BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

CORPUS PARA UN DICCIONARIO DE CIENCIAS DE LA COMPUTACIÓN. APLICACIÓN DE PRUEBAS PILOTO

TESIS PARA OBTENER EL GRADO DE LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA: ROSA MARÍA ROSAS VÁZQUEZ

DIRIGE: **DRA. MARÍA ELENA FRANCO CARCEDO**
DR. HÉCTOR JIMÉNEZ SALAZAR
DR. DAVID PINTO AVENDAÑO

PUEBLA, PUE., ENERO 2009.

AGRADECIMIENTOS

A mis padres y hermanos a quienes les debo mi formación y apoyo.

A mi esposo, mi aliado en la vida quien me alienta constantemente con amor y paciencia.

A mis directores de tesis: Dra. María Elena Franco Carcedo, Dr. Héctor Jiménez Salazar y Dr. David Pinto Avendaño por su apoyo y guía incondicional.

ÍNDICE

I. Introducción

Los avances tecnológicos de equipo de cómputo y programas retroalimentan su propio desarrollo, investigación, uso y aplicación; ello implica manejar una terminología específica o un subsistema del lenguaje verbal por área de conocimiento que se denomina *jerga*; estaríamos hablando de jerga computacional donde, por cierto se utilizan numerosos anglicismos(la mayoría de las novedades en literatura y productos del área se producen en el idioma inglés) en lugar de su traducción española; así pues, al referirse a conceptos de este tipo se produce un caos debido a la falta de manejo de dicha jerga computacional, por ejemplo: "lo consulté en un *blog*", "¡pásalo a la *caché*!", etc., de ahí surgen preguntas obligadas como: ¿qué es un *blog*?, ¿qué es la *caché*?. Por fortuna existen diccionarios, un diccionario es útil, ya que recoge y explica de forma ordenada voces de una o más lenguas, de una ciencia o de una materia determinada (definición tomada del DRAE, Diccionario de la Real Academia Española).

En el caso del español existe una normativa académica que nos cohesiona a todos los hispanohablantes tanto a nivel léxico como a nivel de las reglas que constituyen el sistema verbal y que nos orienta a cerca de si una determinada expresión o término pertenece a un subsistema coloquial o a uno jergal, o es estándar pero adquiere significación propia en un ámbito jergal determinado (por ejemplo *raíz* que posee una significación estándar pero acepciones específicas para la jerga matemática, odontológica, botánica, etimológica, etc.). Al consultar términos como los del ejemplo en el DRAE[1], no se obtiene un apoyo significativo, ya que las definiciones para éstos no existen o no son satisfactorias. Se necesita, pues, un DICCIONARIO ESPECIALIZADO DE CIENCIAS DE LA COMPUTACIÓN, al igual que existen en otras áreas, con el *Vacabulario Científico y Técnico* de la Real Academia de Cs. Exactas, Físicas y Naturales, o de Estomatología, o de Medicina, etc.

[1] Ya sea en la edición en papel o en la página web http://www.rae.es

Este trabajo contribuirá con el/los programas para las distintas etapas de un proyecto más amplio realizado por la Dra. Mª Elena Franco Carcedo y el Dr. Héctor Jiménez Salazar, presentado en el *Congreso Lingüística 2007* titulado: “Un modelo metodológico –lexicográfico para un *Diccionario de la Computación*”.

Se desarrollarán las herramientas de cómputo que tomen como base de conocimiento textos del dominio de computación, extraigan la terminología e identifiquen contextos definitorios de los mismos. Se constituirá un *corpus* para formar una taxonomía del dominio, con la cual sea posible laborar de forma separada cada subdominio, y facilitar así las pruebas de los algoritmos.

El presente trabajo constituye la primera parte de una investigación más amplia cuya segunda parte aparecerá en la tesis de maestría que se presentará en breve.

1.1. Justificación.

No existe en nuestro medio un diccionario jergal de nuestro ámbito que facilite la precisión y exactitud requeridas en un texto científico-técnico, por lo que se recurre a menudo al uso de anglicismos en detrimento de la riqueza lingüística del español y a la incomprensión terminológica o su vaguedad.

Los resultados del presente proyecto servirán como base y primer paso para la elaboración de otras tesis tanto de la Facultad de Ciencias de la Computación de la BUAP como de otros interesados.

1.2. Límites de licenciatura y maestría

Se establecieron las partes que constituyen el trabajo de licenciatura así como de maestría:

Licenciatura:

- Establecimiento del marco teórico.
- Búsqueda y compilación del *corpus*.

Maestría:

- Algoritmo y elaboración del programa.
- Prueba piloto.
- Reajuste y resultados.

Por circunstancias personales conviene presentar ambas partes de forma separada, este trabajo constituye la primera parte del proyecto general que incluye una segunda parte para tesis de maestría.

II Objetivos

2.1. General

Desarrollar un programa (*software*) para la elaboración de un *Diccionario de Ciencias de la Computación*. Proporcionar un instrumento de apoyo a trabajos de tesis de la Facultad de Ciencias de la Computación, basado en un modelo metodológico-lexicográfico para el diccionario.

2.2. Específicos

- Realizar un acopio de documentos de computación.
- Aplicar y ajustar un algoritmo que identifique términos jergales de la computación, a partir de la extracción de contextos para la definición del término.
- Conformar, analizar y verificar el *corpus* progresivamente.
- Efectuar prueba piloto inicial para, en su caso, ejecutar los ajustes pertinentes a que hubiera lugar.
- Reconfirmar el modelo mediante el contraste de resultados.
- Coadyuvar a la elaboración de un diccionario jergal de computación que aún no existe.
- Obtener el grado académico de maestría mediante la aportación de un trabajo que esperamos de excelencia.

III Metodología

A) Aspectos lingüísticos

3.1. Establecimiento del marco teórico

El primer elemento a considerar fue el tipo de *producto final* y el marco teórico que deseábamos como idóneo para, entre otros, cubrir las necesidades de estudiantes, profesores, investigadores y usuarios en el ámbito de la computación; es decir, el alcance que se pretende, por una parte, y por la otra, el establecimiento de un cierto *modelo* metodológico y que normará el aspecto lexicográfico. Tras algunas reuniones acordamos lo siguiente[2]:

- Será un **Diccionario[3] jergal**, es decir, específico de un área de conocimiento.
- De tipo **onomasiológico**, es decir, proporcionará las entradas y su definición[4].
- Ordenado **alfabéticamente**[5].
- Con definiciones **descriptivas** más que normativas[6], sin llegar a enciclopédicas[7].
- **Nominal**, es decir, de hiperónimo, si lo hay, y los semas esenciales[8].
- El ***corpus*** estará constituido por tesis, manuales, traducciones, artículos, de la propia Facultad y Universidad y nacionales y españolas[9].

---

[2] En esta parte seguimos lo establecido en el proyecto general de los directores de esta tesis, ya mencionado.
[3] No un glosario de términos, ni un lexicón, ni un vocabulario, sino algo de mayor alcance.
[4] Frente a los de tipo *semasiológico*, en que el usuario conoce la definición, pero busca la palabra o lema, como en los Diccionarios de refranes, por ejemplo.
[5] Excepto *ch ,ll*, que irán respectivamente en *c, l*. La ordenación temática suele aparecer en los glosarios uso de los manuales de computación: *actualizar -desde OS/2; -desde una versión de DOS 3 -solución de problemas*, etc.
[6] En su momento trataremos de problemas y procedimientos concretos, baste decir que, en lo relativo a 'usos', que contemplamos incluir entradas de los términos usados en la Península y en México.
[7] Lo cual plantea otra opción a determinar: incluir o no dibujos, figuras, gráficos, etc., que se atenderá en su lugar.
[8] Se verá con mayor precisión en el apartado de *definición.*
[9] Propongo esto último por manejar tanto las variedades léxico-semánticas propias de México, cuanto las del español peninsular, que ofrecen diferencia, como *corpus*, sugiero *El País* de los jueves, que trae amplias secciones sobre *informática*. En su apartado metodológico y de procedimiento lexicográfico se estipulará el tratamiento de ellas.

3.2. Búsqueda y compilación del *corpus*

Se constituirá un *corpus* para pruebas piloto y reajustes del modelo, usando documentos con temas propios de computación como pueden ser: tesis del área, artículos, revistas, libros, entre otros. Se espera que el programa (*software*) forme una clasificación del dominio con la cual sea posible crear subdominios.

A partir de este *corpus* se determinará la cantidad de términos resultantes de la extracción terminológica.

Del resultado de las pruebas piloto sobre este *corpus*, se forjarán criterios para aplicar reajustes tanto al modelo como al algoritmo implementado.

3.3. Selección de entradas

Sólo se incluirán unidades léxicas (simples y sintagmas léxicos) **plenas y variables o de flexión**, no las *gramaticales* (caso especial son algunos adverbios[10]) ni las no específicas del área, lo cual nos eximió de establecer criterios acerca de voces malsonantes, o susceptibles de tabú o censura. Como **léxico jergal** aparecen las de uso estándar, pero con una acepción específica. En los compuestos sintagmáticos se aportará una **definición genérica seguida de sus clases**

sistema sust. masc. Conjunto de elementos que interactúan bajo ciertas reglas

> **~ *de acceso múltiple*** Sistema que permite a varios usuarios la utilización aparentemente simultánea del ordenador. Cada usuario tiene un terminal y una unidad de representación visual, conectados a aquel.
>
> **~ *de cómputo*** Sistema que comprende el *hardware* y el *software*, ya sea con características específicas de cada uno de éstos o con capacidades suficientes para resolver un problema particular.

[10] Las razones me parecen obvias, en particular los caso de adjetivo-sustantivo -*mente*.

3.4. Categorización

Decidimos incluir la categoría gramatical; las razones, entre otras, se basan en la consideración de casos de doble adscripción, o de inseguridad (por ejemplo, participios con valor nominal exclusivo _generalmente, los fonéticos o 'fuertes' o 'irregulares'_ o con doble o triple valor: sustantivo, adjetivo y verbal, etc.) y también para facilitar la comprensión y desarrollar en los usuarios la destreza en forjar definiciones. Cabe señalar que, al menos por mi parte, me valgo de dos grandes modelos y guías en las elecciones de criterios, además, naturalmente, de las internas a la obra, sus presupuestos y objetivos: el *Diccionario de la lengua española* (DRAE) de la Real Academia Española y el *Vocabulario Científico y Técnico* (VCT) de la Academia de Ciencias Exactas, Físicas y Naturales[11].

3.5. Lematizar

Redujimos las unidades léxicas del *corpus* a su forma paradigmática[12]. Ya se contaba con un programa propio para *lematizar*, si bien no funciona para cubrir todas las necesidades, quizá; AGME es un lematizador proporcionado por el Laboratorio de Lenguaje Natural del CIC-IPN (Velásquez y otros 2002) como para lematizar *archivo; dibujo* como sustantivo y como verbo (*archivar; dibujar*), es decir, alcanzar una doble lematización.

Tratamos de prever todos los casos posibles, aunque quizá algunos no se produzcan, puesto que las entradas están restringidas a la jerga, en todo caso, forjamos los siguientes **criterios de inclusión**:

Derivadas por dos procedimientos:

- por sufijo aspectual (*-ero, -ez,-ble, tema-rio*) constarán con entrada propia

[11] Publicados ambos por la editorial Espasa, de Madrid, en 2001, 21ª ed. y 1996 3ª ed., respectivamente.

[12] Desde luego hay casos especiales, como los sustantivos 'de un solo género' (ambiguos y epicenos), o los 'falsos plurales' (duales), o adjetivos que no cambian por género, etc. que se verán en su momento; tratamos de ir de lo más amplio y genérico a lo más restringido y particular.

- potestativos

a) con significación propia, lexicalizados (*ventanilla*) constarán como entradas

b) no lexicalizados (aumentativos, peyorativos, etc.) no se registrarán

- los prefijos, sufijos, prefijoides, sufijiodes significativos (*cito-; -itis*), no, por no ser probables, pues la taxonomía no está desarrollada o es inexistente.

Compuestos:

- ortográficos (que conforman una unidad: *paraguas*), sí
- sintagmáticos: aparecerán tras la entrada simple
- modismos, no.

Abreviamientos (*compu-computadora*, etc.) no, en principio[13].

Comparativos, superlativos y adverbios: sólo si son irregulares y jergales

Contracciones: no es previsible; en su caso, se vería su tratamiento[14]-

Siglas y acrónimos, marcas de uso estandarizado[15], sí

Neologismos semi-incorporados y *usos*[16] estandarizados y representativos sí, pues se trata de proveer de una herramienta útil al usuario

Ilustraciones, dibujos, esquemas, gráficos, etc., sí, en la medida en que complementen[17] la definición y en casos particulares

Fórmulas en ciertos casos se planteará, tras la definición, la fórmula matemática que complemente[18]. Un ejemplo podría ser ***tfidf*** que es el

[13] La razón, al igual que en la exclusión de modismos, se fundamenta en que se trata de un Diccionario de tipo jergal, no de subsistemas *coloquial o geolectal*, aunque sí daremos dobles entradas en ciertos casos.

[14] PROgramación en LOGica (PROLOG); lenguaje de programación
- programación orientada a objetos; paradigma de programación
- abrir un archivo; habilitar un archivo para operaciones de lectura o escritura.
- recuadro; selección geométrica para dibujar una imagen, normalmente rectángulo
- blog; sitio web compartido
- escalable; proporción decreciente entre efectividad y recursos usados por un procedimiento.

[15] Somos conscientes de que deberemos señalar con precisión los parámetros que miden la 'estandarización'; uno de los indicadores podría ser su frecuencia de aparición; otro, el tratamiento ortográfico.

[16] Al hablar de *uso* habremos de afinar más adelante(como en marcar geolectalismo), ya que nos proponemos ofrecer al menos las variantes léxico-semánticas entre el español peninsular y el de México, como *ordenador-computadora; Informática-Ciencias de la Computación*, etc.

[17] *A priori* pensamos que sí se producirán casos en que la inclusión precisará y redondeará la información verbal.

[18] Se trataría de casos análogos, en que la formulación matemática precisa el sentido y lo cierra, por así decir.

nombre de un esquema de ponderación de índices para referencia a textos, que a la vez representa una fórmula

Clasificación jergal del término: en una etapa posterior, la 3ª, indicaremos, en su caso, su especificidad dentro de las ramas del ámbito de la computación, como subclase jergal.

3.6. Procedimientos formales

- Lema o entrada: minúscula (mayúsculas para siglas, como *EPROM*, aunque en este caso, iría primero en su lugar alfabético remitiendo a la entrada *memoria*) y negrita, seguido de la categoría (en abreviatura).
- Los homónimos, numerarlos, según el procedimiento del DRAE.
- Las marcas o siglas, tanto comerciales como no comerciales, españolas o inglesas, de uso común se incluyen y se remiten al sustantivo correspondiente
- En voces comunes con diversas acepciones (*bandeja, menú*[19]), sólo la jergal
- Ordenación de acepciones: el léxico jergal tiende a la univocidad, por lo que los casos de polisemia, sinonimia y pluralidad de acepciones serán infrecuentes. Ejemplos de ambigüedad, ciertamente habrá pocos. He aquí uno:

  - ventana, zona rectangular de la pantalla de un monitor.

  - ventana, contexto de un término; "ventana de tamaño 3 alrededor del término 'trabajos':

  - ...los más importantes trabajos de los que ..."

- Por las mismas razones y ser ésta una ciencia reciente, no hallamos formas *en desuso, antiguas* o, en general, obsoletas[20],[21]

---

[19] En el primer ejemplo, el DRAE no aporta acepción propia, en el segundo, la acepción 4:"*Informática*. Colección de opciones que aparece en la pantalla de un ordenador." El VCT no registra ninguna de las dos.

[20] Hay una serie de consideraciones sobre las *marcas* diastráticas, diafásicas y de estilo, que se plantean en un manualito publicado en 1998 por la UIA (2ª ed., corregida y aumentada, en 2004 en la ed. UAP), *Lenguaje científico y técnico y elaboración de tesis de posgrado*, y en *Los géneros y su práctica. Con una guía gramatical*, donde se tratan estas cuestiones.

[21] Insistimos en que se plantea un diccionario jergal de computación, restringido, las diferencias diastráticas, cuyos límites suelen ser imprecisos, difuminados, dado el evento que lo rige _subsistema diafásico en cuanto adscrito a un ámbito del saber determinado_ y el *corpus* mismo, no se prestan al hallazgo; en situación algo más informal, oral (que no incluimos), se usará el estándar alto o culto.

- Los extranjerismos (anglicismos en su mayoría) remiten a la forma *españolizada*[22], cuyo uso consideramos preferible
- En su caso, y siguiendo el ejemplo del VCT[23], aportaremos los sinónimos correspondientes.

[22] Para los hábitos articulatorios de un hispanohablante, pronunciar (por las reglas del español, no del inglés, pues nadie conoce TODOS los idiomas y sus reglas de equivalencia pronunciación-ortografía, pero sí el propio, por lo que nos decantamos por esta opción) palabras como *software* es una proeza, en tanto que *programa; soporte lógico*, resultan sencillas; ¿por qué *mouse*, cuando tenemos una traducción perfecta: *ratón*?

[23] *Vocabulario Científico y Técnico*, Real Academia de Cs. Exactas, Físicas y Naturales, Madrid, Gredos, 1996.

B) Aspectos computacionales

3.7. Herramientas computacionales

La compilación de los *corpora* se apoyará en la reunión de textos de subdominios específicos de la computación. Este proceso utilizará el cálculo del punto de transición (frecuencia media de ocurrencia de las palabras de un texto) cada vez que se agregue un texto al *corpus*; de modo que se controle el incremento del vocabulario sin que éste cambie el subdominio. En cada *corpus* de un subdominio, se extraerán los términos más 'representativos' y se expanderán para aumentar la evocación del método. A partir de la terminología se identificarán las definiciones contenidas en los textos (Reyes, 2004). Todos estos pasos tienen el propósito de ayudar al trabajo manual que a la vez servirá como una manera de evaluar la efectividad de los métodos. Se aplicarán métodos clásicos para la determinación de la terminología, y el trabajo realizado en la UNAM y la Universidad Pompeu Fabra sobre contextos definitorios.

Los algoritmos C-value que son presentados en Barrón *et al.* (2006) han sido estudiados para determinar el grado de especificidad de un término. En Alarcón y Sierra (2005) se identifican los contextos definitorios utilizando patrones morfo-sintácticos y usando categorías gramaticales que etiquetan las palabras del texto.

Como se menciona en Alarcón, Bach y Sierra (2006), ha habido algunos trabajos sobre la determinación de la terminología y la identificación de los contextos definitorios (CD); destacan aquellos trabajos enfocados a la terminografía, que ocupa un lugar importante en la resolución de este problema, ya que se encarga, por un lado, de la elaboración de ontologías que representen la red conceptual de un área específica, y por otro lado, de la elaboración de diccionarios donde se explique el significado de los términos.

En el problema de la extracción automática de CDs desde una perspectiva teórico-descriptiva, uno de los más importantes trabajos es el de Pearson (1998), en el que se describe el comportamiento de los términos en el contexto real donde aparecen y donde se menciona que cuando un autor define un término, suele recurrir a patrones tipográficos para resaltar visualmente la presencia del término y/o la definición, y a patrones léxicos y metalingüísticos para ligar los dos elementos anteriores mediante estructuras sintácticas. Esta última idea fue reforzada más tarde por el estudio de Meyer (2001).

Habrá, por tanto, que desarrollar herramientas de cómputo que permitan extraer la terminología e identificar contextos definitorios en textos del dominio de computación. Esto, además, conlleva a compilar un *corpus* adecuado. Se pretende, abordar este último problema formando, en primer lugar, una taxonomía del dominio con la cual pueda tratarse separadamente cada subdominio, y facilitar así las pruebas de los algoritmos.

AWK es un lenguaje de programación potente en el procesamiento de datos de tipo texto y en la aplicación de expresiones regulares, además de que se escriben pocas líneas de código. Se ejecuta bajo el sistema operativo OPEN SUSE LINUX versión 10.3. AWK será utilizado para programar la elaboración del *corpus*, posteriormente se evaluará su utilización en los demás requerimientos. Por otra parte llama mucho la atención Python, ya que al igual que AWK es potente con los datos, además de que ofrece mucho más estructura y soporte para programas más extensos.

Los textos del dominio de computación recopilados corresponden a algunas tesis profesionales de licenciatura, los temas son diversos y todos están en formato “pdf” de Acrobat Reader, que contienen texto y, además, tipos de caracteres e imágenes, lo que permite que la información se lea con eficiencia, pero no es apropiado si los necesitamos con un objetivo computacional, ya que las imágenes y caracteres especiales tendrían que ser filtrados, lo que representaría una tarea adicional.

Los documentos se han convertido a texto plano (sin caracteres especiales ni imágenes) para su utilización como entrada de programas en AWK.

3.7.1. Textos base para el *corpus.*

Como señalamos en el párrafo anterior, tomamos como textos de base iniciales algunas tesis de la Facultad de Cs. de la Computación de la BUAP, cuyos autores y títulos reseñamos a continuación, en el entendido de que progresivamente habrá de aumentar el acervo con otros textos, entre los que tenemos proyectado el diario *El País*, en particular de los días jueves, que trae una páginas de la sección "El País cibernético" con las últimas novedades, artículos, reseñas, etc. del ámbito de la computación.

Se consideraron los siguientes documentos organizados en cuatro clases:

Clase: Complejidad de Algoritmos

DOCUMENTO 1:
Bronca Mazzocco I. (2007): *Planificación de Movimientos Utilizando Técnicas de Simple Consulta.*

DOCUMENTO 2:
Mirón Enríquez A. (2008): *Una Biblioteca de Clases para la Implementación de Algoritmos Genéticos Paralelos con MPI y C++.*

DOCUMENTO 3:
Salcedo Haro M. (2007): *Diseño, Análisis e Implementación de un Algoritmo Paralelo para la Resolución de Problemas de Programación Lineal en Enteros Puros.*

DOCUMENTO 4:
Sánchez López J.A. (2007): *Método De Búsqueda Dispersa Aplicado Al Problema De La Mochila.*

Clase: Graficación

DOCUMENTO 1:
Olivares Morales J. C. (2008): *Sistema Analizador y Métrico de una Fotografía a Nivel Celular.*

DOCUMENTO 2:
Toral Lima F. (2007): *Sistema de Captura y Procesamiento de Imágenes de Radiografías Obtenidas por Resonancia Magnética.*

DOCUMENTO 3:
Xalteno Altamirano J. E. (2007): *Sistema de Apoyo para la Conducción de Automóviles.*

Clase: Redes

DOCUMENTO 1:
Red Digital de Servicios Integrados.

DOCUMENTO 2:
Monge Rebollar J. I. (2007): *Desarrollo de Aplicaciones para Cómputo Móvil.*

DOCUMENTO 3:
Fernández de Lara Sanabria A. (2007): *Sistema Distribuido de Información Visual Hospitalaria.*

DOCUMENTO 4:
Ramírez Vargas J. C. (2007): *Implementación de Protocolos de Comunicación y del Agente Consumidor en un Mercado de Objetos de Aprendizaje.*

Clase: Sistema de Base de Datos

DOCUMENTO 1:
Espinoza Castellanos M.C. (2008): *Sistema Agenda para la Administración de Eventos.*

DOCUMENTO 2:
López Luna D. S. (2008): *Biblioteca Digital de Apoyo a los Cursos de la FCC.*

DOCUMENTO 3:
González Nieto E.: *Control de Ingresos y Egresos de una Dependencia de la BUAP.*

DOCUMENTO 4:
Chávez Contreras E. (2008): *Transferencia de Información entre una Base de Datos y una Aplicación de Comercio Electrónico Mediante XML.*

DOCUMENTO 5:
Balderas Espinosa Marco Antonio: *Sistema de Encuestas Telefónicas (SET).*

DOCUMENTO 6:
Vique Tepango Urzulo Nahum: *Marcadores Sociales y Favoritos.*

DOCUMENTO 7:
Erick Albertho Euan Waldestrand: *Sistema de Registro de Incidencias como parte del Sistema Integral de Seguridad Universitaria.*

DOCUMENTO 8:
Manuel Fuentes Solar: *Sistema de Inscripción Remota (SIR).*

DOCUMENTO 9:
Gilberto Sánchez Cervantes: *Diseño y Construcción de un Sistema de Control Escolar para la Preparatoria Gral. Lázaro Cárdenas del Río.*

DOCUMENTO 10:
Roberto Emmanuel Soriano Muñoz: *Módulo de Agenda y Citas para el Sistema de Información Hospitalario Modular.*

DOCUMENTO 11:
SICAFIL (Sistema de Control de Activo Fijo en Linux)

DOCUMENTO 12:
Jhonatan Ramírez Domínguez: *Sistema Web: Entretenimiento para Estudiantes Universitarios.*

DOCUMENTO 13:
Beatriz Flores Cortés: *Administración de Scripts de PHP y MySQL.*

DOCUMENTO 14:
Edgar Alfonso Díaz Díaz: *Sistema Web para el Proceso de Tutorías de la Facultad de Ciencias de la Computación.*

3.7.2 Ley de Zipf

Zipf (1949) formuló la **ley de frecuencia de las palabras** en un texto (*ley de Zipf*), que establece que si se cuenta el número de ocurrencias de cada palabra diferente en un texto T y se ordenan las palabras encontradas en ese texto de forma descendente en una tabla de acuerdo a su frecuencia, es decir, la primera palabra es la más frecuente, la segunda palabra es la segunda más frecuente y así sucesivamente, entonces

$$r * f = k \qquad (1.1)$$

r es el orden de la palabra en la lista (rango), *f* es la ocurrencia de la palabra (frecuencia), y *k* es una constante para *T*.

Así, si tenemos un texto y si obtenemos la frecuencia de las palabras usadas en ese texto, y si ordenamos esas palabras descendentemente por su frecuencia, identificaremos los términos que representan de manera apropiada el contenido del texto.

Goffman afirma que la *ley de Zipf* considera solamente las palabras de alta frecuencia, y que dos palabras de alta frecuencia no pueden tener la misma frecuencia, en un determinado texto. Además, debido a la *ley de Zipf*, unas pocas palabras tienen frecuencia alta, mientras que la mayoría de las palabras tienen baja frecuencia, es decir, esta ley predice y describe los dos extremos de la distribución de las palabras en un determinado texto (Pao, M.L., 1977).

3.7.3 El punto de transición

La distribución de las palabras está ordenada en dos extremos, es posible identificar una región crítica en la que ocurra la transición de palabras de alta frecuencia a palabras de baja frecuencia. Para llegar a éste punto de transición (PT), partimos de la ley de ocurrencias para palabras de baja frecuencia, propuesta por Booth, A. (1967). Denotamos con *p(r)* la probabilidad de que ocurra una palabra de rango *r*. Si el texto contiene *V* palabras diferentes entonces, para una palabra con rango *r* y frecuencia *f*, *Vp(r)* = *f*. George Kinsley Zipf observó que una palabra ocurre con frecuencia 1 si cumple:

$$2 > Vp(r) \geq 1 \qquad (1.2)$$

La *ley de Zipf* sugiere que *p(r)* = *k*/*r*, donde *k* es la constante para el texto T. Así, substituyendo en la ecuación (1.2):

$$2 > Vk/r \geq 1 \qquad (1.3)$$

De la ecuación (1.3) podría decirse entonces que hay dos valores para *r*, uno mínimo y otro máximo:

$$r_{min} = 1/2\ kV, \qquad r_{max} = kV \qquad (1.4)$$

es decir, existen varias palabras con frecuencia 1 para las cuales se debe satisfacer que su rango está entre estos valores. Se considera, al igual que en la ecuación (1.2), que los valores de los rangos para palabras con frecuencia 1 son los dados por las ecuaciones (1.4). Si I_1 representa el número de palabras con frecuencia 1, entonces $I_1 = r_{max} - r_{min}$, por tanto:

$$I_1 = 1/2\ kV \qquad (1.5)$$

Al igual que en la ecuación (1.2), algo semejante puede hacerse para el número de palabras con frecuencia *n*. Una palabra ocurre *n* veces en un texto si cumple con:

$$n + 1 > Vp(r) \geq n \qquad (1.6)$$

Entonces, el número de palabras con frecuencia *n*, sería:

$$I_1 = \frac{1}{n\,(n+1)}\, kV \qquad (1.7)$$

De las ecuaciones (1.5) y (1.7) se obtiene la siguiente proporción independientemente de las constantes del texto *k* y *V*, es decir, válida para cualquier texto.

$$I_n / I_1 = \frac{2}{n\,(n+1)} \qquad (1.8)$$

Como se menciona en Urbizagástegui-Alvarado (1999), se considera que $I_n = 1$ ($n > 1$), para un valor fijo de *n*, lo cual indica una de las palabras cuya frecuencia de ocurrencia en el texto es *n*. Tal palabra no es rara en el texto, debido a que su frecuencia ya no es baja a causa de la distribución de las palabras. Este valor *n* señala precisamente la posición, en la tabla de palabras ordenadas por su frecuencia, del PT, ya que arriba de él habrá frecuencias de palabras muy comunes y abajo frecuencias de palabras poco comunes. Como lo mencionan Luhn, H.P. (1958) y Urbizagástegui-Alvarado (1999), las palabras con mayor contenido semántico de un texto se encuentran en la zona de transición entre las palabras de frecuencia alta y las palabras de frecuencia baja. Por esta razón interesa conocer el valor de *n*. Aplicando la condición $I_n = 1$ en la ecuación (1.8) podemos despejar *n* y obtener:

$$n = \frac{\sqrt{1 + 8I_1} - 1}{2} \qquad (1.9)$$

Utilizando está ecuación es posible calcular la región crítica, es decir, calcular el PT en el texto. Una vez calculado el PT, se selecciona una lista de palabras alrededor de él y que son las que mejor representen jerga computacional.

IV. Compilación del *Corpus*

4.1 Método de compilación del *corpus* utilizando el punto de transición.

Una técnica para facilitar la compilación de un *corpus* es el punto de transición derivado de la *ley de Zipf*. Este trabajo propone la utilización del punto de transición para la compilación del *corpus* que necesitamos, aplicándolo a documentos de tesis de licenciatura de la Facultad de Ciencias de la Computación.

Debido a la *ley de Zipf*, cualquier *corpus* contiene muy pocas ocurrencias para la mayoría de palabras contenidas en éste, por lo que se sugiere considerar un número óptimo de frecuencias de aparición alrededor del cual se construye un intervalo de aceptación llamado punto de transición (PT), explicado ya anteriormente; después de algunas operaciones se obtiene:

$$n = \frac{\sqrt{1 + 8I_1} - 1}{2} \qquad (1.9)$$

Jiménez Salazar H. & Pinto D., Rosso P. (2005) observaron que, al aplicar el PT, se alcanza su máxima eficiencia en la selección de términos con 40% de los términos más cercanos al PT.

Etapas empleadas por el método:

1. De manera manual se eliminaron caracteres especiales y símbolos, generados al migrar el documento origen de Adobe Reader (pdf) al formato de texto simple (txt).

2. Conversión del texto a minúsculas, ya que en algunos casos lo requería.
3. Eliminación de palabras cerradas utilizando un diccionario de palabras

cerradas y las ruidosas; aquellas que no son relevantes para la representación del documento.

4. Obtención del punto de transición utilizando la ecuación (1.9), frecuencia más alta y el término que la consiguió, términos con frecuencia 1, los umbrales extremos, y el conjunto de términos que se encuentran en el 40% de los términos más cercanos al PT.

4.2 Experimentos

Para examinar la efectividad del método propuesto sobre el enriquecimiento del *corpus*, se empleó frecuentemente el punto de transición en documentos de tesis. Estos documentos fueron organizados manualmente en cuatro clases diferentes, en las cuales se consideró incurrían de manera más natural los documentos.

Se realizaron 3 experimentos empleando 25 documentos de tesis de licenciatura, cada documento en promedio tiene 1450 términos diferentes. Se realizó un preproceso a la colección de documentos para: la separación de palabras utilizando el carácter espacio y los caracteres de puntuación, conversión de mayúsculas a minúsculas, y eliminación de palabras cerradas y ruidosas, esto último para excluir aquellos términos que no representan de manera significativa el texto.

Experimento 1.

Gradualmente se aplicó el método a 14 documentos de tesis de la clase Sistemas de Base de Datos, que formaron un *corpus* no etiquetado de 1,475,146 palabras, con un vocabulario de 90,926 términos. Para realizar los programas que permitieran calcular el método se utilizó el lenguaje de programación AWK.

La ejecución de los programas (Anexo) se realizó de la siguiente manera:

1. ./minusculas SBD1ok.txt>sbd1.txt
2. ./ptran cerra3.txt ruidosas sbd1.txt>sal_sbd1.txt

Ejecución del método para el primer documento de la clase Sistema de Base de Datos (DOCUMENTO 1):

```
------------------------------>
Frecuencia mas alta: 211 , Termino=
datos
Terminos con frecuencia 1= 871
Punto de Transicion= 41.2403
Umbral1= 24.7442 , Umbral2= 57.7364
------------------------------>
Terminos dentro del PT
------------------------------>
51 software
49 modelo
48 registro
48 figura
46 sql
41 administración
39 baja
38 solo
37 constancias
37 bases
36 web
36 modificación
35 realizar
35 opciones
35 diseño
35 desarrollo
33 reservación
31 tipo
31 mysql
31 información
31 cliente
30 son
30 php
29 mostrará
29 descripción
28 servidor
28 null
28 consulta
27 regresar
26 mensaje
25 vez
25 gestión
------------------------------>
```

Ejecución del método para el segundo documento de la clase Sistema de Base de Datos (DOCUMENTO 2):

```
------------------------------>
Frecuencia mas alta: 358 , Termino= datos
Terminos con frecuencia 1= 1573
Punto de Transicion= 55.5914
Umbral1= 33.3549 , Umbral2= 77.828
------------------------------>
Terminos dentro del PT
------------------------------>
77 son
76 opción
76 información
73 diagrama
71 biblioteca
68 solo
67 documentos
61 registro
59 muestra
58 relación
58 bases
56 entidad
55 modificar
55 eliminar
55 diseño
54 software
53 bibliotecas
51 desarrollo
51 consulta
50 web
50 digital
48 tipo
47 atributos
46 sql
44 mysql
43 php
43 administración
42 realizar
42 nombre
42 campos
41 modificación
39 opciones
39 baja
37 descripción
37 constancias
36 llave
36 elementos
36 cliente
36 casos
36 agregar
34 principal
------------------------------>
```

Ejecución del método para el tercer documento de la clase Sistema de Base de Datos (DOCUMENTO 3):

```
------------------------------>
Frecuencia mas alta: 561 , Termino= datos
Terminos con frecuencia 1= 1873
Punto de Transicion= 60.7066
Umbral1= 36.424 , Umbral2= 84.9893
------------------------------>
Terminos dentro del PT
------------------------------>
80 solo
78 opción
77 diagrama
74 diseño
73 desarrollo
71 biblioteca
70 muestra
68 registro
67 documentos
66 tipo
65 entidad
63 modificar
60 consulta
60 cliente
59 atributos
56 eliminar
53 nombre
53 bibliotecas
52 lenguaje
52 conjunto
50 web
50 relaciones
50 realizar
50 digital
49 nivel
49 administración
48 sql
47 modificación
46 ingeniería
46 función
46 esquema
46 descripción
46 campos
45 sistemas
45 relacional
45 cuenta
44 mysql
43 programación
43 php
43 objetos
42 programas
42 clave
41 parte
41 opciones
41 elementos
41 análisis
40 tiempo
40 resultado
40 operación
40 entidades
40 casos
40 acceso
39 vez
39 baja
38 principal
38 pantalla
38 dependencia
37 constancias
------------------------------>
```

Ejecución del método para el cuarto documento de la clase Sistema de Base de Datos (DOCUMENTO 4):

```
------------------------------>
Frecuencia mas alta: 732 , Termino= datos
Terminos con frecuencia 1= 2274
Punto de Transicion= 66.9407
Umbral1= 40.1644 , Umbral2= 93.717
------------------------------>
Terminos dentro del PT
------------------------------>
93 tipo
93 tabla
93 canciones
91 género
89 documentos
87 solo
86 canción
84 realizar
84 lenguaje
82 pago
82 opción
82 entidad
81 proceso
81 administración
71 biblioteca
70 vez
70 electrónico
70 catálogo
69 parte
69 modificar
67 nombre
67 mysql
66 carrito
66 atributos
64 presenta
64 eliminar
64 elementos
63 compra
63 casos
62 consulta
61 resultado
61 navegación
60 tienda
59 diagramas
58 conjunto
57 análisis
56 permite
56 html
56 cuenta
55 servidor
55 nuevo
55 gestión
54 través
54 internet
54 búsqueda
53 relaciones
53 orden
53 opciones
53 objetos
53 ingeniería
53 formato
53 bibliotecas
53 aplicaciones
52 realiza
52 campos
51 sql
51 sistemas
51 modificación
51 estructura
51 descripción
51 actividades
51 acceso
50 nivel
50 función
50 esquema
50 digital
50 agregar
49 tiempo
49 entidades
49 ejemplo
48 programación
48 capítulo
47 relacional
47 conceptual
46 uml
45 programas
45 principal
45 operaciones
44 manera
```

44 implementación
44 describe
43 php
43 modelos

43 autenticación
42 clave
---------->

Ejecución del método para el quinto documento de la clase Sistema de Base de Datos (DOCUMENTO 5):

```
------------------------------>
Frecuencia mas alta: 791 , Termino= datos
Terminos con frecuencia 1= 2783
Punto de Transicion= 74.1073
Umbral1= 44.4644 , Umbral2= 103.75
------------------------------>
Terminos dentro del PT
------------------------------>
102 bases
100 encuesta
100 class>>
99 <<navigation
98 alta
97 solo
95 servidor
94 vez
94 tabla
94 registrado
94 proceso
93 lenguaje
93 canciones
91 género
90 parte
89 documentos
89 administración
86 opción
86 canción
83 pago
83 entidad
77 objetos
77 java
77 atributos
76 cuestionario
75 electrónico
74 tiempo
73 nombre
73 modificar
73 encuestas
73 análisis
72 biblioteca
71 módulo
70 catálogo
69 mysql
68 través
68 presenta
68 elementos
67 cuenta
67 casos
66 resultado
66 carrito
65 respuestas
65 eliminar
64 sistemas
63 internet
63 diagramas
63 consulta
63 compra
62 navegación
62 aplicaciones
60 tienda
60 programación
60 permite
60 conjunto
59 resultados
59 función
59 control
58 relaciones
58 realiza
58 nuevo
57 orden
57 operaciones
57 html
57 gestión
57 formato
57 ejemplo
57 acceso
56 etc
56 estructura
55 opciones
55 manera
54 ingeniería
54 implementación
54 búsqueda
53 sql
53 esquema
53 digital
53 creación
```

53 bibliotecas
53 actividades
52 principal
52 nivel
52 general
52 campos
51 relacional
51 modificación
51 descripción
51 agregar
50 respuesta
50 entidades
50 encuestado
49 programas
48 tipos
48 objeto
48 capítulo
47 conceptual
47 clave
46 uml
46 proyecto
45 requerimientos
---------->

Resultados:

Al aplicar el método (ver la Tabla 1) paulatinamente a todos los documentos del entrenamiento, se observó que al usar el punto de transición y más específicamente la *ley de Zipf*, se revela con sobresaliente precisión la ocurrencia de las palabras en un texto. Los resultados del experimento muestran el aumento del punto de transición para todos los cálculos una vez que se fueron agregando de forma gradual cada documento.

Al examinar las palabras que se encuentran en el 40 % de los términos más cercanos al PT (anteriormente mostradas como resultado de la ejecución del método), se hallan algunos términos jergales del área de Ciencias de la Computación, y específicamente de la clase Sistema de Base de Datos.

Docs. Entrenamiento	#Tèrm.	#Tèrm. Vocab.	#Tèrm. PT	Frec. mas alta/Tèrm.	Térm. Frec = 1	PT	%	Umbral 1	Umbral 2
sbd1	14552	1623	32	211/datos	871	41.2403	-	24.7442	57.7364
sbd2	29206	2996	41	358/datos	1573	55.5914	34.80%	33.3549	77.828
sbd3	39693	3715	58	561/datos	1873	60.7066	9.20%	36.424	84.9893
sbd4	59721	4732	85	732/datos	2274	66.9407	10.27%	40.1644	93.717
sbd5	73033	5734	101	791/datos	2783	74.1073	10.71%	44.4644	103.75
sbd6	84844	6174	115	971/datos	2940	76.1828	2.80%	45.7097	106.656
sbd7	102071	6825	144	1243/datos	3192	79.4015	4.22%	47.6409	111.162
sbd8	114429	7348	150	1408/datos	3381	81.7329	2.94%	49.0397	114.426
sbd9	126225	7707	172	1597/datos	3509	83.275	1.89%	49.965	116.585
sbd10	139970	8030	175	1923/datos	3608	84.4485	1.41%	50.6691	118.228
sbd11	153655	8601	178	2164/datos	3824	86.9543	2.97%	52.1726	121.736
sbd12	164586	8816	191	2387/datos	3833	87.0571	0.12%	52.2343	121.88
sbd13	182053	9199	214	2809/datos	3979	88.709	1.90%	53.2254	124.193
sbd14	191108	9426	223	2924/datos	4061	89.6235	1.03%	53.7741	125.473

Tabla1. Aplicación del punto de transición a 14 documentos de la clase Sistema de Base de Datos.

Experimento 2.

En este experimento se llevó a cabo la aplicación del método a 15 documentos de tesis de cuatro clases: Complejidad de Algoritmos, Graficación, Redes, y Sistemas de Base de Datos, lo que formó un *corpus* no etiquetado de 280,921 palabras, con un vocabulario promedio de 5662 de cada clase. Además se calculó el punto de transición

Ejecución del método a 4 documentos, formando así la clase Complejidad de Algoritmos. La clase contiene 78,796 palabras con un vocabulario de 6,236:

---------->
Frecuencia mas alta: 389 , Termino= algoritmo
Terminos con frecuencia 1= 2793
Punto de Transicion= 74.2412
Umbral1= 44.5447 , Umbral2= 103.938
---------->
Terminos dentro del PT
---------->

103 funciones	85 q
99 parámetros	83 probabilidad
99 muestra	83 cromosomas
98 robot	82 lineal
98 mochila	81 simple
97 objetivo	81 memoria
95 variable	78 local
94 resultados	77 libre
93 modelo	76 mutación
92 ejecución	74 información
90 figura	73 roadmap
89 nodos	73 operaciones
88 proceso	73 migración
88 milestones	73 consulta
86 sección	73 contiene
86 ejemplo	72 operadores
86 procesadores	71 milestone
86 manera	70 usuario
85 resolver	69 mismo

67 referencia
67 árbol
66 objetos
65 características
65 configuraciones
64 múltiples
64 parte
64 m
64 trabajo
63 factible
63 nombre
62 tamaño
62 comunicaciones
62 implementación
62 punto
62 esquema
62 planificador
61 tipos
61 sistemas
61 paralelo
60 campo
59 puntos
59 procesador
59 ramificación
58 f
58 b
58 nuevo
57 técnicas
57 valores
57 realizar
56 consiste
55 simplex
55 generación
54 dispersa
54 biblioteca
54 n
54 clases
---------->
53 general
53 posible
52 optimización
52 gran
51 enteros
51 vez
51 poblaciones
50 diseño
50 procesos
50 ciclo
50 base
49 existe
49 técnica
49 locales
49 configuración
49 mejora
49 solo
49 uso
49 usando
48 cota
48 permite
48 continuación
47 resolución
47 planificación
47 movimientos
47 definir
47 regresa
47 mayor
47 x
46 muestreo
46 complejidad
46 programa
46 embargo
46 utilizando
46 obtener
45 cromosoma
45 segmento

Ejecución del método a 3 documentos, formando así la clase Graficación. La clase contiene 72,440 palabras con un vocabulario de 5,653 términos:

---------->
Frecuencia mas alta: 826 , Termino= imagen
Terminos con frecuencia 1= 2775
Punto de Transicion= 74
Umbral1= 44.4 , Umbral2= 103.6
---------->
Terminos dentro del PT
---------->

97 permite
95 final
92 tabla
92 formato
90 resultados
90 grises
87 clic
86 función
86 tipo
86 través
84 punto
81 mismo
79 proceso
79 colores
79 trabajo
78 tamaño
78 información
78 conjunto
77 matriz
77 obtener
75 tesis
75 b
75 manera
73 capítulo
72 plantilla
71 resultado
71 procesos
69 estudio
68 ventana
68 tiempo
67 operaciones
67 microscopio
67 bits
66 archivo
64 calibración
64 nombre
64 región
64 menú
64 caso
63 interfaz
61 aplicación
60 realizar
60 serie
59 zoom
59 características
58 tejidos
58 modelo
58 área
58 funciones
55 diseño
55 libre
53 guardar
52 estructura
52 curva
51 botón
50 delphi
50 computadora
50 filtros
49 método
49 señales
49 pruebas
48 puntos
48 tejido
48 existen
47 texto
47 opciones
47 seleccionar
46 células
46 form
45 fisiología

---------->

Ejecución del método a 4 documentos, formando así la clase Redes. La clase contiene 69,964 palabras con un vocabulario de 6,027 términos:

---------->
Frecuencia mas alta: 396 , Termino= datos
Terminos con frecuencia 1= 3024
Punto de Transicion= 77.2705
Umbral1= 46.3623 , Umbral2= 108.179
---------->
Terminos dentro del PT
---------->

108 cliente	68 solo
108 paciente	67 usuarios
108 tecnología	66 ap
107 aplicaciones	63 redes
106 servicios	62 comunicaciones
102 aplicación	62 objetos
101 consumidor	61 pc
100 vez	61 crear
99 parte	60 proceso
98 manera	60 archivo
97 perfil	60 ejemplo
96 enviar	59 administrador
94 proveedor	58 acl
91 través	57 llamada
90 realizar	57 uso
90 rdsi	56 estudios
89 agentes	56 modelo
88 canal	55 línea
88 interfaz	55 fecha
85 baja	52 ac
84 enlace	52 comportamiento
82 permite	52 pacientes
81 performativa	51 conexiones
81 registro	51 nivel
78 sistemas	51 código
78 búsqueda	50 internet
77 módulo	49 id
77 protocolos	49 mismo
77 canales	49 estudio
73 mensajes	49 funciones
73 archivos	49 imágenes
71 nombre	48 móviles
71 función	48 servidores
70 control	48 l

48 cap
47 bt
47 petición
47 doctor
---------->

47 solicitud
47 personal
47 s

Ejecución del método a 4 documentos, formando así la clase Sistema de Base de Datos. La clase contiene 59,721 palabras con un vocabulario de 4,732 términos:

---------->
Frecuencia mas alta: 732 , Termino= datos
Terminos con frecuencia 1= 2274
Punto de Transicion= 66.9407
Umbral1= 40.1644 , Umbral2= 93.717
---------->
Terminos dentro del PT
---------->

93 tipo
93 tabla
93 canciones
91 género
89 documentos
87 solo
86 canción
84 realizar
84 lenguaje
82 entidad
82 opción
82 pago
81 proceso
81 administración
71 biblioteca
70 vez
70 catálogo
70 electrónico
69 parte
69 modificar
67 nombre
67 mysql
66 carrito
66 atributos
64 eliminar
64 elementos
64 presenta
63 casos
63 compra
62 consulta
61 resultado
61 navegación
60 tienda
59 diagramas
58 conjunto
57 análisis
56 html
56 cuenta
56 permite
55 servidor
55 gestión
55 nuevo
54 través
54 búsqueda
54 internet
53 aplicaciones
53 orden
53 relaciones
53 opciones
53 ingeniería
53 objetos
53 formato
53 bibliotecas
52 realiza
52 campos
51 sql
51 estructura
51 modificación
51 actividades
51 descripción
51 sistemas
51 acceso
50 función
50 agregar
50 nivel
50 esquema
50 digital
49 entidades

49 tiempo
49 ejemplo
48 capítulo
48 programación
47 relacional
47 conceptual
46 uml
45 operaciones
45 programas
45 principal
44 implementación
44 describe
44 manera
43 php
43 autenticación
43 modelos
42 clave
---------->

Resultados:

Aplicamos el método a cada *corpus* resultante del total de los documentos organizados por clase. Observando las palabras que se encuentran en el 40 % de los términos más cercanos al PT, nos damos cuenta que se hallan términos jergales aún más representativos de su propia clase que los obtenidos en el experimento 1. La Tabla 2 muestra de manera resumida los resultados.

Clase	#Docs.	#Térm.	#Térm. Vocab.	Térm. PT	Frec. mas alta/Térm.	Térm. Frec=1	PT	Umbral 1	Umbral 2
Complejidad Alg.	4	78796	6236	112	389/algoritmo	2793	**74.2412**	44.5447	103.938
Graficación	3	72440	5653	70	826/imagen	2775	74	44.4	103.6
Redes	4	69964	6027	75	396/datos	3024	77.2705	46.3623	108.179
Sistemas B.D.	4	59721	4732	85	732/datos	2274	66.9407	40.1644	93.717

Tabla 2. Aplicación del punto de transición a 4 clases del área de Ciencias de la Computación.

Con el objetivo de comparar el resultado del punto de transición al unir los *corpora* de dos clases diferentes, se realizaron los cálculos mostrados en la Tabla 3, acompañados por el porcentaje que muestra sus diferencias. Por ejemplo (ver Tablas 2 y 3), en la clase Complejidad de Algoritmos, el punto de transición es 74.2412; al unir el *corpus* de esta clase con la clase Graficación, se obtuvo 91.447, es decir, que el punto de transición sufrió un aumento del 23.18%.

Clase	Complejidad Alg.		Graficación		Redes		Sistemas B.D.	
Complejidad Alg.	-	-	**91.447**	23.18%	93.5226	25.97%	87.137	17.37%
Graficación	91.447	23.58%	-	-	93.3203	26.11%	87.0229	17.60%
Redes	93.5226	21.03%	93.3203	20.77%	-	-	88.3271	14.31%
Sistemas B.D.	87.137	30.17%	87.0229	30.00%	88.3271	31.95%	-	-

Tabla 3. Aplicación del punto de transición entre clases diferentes.

Reconocemos algunas peculiaridades acerca de los términos elegidos por la aplicación del método de compilación, que serán determinantes en la construcción del diccionario jergal[24] :

1. Surgen jergales significativos: sql, web, mysql, php, null, etc.
2. Palabras de más de un elemento: consulta - “consulta sql”, mochila - “algoritmo de la mochila”, través - “a través de”,
3. La abreviatura id, que tendrá que ser remitida a la entrada en español.
4. Anglicismos como roadmap y form, remitir a la entrada del término en español.

[24] Tomando en cuenta los alcances acordados en el establecimiento del marco teórico.

Experimento 3.

Se aplicó el método a los 4 *corpus* de las cuatro clases consideradas: Complejidad de Algoritmos, Graficación, Redes, y Sistemas de Base de Datos. Los resultados finales coinciden con los resultados obtenidos por el experimento 2, sin embargo aquí se realizó el cálculo del punto de transición entre el *corpus* de una clase (por ejemplo Redes) y el *corpus* del primer documento de una segunda clase (por ejemplo Complejidad de Algoritmos) posteriormente, al *corpus* resultante se le agregó el segundo documento (de la misma clase Complejidad de Algoritmos, en nuestro ejemplo) y así sucesivamente.

Clase	Doc.	Comp. Alg.	%	Graficación	%	Redes	%	Sistema B.D.	%
Comp. Alg.	DOC1	-	-	78.229	-	81.453	-	72.536	-
	DOC2	-	-	84.7306	8.31%	87.24	7.10%	79.8881	10.14%
	DOC3	-	-	88.8546	4.87%	91.087	4.41%	83.929	5.06%
	DOC4	-	-	**91.447**	2.92%	**93.523**	2.67%	**87.137**	3.82%
Graficación	DOC1	87.456	-	-	-	89.357	-	82.3508	-
	DOC2	89.8673	2.76%	-	-	91.632	2.55%	84.5191	2.63%
	DOC3	**91.447**	1.76%	-	-	**93.32**	1.84%	**87.0229**	2.96%
Redes	DOC1	77.1032	-	76.3261	-	-	-	70.2831	-
	DOC2	88.0903	14.25%	87.9096	15.18%	-	-	82.8322	17.86%
	DOC3	91.0874	3.40%	90.9344	3.44%	-	-	85.9769	3.80%
	DOC4	**93.5226**	2.67%	**93.3203**	2.62%	-	-	**88.3271**	2.73%
Sistema B.D.	DOC1	77.3605	-	76.8967	-	79.201	-	-	-
	DOC2	82.3387	6.44%	81.8301	6.42%	83.834	5.85%	-	-
	DOC3	84.1891	2.25%	83.5491	2.10%	85.467	1.95%	-	-
	DOC4	**87.137**	3.50%	**87.0229**	4.16%	**88.327**	3.35%	-	-

Tabla 4. Aplicación del punto de transición entre el *corpus* de una clase y el *corpus* de los documentos de otra.

Resultados:

En el experimento se calculó el punto de transacción de dos *corpus*: el primero el de la clase y un segundo *corpus* que corresponde a un nuevo documento de una clase diferente. Al analizar los datos que resumen los resultados del experimento (Tabla 4) se observaron dos casos:

5. Aquellos en los cuales el aumento fue importante, como por ejemplo, el caso de la clase Sistema de Base de Datos con PT = 66.9407 que, al agregarle el *corpus* del documento 1 correspondiente a la clase Redes, PT = 70.2831, con el segundo PT = 82.8322, con el tercero PT = 85.9769 y por último el cuarto PT = 88.3271, alcanzó un aumento de 31.95% de los términos.
6. Aquellos en los cuales el aumento fue menor, como por ejemplo, la clase Complejidad de Algoritmos con PT = 74.2412, cuyo punto de transición al agregarle el documento 1 de la clase Sistema de Base de Datos es PT = 77.3605, documento 2 PT = 82.3387, documento 3 PT = 84.1891 y documento 4 PT = 87.137 aumentó a 17.37%.

Se usó el punto de transición para la selección de términos índice para beneficiar la categorización de textos.

Conclusiones y trabajo futuro

Como se menciona en Reyes 2004, un documento que aumente el punto de transición significa que el documento es ajeno a nuestro dominio (clase), mientras que un documento que disminuya el punto de transición indica que el contenido del documento ya se encuentra dentro de nuestra colección y debería ser considerado en caso de querer aumentar la frecuencia de términos que puedan estar sobrepresentados. Así, en este trabajo se concluye que, para aquellos documentos que han aumentado el punto de transición de manera considerable (en base a los porcentajes obtenidos), el documento no ajusta en el dominio al cual se intentó probar que perteneciera. Por el contrario, aquellos en los cuales el aumento no ha sido tan drástico indica que, el documento contiene términos que ya pertenecían al *corpus*, incluso éstos aumentan la frecuencia de los términos.

Los documentos utilizados en este trabajo pertenecen a tesis de licenciatura, sin duda alguna los términos que en ellos se encuentran poseen muchas acepciones, por ejemplo, la palabra “modelo” en la mayoría significa el dibujo, diagrama o representación de algo sin que este significado tenga relevancia para la representación del documento, pero en otros pertenecería al conjunto de términos que representan el documento como “modelo entidad-relación” en la clase o categoría Sistema de Base de Datos.

Se ha presentado una aplicación sencilla de la compilación de un *corpus*, que forma parte de la metodología para la elaboración de un diccionario de Ciencias de la Computación. El *corpus* se conformó con documentos de tesis de licenciatura de la Facultad de Ciencias de la Computación BUAP.

Referente al trabajo futuro, se plantea continuar con la segunda parte de esta investigación, la cual queda abierta para la subsiguiente parte. Es importante señalar que el proyecto en cual está inserto este trabajo es mucho más amplio y ambicioso, a saber, el Diccionario de Ciencias de la Computación de la Doctora María Elena Franco Carcedo y el Doctor Héctor Jiménez Salazar.

Es necesario, seleccionar con especial empeño los documentos que se van agregando al *corpus*, ya que, los documentos de tesis aquí usados pueden pertenecer a más de un subdominio específico de computación, en consecuencia, los subdominios pierden su naturaleza. Del vocabulario resultante en la aplicación del método, falta identificar los términos jergales de computación (entre estos anglicismos). También, habrá que determinar la terminología aplicando métodos clásicos, e identificando contextos definitorios (CD) (Alarcón Sierra 2005 y Alarcón, Bach, Sierra 2006).

Anexo

Programa con el cual se apoyó el método en lenguaje AWK:

ptran.awk

```
awk '
BEGIN { NTP=0.4;}
FILENAME==stp{
  stopwords[tolower($1)]=1;
  next;
}
FILENAME==ruido{  stopwords[tolower($1)]=1;
  next;
}
{  gsub(/[,;:_-?()\x2E[\]{}=*+$0-9\/"\x27%&]/, " ", $0);
  for(i=1;i<=NF;i++)
     if (!($i in stopwords)) freq[$i]++;
}
END {
  print "------------------------------->";
  n=asort(freq, a);
  for(x in freq) {
    if (freq[x] == a[n])
    print "Frecuencia mas alta: " freq[x], ", Termino= " x;
    if (freq[x] == 1) T1++;
  }
  print "Terminos con frecuencia 1= "T1;
  PT = (sqrt(1+8*T1) - 1)/2;
  print "Punto de Transicion= " PT;
  Umbral1= (1-NTP)*PT;
  Umbral2= (1+NTP)*PT;
  print "Umbral1= "Umbral1, ", Umbral2= "Umbral2;
  print "------------------------------->";
  print "Terminos dentro del PT";
  print "------------------------------->";
  for (x in freq) {
    if ((freq[x] >= Umbral1) && (freq[x] <= Umbral2)) print freq[x], x;
  }
  print "------------------------------->";
} ' stp=$1 ruido=$2 $*
```

Bibliografía

- Abbagnano, Nicola:
    - (1996): ***Diccionario de filosofía***. FCE. México
- Alarcón, Rodrigo; Sierra, Gerardo:
    - (2005): ***Reglas léxico-metalingüísticas para la extracción automática de contextos definitorios.*** Lingüística Aplicada, Universidad Pompeu Fabra.
- Alarcón, Rodrigo; Bach, Carmen; Sierra, Gerardo:
    - (2006): ***Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica***, UNAM.
- Amorós Rica y Merlin Walch:
    - (1993): ***Dictionnaire Juridique Français-Espagnol, Español-Francés***. LGDJ. París
- Barrón, Alberto; Sierra, Gerardo y Villaseñor, Elio:
    - (2006): **"C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español"**, *3er Taller de Tecnologías del Lenguaje Humano*, San Luis Potosí.
- Booth, A.:
    - (1967): "**A Law of Ocurrences for Word of Low Frequency".** Information and Control, 10(4) pp 386-93.

- Campos, Juana y Barella, Ana:
    - (1993): ***Diccionario de refranes***. Espasa. Madrid
- Carcedo, Elena F.:
    - (1994): "Bibliografía descriptiva", "Índice onomástico", "Índice de topónimos", "Transcripción y documentos originales" [reproducción]" en ***La personalidad literaria de Gabriel Lobo Laso de la Vega (1555-1615),***

con la edición de los Elogios y las Tragedias. Univ. Complutense (ed. digital: ISBN 84-8466-312.4). Madrid

- (1995/nov.): "Sobre el lenguaje científico y técnico y su traducción" en ***Encuentro de la Traducción Científica y Técnica***. Puebla

- (1996/oct.): "Reflexiones sobre la traducibilidad de las lenguas" en ***III Encuentro sobre la Traducción Cinentífico-Técnica***. Univ. Veracruzana. Jalapa

- (1999/sept.) "Calidad de la Enseñanza y Educación Continua" en ***COPEI 99***, Puebla

- (2000/mar.): "Juicio, Creación, Lenguaje" en ***Taller de pensamiento crítico y creativo UIA***. Puebla

- (2001/jul.): "Estudio del nivel de domino del lenguaje verbal I: léxico y lenguajes especializados" en ***Magistralis*** nº 21 UIA. Puebla

- (2001/nov.): "Sobre formación educativa. Una propuesta" en ***Segundo Encuentro de Investigación y Educación.*** Memoria del Colegio de Bachilleres. Puebla

- (2002/oct.): "Lenguaje, ¿instrumento de discriminación y marginación?" en ***I Congreso Internacional de Psicología Social.*** (Memoria en formato digital). Puebla

- (2003): ***Los géneros y su práctica. Con una guía gramatical***. UAP. Puebla

- (2004): ***Lenguaje científico y técnico y elaboración de tesis de posgrado***. UAP. Puebla

(2006/oct.): "Lenguaje natural, identidad y exclusión/inclusión social" en ***III Congreso Internacional de Psicología Social.*** BUAP. Puebla

- Casares, Julio:

  - (1959): ***Diccionario ideológico de la lengua española***. Barcelona

- Corominas, J. y Pascual, J.A.:

  - (1980): ***Diccionario crítico etimológico castellano e hispánico***. (6 vols.) Gredos. Madrid

- Covarrubias, Sebastián:

- (1998): ***Tesoro de la Lengua Castellana o Española***. ed. *ad litteram* de la de 1611, facsímil de la de 1943. Alta Fulla. Barcelona

- [Damms, ed.]:
  - (1973): ***Damms Lommeordböker spansk-norsk, norsk-spansk***. (2 vols.). Oslo
- [EDAF, ed.]:
  - (1971): ***Diccionario de la Mitología Mundial***. EDAF. Madrid
- Facultad de Medicina-Univ. de Navarra:
  - (1999): ***Diccionario Espasa de Medicina***. Espasa. Madrid
- Franco Grande, Luis:
  - (1972): ***Diccionario Galego-Castelan e Vocabulario Castelan-Galego***. Galaxia. Vigo
- García, Antonio
  - [sin fecha]: ***Dizionari del turista italiano-spagnolo, spagnolo-italiano***. Vallardi. Milán
- [Hamlyn, ed.]:
  - (1977): ***Hamlyn Spanish Dictionary***. Hamlyn. Londres
- Jiménez Salazar H., Castro M., Rojas Franco, Miñón E., Pinto D., Carcedo, Elena F,:
  - (2005) **Unsupervised Term Selection using Entropy**, Research on Computing Science 14, ISSN 1665-9899, pp 163-172, México.

- Jiménez Salazar H. & Pinto D., Rosso P.:
  - (2005) **Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos**, Procesamiento del Lenguaje Natural, núm. 35, pp. 383-390.
- [Kunnskapsforlaget, ed.]:
  - (1990): ***Spansk-Norsk, Norsk-Spansk***. Kunnskapsforlaget. Oslo
- [Langenscheidt, ed.]
  - (1960): ***Diccionario Universal Langenscheidt inglés-español, español-inglés***. Langenscheidt. Gran Bretaña

- (1965): ***Diccionario Universal Langenscheidt italiano-español, español-italiano***. Langenscheidt. Gran Bretaña

- Lavid, Julia:

- (2005): ***Lenguaje y nuevas tecnologías***. Cátedra. Madrid

- Librairie Larrousse:

- (1977): ***Petit Larousse illustré***. Larousse. París

- Luhn, H.P.:

- (1958): ***The Automatic Creation of Literature Abstracts***. 1er. National Convention, IBM Journal.

- Martí, Antonia y Llisterri, Joaquim (editores):

- (2002): ***Tratamiento del lenguaje natural***. Univ. de Barcelona. Barcelona

- Meyer, I.:

- (2001): "Extracting Knowledge-rich contexts for Terminography". En Bourigault: ***Recent Advances in Computational Terminology***. John Benjamin's, Amsterdam, págs. 279-302.

- Pearson, J.:

- (1998): ***Terms in context, John Benjamin's***, Amsterdam.

- Pao, M.L.:

- (1977): ***Automatic Indexing Based of Goffman's Transition of Word Ocurrences***, Proc. of the ASIS Annual Meeting.

- Real Academia de Ciencias Exactas, Físicas y Naturales:

- (1996): ***Vocabulario Científico y Técnico*** Espasa. Madrid

- Real Academia Española

- (1975): ***Diccionario manual e ilustrado de la lengua española***. Espasa. Madrid

- (1985): ***Esbozo de una nueva gramática de la lengua española***. Espasa. Madrid

- (1992): ***Diccionario de la lengua española***. (2 vols.) Espasa. Madrid

- (2001): ***Diccionario de la lengua española***. Espasa. Madrid

- Reyes, Berenice:

- (2004): ***Un método para la compilación de un corpus***, Tesis de Lic. en Ciencias de la Computación, FCC-BUAP, Puebla.

- Rojo Pérez, Elena:
  - (2000): ***Los diccionarios. Introducción a la lexicografía del español.*** Trea. Gijón

- Urbizagástegui-Alvarado, R:
  - (1999): ***Las Probabilidades de la ley de Zipf en la Indización Automática.*** Reporte de la Universidad de California Riverside.

- Velásquez, Francisco; Gelbukh, Alexander y Sidorov, Grigori:
  - (2002): ***AGME, un sistema de análisis y generación de la morfología del español***, IBERAMIA 2002, España.

- Zipf, G. K.:
  - (1949): ***Human behaviour and the principie of least effort***, Addison-Wesley.