BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA.
Facultad de Ciencias de la Computación

Algunas variantes del método de Grefenstette para generar thesauri.

Tesis que presenta
Lic. Elda Mendoza Martínez

Para obtener el grado
Maestro en Ciencias de la Computación

Director de la tesis: **Dr. Héctor Jiménez Salazar**

Puebla, 2005.

Agradecimientos

- A la Vicerrectoría de Investigación y Estudios de Posgrado de la BUAP por haber apoyado la difusión de algunos resultados del presente trabajo.

- Al Dr. Joel Lavalle Martínez por su gentil colaboración en la revisión de los términos del dominio de la Salud.

- A mi asesor Dr. Héctor Jiménez Salazar, por todo el apoyo brindado durante el desarrollo de este trabajo, y por su valiosa contribución para la realización de este trabajo.

- A mis jurados: M.C. David Eduardo Pinto Avendaño y M.C. José Andrés Vázquez Flores , por sus contribuciones efectivas a este trabajo y por su gran disposición.

Dedicatoria

A mis papás...

Sabiendo que jamás existirá
una forma de agradecer
toda una vida de sacrificios y esfuerzos,
quiero que sientan que el objetivo logrado
también es suyo y que la fuerza que me ayudó
a conseguirlo fue su apoyo.

Con cariño y admiración...

Elda

Índice general

Introducción **1**

1. Bases para la construcción automática de *thesauri* **5**

1.1. Colocaciones 6

1.2. Los *corpora* 8

1.3. Términos multipalabra 10

1.4. Inducción léxica 10

1.5. Construcción de *thesauri* 12

1.6. Método de Grefenstette 16

2. Identificación de términos **18**

2.1. Determinación del umbral 18

2.2. Términos sencillos 21

2.3. Algoritmo para identificar términos multipalabra 24

2.4. Resultados 28

3. Construcción automática de *thesauri* **32**

4. Conclusiones **39**

A. Términos de Computación **45**

B. Términos importantes de Salud **47**

C. *Thesaurus* **49**

Lista de figuras

2.1. Vocabulario con frecuencias y el PT. 19
2.2. Preprocesamiento del texto 22
2.3. Vocabulario con frecuencias y distancia al PT. 23
2.4. Bolsa resultante del proceso 24

Lista de tablas

1.1. *Corpora* utilizados. 9

2.1. Determinación del mejor umbral (CorpusPor). 20
2.2. Identificación de términos del dominio (CorpusSal). 28
2.3. Determinación de mejor combinación de métodos (CorpusMin). 29
2.4. Desempeño final para identificación de términos del dominio (CorpusCom). 30

3.1. Algunas entradas del *thesaurus* constituido usando una oración como contexto (CorpusCom). 36
3.2. Algunas entradas del *thesaurus* construido usando tres oraciones como contexto (CorpusCom). 36
3.3. Comparación de las dos formas de elegir contextos (CorpusCom). 37

Introducción

Los *thesauri* vienen desarrollándose como información documental. Parece interesante recordar algunos hitos de su evolución, desde 1852 fecha de publicación de la primera edición del reconocido *thesarus* literario de Roget [26], y de 1878 cuando Poole discutiera las características que Roget otorgó al índice de su obra. Pasando por 1948, fecha en que Bernier [25] definió al *thesarus* documental como "herramienta conceptual de relaciones entre términos de tipo postcoordinado". Sin embargo hasta 1951, Hans P. Luhn [16] fue el primero en emplear el término *thesaurus*. Uno de los objetivos es utilizarlos para la recuperación documental en bases de datos escasamente estructuradas (no relacionales) que no hayan sido indizadas previamente con ningún otro tipo de lenguaje controlado (ej. las bases de datos de prensa) [23].

En general, un *thesaurus* interrelaciona los términos de una lengua para construir un lenguaje documental estructurado. Su fin es servir a los indizadores para representar los documentos en las bases de datos, y actuar como guía terminológica para la estandarización de las entradas de los encabezamientos de materias en la clasificación de dichos documentos. Además, posibilita las tareas de recuperación documental: ecuaciones de búsqueda, navegación,

asociación terminológica, comprensión del entorno de los descriptores del *thesaurus*, etc.

Con la base de datos léxica WordNet para el inglés (y EuroWordNet para otras lenguas) ha sido posible explotar las relaciones semánticas para proponer mejoras en las herramientas para el procesamiento de textos. Sin embargo, estas bases de datos son de índole general, y el objetivo de abarcar los diferentes sublenguajes queda por resolverse aún ahora aprovechando los métodos y recursos existentes. Es de particular interés construir *thesauri* para sublenguajes como el científico, en el que hay abundancia de textos de utilidad. Algunas de las características de este sublenguaje científico son:

- La abundancia de términos compuestos que definen los conceptos o entidades científicas (a mayor especialización, mayor complicación terminológica).
- Frente al lenguaje general de variada "riqueza terminológica" (connotación), los conceptos científicos deben expresarse a través de claras referencias (denotación).
- Sus términos no tienen cabida en los diccionarios generales, o cuando la tienen no incluyen la acepción deseada para el dominio científico. Deberemos, entonces, recurrir a diccionarios o enciclopedias especializadas.
- La utilización de una sintaxis propia de este sublenguaje, no claramente diferenciada de la del lenguaje natural.

- Y la existencia de un estilo diferenciado, tomado de la doctrina científica habitual.

Los primeros desarrollos prácticos enfocados a la "recuperación conceptual", datan de los años ochenta en Estados Unidos. Se desarrollaron dentro de los sistemas integrados de gestión de la información de grandes instituciones como la API-CAIS (American Petroleum Institute), la NASA, o la NLM (National Library of Medicine)[23].

Aunque las investigaciones se iniciaron para mejorar la indización, después se enfocaron hacia la recuperación automatizada en las grandes bases de datos documentales. Mediante las herramientas de la Inteligencia Artificial (IA), se aplicaron al desarrollo de agentes inteligentes en campos específicos del conocimiento científico. Autores como Bates [22], Schmitz-Esser [31] o Milstead [19] nos han presentado las características de este tipo de *thesauri* conceptuales, y la indeterminación de las búsquedas en lenguaje natural de los usuarios no expertos, con el sublenguaje científico de los profesionales de un área específica del conocimiento.

Cuando el *thesaurus* se construye automáticamente, es decir, sin información adicional de relevancia del usuario, se distinguen varios enfoques [6]:

- *Thesauri* construidos a partir de la medida simple de coocurrencias de términos [20]. La similitud entre términos se realiza basándose en la Hipótesis de Asociación: "si un término es buen discriminante de documentos relevantes y no relevantes, sus términos asociados también lo serán"[8].

- *Thesauri* construidos a partir de *clustering* de documentos [3]. Primero se clasifican los documentos, y los términos poco frecuentes en una clase, se utilizan para construir el *thesaurus* de términos relacionados.

- *Thesauri* basados en información sintáctica. La relación entre términos se realiza en base a conocimiento sintáctico y análisis de coocurrencias. Se emplean gramáticas y diccionarios para obtener los términos relacionados con uno dado [11].

En este trabajo experimentamos con el primer enfoque, pues es relativamente simple y efectivo aplicado al dominio de la computación. En el primer capítulo se asentaron las bases para la obtención de términos multipalabra, del método de inducción léxica y de la construcción de *thesauri*. En el segundo capítulo se explica cómo se obtuvieron los términos importantes del dominio, el uso del punto de transición, así como el método de inducción léxica, con experimentos que sustentaron las decisiones tomadas. En el tercer capítulo se realiza la construcción de un *thesaurus*, apoyándose en el método de Grefenstette. Por último, en el capítulo cuatro se exponen las conclusiones del presente trabajo.

Capítulo 1

Bases para la construcción automática de *thesauri*

Para construir un *thesaurus* requerimos material que permita determinar sus componentes: las entradas, que son términos de un vocabulario específico; y, las palabras relacionadas, que también serán parte de este vocabulario.

Hay varias formas de conformar el vocabulario de un dominio. Por supuesto, una es hacerlo manualmente, o bien, basándose en una lista de términos importantes obtenidos automáticamente.

En este capítulo se ofrece un panorama de los métodos utilizados.

1.1. Colocaciones

Estamos interesados en llevar a cabo este proceso de manera automática, y basamos nuestro enfoque en la extracción de colocaciones. Una colocación es una expresión que consiste de dos o más palabras las cuales corresponden a una manera convencional de decir las cosas, o en otras palabras según Firth: [17] "Las colocaciones de una palabra dada son frases de la palabra en lugares habituales o acostumbrados". O bien, una colocación es definida, por Choueka [32], como una secuencia de dos o más palabras consecutivas que tienen características de una unidad semántica y sintáctica con un significado exacto y no ambiguo o connotación que no puede ser derivada directamente del significado de sus componentes. Las colocaciones incluyen frases nominales, frases comunes y otras más.

Las colocaciones son importantes para un gran número de aplicaciones: en la generación de lenguaje natural, para estar seguros de que la salida no contenga un error; en lexicografía computacional, para identificar automáticamente las colocaciones importantes a ser listadas como entradas en un diccionario; y en la búsqueda de términos, para aumentar la precisión de los resultados.

El interés por las colocaciones es debido a que muestran diferentes maneras en la cual una palabra es usada. Al formar colocaciones se encuentran términos multipalabra los cuales tienen independencia y probablemente una aparición en un diccionario.

Claramente, el método más fácil para encontrar términos multipalabras, en un *corpus* es contando; es decir, si dos términos ocurren frecuentemente juntos, entonces es evidente que tienen una función especial. El sólo seleccionar los bigramas (secuencia de dos palabras adyacentes de ocurrencia consecutiva) no es muy adecuado ya que encuentra pares de palabras que no forman una unidad referencial (un solo referente). Un experimento [18] mostró que utilizando una heurística simple se mejoran estos resultados: etiqueta el *corpus* y utiliza un filtro con partes del discurso para obtener las colocaciones. Este método, pese a su simplicidad, arrojó buenos resultados. Los métodos hacen referencia a los bigramas que aparecen en el *corpus*. A partir de los bigramas se pueden generalizar los métodos considerando trigramas, etc.

Hay diversos enfoques para garantizar que un bigrama sea efectivamente una colocación, por ello se utiliza una prueba estadística como lo es la *información mutua*. La información mutua es una medida no negativa y simétrica de la información en común entre dos variables; en otras palabras, qué tanto una palabra nos dice de la otra [14].

En resumen, tenemos que acudir a métodos variados para la determinación de la terminología de un dominio. Estos métodos, y en general los empíricos, se apoyan en *corpora* del dominio que se trata.

1.2. Los *corpora*

En esta sección se introducen las características generales de los *corpora*, y presenta los *corpora* sujetos a nuestra experimentación.

La lingüística tiene el concepto de *corpus* como una muestra de textos reales suficientemente grande de una lengua determinada. Por su parte, la lingüística computacional lo define como una colección de textos codificados electrónicamente, una base de datos o archivo textual que se integra en un sistema de almacenamiento y recuperación de la información, un conjunto de bases de datos textuales unidas en un sistema de estructuración de datos, textos, referencias y utensilios informáticos para su tratamiento en conexión directa a una computadora [23]. Los textos se archivan fundamentalmente para que constituyan un gran depósito ordenado que sirva para satisfacer necesidades de información en la realización de proyectos como diccionarios o enciclopedias electrónicas, sistemas de traducción por computadora, de consulta bases de datos en lengua natural, o como banco de pruebas para la comprobación de hipótesis o análisis lingüísticos expresados mediante una gramática formal. El primer proyecto de *corpus* que se tiene noticia fue de la lengua inglesa británica escrito y hablado elaborado por Randolph Quirk "Survey of English Usage"(SEU) en 1959 [27]. Poco después, Nelson Francis y Henry Kucera [24], en la Universidad de Brown, empiezan a trabajar en la creación del Brown *corpus* que se define como una muestra estandarizada del inglés americano en forma impresa destinada al procesamiento en computadora, se encuentra etiquetado y balanceado, desafortunadamente implica un monto económico el obtenerlo. Después, el *corpus* Lancaster-Oslo-Bergen

(LOB) [29]fue construido como una réplica en Inglés Británico del *corpus* Brown, el cual sigue estando en el idioma Inglés. El *corpus* Canadian Hansards [5]es el mejor ejemplo de un *corpus* bilingüe, éste contiene textos paralelos en dos o más idiomas que en realidad son traducciones de cada uno. A partir de este momento aparecen trabajos (principalmente para el inglés) de creación y explotación de *corpus* lingüísticos. Hasta ahí llegamos en ejemplos de *corpus* disponibles en Internet, claro con un costo.

A continuación se muestra una tabla de *corpora* utilizados en los experimentos de este trabajo; en ella se muestra el nombre con el que nos referiremos a cada *corpus* (columna 1), su tamaño (columna 2, columna 3 y columna 5, en palabras, número de textos y número de caracteres, respectivamente), y el dominio al que pertenecen (columna 4).

Nombre	**Palabras**	**Textos**	**Dominio**	**Tamaño**
CorpusSal	149,219	8	Salud (Nutrición, Planificación, Asma, etc.)	965 Kb
CorpusCom	162,931	33	Computación (S.O., B.D., Robótica, I.A., etc.)	932 Kb
CorpusMin	118,212	18	Computación (Minería)	736 Kb
CorpusPor	10,106	3	Computación (Seguridad, Algoritmos Genéticos e Internet	64 Kb

Tabla 1.1: *Corpora* utilizados.

1.3. Términos multipalabra

De la colección de textos lo que mas nos interesa son los términos relevantes del dominio para identificar las palabras que representan a un texto. Lo anterior se hará con base en un resultado derivado de la Ley de Zipf [2]. El experimento realizado empleó el, así llamado, punto de transición. El punto de transición es la frecuencia de un término del texto que divide a los términos en los de alta y baja frecuencia.

$$PT = \frac{\sqrt{1 + 8I_1} - 1}{2} \tag{1.1}$$

Una vez obtenido el punto de transición, se seleccionó un conjunto de términos alrededor de él para conformar el conjunto de palabras que representan al texto. Los experimentos reportados en [28] indican que al tomar una banda de frecuencias del 25 % alrededor del punto de transición se obtienen buenos resultados. Los detalles de esta determinación se explica en la sección 2.3.

1.4. Inducción léxica

Gierl y Frost [4],identifican terminología de un dominio específico basándose, principalmente, en la medida de información mutua y la inducción léxica. Se utiliza un índice de asociación para cuantificar las asociaciones entre palabras y las ocurrencias léxicas dentro de un texto del *corpus*. Esta medida está basada en el concepto teórico de Información Mutua. Este índice de

asociación mide una ocurrencia ordenada linealmente de las palabras. Entre palabras contiguas es usada la siguiente fórmula:

$$IM(X,Y) = \log_2 \frac{f(X,Y)}{f(X)f(Y)} \tag{1.2}$$

donde f(x,y) es la frecuencia en el texto origen del compuesto ordenado del bigrama (x,y) y de f(x), f(y) la frecuencia de las palabras que lo constituyen.

El enfoque descrito de la información mutua puede ser usado para seleccionar todas las asociaciones de bigramas con un valor de información mayor dentro de un *corpus*. Para maximizar el dominio especificado del conjunto seleccionado de términos, la información mutua puede ser combinada con una técnica inductiva en la cual hay un conjunto de palabras que restringen los términos seleccionados de información mutua alta tanto como sean generados a partir de términos previamente elegidos.

En el experimento presentado en [13] el *corpus* fue normalizado cambiando mayúsculas por minúsculas, no se preprocesan errores ortográficos o de plurales, y cada ciclo empezó con un nuevo conjunto de palabras. Para el ciclo inicial fueron tomadas las palabras de los encabezados. Un conjunto de pares sintácticos fueron seleccionados tomando en cuenta que cada par:

- contuviera al menos una palabra clave de dónde salieron (esta es una ventaja del método que se implementó).
- ocurriera al menos dos veces en el *corpus*.

- tuviera información mutua mayor a un umbral seleccionado previamente.

Para maximizar la asociación en el dominio específico, un subconjunto de los pares seleccionados fueron producidos usando el umbral de información mutua más restringido ($IM > 6$, así como se muestra en el algoritmo de la sección 2.3).

La exploración de este método, ocasionó varios efectos al variar los valores de los cuatro parámetros experimentales siguientes:

1. El conjunto inicial de las palabras clave.
2. El umbral más bajo de información mutua para recolectar pares.
3. El umbral más alto de información mutua para pares usados para generar las nuevas palabras clave.
4. El número de ciclos iterativos.

Con un *corpus* finito, el proceso siempre terminará cuando no se produzcan más palabras clave, como se muestra en el ejemplo de la sección 2.3.

1.5. Construcción de *thesauri*

Se pueden distinguir dos enfoques principales en la recuperación de información, los que se basan en técnicas simbólicas y los que utilizan técnicas empíricas.

Las técnicas de Inteligencia Artificial para el tratamiento del lenguaje natural han sido etiquetadas [7] como de *conocimiento rico* debido a que requieren una gran inversión para conocer las estructuras de dominio específico antes de que puedan ser aplicadas al tratamiento del texto. El costo de crear y mantener este conocimiento ha sido reconocido, lo cual ha alentado a algunos investigadores a explorar otras posibles maneras de acelerar o automatizar su adquisición. Esta perspectiva ha motivado un sin número de enfoques de *conocimiento pobre* para extraer información semántica de dominios específicos partiendo de fuentes existentes. Un enfoque de conocimiento pobre para la extracción automática de información semántica es explotar la aparición de patrones de texto en los documentos cuya estructura semántica es conocida. Tal técnica puede ser vista como reciclar los juicios humanos de cómo las palabras están relacionadas. La tarea de la técnica de conocimiento pobre es reconocer patrones que mecánicamente pueden ser explotados sin establecer relación a un nivel más profundo de comprensión.

La mayoría de los métodos de extracción semántica usando conocimiento pobre fue basada en estadísticas de la ocurrencia de palabras dentro de una misma *ventana* de texto; donde una ventana puede ser un cierto número de palabras, sentencias, párrafos o un documento entero alrededor de un término. El enfoque de la coocurrencia de términos en un documento demuestra el poder de las técnicas de conteo para el descubrimiento semántico.

El enfoque más clásico de conocimiento pobre en la extracción semántica usando coocurrencia es utilizando una pequeña ventana (cuatro o cinco

palabras) para extraer las palabras que comúnmente se encuentran alrededor del término [12]. Es fácil de implementar debido a que no requiere de información léxica, aunque usualmente utiliza una lista de palabras cerradas, tales como artículos, preposiciones, etc. las cuales son eliminadas ya que se consideraban irrelevantes. El contexto alrededor de la palabra es usado de dos maneras: para calcular qué palabras aparecen frecuentemente juntas y para determinar qué palabras comparten el mismo contexto. Este enfoque lo usaron Church & Hanks [21] basándose en una definición teórica de información mutua, la cual compara la probabilidad de observar dos palabras por separado y de observar cada palabra independientemente; $P(xy)$ y $P(x)P(y)$. Las palabras que tienen el valor de información mutua alto en un *corpus*, usualmente tienen una relación semántica.

Estas técnicas de conocimiento pobre que usan números de ocurrencia o frecuencia de palabras dentro de un documento o una ventana son aplicables ciertamente a *corpora* de cualquier dominio. Para la coocurrencia de los documentos se usa comúnmente métodos estadísticos, aunque se presentan tres problemas con ello:

1. Cada palabra en un documento es considerada potencialmente relacionada con otra, no importa la distancia entre ellas, por ej. Las palabras del principio y del final del documento pueden presentarse como un par para dicha técnica, aunque no exista ninguna conexión entre los temas discutidos en esos dos puntos, lo mismo sucede cuando la ventana es pequeña.

2. Por más que se utilicen técnicas de agrupamiento semántico, para la

coocurrencia en los documentos, sólo serán similares si aparecen físicamente en el mismo documento cierto número de veces. En general, las palabras diferentes usadas para describir conceptos similares, nunca pueden ser usadas en el mismo documento y, por lo tanto, estos métodos no las toman en cuenta.

3. Las medidas de similitud generalmente son de orden cuadrático o cúbico, usando "aparición en el mismo documento "como un atributo, significa que se restringirá a un *corpus* pequeño de unos cuantos documentos.

También, se ha explorado un campo intermedio entre el conteo simple de palabras y los enfoques de conocimiento rico. Se acepta que es necesario un cierto nivel de análisis sintáctico, y es posible realizar la tarea sin utilizar estructuras de conocimiento rico asociadas con un objeto léxico.

Los ejes semánticos alrededor de una palabra dada. Se definen modificando el concepto introducido en [10] donde se definieron las palabras como "vecinos recíprocamente cercanos ": X es vecino cercano de Y si la palabra X aparece en la lista de similitud de Y dentro de las N primeras palabras (Grefenstette usó $N = 10$ [13]). Estas palabras pueden servir como semillas para la definición de eje semántico. Considérese que una palabra A fue encontrada cerca de B, C, D, E y F, se supone que B fue cercano recíprocamente a A; esto es, si suponemos que A también fue una de las palabras más cercanas a B. Podemos confiar que A-B forman un eje semántico y tratan de conectar a las otras palabras C, D, E y F a este eje. Esto se justifica porque cualquiera de esas palabras son también vecinos cercanos de B, independiente de

A. Esto define un conjunto de palabras las cuales son cercanas a A, vecinos cercanos de B, y cercanos a este eje suponiendo que A-B sea un eje semántico.

1.6. Método de Grefenstette

El enfoque más simple para construir un *thesaurus* automáticamente lo propuso Grefenstette [13] basándose en trabajos previos, p.e. Hindle [10]. Las bases en los que se apoya dicho enfoque se presenta a continuación.

Grefenstette aplicó en 1996, técnicas de conocimiento pobre para generar automáticamente *thesaurus* de textos sin procesar. Utilizó como entrada un megabyte como *corpus* del dominio de medicina. Para cada término se tiene:

- Palabras Relacionadas.
- Verbos comúnmente asociados.
- Expresiones comunes.
- Palabras de la misma familia.

Para cada información utiliza diferentes técnicas, particularmente nos vamos enfocar en las técnicas que utiliza para la extraer las palabras relacionadas.

Las palabras relacionadas son extraídas por un paquete llamado SEXTANT [12]. Y sigue los siguientes pasos:

1. Los textos del *corpus* son separados por *tokens* utilizando una gramática regular.

2. Los *tokens* son analizados morfológicamente y son etiquetados con su parte del discurso.

3. El texto etiquetado es desambiguado por un desambiguador estocástico que provee una etiqueta simple para cada *token*.

4. El texto desambiguado es analizado sintácticamente y se extraen las dependencias entre las palabras.

5. La representación de un término toma como atributos a los sustantivos, todos los adjetivos, los verbos de los cuales son sujetos u objetos y otros sustantivos que los modifican, o cláusulas preposicionales.

6. La medida de Jaccard se utiliza para calcular la similitud entre atributos de sustantivos.

7. Las entradas de los sustantivos de mayor similitud para formar a los vecinos recíprocamente cercanos; este concepto es una ligera extensión de la idea descrita en [10].

Por otra parte, en un enfoque similar [9], se concluyó que al extraer frases nominales que incluyan preposiciones y conjunciones es posible producir frases largas y complejas que enriquecen al método.

Capítulo 2

Identificación de términos

En esta sección se describe la forma en que se obtuvieron los términos multipalabra. En este trabajo de tesis tales términos se tomarán como base en la construcción de un *thesaurus*. Primero se describe cómo se extraen los términos sencillos representativos del dominio utilizando el punto de transición. Después se aplica un método de inducción léxica, el cual itera hasta encontrar los términos multipalabra más grandes posibles.

2.1. Determinación del umbral

Para determinar las palabras relevantes de un texto, se utilizó el punto de transición, usando una fórmula derivada de la Ley de Zipf, y una variante que se basa en el desarrollo de la Ley de términos de baja frecuencia [1]. Esta última define al PT como la mayor frecuencia que no se repite en el vocabulario de un texto ordenado por sus frecuencias. Esto es, alternativamente a

la ecuación 1.1 el PT se puede obtener de la siguiente manera:

Dado el vocabulario de un texto, se obtienen las frecuencias de las palabras, se ordenan descendentemente, y se determina la frecuencia que cumple con la condición antedicha. En la figura 2.1 se muestra un fragmento del vocabulario de un texto ordenado descendentemente y la frecuencia que representa el punto de transición:

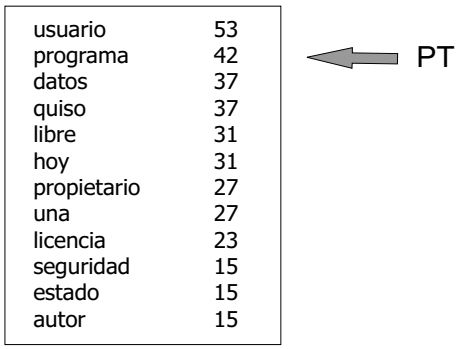

Figura 2.1: Vocabulario con frecuencias y el PT.

Se parte de que la frecuencia de los términos con alto contenido semántico está alrededor del PT. Por tanto, para la determinación del umbral y la selección de los términos importantes, se experimentó variando el porcentaje de términos tomados alrededor del PT.

Se utilizó el corpus CorpusPor (tabla 1.1), se determinó el punto de transición mediante la variante antes mencionada, y para cada texto se tomaron tantos términos cercanos al PT como el 30 %, 40 % y 50 % del valor del PT, dando como resultado los valores de precisión y evocación que se muestran

en la tabla 2.1.

		Primer Texto	**Segundo Texto**	**Tercer Texto**
	Tamaño	(25 KB)	(20 KB)	(19 KB)
Umbral	**PT**	37	17	12
30 %	**Precisión**	0.75	0.67	0.00
	Evocación	0.50	0.17	0.00
40 %	**Precisión**	**0.60**	**0.75**	**0.00**
	Evocación	**0.50**	**0.25**	**0.00**
50 %	**Precisión**	0.47	0.67	0.17
	Evocación	0.50	0.25	0.20

Tabla 2.1: Determinación del mejor umbral (CorpusPor).

El cálculo de la precisión (P) y evocación (E) usa las formulas:

$$P = \frac{\#(R \cap S)}{\#S} \tag{2.1}$$

$$E = \frac{\#(R \cap S)}{\#R} \tag{2.2}$$

donde R son los resultados según el experto pertenecen al dominio y S son los resultados arrojados por el programa.

Así, para obtener los cálculos presentados en la tabla 2.1, se realizó la identificación manual de términos importantes en cada texto por un experto.

En los experimentos realizados, se llegó a la conclusión de tomar el 40 % de los términos alrededor del punto de transición, ya que presenta mejor evocación y precisión que 30 %, y solamente en un caso (texto más pequeño) mejoró con el 50 %.

2.2. Términos sencillos

Una vez determinado el umbral usaremos en nuestra aplicación el procedimiento que a continuación se describe. Dado un texto del corpus Corpus-Com (tabla 1.1), el cual se preprocesa eliminando los símbolos no alfabéticos, así como las palabras cerradas (artículos, conjunciones, etc), las palabras restantes constituyen el vocabulario y se obtiene la frecuencia de las palabras. El siguiente paso es determinar el punto de transición usando el algoritmo mencionado anteriormente (frecuencia que no se repite). Se obtiene así, la distancia de cada frecuencia del vocabulario al punto de transición:

$$DPT_i = |fr_i - PT| \tag{2.3}$$

donde fr_i es la frecuencia del término.

Por último, se ordena la lista de palabras según DPT_i y se toman tantas palabras del inicio de la lista como el 40 % del PT. En resumen, el método para obtener términos alrededor del punto de transición se precisa con el algoritmo siguiente:

Entrada: Texto: T

Salida: Lista de términos: L

1. $T\prime = Preprocesa(T)$.
2. $Voc = ObtieneVocabulario(T\prime)$.
3. $Vocf = Frec(Voc)$.
4. $PT = DeterminaPT(Vocf)$.

5. $Vocf\prime = DistanciaPT(Vocf, PT)$.

6. $L = OrdenaSeleccion(40\,\%, Vocf\prime)$.

A continuación se muestra un ejemplo del proceso de obtención de términos importantes en un texto (fig. 2.2). En el primer recuadro se presenta un texto, al que se le aplica el pre-procesamiento (cambio de mayúsculas a minúsculas, eliminación de algunos símbolos excepto los signos de puntuación .,:;¿?¡!) y, así, obtenemos el texto del segundo recuadro.

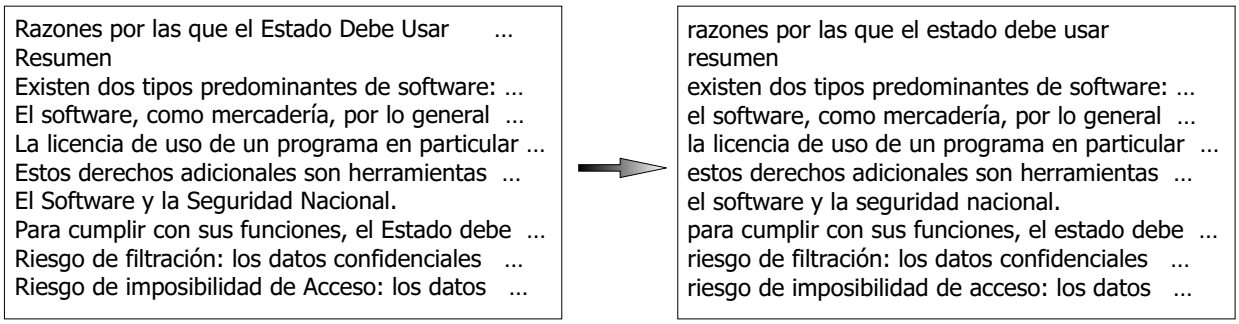

Figura 2.2: Preprocesamiento del texto

Después, como se ha dicho, se determina el punto de transición; esto es, se obtienen las frecuencias de las palabras del vocabulario como se muestra en el recuadro de la izquierda y, en el recuadro de la derecha, el vocabulario ordenado por la distancia al punto de transición para determinar los términos que se sitúan cerca del punto de transición.

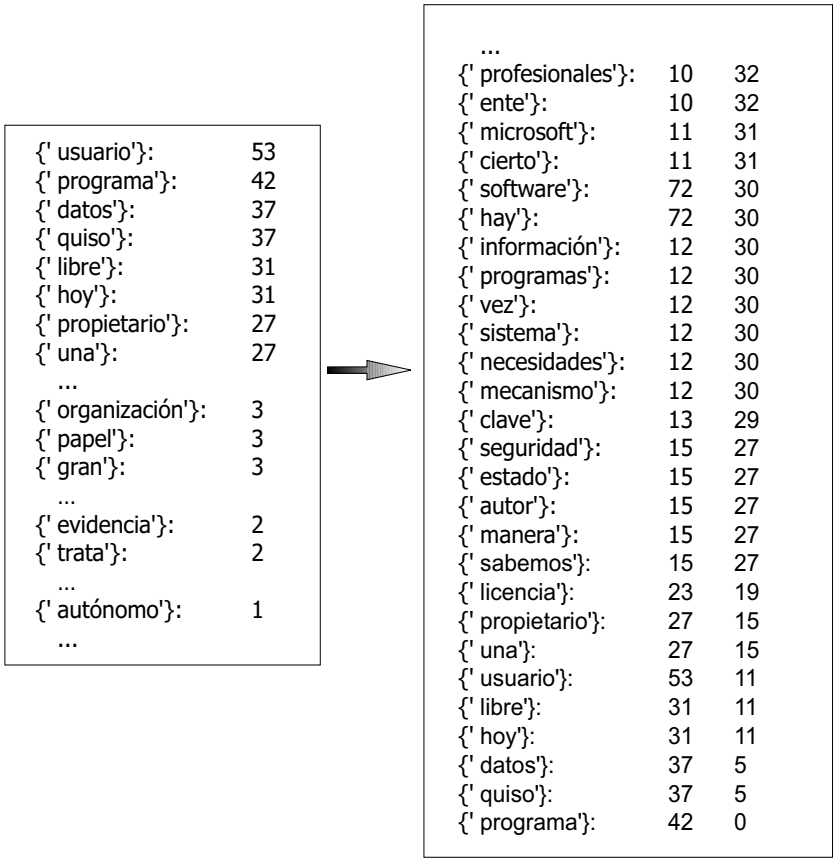

Figura 2.3: Vocabulario con frecuencias y distancia al PT.

Finalmente se obtienen los términos que se presentan en la figura 2.4 usando el umbral que se determino en la sección 2.1.

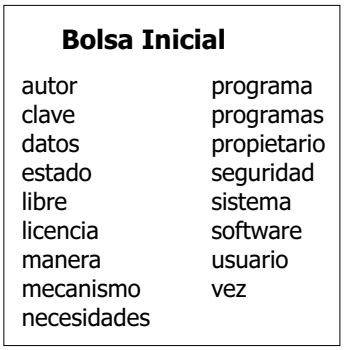

Figura 2.4: Bolsa resultante del proceso

2.3. Algoritmo para identificar términos multipalabra

El método de inducción léxica es un método iterativo para encontrar términos multipalabra, y termina al encontrar un punto fijo (cuando al iterar no se obtienen nuevos resultados).

Dada la bolsa inicial, se toman en cuenta los términos que en ella están para formar bigramas candidatos a partir del texto, siempre y cuando cumplan que:

1. La frecuencia individual de los términos sea mayor a tres.
2. Los términos que conforman el bigrama candidato no sean verbos.
3. La información mutua del bigrama sea mayor que 6.

Se utilizó una medida estadística llamada información mutua, la cual determina qué tanta información tiene un término del otro para coocurrir en

un texto, y se obtiene con la siguiente fórmula:

$$IM(X,Y) = \log_2(\frac{N * fr(X,Y)}{fr(X) * fr(Y)} + 1) \tag{2.4}$$

Una vez que el bigrama candidato supera las restricciones es aceptado. Además, si un término del bigrama no se encontraba en la bolsa se agrega.

La primera condición para formar bigramas fue tomada de [7]; asegura que no sean términos espurios. Los términos multipalabra que buscamos constituyen frases nominales simples o partes de ellas, por tanto no tienen verbos. Y, la tercera condición se inspira en el trabajo de K.Church & P. Hanks [21]; de igual forma, ellos señalan que en *corpora* de cierto tamaño (>1MB) la IM de los términos asociados es mayor que 10. Así, algunas pruebas con nuestro *corpus* nos indicaron tomar el umbral de IM mayor a 6.

El algoritmo siguiente resume las ideas anteriores:

Entrada: Texto: T
Lista de términos: L
Salida: Lista de términos importantes: B_1

$B_0 = \{\}$; $B_1 = L$
While ($B_0 \neq B_1$)
$B_0 = B_1$
FormaBigPos (term1, term2, Big, T, B_0)
If (fr(term1)>3 AND fr(term2))>3

If (Pos(term1) OR Pos(term2)) <> VERB

IM = InformacionMutua (Big)

If IM>6

$B_1 = B_1 +$ Big

If (term2 $\notin B_1$)

$B_1 = B_1 +$ term2

End

El algoritmo utiliza un etiquetador *a priori* (Pos) para identificar los verbos.

A continuación se muestra un ejemplo de la aplicación del método de inducción léxica aplicado a un texto del corpus CorpusCom (tabla 1.1). La lista del 40 % de los términos más cercanos al punto de transición forma la bolsa inicial del método:

Bolsa Inicial (B_0)	{ programa, datos, libre, propietario, usuario, licencia, seguridad, manera, autor, estado, clave, necesidades, vez, programas, mecanismo, sistema }

Después de la obtención de bigramas candidatos y la verificación de las restricciones, en la primer iteración, la bolsa queda de la siguiente manera:

Iteración 1 $B_1 =$ { programa, datos, libre, propietario, usuario, licencia, seguridad, manera, autor, estado, clave, necesidades, vez, programas, mecanismo, sistema, sistema_operativo } $\neq B_0$

Como se ve, se incrementó la bolsa al aceptar un nuevo término multipalabra que es el caso de sistema_operativo. Por este motivo se realiza otra iteración :

Iteración 2 $B_1 =$ { programa, datos, libre, propietario, usuario, licencia, seguridad, manera, autor, estado, clave, necesidades, vez, programas, mecanismo, sistema, sistema_operativo } $= B_0$

Como no hubo aumento en la bolsa, el algoritmo termina. Se aplicó el método de inducción léxica al corpus CorpusSal (tabla 1.1), los resultados fueron validados por un experto en el area de Salud. En la tabla 2.2 se muestran los resultados.

	Porcentaje de Aciertos	Porcentaje de Fronterizos	Porcentaje de Fracasos
Separado por fórmula	38.9	35.5	25.6
Separado por lista	82.4	14.1	3.5

Tabla 2.2: Identificación de términos del dominio (CorpusSal).

2.4. Resultados

En uno de los experimentos realizados se empleó el *corpus* CorpusMin (tabla 1.1). Se aplicaron los dos métodos al *corpus* para calcular el punto de transición (por fórmula y por lista). También, los métodos anteriores se probaron de dos formas; de manera global y para cada texto. En forma global significa que se aplican los métodos directamente al texto resultante de concatenar los textos individuales. En forma individual, los métodos se aplican a cada uno de los textos; el resultado se forma con los términos encontrados en cada caso.

Se obtuvieron así, las siguientes listas:

- Lista de palabras realizando el cálculo del PT mediante fórmula al *corpus* completo.
- Lista de palabras realizando el cálculo del PT mediante fórmula a los textos (unión de las bolsas resultantes de cada texto).
- Lista de palabras realizando el cálculo del PT por lista al *corpus* completo.

- Lista de palabras realizando el cálculo del PT por lista a los textos (unión de las bolsas resultantes de cada texto).

Una vez que se obtienen las listas, se organizan para facilitar el trabajo manual de elegir los términos que a juicio de un experto son representativos del dominio, los términos que no son representativos del dominio pero están "cercanos"a él y, por último, los que no son representativos del dominio.

En la tabla 2.3 se muestra el porcentaje de los términos encontrados al emplear los métodos de las dos formas.

	Porcentaje de Aciertos	**Porcentaje de Fronterizos**	**Porcentaje de Fracasos**
Global por fórmula	36.7	36.7	26.7
Global por lista	36.3	38.4	25.3
Separado por fórmula	58.1	23.9	18.0
Separado por lista	**33.9**	**48.2**	**17.9**

Tabla 2.3: Determinación de mejor combinación de métodos (CorpusMin).

Convenimos que los términos contabilizados en aciertos son los que pertenecen al dominio, los fronterizos aquellos que no son del dominio pero son imprescindibles para él, y, en los fracasos, se cuentan los términos decididamente ajenos al dominio.

Al emplear inducción léxica sobre términos importantes puede obtener-

se un porcentaje alto de términos aceptables; calculando el PT "por lista"y en "textos separados"(82.1 % en nuestro experimento), ya que se tomó en cuenta que los términos fronterizos también son representativos del dominio. La ligera diferencia entre los dos métodos, para calcular el PT en textos separados, condujo a un segundo experimento.

El siguiente experimento se realizó empleando el *corpus* CorpusCom (tabla 1.1). Se aplicaron los dos métodos para calcular el punto de transición (por fórmula y por lista) al *corpus*, en forma global y en forma individual. Al igual que en el caso anterior, manualmente se eligen los términos que a juicio de un experto son representativos del dominio, los términos que no son representativos del dominio pero están "cercanos"a él y, por último, los que no son representativos del dominio. En la tabla 2.4 se muestra el porcentaje de los términos encontrados en este experimento:

	Porcentaje de Aciertos	**Porcentaje de Fronterizos**	**Porcentaje de Fracasos**
Global por fórmula	50.0	17.0	33.0
Global por lista	47.7	18.6	33.7
Separado por fórmula	**61.8**	**16.7**	**21.5**
Separado por lista	43.6	19.9	36.5

Tabla 2.4: Desempeño final para identificación de términos del dominio (CorpusCom).

En conclusión:

1. La mejor combinación de técnicas para determinar los términos importantes es usando la fórmula y aplicándola a cada texto.

2. El uso del PT para determinar términos importantes de un texto es atractivo por su sencillez.

Capítulo 3

Construcción automática de *thesauri*

Un *thesaurus* es un diccionario que muestra palabras relacionadas semánticamente. En esta sección se emplean los términos importantes de un dominio que se obtuvieron en la sección anterior. La técnica desarrollada por Grefenstette ha sido presentada en el capítulo 1 en la sección 1.5 y fue la que aquí tomamos como base.

Se hicieron algunos cambios al método de Grefenstette para conocer su alcance empleando otros recursos. Brevemente, se procede de la siguiente manera. Primero se obtienen los contextos de los términos, después se aplica una fórmula de similitud a los contextos para determinar los vecinos cercanos de los términos. Por último, se forman parejas relacionadas apoyándose en criterios de umbralización sobre los vecinos cercanos.

Las variantes con respecto al método de Grefenstette son:

1. Los contextos de cada término t fueron determinados por las oraciones (secuencia de términos entre puntos) que contuvieran a t mientras que Grefenstette usa un contexto sintáctico (el verbo principal de la oración, la cabeza del objeto directo, etc).

2. Como atributos de los términos de entrada se utilizan los términos importantes de los textos, a diferencia del método de Grefenstette que usa los más frecuentes

Dados los términos importantes y sus vecinos cercanos, se procede a la construcción del *thesaurus*, para lo que se realizará otro pre-procesamiento del texto, esta vez se eliminarán palabras cerradas y los signos de puntuación excepto el punto. El método para construir el *thesaurus*, ThesGref, se precisa con el algoritmo siguiente:

Entrada: Texto: T
Lista de términos importantes: Bo
Salida: *thesaurus*: D

Para cada $X \in Bo$
Contextos = ObtenerContextos (T,X).
Contextos$\prime$ = Ordena_Frec [Vocabulario(Contextos)].
Vecinos = DeterminaVecinosC (Contextos$\prime$).
D = D $\cup\{(X$, Vecinos) $\}$.

A continuación se muestra un ejemplo de la traza de ThesGref. Partiendo de $B0$ = { programa, datos, libre, propietario, usuario, licencia, seguridad, manera, autor, estado, clave, necesidades, vez, programas, mecanismo, sistema, sistema_operativo } para cada $X \in Bo$ el resultado (fragmento) de ObtenerContextos es:

X	ObtenerContextos
licencia	modificación software propietario restringe derechos usuario mero ...
seguridad	seguridad nacional imperativo uso forma exclusiva áreas pública...
sistema	estratégicos confianza rota repetidas veces valgan ejemplos ...
información	procesar información relativa instituciones autorizadas garantizado ...
datos	almacenamiento estándar usuario seguro futuro seguir descifrando ...
...	...

Después se ordenan los sustantivos de los contextos de acuerdo a su frecuencia, haciendo uso de Ordena_Frec :

X	Ordena_Frec
licencia	software(16) propietario (8) derechos(3) usuario(16) uso(9) nt(3) ...
seguridad	seguridad(15) nacional(4) software(5) libre(3) permite(2) usuario(2) ...
sistema	puerta(2) trasera(2) interbase(2) sistema(9) datos(3) borland(2) ...
información	información(11) útil(2) datos(4) programa(5) usuario(4) manera(2) ...
datos	datos(37) tres(2) riesgo(3) confidenciales(2) manera(6) acceso(5) ...
...	...

Se toman los primeros 10 sustantivos de cada término, se conforman los vecinos, y por último obtenemos el *thesaurus*:

X	Vecinos
licencia	usuario 0.270 programa 0.227 propietario 0.277 libre 0.239 ...
seguridad	datos 0.109 mecanismo 0.183
sistema	usuario 0.09 mecanismo 0.256
información	manera 0.178
datos	seguridad 0.109 usuario 0.220 programa 0.183 mecanismo 0.140 ...
...	...

En los experimentos realizados se construyó un *thesaurus* utilizando el método explicado anteriormente usando como contexto del término una oración (donde aparecía el término), así como una variante del método utilizando 3 oraciones : una antes, donde ocurre el término, y otra después.

Los resultados de esta prueba para el mismo corpus fueron los siguientes:

***Thesaurus* con una oración :** 230 términos, de los cuales 26 términos tenían palabras relacionadas, 23 términos coinciden con *thesaurus* de tres oraciones y 3 términos no coinciden.

Término		Palabras Relacionadas	Evaluación
acceso	–	datos, forma	2/2
clave	–	función	0/1
conocimiento	–		0/0
estado	–	libre	0/1
licencia	–	mecanismo, libre, usuario, propietario, programa	4/6

Tabla 3.1: Algunas entradas del *thesaurus* constituido usando una oración como contexto (CorpusCom).

Término		Palabras Relacionadas	Evaluación
acceso	–	personas	1/1
clave	–		0/0
conocimiento	–	problemas	1/1
estado	–	información, libre, servidor, vez	2/4
licencia	–	mecanismo, libre, usuario, propietario, programa	5/6

Tabla 3.2: Algunas entradas del *thesaurus* construido usando tres oraciones como contexto (CorpusCom).

***Thesaurus* con tres oraciones :** 230 términos, de los cuales 32 términos tenían palabras relacionadas, 23 términos coinciden con *thesaurus* de una oración y 9 términos no coinciden. En las tablas 3.1 y 3.2 aparecen algunos términos para contextos con una y tres oraciones respectivamente. La primer columna es el término, la segunda sus palabras relacionadas y la tercera la evaluación según el experto de qué palabras están realmente relacionadas con el término.

Factor de evaluación	*Thesaurus* (una oración)	*Thesaurus* (tres oraciones)
términos sin palabras relacionadas	204	198
términos con palabras relacionadas	26	32
porcentaje de aciertos en palabras relacionadas	72	70
porcentaje aciertos en palabras relacionadas de términos que coinciden (23)	73	67
términos que no coinciden	3	9
porcentaje de aciertos en palabras relacionadas de términos que no coinciden	33	82

Tabla 3.3: Comparación de las dos formas de elegir contextos (CorpusCom).

Se validaron las palabras relacionadas con cada término considerando la siguiente verificación: la palabra está relacionada con el término, la palabra no tiene ninguna relación con el término.

En la tabla 3.3 se muestra un resumen de los resultados obtenidos de la validación de los *thesauri*. Es decir, de los 230 términos obtenidos en el *thesaurus*, cuántos términos tienen palabras relacionadas, cuántos términos no contienen palabras relacionadas, etc., para los dos métodos: una oración y tres oraciones.

Es notable el hecho de haber obtenido un *thesaurus* formado por un pequeño porcentaje de los términos identificados como importantes en el dominio (menor al 10 %). Aun estos términos se relacionaron aceptablemente (72 %), y, de ellos, solamente el 11.5 % (tres términos) no son del dominio

(“autor”, “vez” y “cosa”); a su vez, sólo para un término (“autor”) se consiguieron relaciones donde la mayoría fueron correctas (los otros dos términos no tuvieron relaciones correctas).

En este sentido, puede afirmarse que, para el *corpus* empleado, el algoritmo descriminó razonablemente los términos. Debe ubicarse este hecho en la elección inicial de los términos del dominio: términos de frecuencia media (por el uso del PT); frente a la propuesta de otros algoritmos que utilizan términos de alta frecuencia.

Capítulo 4

Conclusiones

Se ha presentado una modificación al método de Grefenstette para construir *thesauri*.

La modificación se realizó enfocándose hacia la terminología de un dominio. Así, se eligieron términos importantes, y se propuso la identificación de términos multipalabra con inducción léxica, para aplicar, finalmente, la técnica subyacente en el método de Grefenstette, e identificar términos relacionados, es decir, vecinos recíprocamente cercanos.

En la identificación de términos importantes para el dominio se obtuvo una precisión de 80 %, usando el punto de transición en los textos que integraron el *corpus*, y para las palabras relacionadas se alcanzó una precisión de 73 %.

Es importante señalar que el haber aplicado un método que identifica

términos con frecuencia media (punto de transición) ocasionó que muchos términos que identificaría el método de Grefenstette escapen a este método, por no aparecer en la lista inicial. Esta observación implicaría realizar un cambio en la obtención de contextos; por ejemplo, obtener términos de asociación de segundo orden. Otras pruebas que pueden llevarse a cabo partiendo de los resultados de este trabajo son:

- Comparar los resultados obtenidos en este trabajo con los que se obtendrían aplicando el método de Grefenstette. Es importante señalar que ello requiere un analizador sintáctico; lo cual hace relativamente costosa su aplicación.

- Variar los métodos de obtención de términos importantes apoyándose en algoritmos que proporcionen palabras clave de un texto, además comparar la terminología, y comparar el *thesaurus* generado a partir de estos términos.

- Naturalmente, pueden hacerse más pruebas variando algunas funciones utilizadas en los algoritmos, por ejemplo con la función de similitud (coseno, Dice, similitud complementaria, etc).

- Una variante a la elección de términos importantes es considerar el cálculo de la entropía de cada término para decidir su relevancia, lo cual se ha visto funciona muy bien en la recuperación de información [15].

Bibliografía

[1] A.D. Booth.: A law of occurrences for words of low frequency, *Information and control, 10(4) 386-393*, 1967.

[2] B. Reyes, E. Moyotl & H. Jiménez.: Reducción de términos índice usando el punto de transición, *Facultad de Ciencias de la Computación, B. Universidad Autónoma de Puebla*, 2004.

[3] C.J. Crouch & B. Yang.: Experiments in automatic statistical thesaurus construction, *Proc. Conf. ACM 15 77-88*, 1992.

[4] C. Gierl & D.P. Frost.: Identification of Domain-Specific Terminology by Combining Mutual Information and Lexical Induction, *ECAI92 10th European Conference on Artificial Intelligence*,1992.

[5] Natural Language Group.: Hansards, *36th Parliament of Canada Release 2001-1a, USC Information Sciences Institute, Ulrich Germann (ed)*, 2001.

[6] C. Hand, H. Fujii & W. Croft.: Automatic query expansion for japanese text retrieval, *Technical Report UM-CS-1995-011, Department of Com-*

puter Science, Lederle Graduate Research Center, University of Massachusetts, 1995.

[7] C.D. Manning & H. Schütze.: Foundations of Statistical Natural Language Processing, *Massachusetts: The MIT Press*,1999.

[8] C. van Rijsbergen.: Information Retrieval, *Dept. of Computer Sciencie, University of Glasgow, 2nd Ed.*,1979.

[9] C. Varaschin & V.L. Strube.: Experiments on Extracting Semantic Relations from Syntactic Relations, *CICLing 2003, Faculdade de Informática PPGCC, PUCRS*, 2003.

[10] D. Hindle.: Noun classification from predicate-argument structures, *Proc. Conf. ACL 28 268-275*, 1990.

[11] G. Grefenstette.: Use of syntactic context to produce term association lists for text retrieval, *Proc. Conf. ACM 15 89-97*, 1992.

[12] G. Grefenstette.: SEXTANT: extracting semantics from raw text, implementation details, *Integrated Computer-Aided Engineering, 6*, 1994.

[13] G. Grefenstette.: Automatic thesaurus generation from raw text using knowledge-poor techniques, *Xerox Report*, 1996.

[14] H. Jimenez.: Domain Membership Degrees and Classification Methods, *CIC-IPN Computacion y Sistemas Vol. 5 No. 4, 288-295*, 1996.

[15] H. Jimenez, M. Castro, F. Rojas, E. Miñón, D. Pinto & E. Franco.: Non Supervised Term Selection using Entropy, *en revisión*, 2005.

[16] H.P. Luhn.: Hans Peter Luhn 1896-1964, *Medford: ASIS, http://www.asis.org/Features/Pioneers/luhn.htm*, 1998.

[17] J.R. Firth.: A synopsis of linguistic theory 1930-1955, *In Studies in Linguistic Analysis, 1-32*, 1957.

[18] J.S. Justenson & S.M. Katz.: Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering 1, 9-27*, 1995.

[19] J.L. Milstead.: NISO Z39.19: Standard for Structure and Organization of Information Retrieval Thesauri, *Taxonomic Authority Files Workshop, Washington D.C.*, 1998.

[20] J. Minker, G.A. Wilson & B. Zimmerman.: An evaluation of query expansion by the addition of clustered terms for a document retrieval system, *Information Storage and Retrieval, 329-348*, 1972.

[21] K.W. Church & P. Hanks.: Word association norms, mutual information and lexicography, *Proc. Conf. ACL 27 76-83*, 1989.

[22] M.J. Bates.: Subject Access in Online Catalogues: A Design Model, *Journal al the ASIS 37(6) 361*, 1986.

[23] M.A. López A. & J.A. Moreiro G.: Presente y futuro de los tesauros como herramienta conceptual de precisión para la recuperación, *Reporte de la Universidad Carlos III de Madrid, España*, 2000.

[24] N.W. Francis & H. Kucera.: Frecuency Analysis of English Usage, *Houghton Mifflin, Boston*, 1982.

[25] N. Roberts.: Historical studies in documentation the pre-history of the information retrieval thesaurus, *Journal of documentation, 271-285*, 1984.

[26] P.M. Roget.: Roget's Thesaurus, *http://www.rain.org/ karpeles/rogetdis.html*, 1852.

[27] R. Quirk.: On Corpus Principles and Design, *J. Svartvik (ed.) 457-469*,1992.

[28] R. Urbizagástegui-Alvarado.: Las posibilidades de la ley de Zipf en la indización automática, *Reporte de la Universidad de California Riverside*, 1999.

[29] S. Johansson, G.N. Leech & H. Goodluck.: Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with Digital Computers, *Department of English University of Oslo*, 1978.

[30] W.F. Poole.: The plan of the new Poole's Index, *Library Journal 3, 109-110*, 1878.

[31] W. Schmitz-Esser.: New Approaches in Thesaurus Application, *International Classification, 143-147*, 1991.

[32] Y. Choueka.: Looking for needles in a haystack or locating interesting collocational expressions in large textual databases, *In Proceedings of the RIAO, 38-43*, 1988.

Apéndice A

Términos de Computación

Fragmento de términos obtenidos por el método basado en el punto de transición aplicado a textos separados y usando la fórmula del dominio de Computación.

acceso	archivo	cerebro-individual	consultas
acceso-depende	artificial	ciencias	contraseña
acceso-directo	aspectos	cifrado	correo
acuerdo	atributos	clave	cosa
administración	autenticación	claves	cruza
agente	autor	cliente	cuáles
agentes	autor-original	clientes	datos
algoritmo	basados	columna	datos-documental
algoritmos	base	columnas	datos-protegidos
almacén	bases	complejidad	datos-textual
análisis	bits	computadoras	decisión
aplicación	búsqueda	conectividad	definición
aprendizaje	celdas	conjunto	desarrollo
aptitud	cerca	conocimiento	desarrollo-incremental
árbol	cerebro	conocimientos	detalle

diferentes
diferentes-fuentes
dimensión
dimensiones
dirección
dirección-base
disco
diseño
dispositivos
dispositivos-existentes
dispositivos-reciben
dm
documentos
documentos-susceptibles
ejecución
ejemplo
empresas
enfoque
enlaces
entorno
entrada
entrada-salida
espacio
espacio-libre
espacio-suficiente
espacio-total
especie
especie-humana
establecimiento
establecimientos-educacionales
estado
estado-inicial
evolución
experto
expertos
extracción
extremos
ficheros
final
forma
forma-exhaustiva
forma-parecida
forma-simple
forma-transparente
forwarding
función
función-objetivo
genético
genéticos
gente
gestión
hechos
hechos-conocidos
hechos-forman
hechos-particulares
herramienta
herramientas
humana
humano
in
inferencia
información
información-necesaria
información-oculta
informes
ingeniería
inteligencia
inteligencia-humana
internet
investigación
libre
licencia
line
llave
local
log
manejo
manejo-propio
máquinas
marketing
mecanismo
mensaje
mensajes
mercado
mercado-competitivo
metaconocimiento
minería
modelo
modelo-predictivo
monitoreo
mundo
mysql
necesidades
negocio
nivel
nombres
número
objetivo
objetivos
objetos
olap
operatividad
operativos
página
páginas
patrones
persona
personas
población
port
precisión
predicados
predicción
predictivo
pregunta
problema
problema-particular
problemas
problemas-involucrados
problemas-motor
proceso
proceso-completo
proceso-implica
productos
programa
programación
programas
propietario
razonamiento
recursos
recursos-necesarios
red
redes
reglas
representación
respuestas
rsa
sbc
seguridad
selección

Apéndice B

Términos importantes de Salud

Fragmento de términos obtenidos por el método basado en el punto de transición por lista aplicado al corpus completo del dominio de Salud.

acético	alimentarios-propios	años-setenta	aumento-significativo
actividad	alimento	años-siguientes	autores
actividades	alimentos	apoyo	beneficios
actividades-deportivas	alumno	apropiados	cabo
actividades-dirigidas	alumnos	asociación	cabo-estudios-especiales
actividades-dirigidas-recursos	análisis	aspectos	calidad
actividades-fisicas	análisis-anuales	aspectos-clave	calidad-aprobado
activo	análisis-reciente	aspectos-comunes	calidad-relacionados
adaptación	análisis-relacionados	aspectos-destacados	cáncer
agua	análisis-trimestral	aspectos-relacionados	capítulo
alcohol	año	aspectos-relativos	capítulo-consiste
alimentación	años	atención	cápsula
alimentación-variada	años-muestrales	atención-prenatal	características
alimentaria	años-ochenta	atención-primaria	carga
alimentarios	años-posteriores	aumento	carga-total

casos
casos-notificados
casos-ocurren
células
centinela
cervicouterinas
cirugía
clamidia
clínica
comida
comida-habitual
complicaciones
comunidad
comunidad-educativa
conjuntivo
conservación
consumidor
consumo
consumo-creciente
consumo-diario
consumo-excesivo
consumo-intervienen
contractura
contractura-grave
control
corporal
costo
costo-efectividad
costos
crisis
cuello
cuenta
cuenta-diversos
datos
datos-adecuados
datos-aportados
datos-indicaron
datos-nacionales
datos-obtenidos
datos-permiten
datos-presentados
datos-regional
debido
desarrollo
desarrollo-emocional
desarrollo-ocurren
desarrollo-posterior
desarrollo-reciben
detección
determinación
día
diagnóstico
dieta
dieta-variada
diferentes
diferentes-etapas
diferentes-funciones
diferentes-horas
directiva
displasias
disponibles
dispositivos
dispositivos-implantables
dolor
educación
efectos
eficaz
ejemplo
ejemplo-salmonella
encargados
encontramos
enfermedad
enfermedad-cardiovascular
enfermedad-determinada
enfermedades
epidemia
escamosas
esfuerzo
especificidad
estrategias
estudio
estudio-indicaron
estudios
estudios-especiales
estudios-futuros
estudios-observacionales
estudios-relacionados
etiología
europa
europeo
evaluación
examen
examen-adicional
examen-integral
examen-internacional
exceso
fabricantes
factores
factores-importantes
factores-individuales
fda
física
flexibilidad
forma
forma-activa
forma-atractiva
frecuencia
gel
genital
gonorrea
grado
grado-alto
grado-inferior
grasas
grupo
grupo-asesor
grupo-concreto
grupos
grupos-activos
grupos-considerando
grupos-independientes
grupos-vulnerables
hábitos
horas
implante
importantes
importantes-asociados
importantes-beneficios
incidencia
infancia
infección
infecciones
información
información-apropiada
información-completa
información-convincente
información-correcta
información-formal
información-necesaria
información-obtenida

Apéndice C

Thesaurus

Fragmento de términos relacionados obtenidos por el algoritmo ThesGref aplicado al corpus CorpusCom.

acceso	datos 0.098 forma 0.047	**archivo**	
acceso-depende		**artificial**	
acceso-directo		**aspectos**	
acuerdo		**atributos**	
administración		**autenticación**	
agente		**autor**	usuario 0.126 propietario 0.164 necesidades 0.227
agentes		**autor-original**	
algoritmo		**basados**	
algoritmos		**base**	
almacén		**bases**	correo 0.307 desarrollo 0.125
análisis		**bits**	
aplicación		**búsqueda**	
aprendizaje		**celdas**	
aptitud		**cerca**	
árbol		**cerebro**	

cerebro-individual	
ciencias	
cifrado	
clave	función 0.263
claves	
cliente	
clientes	
columna	
columnas	
complejidad	
computadoras	
conectividad	
conjunto	
conocimiento	
conocimientos	
consultas	
contraseña	
correo	bases 0.307
cosa	tiempo 0.333
cruza	
datos	acceso 0.098 programa 0.183 mecanismo 0.140 propietario 0.152 usuario 0.220 seguridad 0.109
datos-documental	
datos-protegidos	
datos-textual	
decisión	
definición	
desarrollo	bases 0.125 libre 0.084 local 0.2
desarrollo-incremental	
detalle	
dimensión	
dimensiones	
dirección	
dirección-base	
disco	
diseño	
dispositivos	
dispositivos-existentes	
dispositivos-reciben	
dm	
documentos	
documentos-susceptibles	
ejecución	programas 0.166
ejemplo	
empresas	
enfoque	
enlaces	
entorno	
entrada	
entrada-salida	
espacio	