UNIVERSIDAD NACIONAL DEL ALTIPLANO - PUNO
ESCUELA DE POST - GRADO
MAESTRÍA EN INFORMÁTICA

"RECUPERACIÓN SEMANTICA DE LA INFORMACIÓN USANDO LA SIMILITUD DISTRIBUCIONAL"

TESIS

PRESENTADA POR:

EDGAR HOLGUIN HOLGUIN

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGÍSTER SCIENTIAE EN INFORMÁTICA

PUNO - PERÚ

2013

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSTGRADO

MAESTRIA EN INFORMATICA

"RECUPERACION SEMANTICA DE LA INFORMACION USANDO LA SIMILITUD DISTRIBUCIONAL"

TESIS

PRESENTADA POR:

EDGAR HOLGUIN HOLGUIN

PARA OPTAR EL GRADO DE

MAGISTER SCIENTIAE EN INFORMATICA

APROBADA POR EL JURADO REVISOR DE TESIS CONFORMADO POR:

PRESIDENTE : _____

M. Sc. Ernesto Nayer Tumi Figueroa

PRIMER MIEMBRO : _____

M. Sc. Leonel Coyla Idme

SEGUNDO MIEMBRO : _____

M. Sc. Percy Huata Panca

TERCER MIEMBRO Y ASESOR : _____

Dr. Angel Manuel Olazabal Guerra

DEDICATORIA

Con gran afecto a mis padres: Francisca Paula Holguin Yupanqui y Mario Jacobo Holguin Flores, a mi esposa Marina Aceituno Vargas, a mis hijos Valerie Melanieb y Gabriel Fernando y a mis hermanos Adelina, Héctor y Ronald.

AGRADECIMIENTO

Mis más sinceros agradecimientos a la Universidad Nacional del Altiplano, por brindarme la oportunidad de formarme profesionalmente, a la Escuela de Post Grado de la Universidad Nacional del Altiplano por darme la oportunidad de continuar mis estudios de Post Grado, al Dr. Angel Manuel Olazabal Guerra y M.Sc. Juan Antonio Flores Moroco, asesores de esta tesis. Al M.Sc. Ernesto Nayer Tumi Figueroa, M.Sc. Leonel Coyla Idme y M.Sc. Percy Huata Panca por sus consejos y a todos mis amigos.

INDICE

CAPITULO I

EL PROBLEMA DE LA INVESTIGACION

CAPITULO II

MARCO TEORICO

CAPITULO III

METODOLOGIA

CAPITULO IV

RESULTADOS Y DISCUSION

INDICE DE CUADROS

INDICE DE FIGURAS

RESUMEN

Recuperar información con un criterio semántico desde la Web, en donde la información almacenada no es estructurada, se requiere de mecanismos complejos y diversos que consideren el procesamiento de lenguaje natural. En esta tesis se implementó y analizó un método de recuperación semántico de la información, partiendo de la premisa que si existen palabras que coocurren en un contexto determinado, éstas tienen una relación semántica. Para la implementación de un mecanismo de recuperación sobre una colección de documentos se hizo necesario un procesamiento, representación y análisis de relación de los términos. El Modelo Vectorial para la recuperación semántica de la información utilizado, permitió definir las premisas necesarias e importantes para determinar si un conjunto de palabras son relevantes a la necesidad de información, calculando la medida de similitud y establecimiento del ranking de vocablos más semejantes semánticamente. La lejanía o cercanía de dos vocablos se determinó utilizando la similitud distribucional representado por un vector de coocurrencia y se cuantificó mediante el coseno del ángulo que forman sus vectores. Al evaluar el rendimiento del sistema de recuperación de la información, se concluyó que es importante el corpus utilizado en la construcción del mismo así como el pre procesamiento, estructura y técnicas de recuperación.

PALABRAS CLAVE: corpus, funciones, similitud, ontologías, semántica, información, Web.

ABSTRACT

Recall information with a semantic criterion from the Web, where the stored information is unstructured, it requires complex and diverse mechanisms that consider natural language processing. In this thesis we implemented and analyzed a method for semantic retrieval of information, based on the premise that if there are words that co-occur in a given context, they have a semantic relationship. To implement a recovery mechanism for a collection of documents became necessary processing, representation and analysis of relationship between the terms. Vector Model for semantic retrieval of information used, allowed to define the conditions required and important to determine whether a set of words are relevant to the information need by calculating the similarity measure and ranking establishment semantically most similar words. The remoteness or proximity of two words was determined using the distributional similarity represented by a vector of co-occurrence and quantified by the cosine of the angle between their vectors. Assessing the system performance information retrieval, it was concluded that it is important corpus used in the construction thereof as well as the pre-processing, structure and recovery techniques.

KEYWORDS: corpus, functions, similarity, ontologies, semantic, information, Web.

INTRODUCCION

El desarrollo de Internet ha permitido un crecimiento permanente del volumen de la información, este potencial de información puede ser aprovechado por todos los usuarios y en diferentes situaciones y la importancia de éste hecho aumenta en el sentido que disponer o no de la información necesaria y justo a tiempo y forma puede resultar en el éxito o fracaso de una operación en tiempos tan competitivos como el actual.

La información almacenada en Internet carece de una organización o estructuración lógica, a diferencia de las tecnologías de procesamiento de datos, que organiza los datos en tablas y grandes bases de datos y proporcionan mecanismos de extracción de información realmente aceptables, las tecnologías de procesamiento de lenguaje natural operan sobre documentos Web, que en su amplia mayoría están en formatos HTML o en archivos PDF, DOC, RTF, TXT y otros, sin estructura lógica alguna y que no proporcionan mecanismos de extracción de información adecuados.

La recuperación no es un área nueva, sino que se viene desarrollando desde finales de la década de 1950, y que en la actualidad adquiere importancia debido al valor que tiene la información. El modelo de espacio de palabras (WSM por sus siglas en inglés Word Space Model), es una estructura que permite la representación del texto, y permite además determinar la lejanía o cercanía semántica entre un par de vocablos tomando en cuenta su

distribución con el resto de elementos del lenguaje, usando un espacio multidimensional, cuyo número de dimensiones *n*, depende del número de vocablos diferentes encontrados en el corpus del texto usado en la etapa de entrenamiento o construcción. Cada vocablo se representa mediante un vector de *n* dimensiones, que determina la distribución de éste con los demás elementos del sistema.

Existen diversos modelos para recuperar la información tales como: el modelo booleano, vectorial, probabilístico y otros. En esta tesis, se aborda el estudio utilizando el método vectorial que cuantifica la similitud semántica de dos palabras mediante el coseno del ángulo que forman sus vectores, seleccionando los vocablos que comparten el mismo tópico o grupo semántico, cuyos resultados son de gran utilidad en casi todas las áreas del procesamiento del lenguaje natural.

CAPITULO I

PROBLEMA DE INVESTIGACION

1.1. PLANTEAMIENTO DEL PROBLEMA

1.1.1. Formulación del problema

El ser humano utiliza el lenguaje natural para comunicarse, compartir experiencias, explicar el mundo que lo rodea y registrar su evolución. A través del tiempo, el ser humano también ha utilizado la escritura como un medio para registrar y comunicar su existencia; ésta ha evolucionado y en la actualidad, la información se representa digitalmente, desde simples archivos de texto, libros y revistas electrónicas, hasta librerías digitales y en espacios mucho más grandes y complejos como la Web.

Acceder a toda la información existente y disponible en la Web se hace necesario e importante para las personas (usuarios), ya que la gran información almacenada en la Web puede ser aprovechado por todos los usuarios y en diferentes situaciones y la importancia de éste hecho aumenta en el sentido que disponer o no de la información justo a tiempo y forma puede resultar en el éxito o fracaso de una operación en tiempos tan competitivos como el actual. Sin embargo éste hecho plantea retos importantes porque la información disponible está en una forma no estructurada, en formatos y formas diversas, entonces es necesario contar con un método que permita unir preguntas o necesidad de información de los usuarios con respuestas o documentos que están presentes en Internet considerando un lenguaje tan amplio como el español.

La falta de organización en el almacenamiento de la información, requiere la utilización de algún tipo de tecnología que gestione de forma eficaz toda la información disponible, para que tanto las búsquedas como las consultas sean efectivas. Esta problemática ha derivado en la utilización de dos tipos de tecnologías bien diferenciadas: Tecnologías de Procesamiento de Datos y Tecnologías de Procesamiento del Lenguaje Natural. Cada una de estas tecnologías procesa de forma diferente la información. A diferencia de las Tecnologías de Procesamiento de Datos que se ocupan de reducir el espacio ocupado, almacenar de forma óptima los datos, ahorrar tiempos de respuesta en la búsqueda de algún tipo de información, etcétera, las Tecnologías de Procesamiento del

Lenguaje Natural necesitan un conocimiento más profundo del lenguaje para poder procesar la información.

1.1.2. Definición del problema

¿En qué medida se optimiza la búsqueda utilizando la similitud distribucional en la recuperación semántica de la información?

1.1.3. Justificación de la investigación

Recuperar la información en un entorno gigantesco y de crecimiento exponencial como la Web, en donde se almacenan información sin estructuración lógica alguna, requiere de técnicas diferentes a los que nos pueden ofrecer los motores de búsqueda tradicionales basados en índices de búsqueda. Además el usuario necesita recuperar información expresado en su lenguaje natural utilizando el significado de tipo morfológico, sintáctico, semántico y pragmático que le proporciona el lenguaje.

Los motores de búsqueda tradicionales nos permiten realizar búsquedas en entornos formales, principalmente basados en la morfología y sintaxis de la palabra. Sin embargo el lenguaje humano es rico en el sentido que podemos encontrar múltiples expresiones y palabras que pueden tener varios significados, dependiendo de las circunstancias en las que se usa. Por ejemplo, cuando utilizamos la

palabra "Titicaca", hacemos referencia a un Lago que queda cerca de la ciudad de Puno, un lugar Turístico. Entonces es importante contar con una tecnología que permita recuperar información tomando en cuenta esta característica semántica.

1.1.4. Limitaciones y restricciones de la investigación

El ser humano almacena información en diversos medios y formatos, tales como imágenes, videos, sonidos y texto. Éste último es la fuente de análisis para la presente investigación. Además, se utilizará el Corpus de Google, el cual proporciona la frecuencia de las palabras relacionadas más consultadas en la Web.

1.2. OBJETIVOS DE LA INVESTIGACION

1.2.1. Objetivo General

Optimizar la recuperación semántica de la información utilizando la similitud distribucional.

1.2.2. Objetivo Específicos

a. Definir una metodología para el desarrollo de un sistema de recuperación semántica de la información.

b. Implementar un algoritmo que permita calcular la distancia entre vectores y obtener un conjunto de términos con cierta similitud semántica.
c. Evaluar el resultado de la recuperación semántica de la información utilizando la similitud distribucional.

1.3. HIPOTESIS DE INVESTIGACION

La búsqueda mediante la similitud distribucional, optimiza la recuperación semántica de la información.

CAPITULO II

MARCO TEORICO

2.1. ANTECEDENTES DE LA INVESTIGACION

Martínez (2002), en la Universidad de Murcia, en su trabajo de investigación titulada: "Propuesta y desarrollo de un modelo para la evaluación de la recuperación de información en Internet". cuyos objetivos fueron: evaluar la efectividad de la recuperación de la información de un conjunto de motores de búsqueda en Internet que posean unas características similares en el tamaño de búsqueda de los índices y en las posibilidades de búsqueda, determinar el tipo de búsqueda, tomando básicamente la búsqueda por frase literal, la búsqueda por intersección y la combinación de ambas incluyendo operadores booleanos y la formulación de la pregunta más adecuada, estableciendo los criterios para analizar una página web, para ello utiliza la relevancia de Lancaster, que dice: " un documento va a resultar

relevante cuando su contenido tiene que ver con el objeto de la pregunta y, además le resulta útil". La precisión y la exhaustividad, lo realiza mediante la medida simple de Borko, basándose principalmente en los enlaces obtenidos, los enlaces erróneos y los enlaces duplicados. La determinación de la similitud lo realiza utilizando la función de similitud del coseno entre dos vectores. Finalmente agrupa los índices aplicando la técnica del vecino más cercano. En sus conclusiones afirma que el coeficiente de variación de Pearson de la distribución de las exhaustividades medias de cada motor, muestra que los resultados obtenidos por el motor Google lo convierten en el motor de menor dispersión, aunque a medida que ha ido aumentando el número de elementos de la muestra, todos los motores, excepto AltaVista, se han agrupado en torno a un mismo valor. El cálculo de los enlaces duplicados indica que en casi todos los motores poseen porcentajes entre el 1% y el 3%, excepto el motor de búsqueda Terra que fija sus valores entre el 5% y el 6%.

López (2006), en la Universidad de Granada, en su trabajo de investigación titulada: "Modelos de sistemas de recuperación de información documental basados en información lingüística difusa", en ella mejora el diseño de los sistemas de recuperación de información, utilizando técnicas de modelado lingüístico difuso, centrado en mejorar la interacción usuario – sistema de recuperación de información, así como los procesos de evaluación de consultas que realizan dichas consultas. Los objetivos que se plantea son: Revisar los sistemas de recuperación

lingüísticos y técnicas de modelado de información. Diseñar un nuevo modelo de recuperación de información usando aproximaciones lingüísticas de 2-tuplas y no balanceada. Desarrollar técnicas para mejorar la evaluación de las consultas del usuario y finalmente, evalúa las distintas propuestas con respectos a otros modelos de recuperación. En ella concluye que, El modelo lingüístico 2-tupla de sistema de recuperación de la información documental implementado, mejora la interpretación de la semántica de umbral simétrico. El modelo lingüístico no balanceado del sistema de recuperación de información documental implementado, permite flexibilizar las consultas al usuario.

Romero (2012), en el Instituto Politécnico Nacional de México, en su trabajo de investigación titulada: "Diseño e implementación de mecanismos de búsqueda contextualizada y anotado a través de la Web Semántica". En ella presenta un mecanismo de búsqueda que permite encontrar la información acorde al contexto semántico del usuario. Utiliza dos crawlers, uno general y el otro específico para llevar a cabo la extracción de ontologías y documentos de texto y organizar de forma semántica la información. La propuesta de solución que plantea, lo estructura con los siguientes componentes: Definición del contexto, crawler textual, crawler ontológico, analizador de texto (parser), Anotado semántico y poblado de ontologías. En la investigación concluye que el diseño modular que realizó, da solución al problema permitiendo la reutilización de componentes y éstas pueden ser modificadas sin afectar el funcionamiento general de la propuesta. El desarrollo separado de los

crawlers permite realizar exploraciones acorde a cada necesidad, al tratar de manera separada a cada recurso. El módulo de anotado semántico y poblado de ontologías trabaja de manera automática y permite la identificación de las entidades de un documento y la generación de individuos, propiedades y relaciones dentro de la ontología para enriquecerla.

Fresno (2006), en la Universidad Rey Juan Carlos en su trabajo de investigación titulada: "Representación auto contenida de documentos HTML: una propuesta basada en combinaciones heurísticas de criterios", presenta una propuesta que considera dos funciones de ponderación de rasgos. Una de ellas llamada ACC (Analytical Combination of Criteria), se basa en una combinación lineal de criterios heurísticos extraídos de los procesos de lectura y escritura de textos. La otra, FCC (Fuzzy Combination of Criteria), que se construyó a partir de una combinación borrosa o fuzzy. Una de las ventajas que ofrecen ambas funciones es que permite representar un documento HTML, sin necesidad de analizar previamente ninguna colección de referencia. La tesis tiene como conclusiones que los modelos vectoriales son los más utilizados para la representación automática de documentos. Muestra diferentes funciones de ponderación y concluye que las funciones locales resultan más adecuadas para una aplicación en representaciones auto contenidas, porque las globales requieren información externa al documento. En el caso del Clustering, el comportamiento de ACC y FCC resultó más destacable con la colección BankSearch. La evaluación de ambas

representaciones se realizaron mediante procesos de clasificación y clustering de páginas Web, empleando un algoritmo Naïve Bayes, el mismo que resultó satisfactorio.

Seco (2009), en su trabajo de investigación titulada: "Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales", cuyos objetivos fueron: diseñar una estructura de indexación que tenga en cuenta las características textuales y espaciales de los documentos, diseñar una arquitectura para sistemas de recuperación de información geográfica y diseñar una estructura de indexación espacial optimizada para las características de la información geográfica gestionada habitualmente en sistemas GIR. Para ello indexa la información teniendo en cuenta tanto el ámbito textual como, como el ámbito geográfico de los documentos e implementa un sistema geográfico de recuperación de la información. Las conclusiones de las tesis, primero presenta un prototipo desarrollado bajo el paradigma del software libre empleando componentes que éste proporciona. La estructura que utilizó para almacenar la información está basada en el wavelett tree, que es una estructura muy compacta que permite indexar colecciones de puntos en la memoria principal, y de los experimentos realizados demuestra que la relación es muy buena entre la eficiencia de las búsquedas y el espacio necesario para almacenarlo.

2.2. MARCO REFERENCIAL

2.2.1. Problemática de la recuperación de la información.

El ser humano he necesitado registrar toda acerca del mundo que lo rodea, así como sus experiencias en su estadía en la tierra. El medio que inventó fue la escritura el cual ha sido fundamental para soportar su conocimiento en el tiempo. El volumen de la información que ha acumulado y acumula a través del tiempo crece permanentemente y adquiere diferentes formas de representación, desde simples archivos de texto en una computadora personal o un periódico electrónico hasta librerías digitales y espacios mucho más grandes y complejos como la web. Algunos investigadores han planteado que – desde hace varios años – existe un fenómeno denominado "sobrecarga de información" debido a que el volumen y la disponibilidad hacen que los usuarios no cuenten con suficiente tiempo físico para "procesar" todo el cúmulo de medios a su alcance (Tolosa, 2009).

Entonces, resulta importante tratar con toda esa información disponible electrónicamente para que pueda servir a diferentes personas (usuarios) en diferentes situaciones. Esto plantea un desafío interesante: hay importantes volúmenes de información y hay usuarios que se pueden beneficiar de alguna manera con la posibilidad de acceder a ésta, por lo tanto, cómo poder unir preguntas con respuestas,

necesidades de información con documentos, consultas con resultados, ésta es precisamente el desafío que se afronta y se aborda en el estudio de la recuperación de la información (Information Retrieval), proponiendo soluciones al escenario presentado, planteando modelos, algoritmos y heurísticas (Martínez, 2004).

De forma general Baeza-Yates (1999), El problema de la Recuperación de la Información puede ser estudiado desde dos puntos de vista: el computacional y el humano. El primer caso tiene que ver con la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas. El segundo caso corresponde al estudio del comportamiento y de las necesidades de los usuarios.

Si se analiza la problemática de la Recuperación de la Información desde un alto nivel de abstracción podemos establecer que:

a. Existe una colección de documentos que contienen información de interés (sobre uno o varios temas).
b. Existen usuarios con necesidades de información, quienes las plantean al Sistema de Recuperación de Información en forma de una consulta (en inglés, query)
c. Como respuesta, el sistema retorna – de forma ideal – referencias a documentos relevantes", es decir aquellos que satisfacen la necesidad expresada, generalmente en forma de una lista rankeada.

Planteamos que la respuesta "ideal" de un Sistema de Recuperación de Información está formada solamente por documentos relevantes a la consulta, pero en la práctica ésta no es aún alcanzable. Esto se debe a que, entre otros motivos, existe el problema de compatibilizar la expresión de la necesidad de información y el lenguaje y de los documentos. Además, hay una carga de subjetividad subyacente y depende de los usuarios. Entonces, el Sistema de Recuperación de Información recupera la mayor cantidad posible de documentos relevantes, minimizando la cantidad de documentos no relevantes (ruido) en la respuesta. En términos de eficiencia, se plantea la idea de precisión de la respuesta, es decir, cuando más documentos relevantes contengan el conjunto solución (para una consulta dada), más preciso será.

Para cumplir con sus objetivos, un Sistema de Recuperación de Información debe realizar algunas tareas básicas, las cuales se encuentran – fundamentalmente – planteadas en cuestiones computacionales, a saber:

a. Representación lógica de los documentos y – opcionalmente – almacenamiento del original. Algunos sistemas solo almacenan porciones de los documentos y otros lo hacen de manera completa.
b. Representación de la necesidad de información del usuario en forma de consulta.

c. Evaluación de los documentos respecto de una consulta para establecer la relevancia de cada uno.
d. Ranking de los documentos considerados relevantes para formar el "conjunto solución" o respuesta.
e. Presentación de la respuesta al usuario.
f. Retroalimentación o refinamiento de las consultas (para aumentar la calidad de la respuesta)

2.2.2. Recuperación de la información

Las técnicas de recuperación de información empleadas en Internet, proceden de las empleadas en los Sistemas de recuperación de Información (SRI) tradicionales, y es por ello que surgen problemas cuando se realizan operaciones de recuperación de información, en tanto que el entorno de trabajo no es el mismo y las características intrínsecas de los datos almacenados difieren considerablemente.

Los Sistemas de Recuperación de Información (SRI) toman un conjunto de documentos (colección) para procesar y luego poder responder consultas. De forma básica, podemos clasificar los documentos en estructurados y no estructurados. Los primeros son aquellos en los que se pueden reconocer elementos estructurales con una semántica bien definida, mientras que los segundos corresponden a texto libre, sin formato. La diferencia fundamental de un SRI que procese documentos estructurados se encuentra en que puede extraer

información adicional al contenido textual, la cual utiliza en la etapa de recuperación para facilitar la tarea y aumentar las prestaciones. (Lluis, 2009).

A partir de lo expresado anteriormente se presenta una posible clasificación de modelos de RI – la cual no es exhaustiva – de acuerdo a características estructurales de los documentos.

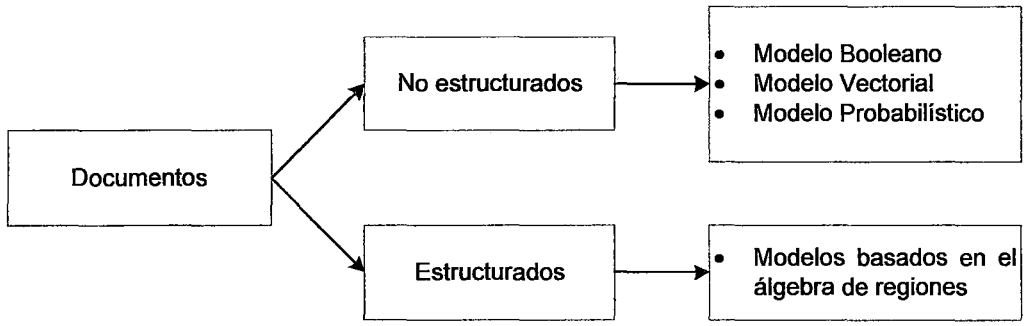


Figura 1. Modelos de recuperación de la información

Fuente: (Tolosa, 2009)

2.2.3. Modelos de recuperación de la información

a. Modelo booleano

El modelo booleano se basa en la teoría de conjuntos y el álgebra de Boole. En éste modelo inicial el usuario especifica en su consulta una expresión booleana formada por una serie de términos ligados mediante operadores booleanos (comúnmente AND, OR y NOT). Dada la expresión lógica de la consulta, el sistema devolverá

aquellos documentos que la satisfacen y que conformarán el conjunto de documentos relevantes. El sistema simplemente particiona los documentos de la colección en dos conjuntos, aquéllos que cumplen la condición especificada (relevantes), y aquellos que no la cumplen (no relevantes), sin ordenación interna alguna, de forma similar a lo que ocurriría con una base de datos tradicional.

Las desventajas importantes asociada al modelo está dado por la dificultad que conlleva la formalización de la necesidad de información del usuario en forma de expresión booleana, sobre todo cuando se trata de usuarios inexpertos y de necesidades complejas. Otro de los grandes inconvenientes del modelo booleano viene dado por su propia naturaleza, de carácter binario. De ésta forma, dada una consulta, un documento simplemente es o no relevante dependiendo de si cumple la condición expresada por la consulta. Por lo tanto, no existe ni el concepto de gradación de relevancia. Al no permitir correspondencias parciales, el sistema podría no devolver documentos que, aun siendo relevantes, no verificasen por completo la condición estipulada. Del mismo, modo, todos los términos de la consulta tienen la misma importancia, cuando es lógico pensar que la semántica de un texto dado se concentre en mayor grado en ciertos términos, por lo tanto al no existir ninguna ordenación por relevancia, el usuario se ve obligado a examinar la totalidad del conjunto de documentos devuelto (Rodriguez, 2011).

b. Modelo Vectorial

Este modelo plantea un marco formal diferente en que se permite tanto la asignación de correspondencias parciales, como la existencia de grados de relevancia en base a los pesos de los términos en consultas y documentos. En éste modelo las consultas y documentos son representados mediante vectores dentro de un espacio multidimensional de finido por los propios términos, de tal forma que cada uno de los términos diferentes del sistema, definen una dimensión. Desde un punto de vista geométrico, si ambos vectores, consulta y documento, están próximos, es factible asumir que el documento es similar a la consulta, el documento es posiblemente relevante (Rodriguez, 2011).

Tolosa (2009), conceptualmente, este modelo utiliza una matriz documento–término que contiene el vocabulario de la colección de referencia y los documentos existentes. En la intersección de un término y un documento se almacena un valor numérico de importancia del término t en el documento d; tal valor representa su poder de discriminación. Así, cada documento puede ser visto como un vector que pertenece a un espacio n-dimensional, donde es la cantidad de términos que componen el vocabulario de la colección. En teoría, los documentos que contengan términos similares estarán a muy poca distancia entre sí sobre tal espacio. De

igual forma se trata a la consulta, es un documento más y se la mapea sobre el espacio de documentos. Luego, a partir de una consulta dada es posible devolver una lista de documentos ordenados por distancia (los más relevantes primeros). Para calcular la semejanza entre el vector consulta y los vectores que representan los documentos se utilizan diferentes fórmulas de distancia, siendo la más común la del coseno. El siguiente ejemplo se muestra un documento y una consulta.

Documento:

"Puno ha sido nominada para la realización del X Congreso Americano de Epidemiología en Zonas de Desastre. El evento se realizará..."

Consulta:

"puno congreso epidemiología"

CUADRO 1

Matriz de relación Documento - Término en un modelo vectorial de recuperación de la información, con pesos normalizados entre 0 y 1

	argentina	...	congreso	...	epidemiologia
d1	0.5		0.3		0.2
...					
d2					
consulta	0.4		0.3		0.3

Fuente: (Tolosa, 2009)

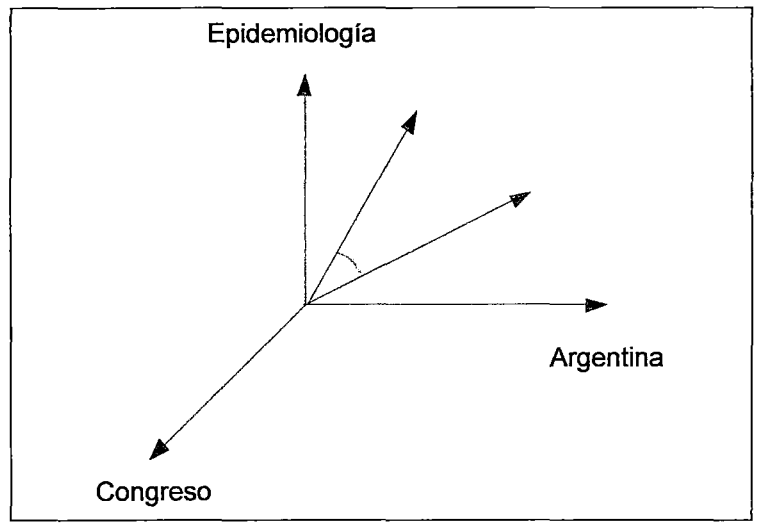


Figura 2. Representación vectorial de relación de palabras

Fuente: (Tolosa, 2009)

c. Modelo probabilístico

Fue propuesto por Robertson y Spark-Jones, El modelo probabilístico formaliza el proceso de recuperación en términos de teoría de probabilidades. A partir de una expresión de consulta se puede dividir una colección de N documentos en cuatro subconjuntos distintos: REL conjunto de documentos relevantes, REC conjunto de documentos recuperados, RR conjunto de documentos relevantes recuperados y NN el conjunto de documentos no relevantes no recuperados. El resultado ideal de a una consulta se da cuando el conjunto REL es igual REC. Como resulta difícil lograrlo en primera intención, el usuario genera una descripción probabilística del conjunto REL y a través de sucesivas interacciones con el SRI se trata de mejorar la performance de recuperación. Dado que una recuperación no es inmediata dado que involucra varias

interacciones con el usuario y que estudios han demostrado que su performance es inferior al modelo vectorial, su uso es bastante limitado (Rodriguez, 2011).

Según el principio de orden de probabilidades, el rendimiento óptimo de un sistema se consigue cuando los documentos son ordenados de acuerdo a sus probabilidades de relevancia. El modelo parte de las siguientes suposiciones:

a. Todo documento es, bien relevante, bien no relevante para la consulta.
b. El hecho de juzgar un documento dado como relevante o no relevante no aporta información alguna sobre la posible relevancia o no relevancia de otros documentos.

Existen múltiples medidas de similaridad utilizadas en éste modelo, siendo el más conocido el sistema Okapi, cuyo esquema de pesos se encuentra entre los más efectivos y, junto al vectorial if-idf es punto de referencia para el desarrollo y evaluación de nuevos modelos y nuevos esquemas de pesos (Rodriguez, 2011).

d. Modelos para documentos estructurados

Para Baeza-Yates (1992), los modelos clásicos responden a consultas, buscando sobre una estructura de datos que representa el

contenido de los documentos de una colección, únicamente como listas de términos significativos. Un modelo de recuperación de documentos estructurados utiliza la estructura de los mismos a los efectos de mejorar la performance y brindar servicios alternativos al usuario (por ejemplo, uso de memoria visual, recuperación de elementos multimedia, mayor precisión sobre el ámbito de la consulta y demás). La estructura de los documentos a indexar está dada por marcas o etiquetas, siendo los estándares más utilizados el SGML (Standard General Markup Language), el HTML(HyperText Markup Language), el PDF (Portable Document Format), el XML (eXtensible Markup Language) y LATEX.

Al poseer la descripción de parte de la estructura de un documento es posible generar un grafo sobre el que se navegue y se respondan consultas de distinto tipo, por ejemplo:

a. Por estructura: ¿Cuáles son las secciones del segundo capítulo?
b. Por metadatos o campos: Documentos de "Editorial UNL" editados en 2000
c. Por contenido: Término "agua" en títulos de secciones
d. Por elementos multimedia: Imágenes cercanas a párrafos que contengan Bush

Existen dos modelos en esta categoría "nodos proximales" y "listas no superpuestas". Ambos modelos se basan en almacenar las ocurrencias de los términos a indexar en estructuras de datos diferentes, según aparezcan en algún elemento de estructura (región) o en otro como capítulos, secciones, subsecciones y demás. En general, las regiones de una misma estructura de datos no poseen superposición, pero regiones en diferentes estructuras sí se pueden superponer. Es necesario mencionar que algunos motores de búsqueda de Internet ya utilizan ciertos elementos de la estructura de un documento, por ejemplo los títulos, a efectos de realizar tareas de ranking, resumen automático, clasificación y otras.

La expansión de estos lenguajes de demarcación, especialmente en servicios sobre Internet, hace que se generen y publiquen cada vez más documentos semiestructurados. Es necesario entonces, desarrollar técnicas que aprovechen el valor agregado de los nuevos documentos. Si bien, en la actualidad éstas no se encuentran tan desarrolladas como los modelos tradicionales, consideramos su evolución como una cuestión importante en el área de RI, especialmente a partir de investigaciones con enfoques diferentes que abordan la problemática (Baeza-Yates, 1992).

2.2.4. Búsqueda semántica

Tolosa (2009), indica que en la comunidad científica, la búsqueda semántica se puede entender de las siguientes tres formas. Primero, búsqueda en la web semántica. Es una variante de la recuperación de información que saca provecho de las características de la web semántica para enriquecer los resultados. Segundo, Indexación y búsqueda con información semántica. Uso de información semántica en la fase de indexación para mejorar los resultados de búsqueda y finalmente, finalmente, búsqueda en lenguaje natural. Recuperación de información mediante consultas realizadas en lenguaje natural en lugar el uso tradicional de palabras clave.

a. Búsqueda en la web semántica

La búsqueda semántica es la nueva generación de algoritmos de búsqueda en la web semántica. Esta definición tiene como premisa que, en la nueva web semántica, una búsqueda no puede limitarse a un simple problema de recuperación de documentos. Hay que sacar provecho de las nuevas interconexiones para dotar de un valor añadido a los resultados. De hecho, poco a poco se va vislumbrando esta nueva tendencia en los buscadores comerciales, los cuales incorporan publicidad orientada, enlaces a elementos multimedia de otras webs por ejemplo fotos en Flickr, vídeos en YouTube y otros como mapas con información geográfica de

localización. Otra de las metas que persiguen los algoritmos de búsqueda orientados a la web semántica consiste en modelar la intención del usuario y facilitarle las tareas.

b. Indexación y búsqueda con información semántica

Los buscadores tradicionales intentan localizar los términos de la consulta dentro de la colección de documentos que tienen indexada; priorizan que las palabras aparezcan en el mismo orden y luego buscan en otros documentos pero flexibilizando la colocación de las mismas. En los _últimos años se han introducido técnicas simples de procesamiento del lenguaje natural, por ejemplo las de corrección ortográfica, que detectan errores sencillos y realizan sugerencias. Otro caso es el de las reglas morfológicas que utilizan los buscadores para devolver los mismos resultados con términos en singular y en plural, diminutivos, etc... Con estas nuevas técnicas de procesamiento de lenguaje natural se abren nuevas posibilidades para mejorar los resultados. Como ya vimos en la introducción a la recuperación de la información existen situaciones en las que un análisis más profundo de la consulta y un conocimiento semántico de los contenidos indexados, permitiría la resolución de distintos problemas como los términos polisémicos, sinonimia, expresiones equivalentes, entre otros.

2.2.5. Medidas y cálculos de similitud

Rodríguez (2011), similitud se define como la proximidad que existe entre unidades lingüísticas y se puede referir a ella utilizando diferentes términos, tales como: similitud, proximidad, afinidad, distancia, etc. Sin embargo hay que ser cuidadosos cuanto se utiliza éste término, ya que es diferente al término distancia de acepción matemática. Consideremos un ejemplo para medir la distancia, definida como medida de divergencia, utilizando el método del espacio vectorial, usado con frecuencia para representar el contenido de los documentos. Un documento d_i de una colección D se representa como un punto en un espacio R^N, siendo N el número de términos relevantes de colección D. El componente k del vector d_i mide la relevancia del término t_K para caracterizar el documento d_i (existen muchas formas de ponderar la relevancia de los términos que pueden utilizarse, el más frecuente es it*idf) Siendo R^N un espacio euclídeo, es posible entonces utilizar la distancia euclidea entre los puntos R^N como medida de distancia entre los documentos.

Para el cálculo de la similitud, se puede considerar tres tipos de información: Contenido, contexto, fuentes de conocimiento externas. El contenido se refiere a la información asociada directamente a la unidad lingüística como una pieza aislada de información. Esta información suele representarse como una lista de pares atributo-valor. Dependiendo del tipo de unidad se utilizan diferentes rasgos o propiedades, tales

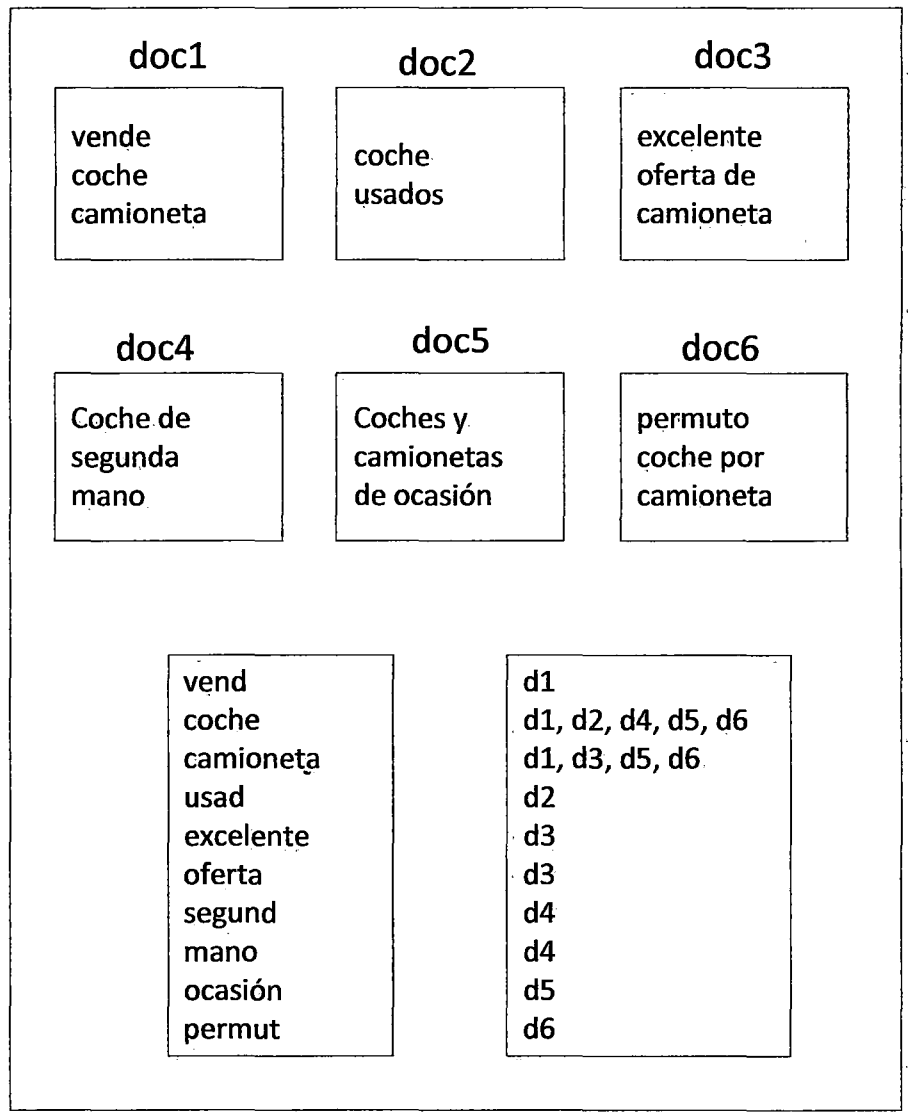
como palabras, unidades semánticas, unidades sintácticas, clusters. El contexto se refiere a la información asociada a la situación en la cual debe calcularse la similitud. Se suelen utilizar fuentes de conocimiento externas para calcular la similitud entre dos unidades lingüísticas, entre las más usadas son lexicones, diccionarios monolingües y bilingües, ontologías y corpus (Rodríguez, 2011).

2.2.6. Métodos de indexación.

Las técnicas de búsqueda conocidas generalmente son secuenciales, es decir, recorren los documentos que forman la base de datos textual secuencialmente buscando las ocurrencias del elemento a localizar. Obviamente, la aplicación de éste tipo de búsqueda es apropiada cuando el texto es pequeño (de pocos megabytes), en otros casos debe recurrirse a la utilización de técnicas de indexación que agilicen las búsquedas, como la técnica de índices invertidos. Un índice invertido es un mecanismo orientado a palabras para indexación de documentos. Es la estructura más elemental para recuperación de palabras. Está formado por dos elementos: el vocabulario (conjunto de términos distintos del texto) y las listas de ocurrencias (para cada término, la lista de documentos donde éste aparece). El espacio requerido para almacenar el vocabulario no es grande. De acuerdo a la ley de Heaps, el vocabulario crece $O(n^B)$ donde B depende del texto, estando entre 0,4 y 0,6 en la práctica (Grossman, 1998).

CUADRO N° 2

Ejemplo de un índice invertido



Fuente: (Baeza-Yates, 1992)

2.3. MARCO CONCEPTUAL

a. Bigrama

Secuencia de dos palabras que aparecen juntas en los documentos (Martínez, 2004).

b. Clasificación.

Rotulación automática de documentos de un corpus en base a clases previamente definidas (Martínez, 2004).

c. Corpus

Serie de documentos referidos a un dominio en particular (Tolosa, 2009).

d. Crawlers

Conocidos también como robots o arañas. Son programas que se caracterizan por explorar de forma automática los sitios Web, analizar su información y hacer exploraciones de las páginas que encuentran referenciadas mediante una URL (Baeza-Yatez, 1992).

e. Desambiguación.

Eliminación de ambigüedades (Martínez, 2004).

f. Filtrado y Ruteo.

Área que permite la definición de perfiles de necesidades de información por parte de usuarios y ante el ingreso de nuevos documentos al SRI, se analiza y se reenvía a quienes estimen que van a ser relevantes (Blair, 2001).

g. Lematización.

Consiste en llevar palabras a su raíz, por ejemplo: carro, carrito, carrazo, etc. Se reduce a carro. Con esto se logra que las palabras que las palabras que provienen de una misma raíz sean tratadas como iguales (Lluis, 2009).

h. Lexema.

Unidad léxica abstracta que no puede descomponerse en otras menores aunque sí combinarse con otras para formar compuestos, y que posee un significado definible por el diccionario, no por la gramática. Por ejemplo: *fácil* es el lexema básico de *facilidad*, *facilitar*, *fácilmente* (Gelbuck, 2010).

i. Modelos de Recuperación.

La tarea de la recuperación de la información puede ser modelada desde distintos enfoques, por ejemplo la estadística, el álgebra de Boole, el álgebra de vectores, la lógica difusa, el procesamiento del lenguaje natural y otros (Berners-Lee, 2001).

j. Morfema.

Palabra de la terminología gramatical moderna con que se designan los elementos lingüísticos que se incorporan a las palabras con significado fijo y forma variable. Morfema puede ser una palabra, prefijo, infijo, sufijo, desinencia, etc. (Rodriguez, 2011).

k. Ontología.

La palabra ontología tiene su origen en la filosofía. Aristóteles la definió como la ciencia del ser o una explicación sistemática de la existencia. En el campo de la Inteligencia Artificial, una ontología define los términos básicos y las relaciones que comprende el vocabulario de un área temática, así como las reglas para combinar términos y relaciones para definir extensiones para el vocabulario (Cimiano, 2006).

l. Parser.

Es un programa que analiza el texto para determinar su estructura sintáctica. (Martinez, 2004).

m. Similitud.

Relación de proximidad entre las unidades linguisticas (y sus opuestas). Similitud, proximidad, afinidad, distancia (Rodríguez, 2011).

n. Subcategorización.

Clasificación rigurosa, sistemática y jerárquica, según rasgos de las unidades léxicas de la lengua, para describir cuántos y de qué tipo son los elementos con los que combina para hacer oraciones completas. Cuando se dice que subcategoriza determinada categoría gramatical, significa que combina con ella (Tolosa, 2009).

o. Sumarización.

Área que entiende sobre técnicas de extracción de aquellas partes (palabras, frases, oraciones, párrafos) que contienen la semántica que determina la esencia de un documento (Tolosa, 2009).

CAPITULO III

METODOLOGIA

3.1. METODOLOGIA DE LA INVESTIGACION

3.1.1. Metodología para el desarrollo de un sistema de recuperación semántica de la información.

Para determinar la similitud semántica de palabras o términos, se hace una comparación de los contextos de ocurrencia de éstas palabras. Para determinar la lejanía o cercanía semántica entre un par de vocablos, tomando en cuenta su distribución con el resto de elementos del lenguaje, se utiliza la estructura conocida como Modelo de Espacio de Palabras (WSM por sus siglas en inglés Word Space Model) el cual determina la afinidad semántica entre dos vocablos usando un espacio multidimensional.

Pre procesamiento. El pre procesamiento se realiza a través de varias etapas y que es peculiar a cada problema. Se inicia con la eliminación de etiquetas XHTML, eliminación de símbolos especiales, eliminación de palabras sin significado, para luego uniformizarlos a palabras minúsculas y sin acentos ortográficos.

Indexación. La indexación es una operación que tiene por función la identificación de los conceptos que representan el contenido de un documento y la traducción de los mismos a una forma que computacionalmente sea manejable. Para la presente investigación se utiliza el método de indexación no lingüístico, es decir se utiliza técnicas estadísticas para el análisis de frecuencias y el cálculo de pesos de los términos.

Análisis lexicográfico. En esta etapa se extraen las palabras y se normalizan.

Extracción de datos. Se buscan en el corpus los bigramas (secuencias de dos palabras) en los que aparecen los diferentes términos.

Filtrado de datos. Se eliminan las unidades más comunes y que por lo tanto son escasamente informativas.

Stemming. Se reducen palabras morfológicamente parecidas a una forma base, con la finalidad de aumentar la eficiencia del sistema de recuperación de información.

Construcción de vectores. Con las unidades restantes se pasa a construir una estructura de datos en la que cada sustantivo queda asociado a una lista de palabras con las que comparte los bigramas, lo cual se representa como un conjunto o vector.

Ponderación de términos. Para los cálculos de similitud se determina la ponderación de cada término brindando peso o valor. El método de ponderación utilizado es TF * IDF (Term Frecuency, Inverse Document Frecuency).

Similitud de vectores. Los vectores se utilizan para comparar los sustantivos entre sí y agrupar aquellos que resultan más similares, calculando esta similitud como la cantidad de palabras que tienen en común. Para la determinación de la similitud entre dos vectores se utilizó la medida del coseno del ángulo que forman ambos vectores.

3.1.1. Implementar un algoritmo que permita calcular la distancia entre vectores y obtener un conjunto de términos con cierta similitud semántica.

Diagramas de entidad – relación. Es utilizado para representar gráficamente la estructura lógica de la de base de datos que soporta el sistema de recuperación semántica de la información.

3.2. POBLACIÓN Y MUESTRA DE LA INVESTIGACION

3.2.1. Población

La población para esta investigación está constituida por las páginas Web existentes en Internet.

3.2.2. Muestra

La muestra se obtuvo mediante un muestreo no probabilístico, tomando los bigramas en español proporcionadas por el Corpus de Google

3.3. TECNICAS E INSTRUMENTOS DE RECOLECCION DE DATOS

3.3.1. Técnicas de recopilación de datos

Las técnicas conceptuales fueron utilizadas para la abstracción, análisis, sistematización y síntesis del planteamiento del modelo a seguir para la elaboración de un modelo de recuperación de la información. Para la implementación del algoritmo, la técnica utilizada para la

recolección de información fue la observación, específicamente la observación heurística y los instrumentos de recolección de datos fueron la guía de observación y el cuestionario.

3.3.2. Método de tratamiento de datos

En la investigación se realizó el tratamiento de la siguiente manera:

- Recopilación y procesamiento de datos.
- Análisis de datos
- Interpretación de los datos
- Validación de la hipótesis

3.4. MATERIAL EXPERIMENTAL

El Corpus de Google, constituido por los bigramas y sus respectivas frecuencias utilizadas en las consultas los mismos que están almacenados en treinta y dos (32) archivos con diez millones (10 000 000) de registros cada uno. El Sistema Gestor de Base de Datos Oracle 11g, Modelo de Espacio de Palabras (Word Space Model WSM). Allfusion ERwin Data Modeler en su versión 7.1, son los materiales utilizados para la presente investigación.

CAPITULO IV

RESULTADOS Y DISCUSION

4.1. DEFINIR UNA METODOLOGIA PARA EL DESARROLLO DEL SISTEMA DE RECUPERACION SEMANTICA DE LA INFORMACION

4.1.1. Metodología para el desarrollo del sistema de recuperación semántica de la información

La presente investigación pretende agrupar palabras en categorías semánticas utilizando una medida de similitud distribucional, para ello se parte de la premisa que existen palabras que muestran una tendencia a aparecer con frecuencia en el corpus junto a otras en un contexto determinado; éstas unidades o palabras que coocurren con frecuencia con cada unidad analizada contienen importante información semántica sobre ellas, ya que ocurrirán casi siempre en contextos similares. Ello nos permite compararlos entre sí y agruparlas en función de la cantidad de

atributos compartidos, y su similitud semántica entre unidades léxicas puede detectarse a través de la búsqueda de coincidencias en el contexto lingüístico.

Entonces, para recuperar información con una similitud semántica, en un contexto dado, se parte de la premisa que si existen palabras que coocurren en un contexto dado, éstas tienen una relación semántica, y son semejantes por que ocurren en contextos similares.

La metodología a seguir para la presente tesis, se puede apreciar con mayor detalle en la siguiente figura.

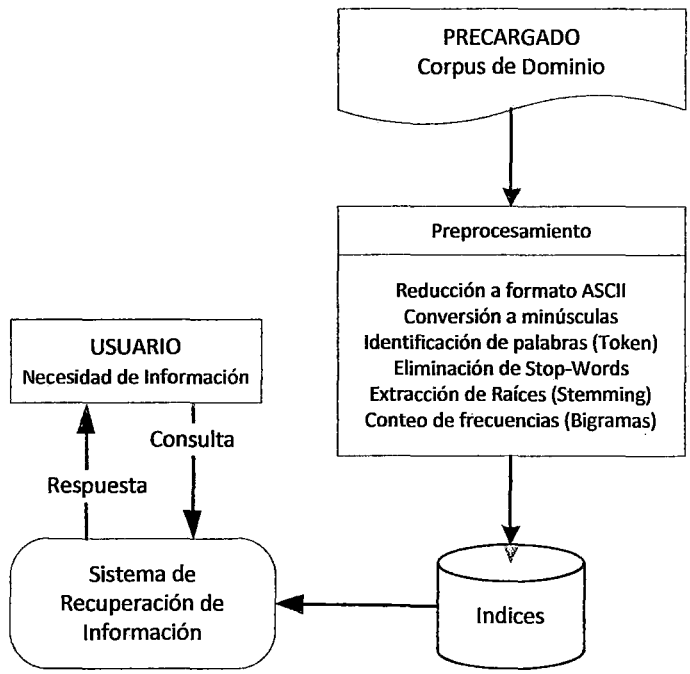


Figura 3. Metodología para la recuperación semántica de la información

Fuente: Elaboración propia

4.1.2. Entrada al sistema el corpus de un dominio.

El corpus de un dominio, es una serie de documentos referentes a un dominio en particular, a este corpus se le hizo un pre-procesamiento. El corpus utilizado es el proporcionado por Google, el mismo que se puede ver en el anexo A.

4.1.3. Reducción a formato ASCII.

En esta etapa se eliminan las etiquetas HTML, caracteres de formato, estilo, signos de puntuación, se eliminan también los puntos al final de una palabra, por ejemplo cuando en el texto se encuentre "sistema." solo se toma en cuenta "sistema". Se eliminan también los acentos ortográficos, la palabra "Perú" queda como "Peru" y las abreviaciones como por ejemplo "O.N.U." queda como "ONU". Dejando solamente las palabras en cada documento.

4.1.4. Conversión a minúsculas.

Para un mejor análisis, no se toma en cuenta las letras mayúsculas, para ello, todo el texto ha sido cambiado a minúsculas y casos por ejemplo "Sistema" ha sido transformado en "sistema" únicamente.

4.1.5. Identificación de palabras (Tokens).

A continuación se realiza un proceso de tokenización, es decir, se identifican las palabras del texto, conjunto de caracteres delimitados por un espacio en blanco. Luego de ello se implementa una relación con cada palabra identificada de manera única con un id. Así una entrada como "cuzco la capital del imperio de los incas" tiene como salida:

Cuadro 3.

Ejemplo de tokenización de palabras

Tokens
1. cuzco
2. la
3. capital
4. del
5. imperio
6. de
7. los
8. incas

Fuente: Elaboración propia

4.1.6. Eliminación de palabras vacías (Stop-Words).

Todas las palabras del texto representa un término, sin embargo algunos de ellos conocidos como los stop-words o palabras vacías y que son de uso frecuente en el idioma, tales como los artículos, pronombres, preposiciones, adverbios, etcétera, no incluyen ni aportan información a

dominio alguno, por ello dichos términos quedan eliminados en este proceso. Para ellos se utiliza un diccionario de palabras vacías como se muestra a continuación y en el Anexo B se presenta una versión más detallada.

CUADRO 4.

Fragmento de lista de Stopwords

Stopwords	Stopwrds
el	un
la	una
lo	unas
las	unos
los	uno
su	sobre
aqui	todo
mio	tambien
tuyo	tras
ellos	otro
ellas	algún
nos	alguno
nosotros	alguna
vosotros	algunos

Fuente: Elaboración propia

Del ejemplo anterior ""cuzco la capital del imperio de los incas", quedaría únicamente:

CUADRO 5.

Lista de palabras después de la eliminación de Stopwords

Palabras sin stopwords
1. cuzco
2. capital
3. imperio
4. incas

Fuente: Elaboración propia

Ejemplo: "puno ciudad a orillas del titicaca el lago sagrado de los incas", quedaría únicamente.

CUADRO 6.

Lista de palabras después de la eliminación de Stopwords

Palabras sin stopwords
1. puno
2. ciudad
3. orillas
4. titicaca
4. lago
5. sagrado
6. incas

Fuente: Elaboración propia

4.1.7. Extracción de raíces (Stemming).

En esta etapa se realiza un proceso de normalización lingüística para lo cual se aplicó extractores de raíces con la finalidad de reducir cada palabra a su raíz o lema eliminando prefijos, sufijos y terminaciones verbales, de este modo palabras que literalmente son diferentes, pero tienen una raíz común, pueden ser consideradas como un solo término en base a su raíz. Como se muestra a continuación, diferentes términos como "asecha", "asechaba", "asechabais", etcétera queda como "asechar", con este proceso se reducen la dimensión del espacio de términos y se mejora la formulación de consultas.

CUADRO 7.

Fragmento del proceso de stemming de palabras

TERMINO	STEM
asecha	asechar
asechaba	asechar
asechabais	asechar
asechaban	asechar
asechabas	asechar
asechad	asechar
asechada	asechar
asechadas	asechar
asechado	asechar
asechados	asechar
asechamos	asechar
asechan	asechar
asechando	asechar
asechar	asechar
asechara	asechar
asecharais	asechar
asecharan	asechar
asecharas	asechar
asechare	asechar
asechareis	asechar
asecharemos	asechar
asecharen	asechar
asechares	asechar
asecharon	asechar

Fuente: Elaboración propia

Para una computación rápida, se utilizó una relación de dos atributos en donde cada registro corresponde a una variante morfológica asociada a su lema o raíz, conocido como stem.

El proceso de stemming o lematización permite tener índices de menor tamaño y una mayor cantidad de respuestas a una consulta dada, debido a que ahora el aplicarse lematización al corpus y a la consulta se recuperan documentos que contengan todas las variantes morfológicas de los términos contenidos en la consulta, aunque ésta última consecuencia puede verse como una desventaja bajo ciertas ocasiones, debido a que aumenta la exhaustividad y disminuye la precisión.

4.1.8. Selección de términos índice.

Los términos resultantes de las transformaciones de texto previas son adoptados como términos índice, en esta etapa se filtraron los lemas en español.

4.1.9. Conteo de frecuencias

El proceso se enriquece con el proceso de lematización, el cual consiste en el conteo de frecuencias en base a los términos que en su información gramatical tienen el mismo lema, de modo que palabras como *programa*, *programar*, *programando*, *programare*, etc. Al momento de hacer el conteo de frecuencias acumulan a la palabra *programa*.

Entonces, sea $\{t_1, t_2, \ldots, t_k\}$ el conjunto de términos y $\{d_1, d_2, \ldots, d_N\}$ el conjunto de documentos, un documento d_i, esta modelado por: $d_i \rightarrow \vec{d}_i = (w(t_1, d_i), \ldots, w(t_k, d_i))$, donde: $w(t_r, d_i)$ es el peso del término t_r en el documento d_i. Al finalizar esta fase tenemos una tabla conteniendo, para cada palabra lematizada su frecuencia por documento (term frecuency)

CUADRO 8.

Fragmento de conteo de frecuencias de palabras en los documentos

Palabra	Documento	Frecuencia
software	7	4
software	9	3
software	15	5
software	21	7
software	24	8
software	25	2
software	30	1
software	43	1
software	44	4
software	47	5
software	49	7
software	50	9
software	55	5
software	57	15
software	58	20
software	59	3
software	70	4
software	80	5
software	81	6
software	90	9
software	91	12
software	95	11

Fuente: Elaboración propia

Para recuperar información en un contexto dado, se parte de la premisa que si existen palabras en un contexto dado, éstas tienen una relación semántica, y son semejantes por que ocurren en contextos similares. Para lo cual se busca en el corpus los bigramas (secuencias de dos palabras) en los que aparecen sustantivos de un contexto determinado. Los términos más comunes y que por tanto resultan escasamente informativas (palabras como mucho, buen, como, etc.) se eliminan del análisis por medio del coeficiente *w* definido en la ecuación:

$$w(i) = \frac{f_o(i)}{f_e(i) + 1}$$

Donde f_o es la frecuencia observada y f_e la frecuencia esperada.

Esta última representa la probabilidad de una palabra de aparecer en un contexto cualquiera, y se calcula registrando la frecuencia de tal unidad en un corpus de referencia, en este caso específico el Corpus de Google.

4.1.10. Indexación

Para agilizar la búsqueda se construyó un índice invertido, el cual está formado por dos elementos: el vocabulario (conjunto de términos distintos del texto) y las listas de ocurrencias (para cada término, la lista de documentos donde aparece).

La calidad de los sistemas de recuperación semántica de la información, se mejora significativamente con las técnicas de tokenización y eliminación de palabras gramaticales (Renteria, 2009), ya que sin este proceso la base de datos de búsqueda serían muy extensa y se realizarían con palabras que no representaría la necesidad del usuario. La disminución del coste computacional, se ha logrado con técnicas centradas en stemming (Hechevarría, 2006), es decir la recuperación de variantes morfológicas de los términos a una forma léxica común, extrayendo solamente la raíz, sin embargo el problema persiste ya que es necesario encontrar un método de análisis lingüístico más profundo y que consiga una representación óptima entre los documentos y la necesidad de información del usuario.

4.2. IMPLEMENTAR UN ALGORITMO QUE PERMITA CALCULAR LA DISTANCIA ENTRE DOS VECTORES Y OBTENER UN CONJUNTO DE TÉRMINOS CON CIERTA SIMILITUD SEMÁNTICA.

4.2.1. Diseño físico del sistema

Se utilizan las siguientes relaciones para implementar el sistema para la recuperación semántica de la información.

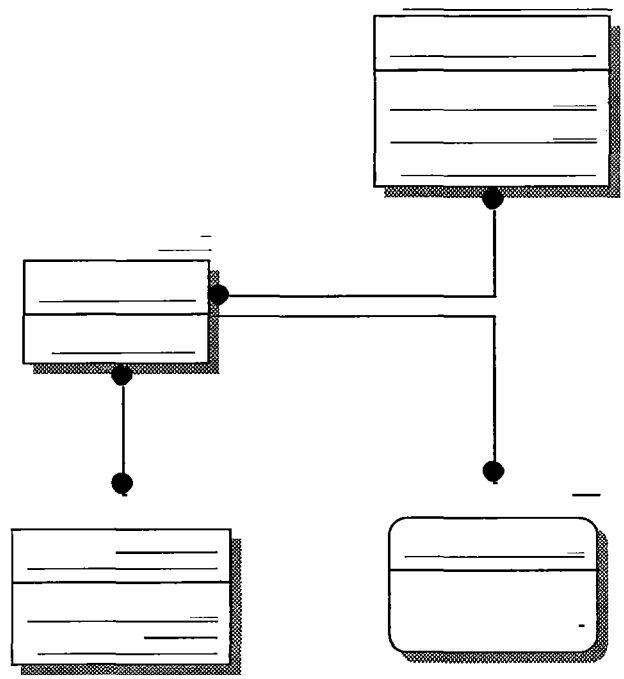


Figura 4. Diagrama entidad - relación del sistema de recuperación semántica.

Fuente: Elaboración propia

4.2.2. Calculo de similitud

La medida de similaridad utilizada en ésta tesis, es la similaridad por coseno del ángulo que forman el vector documento y el vector consulta. Para ello primero se evalúo la frecuencia de aparición de la palabra. Se debe tener en cuenta que el peso o valor de importancia de un término en un documento es inversamente proporcional a su frecuencia en el corpus o colección de documentos, y a su vez es directamente proporcional a su frecuencia en el documento.

Para el cálculo se consideró la frecuencia absoluta de aparición de un término en un documento (***tf***), que es un factor que precisa de una

corrección, porque la importancia de un término en función de su distribución puede llegar a ser desmesurada (por ejemplo, una frecuencia de 2 es 200% más importante que una frecuencia de 1, y la diferencia aritmética es sólo de una unidad). Se consideró también la capacidad de discriminación de un término frente a otro, ya que aquellos términos que aparecen en todos los documentos discriminan poco o nada a la hora de representación del contenido de un documento. Para medir este valor de discriminación se utilizó la medida frecuencia inversa de un documento (***idf***).

El cálculo utilizado para el cálculo de pesos de los términos en los documentos se basa en TF * IDF, y se define como:

$$W_{ij} = f_{ij} * log\frac{N}{n_i}$$

Donde:

N = número de términos distintos en la colección de documentos

n = número de ocurrencias del término en un documento

$$f_{ij} = \frac{freq_{ij}}{maxfreq_{iji}}$$

Para el cálculo de la similitud entre dos vectores, se define como:

$$\cos(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{dy}}{\sqrt{\sum_{j=1}^{t}(w_{qj})^2 * \sum_{j=1}^{t}(w_{dy})^2}}$$

Donde:

$Di = w_{di1}, w_{di2}, w_{di3}, \ldots, w_{dit}$ Vector que representa el documento

$Q = w_{q1}, w_{q2}, w_{q3}, \ldots, w_{qt}$ Vector que representa la consulta

Una suposición que se realiza al adoptar esta representación espacial de t-dimensiones, es que los términos ocurren de manera independiente dentro de los documentos. Esto se expresa en el hecho que los vectores que los representan (cada uno de los cuales define un eje) son ortogonales y por tanto, el coseno del ángulo que forman dichos vectores es igual a cero, ya que $\alpha = 90°$.

El modelo vectorial aporta varios beneficios, que se puede resumir de la siguiente manera:

- El esquema de pesos mejora las prestaciones de la recuperación. Aquí se tiene la flexibilidad de poder aplicar diferentes esquemas de cálculo de los pesos.
- Para mejorar los resultados, se pueden considerar aplicar búsquedas aproximadas.

- La medida de similitud, proporciona un método de ranking de los resultados.
- Mediante esta representación se puede medir la similitud entre diferentes objetos tales como documentos y consultas, documentos y documentos, oraciones y consultas, etc.

Para el diseño de un Sistema de Recuperación de Información existen muchas posibilidades y variantes (Vilares, 2005), de los cuales el modelo vectorial con esquema de pesos tf- idf (Baeza-Yates, 1999) en donde se penalizan los términos que más aparecen en la mayoría de los documentos de la colección se ha consolidado y es el más utilizado, sin embargo los estudios continúan tanto en los modelos vectorial y probabilístico para mejorar los procesos de pesos, indexación y de búsqueda.

4.3. EVALUAR EL RESULTADO DE LA RECUPERACION SEMANTICA DE LA INFORMACION UTILIZANDO LA SIMILITUD DISTRIBUCIONAL

Al ejecutar las consultas al sistema se obtuvieron los siguientes resultados:

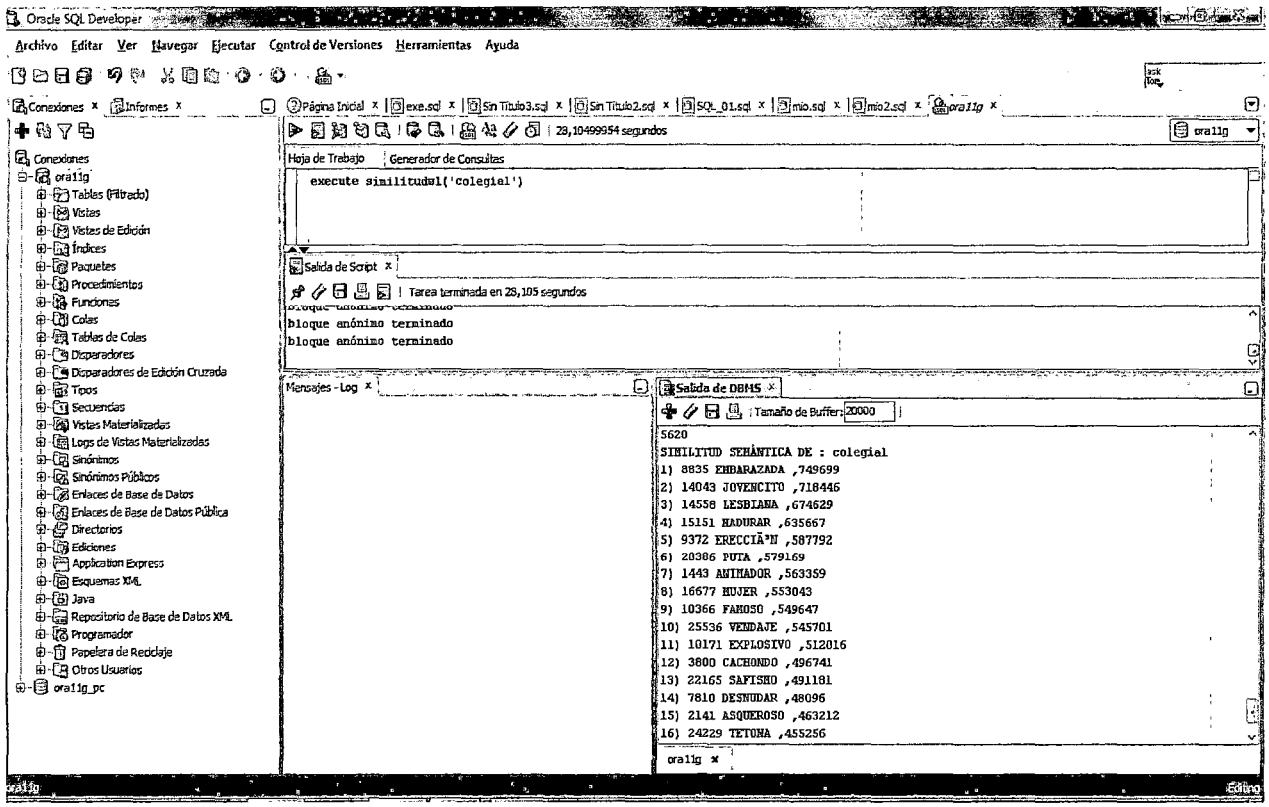


Figura 5. Resultado de búsqueda de la palabra colegial

Fuente: Elaboración propia

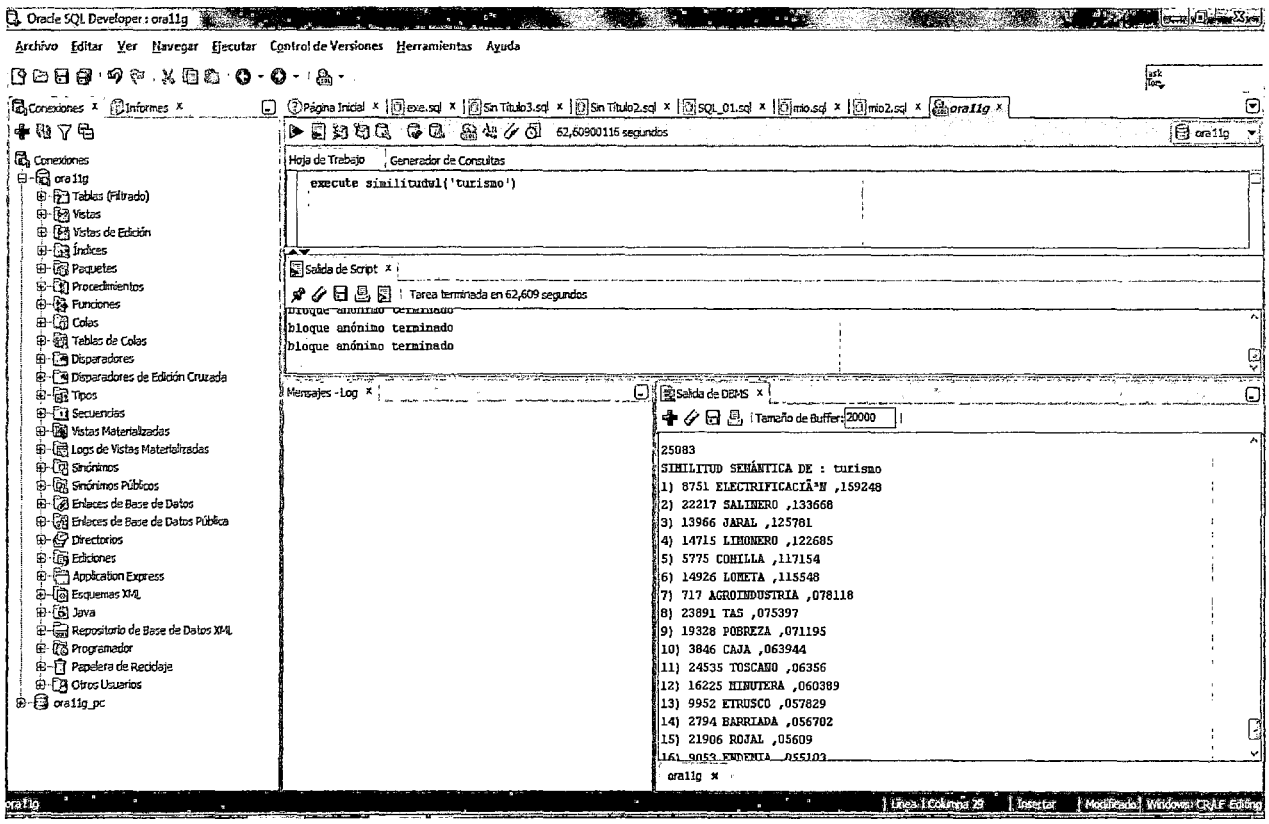


Figura 6. Resultado de búsqueda de la palabra turismo

Fuente: Elaboración propia

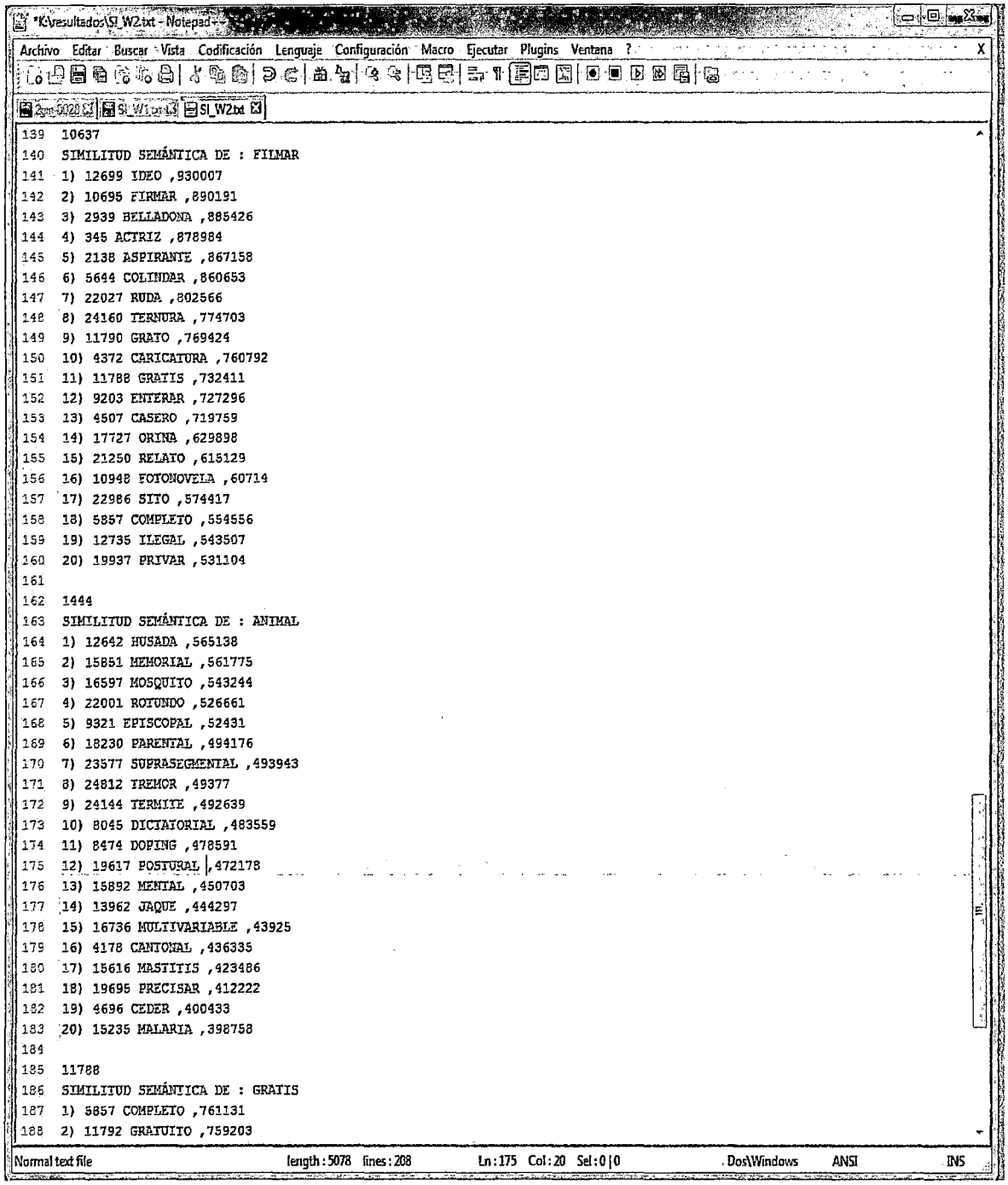


Figura 7. Fragmento de lista de resultados de pruebas de búsqueda

Fuente: Elaboración propia

En la figura 15 se presenta algunos resultados de búsqueda de palabras y sus resultados de acuerdo a la similitud que tienen con ellas. En el primer caso por ejemplo, se ha buscado la palabra colegial y podemos ver que las palabras que

más similitud tienen es embarazada con un 74% , la palabra jovencito con 71%, en el segundo ejemplo la palabra buscada fue turismo y las palabras con mayor similitud son electrificación con un 15%, la palabra salinero con un 13%. También se realizó la búsqueda con la palabra filmar, los resultados fueron que ideo tiene una similitud del 93% y firmar tiene un 89 %. La palabra animal tiene una similitud de 56% con la palabra husada y un 54% con la palabra mosquito.

El rendimiento de un sistema de recuperación de la información depende del tipo de Corpus usado en su construcción (Tejada, 2009), y del tipo de procesamiento del lenguaje natural. La función de similitud tiende a lograr resultados en mayor o menor grado bajo la presencia de ciertas variaciones textuales, conocidos también como situaciones problemáticas (Vilares, 2005) como se puede apreciar en algunos resultados obtenidos por el sistema.

CONCLUSIONES

PRIMERA. Recuperar información semántica se optimiza al aplicar métodos que evalúan los vectores de coocurrencia de dos palabras que tienen cierto grado de similitud, calculando la frecuencia de ocurrencia de los mismos en un contexto determinado.

SEGUNDA. El modelo de espacio de palabras permite determinar la lejanía o cercanía de un par de términos, usando un espacio multidimensional, tomando en cuenta su distribución con el resto de términos del lenguaje y cuyo número de dimensiones, depende del número de vocablos diferentes encontrados en el corpus de Google utilizado.

TERCERA. El algoritmo basado en la medida de similitud por coseno del ángulo que forman entre el vector de las palabras y el vector consulta permiten ponderar y obtener términos con cierta similaridad semántica.

CUARTA. Los términos encontrados por la métrica de semejanza del coseno del ángulo varía considerablemente entre un 0,1 a un 0,9 dependiendo de los bigramas analizados y proporcionados por el Corpus de Google.

RECOMENDACIONES

PRIMERA. Para realizar una comparación de eficiencia en la recuperación de información semántica, se recomienda utilizar corpus alternativo aplicando otras técnicas de cálculo de similitudes.

SEGUNDA. El gestor de base de datos influye en el tiempo de respuesta, por lo que se sugiere analizar resultados aplicando otros gestores de base de datos.

TERCERA. Se recomienda utilizar técnicas de recuperación semántica en el desarrollo de sistemas de información, ya que así se obtendrán mejores resultados.

CUARTA. Se recomienda utilizar técnicas de recuperación semántica en idiomas como el quechua y el aymara, idiomas que debe conocer el mundo.

BIBLIOGRAFIA

Baeza-Yates, R. y Frakes, W.B. (1992). *Information Retrieval: Data Structures and Algorithms Englewood Cliffs*. Chile: Prentice Hall.

Baeza-Yates, R y Ribeiro Neto, B. (1999). Modern Information Retrieval. Estados Unidos: Addison Wesley.

Berners- Lee, Tim. (1989). *Information Management: A Proposal Internal Project Proposal*. Estados Unidos: Scientific American.

Berners-Lee, T y Hendler, J. (2001). *The Semantic Web: A new form of Web.* Estados Unidos: Scientific American.

Blair, D.C. (2001). *Language an Representation in Information Retrieval.* Elsevier Science Publishers.

Bolshakov, I y Gelbukh, A. (2004). *Computational Linguistics. Models, Resources, Applications.* México: Instituto Politécnico Nacional

Brassard, G. y Bratley, P. (2002). *Fundamentos de Algoritmia*. España: Editorial Prentice Hall.

Castells, P. y Macias, J.A. (2001). *An Adaptive Hypermedia Presentation Modeling System for Custom Knowledge Representations*. World Conference on the WWW.

Cimiano, P. (2006). *Ontology Learning and Population from Text, Algorithms, Evaluation and Applications.* Estados Unidos: Springer.

Coates, A.B. (2001). *The Role of XML in Finance.* Estados Unidos.

Codina, L, Marcos, M y Pedraza, R. (2009). *Web Semántica y Sistemas de Información Documental.* España: Ediciones Trea.

Decker, S, Erdmann, M, Fensel, D y Studer, R. (1999). *Ontobroker: Ontology Based Access to Distibuted and Semi Structured Information*. Semantic Issues in Multimedia Systems. Kluwer Academic Publisher.

Eco, H. (2007). *Como Se Hace Una Tesis*. España: Editorial Gedisa S.A.

Fresno Fernandez, V. (2006). *Representación Autocontenida de Documentos HTML: una Propuesta Basada en Combinaciones Heurísticas de Criterios* (Tesis Doctoral). Universidad Rey Juan Carlos. España.

Galicia Haro, S y Gelbukh, A. (2007). *Investigaciones en Análisis Sintáctico para el Español.* México: Instituto Politécnico Nacional.

Gelbuck, A y Sidorov, G. (2010). *Procesamiento Automático del Español con Enfoque en Recursos Léxicos Grandes*. México: Instituto Politécnico Nacional.

Grossman, D.A. y Frieder, O. (1998). *Information Retrieval: Algorithms an Heuristics*. Estados Unidos: Kluwer Academia Publishers.

Hechevarría, A. (2006). *La Lematización en el Preprocesamiento de Textos para la Recuperación de la Información*. IV Congreso de Reconocimiento de Patrones. Cuba.

Jurafsky, D. y Martin, J. (2009), *Speech and Language Processing*. Estados Unidos: Pearson Education, Upper Saddle River.

Lluis Codina, M y Pedraza, R. (2009). *Web Semántica y Sistemas de Informacion Documental.* España: Ediciones Trea.

López Herrera, A. (2006). *Modelos de Sistemas de Recuperación de Información Documental Basados en Información Lingüística Difusa*. (Tesis de Maestría). Universidad de Granada. España.

Manning, C, Raghavan, P. y Schütze, H. (2008). *Introduction to Information Retrieval.* Inglaterra: Cambridge University Press.

Martínez Méndez, F. (2004). *Recuperación de Información: Modelos, Sistemas y Evaluación*. España: Universidad de Murcia.

Paez Warton, J. (2009). *El Plan de Tesis*. Perú: Impresiones Olgraf.

Renteria Agualimpia, W. (2009). *Recuperación controlada de información cualitativa desde repositorio de datos (*Tesis de Maestría). Instituto Politécnico Nacional. México.

Rodríguez Hontoria, H. (2011). *Similitud Semántica*. España: Universidad Politécnica de Catalunya.

Romero Chávez, J. (2012). *Diseño e Implementación de Mecanismos de Búsqueda Contextualizada y Anotado a través de la Web Semántica*. Instituto Politécnico Nacional. México.

Sahlgren, M. (2006). The Word Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Suecia: Universitetsservice US-AB.

Seco Naveiras, D. (2009). Técnicas de Indexación y Recuperación de Documentos Utilizando Referencias Geográficas y Textuales (Tesis Doctoral). Universidad de Da Coruña. España.

Tejada Cárcamo, J. (2006). *Desambiguación de sentidos de palabras usando relaciones sintácticas como contexto local* (Tesis de Maestría). Instituto Politécnico Nacional. México.

Tolosa, G y Bordignon, F. (2009). *Introducción a la Recuperación de Información.* Argentina: Universidad Nacional De Luján

Valderrama Mendoza, S. (2009). *Pasos para Elaborar Proyectos y Tesis de Investigacion Cientifica*. Perú: Editorial San Marcos.

Vilares, J. (2005). *Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español* (Tesis Doctoral). Universidad de la Coruña. España.

ANEXOS

Anexo A : Fragmento del Corpus de Google

Anexo B : Lista de Stopwords

ANEXO A: FRAGMENTO DEL CORPUS DE GOOGLE

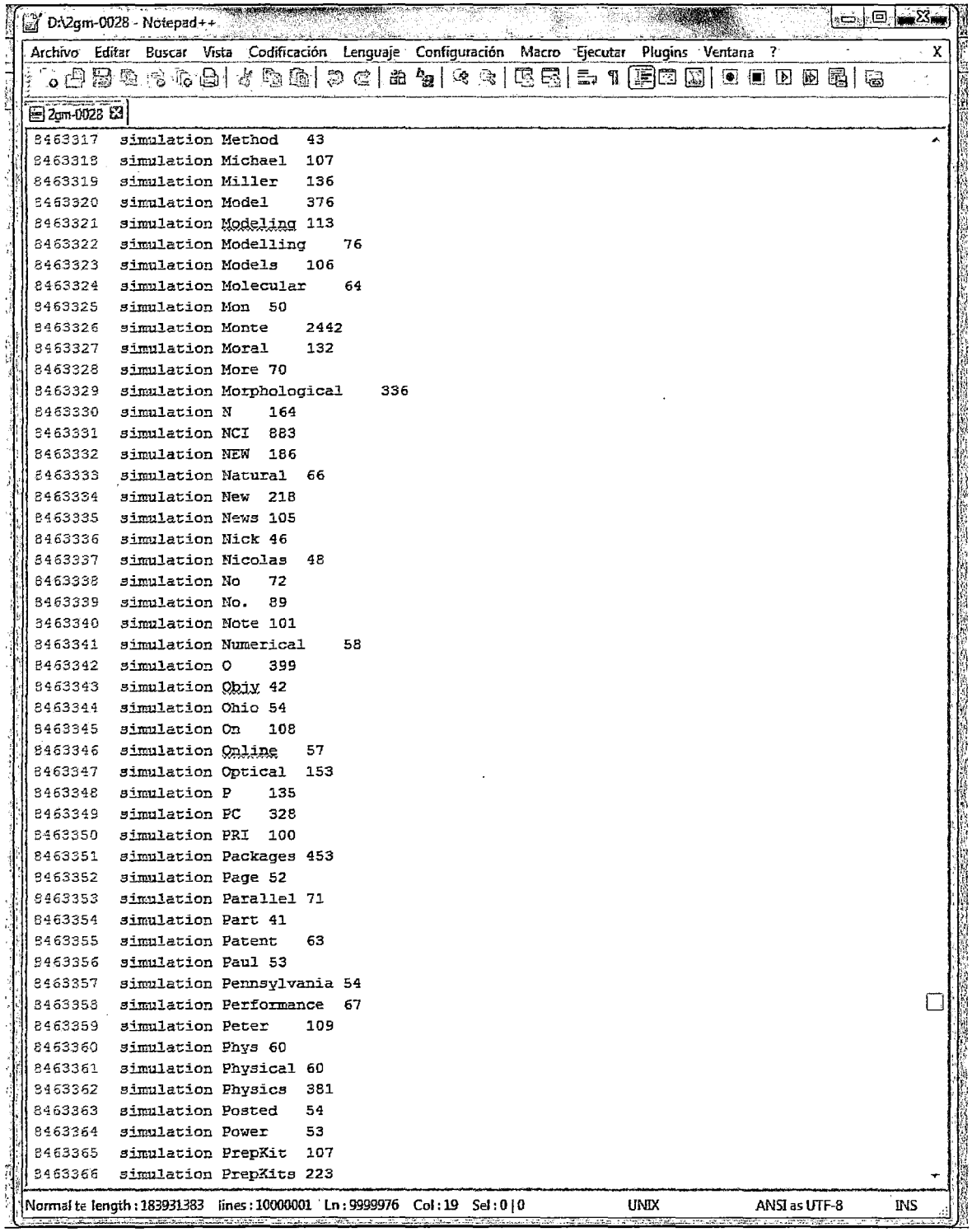


Figura 8. Fragmento del corpus de google

Fuente: Google

ANEXO B: LISTA DE STOPWORDS

CUADRO 9.

Lista de Stopwords

Stopword	Stopwords	Stopword
algún	dentro	fuimos
alguna	desde	gueno
algunas	donde	ha
alguno	dos	hace
algunos	el	haceis
ambos	ellas	hacemos
ampleamos	ellos	hacen
ante	empleais	hacer
antes	emplean	haces
aquel	emplear	hago
aquellas	empleas	incluso
aquellos	empleo	intenta
aqui	en	intentais
arriba	encima	intentamos
atras	entonces	intentan
bajo	entre	intentar
bastante	era	intentas
bien	eramos	intento
cada	eran	ir
cierta	eras	la
ciertas	eres	largo
cierto	es	las
ciertos	esta	lo
como	estaba	los
con	estado	mientras
conseguimos	estais	mio
conseguir	estamos	modo
consigo	estan	muchos
consigue	estoy	muy
consiguen	fin	nos
consigues	fue	nosotros
cual	fueron	otro

Fuente: Code Google

CUADRO 10.

Lista de Stopwords (continuación)

Stopword	Stopwords	Stopword
pero	sus	vais
podeis	también	valor
podemos	teneis	vamos
poder	tenemos	van
podria	tener	vaya
podriais	tengo	verdad
podriamos	tiempo	verdadera
podrian	tiene	vosotras
podrias	tienen	vosotros
por	todo	voy
por qué	trabaja	yo
porque	trabajais	
primero	trabajamos	
puede	trabajan	
pueden	trabajar	
puedo	trabajas	
quien	trabajo	
sabe	tras	
sabeis	tuyo	
sabemos	ultimo	
saben	un	
saber	una	
sabes	unas	
ser	uno	
si	unos	
siendo	usa	
sin	usais	
sobre	usamos	
sois	usan	
solamente	usar	
solo	usas	
somos	uso	
soy	va	
su		

Fuente: Code Google