea, cuando una mejora, en ocasiones la otra declina. Esta última dificultad suele ser resuelta en la IR con la medida *F-measure*. Ésta también ha sido usada en la segmentación; pero, como puede notarse, no es recomendable porque depende de *Precision* y *Recall*, lo que hace que también sea insensible a la proximidad entre los límites de ambas segmentaciones. La expresión de *F-measure* es la siguiente.

$$F\text{ - }measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \qquad (\mathbf{18})$$

Beeferman, Berger y Lafferty en 1997 y 1999, propusieron una métrica llamada P_k para mejorar el proceso de evaluación de la segmentación, esta vez teniendo en cuenta la proximidad entre los límites de la segmentación de referencia y la obtenida por el algoritmo [3], [4]. Estos autores definieron a P_k como la probabilidad de que dos oraciones tomadas aleatoriamente del texto sean correctamente clasificadas, como pertenecientes al mismo segmento o no pertenecientes el mismo segmento. Más formalmente, sean *ref* y *hyp*, la segmentación de referencia y la segmentación del algoritmo respectivamente, se tiene que:

$$P_k(ref, hyp) = \sum_{1 \le i \le j \le n} D(i,j)\left(\delta_{ref}(i,j) \overline{\oplus} \delta_{hyp}(i,j)\right) \qquad (\mathbf{19})$$

donde n representa el número total de unidades textuales en el texto según sea el interés de la segmentación. i y j son dos unidades textuales separadas a una distancia k. δref es una función que tomará el valor de 1 si las unidades textuales i y j pertenecen al mismo segmento en la segmentación de referencia, y toma el valor de 0 el caso contrario. De forma similar, δhyp es una función indicador que toma el valor de 1 si las unidades textuales i y j pertenecen al mismo segmento en la segmentación obtenida por el algoritmo, y toma el valor de 0 en caso contrario. El operador $\overline{\oplus}$ es la función *XNOR*. La función D_k es una distribución de probabilidad de distancia sobre el conjunto posibles distancias entre las unidades textuales seleccionadas aleatoriamente. Los autores demostraron experimentalmente que el valor más adecuado de k se corresponde con la mitad del tamaño promedio de los segmentos en la segmentación de referencia.

La métrica de evaluación P_k fue un primer paso para considerar la proximidad entre los límites de la segmentación de referencia y la segmentación obtenida. No obstante, en esta métrica también se han detectado deficiencias, como el efecto de penalizar más fuerte al algoritmo cuando ignora un límite de segmento que cuando lo pone incorrectamente, entre otras.

Pevzner y Hearst en el 2000 propusieron una métrica llamada *WindowDiff* para mejorar el proceso de evaluación de P_k [33]. *WindowDiff* usa una ventana corrediza de longitud k para recorrer todo el texto y encontrar las discrepancias entre la segmentación de referencia y la que se obtiene como resultado de los algoritmos. Estos autores mantienen a k igual a la mitad del promedio del tamaño que tienen los segmentos en la segmentación de referencia.

En cada posición de la ventana se determina para ambas segmentaciones (la de referencia y la obtenida) el número de límites dentro de la ventana, y si el número de límites no es el mismo se penaliza el algoritmo. Posteriormente, se suman todas las penalizaciones que se encontraron en el texto completo y se normaliza este valor de forma tal que la métrica tome un valor entre 0 y 1. *WindowDiff* toma el valor de 0 si el algoritmo asigna todos los límites correctamente y toma el valor de 1 si difiere con la segmentación de referencia en todas las posiciones de la ventana, por lo que mientras menor sea el valor de *WindowDiff,* mayor será el desempeño del algoritmo. Más formalmente:

$$WindowDiff(ref, hyp) = \frac{1}{N-k}\sum_{i=1}^{N-k}(|\, b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})\,| > 0), \qquad \textbf{(20)}$$

donde $b(i,j)$ representa el número de límites entre la posición i y j en el texto, N representa el número total de unidades textuales en el texto, *ref* es la segmentación de referencia y *hyp* la segmentación del algoritmo.

4.2 Resultados experimentales

En esta sección se exponen los resultados experimentales del método TexLec; para ello, se utiliza la métrica *WindowDiff*, por resultar la más adecuada para mostrar el desempeño de este método. Por otra parte, como se mencionó en la sección anterior, durante la etapa de preprocesamiento se eliminan los *stopwords*, y se lematiza. Las palabras se reducen a su lema (no a su raíz), utilizando el TreeTager; un sistema para marcar con etiquetas y extraer el lema de las palabras en un texto, desarrollado por Helmut Schmid en la Universidad de Stuttgart. El sistema TreeTager se encarga de convertir las letras mayúsculas a minúsculas, así como de reconocer los párrafos. Otro aspecto interesante de TreeTager es que trabaja sobre varios lenguajes como, por ejemplo, Inglés, Francés, Alemán, Italiano, Español, Ruso y otros [39]-[41]. Debe mencionase que la implementación del algoritmo TextLec utilizada en las experimentaciones fue desarrollada en leguaje C. Además, se escogieron varios textos de pruebas con un contenido científico-técnico con vista a ser utilizados como segmentación de referencia, estos textos se describen a continuación.

El primer texto (Texto 1) se construyó uniendo 14 artículos diferentes, tomados de las memorias de "*The 18th International Conference on Pattern Recognition ICPR'2006*"; dicho texto tiene aproximadamente 305 párrafos y un promedio de 22 párrafos aproximadamente por artículo[17]. El segundo (Texto 2) es un texto formado por 8 sub-tópicos de 7 artículos diferentes

[17] A continuación se relacionan los artículos del ICPR'2006 utilizados en la confección del Texto 1.

tomados de la enciclopedia libre Wikipedia (*Solar System, Sun, Geography, Hydrography, Earth, Atmosphere, Animal* y *Soil*); este tiene 29 párrafos[18]. Estos dos textos se crearon con el objetivo de tomar como referencia de comparación los límites entre las piezas de textos que contengan artículos diferentes con un alto grado de certeza.

Además, con el fin de obtener una segmentación de referencia manual se creó un tercer texto (Texto 3) que está formado por el primer epígrafe del capítulo 2 del libro titulado "*Mars*" de Percival Lowell, titulado "*Evidence of it*", y formado aproximadamente por 55 párrafos. Este texto se segmentó manualmente por 5 personas, las cuales discreparon en la posición de los límites de segmentos (ver Figura 4.1), por lo que se escogieron como válidos los 7 límites donde al menos existieron tres coincidencias (3, 10, 15, 28, 36, 43, 52)[19].

1. Perrin, G., Descombes, X., Zerubia, J.: 2D and 3D vegetation resource parameters assessment using marked point processes. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
2. Tong, W. S., Tang, CH. K.: Multiresolution mesh reconstruction from noisy 3d point sets. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
3. Liu, X., Yao, H., Yao, G. Gao, W.: A novel volumetric shape from silhouette algorithm based on a centripetal pentahedron model. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
4. Nishie, K., Sato, J.: 3D Reconstruction from uncalibrated cameras and uncalibrated projectors from shadows. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
5. Berretti, S., Del Bimbo, A., Pala, P.: Partitioning of 3D meshes using reeb graphs. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
6. Hansen, W., Michaelsen, E., Thönnessen, U.: Cluster analysis and priority sorting in huge point clouds for building reconstruction. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
7. Takizawa, H., Yamamoto, S.: Surface reconstruction from stereovision data using a 3-D MRF of discrete object models. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
8. Chen, W. G.: Noise variance adaptive sea for motion estimation: a two-stage schema. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
9. Sun, Z.: A three-frame approach to constraint-consistent motion estimation. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
10. Brandt, S. S.: Robust factorization with uncertainty analysis. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
11. Solem, J. E.: Geodesic curves for analysis of continuous implicit shapes. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
12. Zhou, X., Wang, R.: Symmetric pixel-group based stereo matching for occlusion handling. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
13. Wang, T., Basu, A.: Automatic estimation of 3d transformations using skeletons for object alignment. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.
14. Sibiryakov, A., Bober, M.: Real-time multi-frame analysis of dominant translation. En: Proceedings of the 18th International Conference on Pattern Recognition ICPR2006, volumen 1, 2006.

[18] A continuación se relacionan los artículos de Wikipedia utilizados en la confección del Texto 2.

1. Solar System. Disponible en: http://en.wikipedia.org/wiki/Solar_System [Consultado: 8 de marzo del 2007].
2. Geography Disponible en: http://en.wikipedia.org/wiki/Geography [Consultado: 8 de marzo-abril del 2007].
3. Hydrography. Disponible en: http://en.wikipedia.org/wiki/Hydrography [Consultado: 8 de abril del 2007].
4. Herat. Disponible en: http://en.wikipedia.org/wiki/Soil [Consultado: 8 de abril del 2007].
5. Atmosphere. Disponible en: http://en.wikipedia.org/wiki/Atmosphere [Consultado: 8 de abril del 2007].
6. Animal. Disponible en: http://en.wikipedia.org/wiki/Animal [Consultado: 8 de abril del 2007].
7. Soil. Disponible en: http://en.wikipedia.org/wiki/Soil [Consultado: 8 de abril del 2007].

[19] Las 5 personas escogidas para realizar la segmentación manual son especialistas del Centro, licenciados en ciencias de la computación, ingenieros informáticos y filólogos. Con vista a homogenizar el resultado de este proceso, todas las personas recibieron un conjunto de instrucciones que aparecen como anexos de este trabajo.

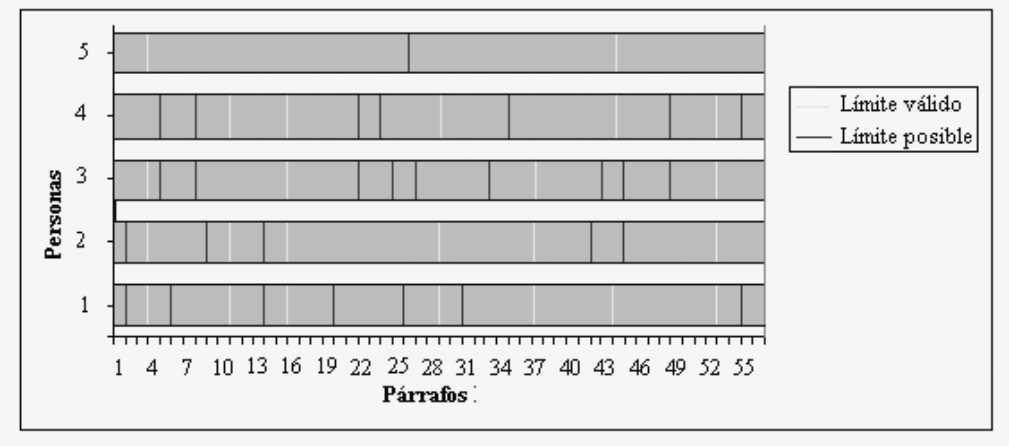


Fig. 4.1. Resultados de la segmentación manual del Texto 3 basada en el juicio humano

Por último, se seleccionaron aleatoriamente 6 textos (texto 4 al texto 9) de artículos de la *Lecture Notes in Computer Science* (LNCS)[20]. Dichos artículos se escogieron con la intención de utilizar como referencia las marcas usadas por los autores para separar los subtópicos que los forman.

4.3 Búsqueda del mejor umbral

Para buscar el mejor umbral que determina si dos párrafos están cohesionados, se utilizó el concepto de ventana inferior definido por el método TextLec para calcular los valores de similitud de cada párrafo con varios párrafos por debajo de él. Se calcularon todos estos valores de similitud en cada uno de los textos de prueba mencionados en la sección anterior. Para cada texto se conformaron cuatro conjuntos con los valores de similitud calculados, utilizando un criterio diferente para cada conjunto. Luego para cada conjunto se consideraron varios posibles

[20] A continuación se relacionan los artículos de las LNCS utilizados en la confección del texto 4 al texto 9.

1. Beyer, D., Henzinger, T. A., Jhala, R., Majumdar, R.: Checking Memory Safety with Blast. En: Proceedings of the 8th International Conference on Fundamental Approaches to Software Engineering, LNCS, volumen 3442, páginas 2-18, 2005.
2. Ding, M., Fenster, A.: Projection-Based Needle Segmentation in 3D Ultrasound Images. En: Proceedings of the 6th International Conference on Medical Image Computing and Computer-Assisted Intervention, LNCS, volumen 2879, páginas 319-27, 2003.
3. Nojima, R., Kobara, K., Imai, H.: Efficient Shared-Key Authentication Scheme from Any Weak Pseudorandom Function. En: Proceedings of 7th International Conference on Progress in Cryptology – INDOCRYPT, LNCS, volumen 4329, páginas 303-16, 2006.
4. Lakshminarayan, Ch.,Yu, Q., Benson, A.: Improving Customer Experience via Text Mining. En: 4th Workshop on Databases in Networked Information Systems, LNCS, volumen 3433, páginas 288-99, 2005.
5. Lee, D., Kim, J.,Seok, J.: Lecture Notes In Computer Science. En : Proceedings of the 6th Asian Computing Science Conference on Advances in Computing Science table of contents, LNCS, volumen 1961, páginas 43-57, 2000.

umbrales, con los cuales se realizaron las corridas del método TextLec. Los umbrales se consideraron teniendo en cuenta la media aritmética de cada conjunto como medida de tendencia central y a la desviación estándar como el promedio de la desviación de los datos respecto a la media aritmética de su conjunto.

Para lograr una mejor comprensión del proceso de búsqueda del umbral, los pasos anteriormente expuestos serán detallados como sigue:

1. Para cada texto de prueba (texto 1 al texto 9):
    a. Se calculan los valores de similitud de cada párrafo con los párrafos dentro de su ventana inferior.
    b. Con todos los valores obtenidos se conforman 4 conjuntos de valores:
        i. Un conjunto formado por la totalidad de los valores determinados.
        ii. Un conjunto formado por el valor máximo determinado en cada ventana.
        iii. Un conjunto formado por el valor mínimo determinado en cada ventana.
        iv. Un conjunto formado por la media de los valores determinados en cada ventana.
    c. Para cada conjunto se determina su media $\bar{x}$ y desviación estándar s.
    d. Para cada conjunto se consideran los siguientes posibles umbrales (ξ):
        i. $\xi = \bar{x}$.
        ii. $\xi = \bar{x} - 2s$.
        iii. $\xi = \bar{x} + 2s$.
        iv. $\xi = \bar{x} - 3s$.
        v. $\xi = \bar{x} + 3s$.
2. Con cada texto de prueba se realizan varias corridas del método TextLec; o sea, con cada conjunto de valores y para cada unos de los umbrales posibles se realiza una corrida del método TextLec.
3. Para cada segmentación obtenida se determina el valor de *WindowDiff*, utilizando las correspondientes segmentaciones de referencia en cada caso.
4. Para cada uno de los cuatro conjuntos de valores de similitud se analizan los valores de *WindowDiff* que se obtienen con los umbrales considerados.
5. Finalmente se toma como umbral el ξ del conjunto para el que el método TextLec alcance el mejor desempeño en la mayoría de los textos de prueba, según la métrica *WindowDiff*.

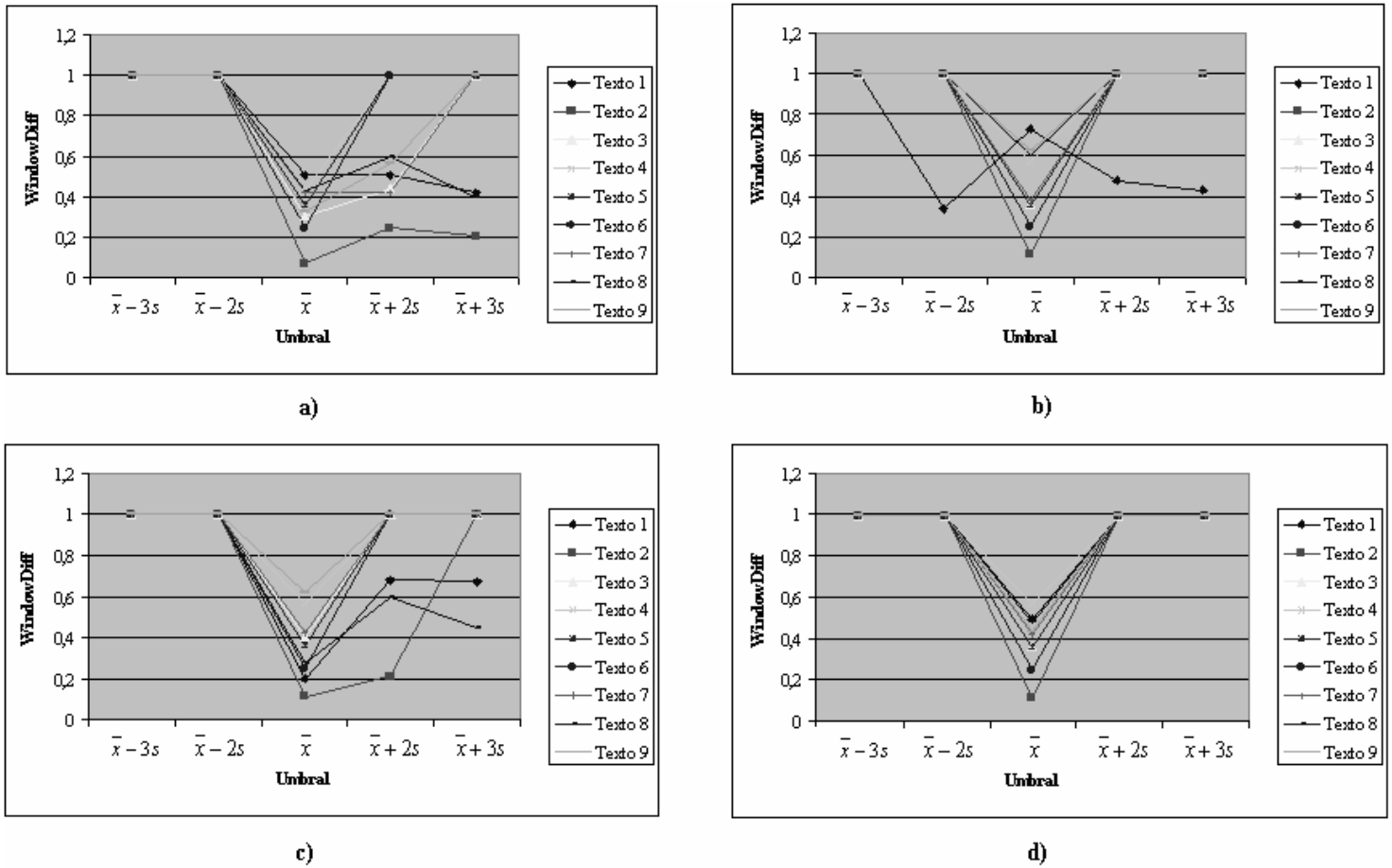


Fig. 4.2 a), b), c) y d). Desempeño del método TextLec, según los valores de *WindowDiff*, en la segmentación de todos los textos de prueba considerando diferentes umbrales

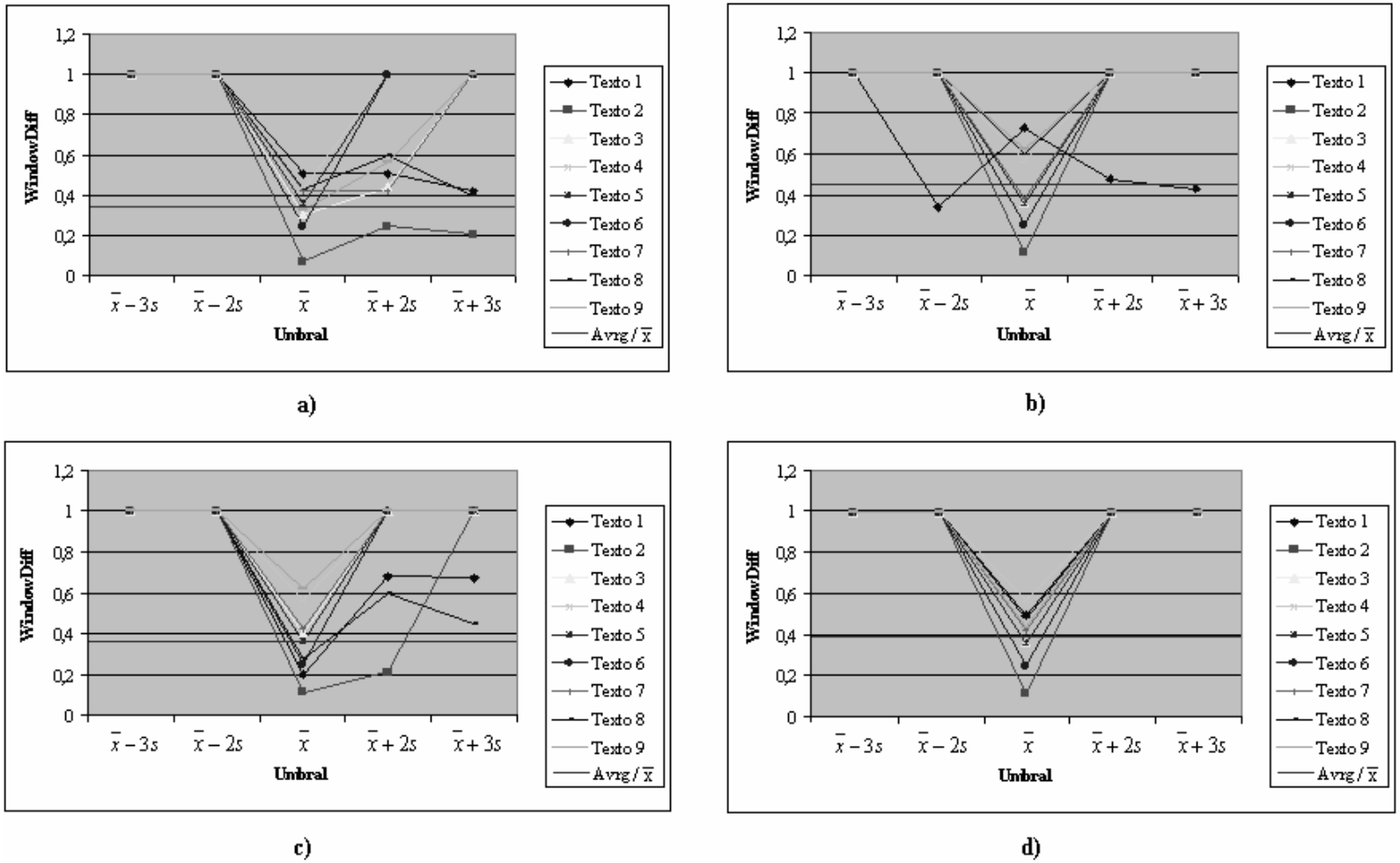


Fig. 4.3 a), b), c) y d). Desempeño del método TextLec, según los valores de WindowDiff, considerando diferentes umbrales y promedio de desempeño cuando el umbral es igual a la media aritmética en cada uno de los conjuntos considerados

En la Figura 4.2 a), b), c) y d, se muestra el resultado del proceso descrito. La segmentación de los textos de prueba, considerando diferentes umbrales sobre el conjunto de todos los valores de similitud determinados, el conjunto de los valores máximos de todas las ventanas, el conjunto de los valores mínimos de todas las ventanas, y el conjunto de las medias de todas las ventanas, respectivamente. Analizando el desempeño del método TextLec a través del comportamiento de los valores de *WindowDiff* que se muestran en esta figura, puede notarse que, de forma general para los cuatro conjuntos de valores seleccionados y para casi la totalidad de los textos, el mejor desempeño del método se observa para un umbral $\xi = \bar{x}$.

Por otra parte, como puede observarse en la Figura 4.3 a), b), c) y d), si bien para este umbral la diferencia entre los promedios de desempeño del método en los cuatro conjuntos de valores es poco significativa, se puede notar un comportamiento más estable del desempeño cuando el umbral se selecciona sobre el conjunto formado por la media de los valores de cada ventana.

Esta experimentación, sin representar una demostración concluyente, sugiere que según lo supuesto por el método TextLec, el conjunto de valores más representativos de la cohesión léxica entre los párrafos de un texto dado es el conjunto formado por la media de los valores de similitud determinados en cada ventana de párrafos, y que la mejor elección del umbral que determina si dos párrafos están cohesionados es la media de este conjunto. Tales criterios fueron considerados en la evaluación del comportamiento del método TextLec comparada con la de otros dos métodos, como se muestra en la sección que sigue a continuación.

4.4 Comparación con otros algoritmos

En esta sección se muestra una comparación del desempeño del método TextLec con el desempeño de los dos métodos que intentan resolver problemas similares a los planteados en este trabajo de investigación, el TextTiling de Hearst y el método de Heinonen. El desempeño de los tres métodos se evalúa a través de la métrica *WindowDiff* en todos los textos de prueba, como se muestra en la Figura 3.4, en la que también se brinda información sobre el promedio de desempeño alcanzado por cada método.

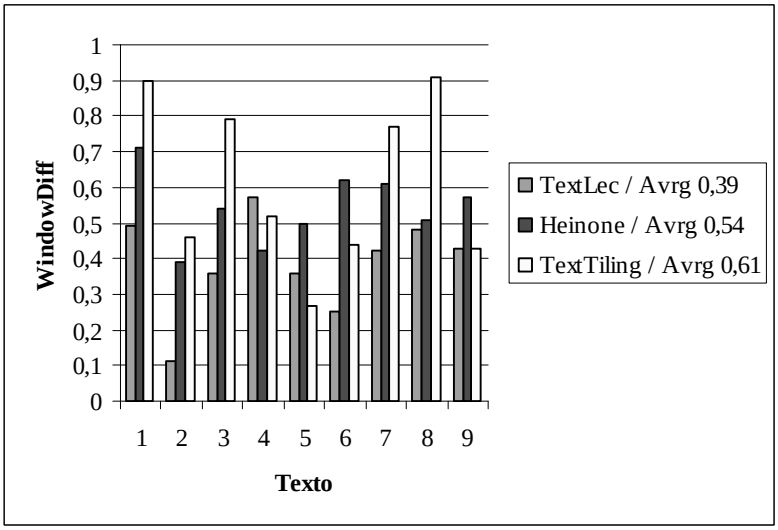


Fig. 3.4. Desempeño de los métodos TextLec, Heinonen y TextTiling en todos los textos de prueba

En los resultados de estas experimentaciones se puede notar un mejor desempeño de TextLec cuando se compara con los otros dos métodos, observándose a TextLec con un valor promedio de desempeño superior. Lo que valida los supuestos de TextLec con respecto a superar las limitaciones detectadas en dichos métodos y a obtener, de forma general, un método más adecuado para resolver la segmentación por tópicos en textos científico-técnicos.

4.5 Otros experimentos

Con la intensión de probar que el método TextLec puede aplicarse a otros tipos de documentos con características similares a las asumidas para los documentos científico-técnicos, se escogió una colección de 29 documentos formados por noticias continuas. Debe mencionarse que resultaron dos segmentaciones de referencia, en la primera se consideró que cada noticia correspondía a un subtópico, y en la segunda se decidió respetar las agrupaciones de noticias que consideraban algunos documentos; por ejemplo, algunos documentos agrupan noticias de un mismo país, de literatura, sociales, entre otros tópicos.

En este caso, para la evaluación del método, se consideraron las medidas *Precision* y *Recall*, y *F-measure*, dado que es necesario medir la coincidencia exacta de los límites de referencia con los límites que obtiene el método. Los valores promedios de estos experimentos se muestran el la Tabla 4.1.

Segmentación de referencia	***Precision***	***Recall***	***F-Measure***
1	58.62	73.90	61.89
2	41.34	70.45	48.97

Tabla 4.1. Resultados de *Precision*, *Recall*, y *F-measure* de TextLec para la segmentación de noticias continuas

En los experimentos se puede observar un desempeño aceptable del método cuando se compara con otros reportados en la literatura, la segmentación de una de transmisión continua de noticias, como, por ejemplo, los reportados por con el método SeleCT, para noticias de CNN y Reuters, los que se muestran en la Tabla 4.2 [45].

Segmentación de referencia	***Precision***	***Recall***
CNN	55.8	53.4
Reuters	79.1	60.6

Tabla 4.2. Resultados de *Precision* y *Recall* de SeleCT para la segmentación de noticias continuas

5 Conclusiones y trabajos futuros

El uso de métodos de segmentación por tópicos puede mejorar los resultados de muchas tareas de procesamiento de textos; por ejemplo, la recuperación de información, la confección

automática de resúmenes, la detección y seguimiento de tópicos, entre otras. En este reporte se muestra una investigación sobre el tema con el objetivo de exponer la elaboración de un método de segmentación que permita identificar los cambios de tópicos en documentos científico-técnicos, teniendo en cuenta las características principales de estos documentos, con vista a satisfacer las necesidades del departamento de Minería de Datos del CENATAV.

Los objetivos se cumplieron satisfactoriamente ya que se mostró el estudio realizado del tema, lográndose conocer varios importantes aspectos de la segmentación de textos por tópicos como, por ejemplo, que esta tarea comprende la segmentación en tópicos globales y la segmentación en subtópicos o segmentación del discurso, pudiendo ser ésta última jerárquica o lineal; que una adecuada selección de las señales o indicadores lingüístico que indican los cambios de tópico deriva en eficacia de la segmentación; también, se detectaron métodos que pudieron ser utilizados según las intenciones, profundizándose en el funcionamiento de éstos, así como en sus principales deficiencias.

A partir de la elaboración del marco teórico de la investigación se obtuvo como resultado un nuevo método de segmentación por tópicos, nombrado TextLec, que mejora las propuestas encontradas. Puede decirse que este método, como aporte fundamental, define para cada párrafo del texto una ventana de párrafos inferiores (por debajo), la cual se emplea para determinar la cohesión léxica del párrafo en cuestión con los párrafos de su ventana, así como para localizar el párrafo cohesionado más lejano de él. De esta forma se logró disminuir la posibilidad de interrumpir la continuidad de un tópico. Para determinar si dos párrafos están cohesionados se considera un umbral de similitud, el cual fue seleccionado experimentalmente teniendo en cuenta el comportamiento del desempeño del método TextLec sobre diferentes referencias de comparación.

Como última conclusión puede decirse que se validó el nuevo método a partir de corpus textuales representativos del universo investigado y su comparación con los métodos más significativos, resultando el método TextLec el de mejor desempeño en casi la totalidad de los casos, como vía de comprobación y fiabilidad de la investigación realizada, a partir de los resultados obtenidos con la implementación del mismo. Para esto se realizó un estudio de las problemáticas de la evaluación de los métodos de segmentación. Debe mencionarse que se realizaron experimentos con otro tipo de documentos, una colección de noticias continuas, obteniéndose resultados aceptables, lo cual muestra, sin ser una demostración concluyente, que el método propuesto puede aplicarse a otros tipos de documentos con características similares a las asumidas para los documentos científico-técnicos. En este estudio se detectó que dichas problemáticas fundamentalmente están asociadas a la subjetividad implícita en decidir cuáles son los límites adecuados de los segmentos y a las áreas de aplicación de los métodos.

Con vista a lograr que la identificación de los cambios de tópicos a través de la alternativa del método TextLec alcance la mayor precisión y con el propósito de aumentar al valor práctico de dicha propuesta, se recomiendan como trabajos futuros los siguientes:

1. Proponer nuevos modelos de represtación de las unidades textuales o explorar otros.
2. Proponer un indicador más robusto para medir la cohesión léxica de las unidades textuales, así como valorar el empleo de otras señales de continuidad o cambio de tópicos. En conjunto, debe estudiarse la selección del umbral a partir del cual se determina si dos párrafos están cohesionados.

3. Analizar la influencia que pueden tener los segmentos logrados a través de la unión de los segmentos que se consideran espurios por su corta longitud en la eficacia del método propuesto.
4. Extender la aplicación del método a otros dominios de documentos.
5. Aplicar el método TextLec en diferentes tareas de procesamiento de textos.

Referencias bibliográficas

1. Alfonso, I. R.: La importancia social de la información. En: ACIMED, volumen 9, número 3, páginas 221-23, 1997.
2. Angheluta, R., Busser, R., Moens, M.F.: The Use of topic segmentation for tutomatic summarization. En: Proceedings of the ACL-2002, Post-Conference Workshop on Automatic Summarization, 2002.
3. Beeferman, D, Berger A., Lafferty, J.: Text segmentation using exponential models. En: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, p'aginas 35–46, 1997.
4. Beeferman, D., Berger, A.,Lafferty, J.: Statistical models of text segmentation. En: Machine Learning, volumen 34, páginas 1-3, 1999.
5. Bernárdez, E.: Introducción a la lingüística del texto. Madrid, Espasa-Calpe, 1982.
6. Bolshakov, I.A., Gelbukh A.: Text segmentation into paragraphs based on local text cohesion. En: Proceedings of the 4th International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence, ISBN:3-540-42557-8, páginas 158-66, 2001.
7. Bunge M.: La concentración mediática, peligro para la democracia. México DF. En: Etcétera, 2003.
8. Burger, S., MacLaren, V., Yu, H.: The ISL meeting corpus: The impact of meeting type on speech style. En: Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002), 2002.
9. Filippova, K., Strube, M.: Using linguistically motivated features for paragraph boundary identification. En: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), páginas 267-74, 2006.
10. Genzel, D.: A paragraph boundary detection system. En: Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), páginas 816-26, 2005.
11. Grosz, B., Sidner, C. Attention, intention, and the structure of discourse. En: Computacional Linguistics, volumen 12, número 3, páginas 172-204, 1986.
12. Gruenstein, A., Niekrasz, J., Purver, M.: Meeting structure annotation: data and tools. En: SIGdial6-2005, páginas 117-27, 2005.
13. Halliday, M.A.K., Hasan, R.: Cohesion in English. Longman Group, New York, 1976.
14. Halliday, M.A.K.: Introduction to functional grammar, London: Arnold, 2004.
15. Hearst, M. A.: TextTiling: A quantitative approach to discourse segmentation. En: Technical Report Sequoia, Computer Science Division, University of California, Berkeley, 1993.
16. Hearst, M. A., Plaunt, C.: Subtopic structuring for full-length document access. En: Proceedings of the 16th Annual International ACM/SIGIR Conference, páginas 59-68, 1993.
17. Hearst, M. A.: Context and structure in automated full-text information access. Thesis Doctoral, University of California at Berkeley (Computer Science Division Technical Report), 1994.
18. Hearst, M. A.: Multi-paragraph segmentation of expository text. En: Proceedings of the 32nd Meeting, Association for Computational Linguistics, páginas 9-16, 1994.
19. Hearst, M. A.: TileBars: Visualization of term distribution information in full text information access. En: Proceedings of the ACM SIGCHI, Conference on Human Factors in Computing Systems, 1995.

20. Hearst, M. A.: Improving full-text precision using simple query constraints. En: Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR), 1996.
21. Hearst, M. A., Pedersen, J., Pirolli, P., Schietze, H., Grefenstette, G., Hull, D.: Four TREC-4 Tracks: The Xerox site report. En: Proceedings of the Fourth Text Retrieval Conference (TREC-4), National Institute of Standards and Technology Special Publication, 1996.
22. Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. En: Computational Linguistics, volumen 23, número 1, 1997.
23. Heinonen, O.: Optimal multi-paragraph text segmentation by dynamic programming. En: Proceedings of COLING-ACL '98, Montreal, Canada, Cite as: arXiv:cs/9812005v1 [cs.CL], páginas 1484-86, 1998.
24. Hernández, J.: Introducción a la lingüística textual. En: Clave Contra Clave, Revista educativa, I.S.S.N.: 1988-4559.
25. Hirschberg, J., Litman, D.: Empirical studies on the disambiguation of cue phrases. En: Computational Linguistics, volumen 19, número 3, páginas 501-30, 1993.
26. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, Ch.: The ICSI meeting corpus. En: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), páginas 364–67, 2003.
27. Jinxi, X., Croft, X.b.: Query expansion using local and global document analysis. En: Proceedings of the Nineteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, páginas 4-11, 1996.
28. Kozima, H., Furugori, T.: Similarity between words computed by spreading activation on an English dictionary. En: Proceedings of EACL-93, páginas 232-39, 1993.
29. Kozima, H.: Text segmentation based on similarity between words. En: Proceedings of the 31th Annual Meeting (Student Session), páginas 286-88, 1993.
30. Linguistic Data Consortium – LCTL Team: Simple named entity guidelines. En: SAY project of computing research laboratory, Version 6.5, 2006.
31. Morris, J., Hirst, G.: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. En: Computational Linguistics, volumen 17, número 1, páginas 21-48, 1991.
32. Niekrasz, J., Gruenstein, A.: NOMOS: A semantic web software framework for annotation of multimodal corpora. En: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), 2006.
33. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. En: Computational Linguistics, volumen 16, número 1, 2000.
34. Ponte, J.M., W. Bruce Croft: Text segmentation by topic. En: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, Lecture Notes In Computer Science, ISBN:3-540-63554-8, volumen 1324, páginas 113 – 25, 1997.
35. Regalado, E. M., Regalado E.: Internet: la red de redes en Cuba. En: Revista Educación Médica Superior, volumen 11, número 1, páginas 39-46, 1997.
36. Reynar, J.C.: Topic segmentation: algorithms and applications. Thesis Doctoral, Presented to the Faculties of the University of Pennsylvania, 1998.
37. Reynar, J.C.: Statistical Models for Topic Segmentation. En: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ISBN:1-55860-609-3, páginas 357 – 64, 1999.
38. Salton, G., Wong, A., Yang, C. S.: A Vector Space Model for automatic indexing. En: Communications of the ACM, volumen 18, número 11, páginas 613–20.
39. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. En: Proceedings of International Conference on New Methods in Language Processing, 1994.
40. Schmid, H.: part-of-speech tagging with neural networks. En: Proceedings of the 15th International Conference on Computational Linguistics (COLING-94). 1994.

41. Schmid, H.: Improvements in part-of-speech tagging with an application to german. En: Proceedings of the ACL SIGDAT-Workshop, 1995.
42. Skorochod'ko, E.: Adaptive method of automatic abstracting and indexing. Proceedings of the IFIP Congress 71, páginas 1179–1182, 1972.
43. Soto, G., Zenteno C.: La subtopicalización en el discurso científico escrito. En: Lenguas Modernas, volumen 28, n'umero 29, páginas 29-52, 2001-2003.
44. Soto, G.: La estructuración jerárquica de la información en el discurso científico escrito: segmento de orientación y núcleo informativo. En: Lenguas Modernas, volumen 30, páginas 7-24, 2004-2005.
45. Stokes, N., Carthy, J., Smeaton, A-F.: SeLeCT: A lexical cohesion based news story segmentation system. En: AI Communications, volumen 17, número 1, ISSN:0921-7126, páginas 3 -12, 2004 .
46. Stokes, N.: Applications of lexical cohesion analysis in the topic detection and tracking domain. Thesis Doctoral, Department of Computer Science Faculty of Science, National University of Ireland, Dublin, 2004.
47. Uribe, M. R: El camino de la lectura entre 'topics' y marcas de cohesión. En: Led on Line, ISBN 88-7916-197-0, 2002.
48. Van Dijk,T.: Texto y Contexto. Madrid, Cátedra, 1993.
49. Van Dijk,T.:Estructura y funciones del discurso, México, Madrid, Siglo Ventiuno 1996.
50. Van Dijk,T.:La ciencia del texto, Mexico, Paidós, 1996.

Anexos

Anexo 1

Instrucciones de segmentación para una experimentación sobre segmentación de documentos:

1. Usted recibirá un texto para su segmentación en subtópicos.
2. Marque donde parezca que los subtópicos cambian.
3. Se recomienda que usted lea rápidamente; no es necesario que entienda todos los detalles, aunque puede releerlo si lo necesita.
4. Los subtópicos deben comenzar y finalizar en un párrafo completo, no a mitad de éste.
5. El fondo del párrafo donde usted considere que comience un subtópico debe marcarse en verde manteniendo el texto en negro. No es necesario indicar el inicio del primer subtópico del texto.
6. El fondo del párrafo donde usted considere que finalice un subtópico debe marcarse en azul manteniendo el texto en negro. No es necesario indicar el final del último subtópico del texto.
7. Si en alguna ocasión no puede decidir entre dos párrafos para finalizar el subtópico, distinga en azul el fondo del que considere más apropiado; no obstante, indique en amarillo el fondo del otro que usted consideró también como posible.
8. Se le permite y agradece hacer cualquier comentario o indicación en el texto como, por ejemplo:
    - Redactar con una frase la idea principal del subtópico.
    - Indicar cuáles son los párrafos que contienen la idea principal del subtópico.

RT_007, Noviembre 2009
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: DCG Matilde Galindo Sánchez
RNPS No. 2143
ISSN 2072-6260
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba