UAEM | Universidad Autónoma del Estado de México

CENTRO UNIVERSITARIO UAEM
NEZAHUALCOYOTL

”BUSCADOR SEMÁNTICO APLICANDO TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN”

ARTÍCULO ESPECIALIZADO

PARA OBTENER EL TÍTULO DE

LICENCIADO EN INGENIERÍA EN SISTEMAS INTELIGENTES

PRESENTA

CLAUDIA EVELYN BUENDIA VALLEJO

ASESOR

DRA. DORA MARÍA CALDERÓN NEPAMUCENO

Buscador Semántico Aplicando Técnicas de Recuperación de Información

Claudia Evelyn Buendía Vallejo
Centro Universitario Nezahualcóyotl
Universidad autónoma del Estado de México
México
buvace@gmail.com

Dra. Dora María Calderón Nepamuceno
Centro Universitario Nezahualcóyotl
Universidad autónoma del Estado de México
México
dmcalderonn@uaemex.mx

Resumen—Se presenta un buscador semántico basado en un diccionario de Español, con la finalidad de darle significado a las búsquedas que realizan los usuarios ya que los buscadores comunes de la web lanza muchos resultados de los cuales el posicionamiento y su relevancia se basan en el número de veces que aparece las palabras que ingresamos para realizar la búsqueda; para lograr esto es necesario utilizar técnicas de recuperación de información (IR) para dar resultados útiles o con más relevancia a lo que el usuario busca, para ello se propone el uso del modelo de espacio vectorial qué calcula la cercanía entre las palabras y las definiciones del diccionario, así como también emplea la lematización del diccionario en español, para poder conectar las palabras entre sí de forma que la búsqueda quede en el contexto al que se refiere el usuario. En este artículo se explicará la metodología usada para lograr una búsqueda más eficaz.

Palabras Claves— Semántica, Motor de búsqueda, Recuperación de información.

I. INTRODUCCIÓN

En la actualidad los buscadores están basados en realizar búsquedas bajo las indexación de miles de páginas; a su vez utilizan datos (metadatos) relevantes de la páginas, sin dejar de lado la publicidad relacionada a la búsqueda, estos métodos nos lanzan como resultados muchas páginas que no son útiles para lo que se necesita. [6]

Los buscadores intentan dar como resultado a las páginas que tienen más relación con el tema de búsqueda sin embargo a veces esto lo hacen lanzando en las primeras posiciones las páginas que contienen más veces las palabras que se usan para la búsqueda, pero esto no quiere decir que sea de completa utilidad. [8]

Para tener búsquedas más útiles es necesario hacer uso de las mismas palabras con las que se describe la página o documento o preguntas muy específicas, por esto los buscadores utilizan metadatos los cuales son palabras específicas que pueden llegar a determinar el contenido de la página o documento.

Sin embargo la cantidad de información que se encuentra en la web esta cada día en aumento, esta también es una causa del por qué las búsquedas son tediosas y requieren del juicio humano para determinar si es útil la información encontrada [7], por esto se busca la implementación de buscadores semánticos o buscadores que logren darle contexto o significado a las búsquedas.

II. METODOLOGÍA

El lenguaje natural, entendida como una herramienta que utilizan las personas para expresarse, tiene propiedades específicas que reducen la eficacia de los sistemas de recuperación de información textual [4]

Recuperación de información (IR) es la ciencia de la búsqueda de información en documentos, textos, imágenes, sonido o datos de otras características, de manera pertinente y relevante.

Para la IR tradicional se requiere considerar los siguientes elementos:

- Un vocabulario (lista de términos en lenguaje natural).
- Un algoritmo que incluya las reglas lógicas de la búsqueda (tabla de verdad)

- Una valoración de los resultados o cantidad de información lograda o posible.

Además, se requiere de un motor de búsqueda, el cual permita planear una pregunta con no menos de dos términos y mostrar los resultados mínimos de ponderación para el despliegue de resultados al usuario. [2]

Kobayashi y Takeda detallan como medir el desempeño de un sistema IR utilizando estadísticas y los tres parámetros tradicionales: velocidad, precisión y recall. [3]

Precisión: La cual se calcula con la ecuación (1).

$$P = \frac{documentos\ relevantes \cap documetos\ recuperados}{documentos\ recuperados} \quad (1)$$

Recall: La cual se calcula con la ecuación (2)

$$R = \frac{documentos\ relevantes\ \cap documetos\ recuperados}{documentos\ relevantes} \quad (2)$$

También es necesario definir el contexto, el cuál es un entorno físico o de situación a partir del cual se considera un hecho. Está constituido por un conjunto de circunstancias (como el lugar y el tiempo) que ayudan a la comprensión se un mensaje. Por ejemplo: un paródico titula "Rafael viajo". Esto no aporta la información necesaria para que el lector decodifique el mensaje. En cambio, el titular "Rafael Nadal viajo ayer a Italia para jugar al Abierto en Roma" sí puede ser interpretado que incluye información sobre el contexto.

Existen técnicas para capturar el contexto de una búsqueda y con ello mejorar la recuperación de información. La búsqueda contextual intenta capturar de mejor manera las necesidades de información del usuario, resolviendo posibles ambigüedades en los términos, y generando un compendio contextual que consiste de sus conceptos claves. [5]

La coincidencia de palabra es una de estas técnicas, a la fecha los sistemas de búsqueda y de recuperación de información, resuelven consultas basándose en la técnica de coincidencia de palabra (keyword-matching). Se basa en encontrar palabras idénticas a través del procesamiento de texto. En consecuencia, se obtiene resultados irrelevantes porque la palabra es ambigua.

Las ontologías son estructuras que representan y permiten almacenar conocimiento de acuerdo a un dominio específico. Tomando documentos de Wikipedia (terminología de uso popular). Estructura semántica de WordNet (almacena relaciones semánticas del sentido de una palabra). [9]

Para integrar la información se requiere de diccionarios lo cual permite de forma significativa establecer de forma certera los mejores parámetros de búsqueda. Relacionando los conceptos con el significado para ver si este es el sentido que se le quiere dar a la búsqueda.

Para todo lo anterior se requiere de un modelo para recuperar la información. Uno de los modelos más utilizados para recuperar información es el modelo de espacio vectorial. En él las consultas, los términos y los documentos se presentaban como vectores en un espacio con múltiples dimensiones. Un vector tiene tantas dimensiones como términos en el espacio del documento. [19]

Este modelo entiende que los documentos pueden expresarse en función de unos vectores que recogen la frecuencia de aparición de los términos en los documentos. Los términos que forman esa matriz serían términos vacíos, es decir, dotados de algún significado a la hora de recuperar información y por otro lado, estarían almacenados en formato "stemmed" (reducidos los términos a una raíz común, tras un procedimiento de aislamiento de la base que agruparía en una misma entrada varios términos).

El Sistema de Recuperación de Información (SRI) propuesto contiene los siguientes cuatro documentos:

D1: el río Danubio pasa por Viena, su color es azul

D2: el caudal de un río asciende en invierno

D3: el río Rhin y el río Danubio tienen mucho caudal

D4: si un río es navegable, es porque tiene mucho caudal.

La matriz correspondiente dentro del modelo del Espacio Vectorial podría ser la siguiente:

	Río	Danubio	Viena	color	azul	caudal	invierno	Rhin	navegable
D1	1	1	1	1	1	0	0	0	0
D2	1	0	0	0	0	1	1	0	0
D3	2	1	0	0	0	1	0	1	0
D4	1	0	0	0	0	1	0	0	1

Tabla 1. Ejemplo de Matriz de términos y documentos en el Espacio Vectorial. Fuente: Elaboración propia

Por medio de un proceso denominado stemming, quizá el SRI hubiera truncado algunas de las entradas para reducirlas a un formato de raíz común, pero para continuar con la explicación resulta más sencillo e ilustrativo dejar los términos en su formato normal. En cuanto a las palabras vacías, hemos supuesto que el SRI elimina los determinantes, preposiciones y verbos ("el", "pasa", "por", etc.), presentes en los distintos documentos. Para entregar la respuesta a una determinada

pregunta se realizan una serie de operaciones. La primera es traducir la pregunta al formato de un vector de términos. Así, si la pregunta fuera "¿cuál es el caudal del río Danubio?", su vector de términos sería Q = (1,1,0,0,0,1,0,0,0). El siguiente paso es calcular la similitud existente entre el vector pregunta y los vectores de los documentos (existen varias funciones matemáticas diseñadas para ello) y ordenar la respuesta en función de los resultados de similitud obtenidos.

En la mayor parte de los sistemas de recuperación de información se tiene que realizar de una forma u otra, la frecuencia de las palabras que aparecen en los documentos , como mayúsculas, minúsculas, acentos entre otros, ya que muchas palabras podrán ser agrupadas al contener una forma similar al inicio de esta. En el área de recuperación de información esta se encarga principalmente del estudio de los sistemas y procesamiento de texto para asignar índices, buscar y devolver datos al usuario.

Para poder reducir las palabras que se derivan de otras se utilizan el algoritmo de Stemming, este es utilizado para el inglés y para el castellano se utiliza el algoritmo de porter que es un derivado de Stemming.

Stemming es un método para reducir una palabra a su raíz o (en inglés) a un stem o lema. Hay algunos algoritmos de stemming que ayudan en sistemas de recuperación de información. Stemming aumenta el recall que es una medida sobre el número de documentos que se pueden encontrar con una consulta.

Numerosos algoritmos de stemming se vienen desarrollando desde haces años. Tres de los más conocidos son los construidos por Lovins en 1968, Porter en 1980 y Paice en 1990. El algoritmo más común para stemming es el algoritmo de Porter. Todos estos algoritmos van eliminando consecutivamente los finales de las palabras, para arribar a su raíz.

El algoritmo de Porter permite hacer stemming, esto es extraer los sufijos y prefijos comunes de palabras literalmente diferentes pero con una raíz común que pueden ser consideradas como un sólo término.

El algoritmo de Porter tiene la ventaja de ir quitando sufijos por etapas, en cambio Lovins requiere de la definición de todas las posibles combinaciones de sufijos.

El algoritmo de Porter se publicó en 1980. Básicamente lee un archivo, toma una serie de caracteres, y de esa serie, una palabra; luego valida que todos los caracteres de la palabra sean letras y finalmente aplica la lematización.

Lematización es el proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc.), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra, en palabras más simples el lema de una palabra es la palabra que nos encontramos en el tradicional diccionario como entrada: Singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos.

La lematización puede realizarse automáticamente mediante programas de análisis morfológico.

Un lematizador es capaz de reconocer la riqueza morfológica del español, ya que al reconocer una palabra presenta sus lemas junto con toda su información morfológica correspondiente. El lematizador, también se encarga de la recuperación de información. [18]

El lematizador es un software que permite, a partir de un texto etiquetado gramaticalmente, conocer el lema como ya se ha dicho.

Algunos lematizadores que existen son:

- Tree Tagger
- Tnt Tagger
- Mbt Tagger

La morfología es el estudio de las palabras, su estructura interna, modificaciones que sufren para expresar las diferentes categorías gramaticales como: género, número, tiempo, modo, etc.

Las palabras están formadas por unidades pequeñas que tiene significado (morfemas), ejemplos:

- Niñ-o
- Cas-a
- Flor-ero
- Roj-os
- Comi-ò
- Deport-ista

Las palabras tienen dos morfemas:

1. Morfema Raíz: raíz
2. Morfemas flexivos: sufijo, prefijos, infijos.

Procesos importantes:

- Flexión: escritor-a escritor-as
- Derivación: lava-ble compra-ble
- Composición: agua-miel saca-puntas

Las palabras se agrupan en categorías gramaticales o clases de palabras de acuerdo a:

- Estructura
- Función que desempeña en la oración
- Su significado

E n el español se clasifican en 8 categorías:

1) Sustantivo
2) Adjetivo
3) Artículo

4) Pronombre
5) Verbo
6) Adverbio
7) Proposición
8) Conjunción

El análisis morfológico clasifica una palabra en la raíz más las categorías que se encuentran en la palabra, por ejemplo:

Gatos = Gato+ Sus+Plur.

Bebo = beber+Ver+Pres. Ind.+ 1 Per.+ Sing.

III. DESARROLLO

Se implementó el modelo de espacio vectorial para la recuperación de información, calculando la cercanía entre las palabras para realizar la búsqueda a las definiciones del diccionario teniendo así como resultado principal de la búsqueda la definición más cercana, también se utilizara el algoritmo de Porter con el cual se calcula la frecuencia de las palabras en un documento; y se empleara la lematización del diccionario en Español, para poder conectar las palabras entre sí de forma que la búsqueda quede en el contexto al que se refiere el usuario.

El modelo de espacio vectorial, en la actualidad es el más utilizado en los sistemas de recuperación de información o SRI. En el modelo, las consultas, los términos y los documentos se representan como vectores en un espacio con múltiples dimensiones. Un vector tiene tantas dimensiones como términos en el espacio del documento.

Se incluirán diccionarios en los cuales se aplicara el algoritmo para indexar las palabras con sus definiciones. [14]

A. Herramientas técnicas

Se utilizara el paquete FreeLing, el cual es una biblioteca de prestación de servicios lingüísticos de análisis. FreeLing está diseñado para ser utilizado como una biblioteca externa de cualquier aplicación que requiera de este tipo de servicios. [10]

Los principales servicios que ofrece:

- El análisis morfológico
- Texto Tokenizacion

Tokens son las palabras de un lenguaje natural: cada tokens es una secuencia de caracteres que representa una unidad de información en el programa fuente. [12]

- Flexible reconocimiento multipalabra
- Cuadro basado en análisis superficial
- WordNet de anotación basado en el sentido y la desambiguación
- El reconocimiento de fechas, números, índices, divisas, y las magnitudes físicas (velocidad, peso, temperatura, densidad, etc)

B. Lenguajes de programación

Se utilizara Perl, lenguaje de programación originalmente desarrollado para la manipulación de textos y que actualmente es utilizado para un amplio rango de tareas incluyendo administración de sistemas, desarrollo web, programación en red y más. [11]

IV. TRABAJOS RELACIONADOS

A. WOLFRAM ALPHA

Wolfram Alpha es un servicio en línea que responde a las preguntas directamente, mediante el procesamiento de la respuesta extraída de una base de datos estructurados, en lugar de proporcionar una lista de los documentos o páginas web que podrían contener la respuesta, tal y como lo hace Google.

Las consultas y procesamientos de cálculos también se hacen en un campo de texto, pero en este se procesan las respuestas y visualizaciones adecuadas dinámicamente en lugar de producirlas como resultado de la obtención de un banco de respuestas predefinidas. Por lo tanto difiere de los motores de búsqueda semántica, los cuales indexan una gran cantidad de respuestas y luego tratan de hacer coincidir éstas con la pregunta hecha.

B. HAKIA

Hakia es un buscador de recursos de la información que trabaja sobre la base del lenguaje natural en la que las búsquedas no son palabras o frases sino interrogantes concretos. Su software interpreta el significado de una pregunta desarrollada de manera coloquial, es decir, Hakia entiende el contenido ya que necesita información léxica y sintáctica además de la semántica.

Hakia ofrece unos resultados en base a unos criterios.

1- En sus resultados aparecen webs creíbles y recomendadas por bibliotecarios.

2- En los resultados se muestran sitios con la última información disponible.

3- Únicamente se muestran páginas que son absolutamente relevantes.

C. IDEAS AFINES

Se tiene una poderosa base de datos que relaciona automáticamente distintas palabras y términos con conceptos similares o afines.

Esta herramienta es muy útil para distintos fines creativos, ya que puede ser utilizada como un generador de ideas, relacionando distintos términos entre sí. Partiendo de una idea clave se puede llegar a distintos conceptos, todos relacionados con el término principal.

D. QUINTURA

Es un pequeño buscador ruso, este buscador genera nubes conceptuales con resultados relacionados con los términos de búsqueda. Es un buscador visual, que además acaba de incorporar la posibilidad de incluir iconos en los resultados. Se encuentra en la línea de evolución que sigue internet, que tiende a volverse cada vez más visual.

E. MNEMOMAP

Es una aplicación web que crea una especie de mapa conceptual formado por las grandes categorías (palabras clave, sinónimos, traducciones, etiquetas o tags) en que ha agrupado los resultados y que a su vez pueden desplegarse en subcategorías con un solo clic.

En segundo lugar (en la parte inferior) distribuye los resultados en pestañas: una para el propio mnemomap, otra para Youtube, una tercera para Filckr y la última para Yahoo!.

F. SABIOS: UNA APLICACIÓN DE LA WEB SEMÁNTICA PARA LA GESTIÓN DE DOCUMENTOS DIGITALES

SABIOS fue desarrollado para la aplicación de la web semántica, ya que este se basa en el marco semántico de documentos y datos. Se vio que en la Escuela de Artes Plásticas se manejaba una gran cantidad de documentos e información. SABIOS planteo una solución a problemas de la Escuela de Artes Plásticas, como búsqueda, recuperación y clasificación de la información. Este se basó en la recuperación de información y la web semántica, poder obtener la información más relevante e importante de los documentos con que el usuario está trabajando.
SABIOS se basó en la metodología MASCommonKADS, que consta de las siguientes fases.

- Conceptualización, en esta se identifica el problema a resolver desde el punto de vista del usuario y determinar los casos de uso.
- Análisis. Propone seis modelos: agentes, tareas, experiencia, coordinación, comunicación y organización.
- Diseño. MASCommonKADS, propone el modelo de diseño clásico, de agentes conocimiento, red y protocolos de comunicación.

Para poder llevar a cabo el proceso de búsqueda de información en SABIOS se definieron actores:

- Actor usuario. Este realiza la búsqueda semántica, donde realiza una búsqueda donde hay una lista de conceptos, y busca si hay un metadato en los documentos que contenga este.
- Actor autor. Persona que ingresa los documentos y relacionarlo con los metadatos, y estos pueden ser modificados
- Actor documentalista. Es el administrador del centro documental, se le asocian tres funciones, editar documento, borrar documento y agregar documento.
- Actor administrador. Este es el que administra el sistema y puede acceder a la base de contexto para editar las ontologías. [17]

Los buscadores ya antes mencionados, realizan sus búsquedas bajo preguntas o temas concretos, en algunos casos solo realizan las búsquedas bajo preguntas, mientras más palabras o contexto tengan la pregunta u oración, serán más precisos los resultados; para el buscador desarrollado solo es necesario tener dos palabras, claro tiene que tener cierto contexto para que nos arroje resultados relevantes, así mismo se utilizan técnicas como la lematizaciòn y morfología que permiten cubrir más el rango de coincidencia de palabras al llevarlas a la raíz y esto mejorando el desempeño del modelo de espacio vectorial.

V. RESULTADOS

Se realizaron 15 consultas; eligió estas consultas buscando tuvieran significado o contexto; ya que se relaciona la consulta con el significando; el cual es nuestro documento en el que se aplicara el modelo del espacio vectorial, lanzando así las palabras que más se asemejan a lo buscado.

De estas consultas, se obtuvieron 12 consultas exitosas y 3 sin resultados relevantes. Para ver la efectividad del buscador se tomaron las 15 consultas al 100% obteniendo la formula (3).

$$Efectividad = \frac{(12)(100)}{15} = 80\% \qquad (3)$$

A. Consultas con éxito

- Aparatos electrónicos
- Respuestas de preguntas
- Malas palabras
- Prenda de vestir
- Sin duda
- Hablar de elegancia
- Sobre programación
- Frutos del bosque
- Dar a luz
- Dar a luz un niño
- Hacer un viaje
- Herramientas de trabajo

B. Consultas sin éxito

- Mejor amigo del hombre
- Utensilios para comer
- Comida saludable

Ejemplos de búsquedas

- Exitosos

Aparatos electrónicos, el resultado se muestra en la figura (1).

Figura 1. Resultado de la búsqueda de Aparatos Electrónicos. Fuente: Elaboración propia

Dar a Luz, el resultado se muestra en la figura 2.

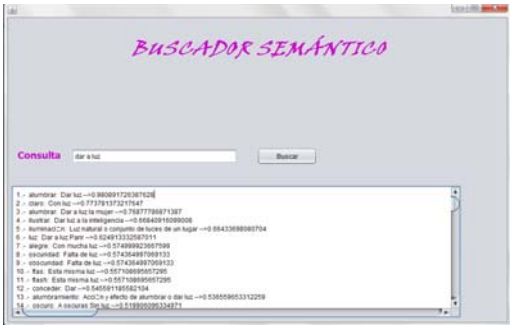

Figura 2. Resultado de la búsqueda de Dar a luz.

- Sin éxito

Comida saludable, el resultado se muestra en la figura 3.

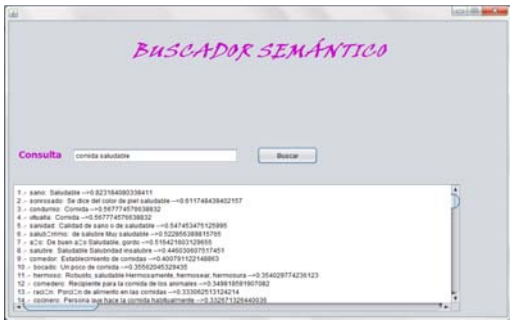

Figura 3. Resultado de la búsqueda de Comida Saludable.

Utensilios para comer, el resultado se muestra en la figura 4.

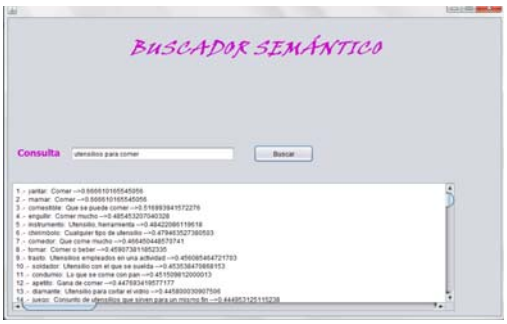

Figura 4. Resultado de la búsqueda de Utensilios para comer

VI. DISCUICIÒN DE LOS RESULTADOS

Como se muestran en las figuras anteriormente mostradas se observa que la coincidencia entre la búsqueda y los primeros resultados lanzados son altos, así como son resultados tienen mucha relevancia; esto es lo principal que se busca lograr, que no solo se mandaran en primer plano los resultados donde se

encontraran más las palabras con las que se realiza la búsqueda, si no que el buscador entendiera el contexto de la búsqueda.

Se puede observar que el algoritmo de Porter, la lematización y el análisis morfológico son de gran ayuda para tener una mejor recuperación de información, ya que al someter las palabras a estos algoritmos se logra tener mayor coincidencia de palabras, y esto se refleja en la relevancia del resultado.

Lo que Kobayashi y Takeda plantean para verificar la efectividad del buscador, nos muestra que tiene un alto grado de ella; es claro que la precisión de este puede mejorar, sin embargo, por los resultados obtenidos se puede decir que la efectividad aceptable.

VII. Conclusiones

La recuperación de información sigue cobrando un gran auge debido al crecimiento de Internet. Los programas de RI se siguen basando en los tres modelos booleano, probabilístico y vectorial.

Se logró desarrollar el buscador semántico utilizando un diccionario en español, la aplicación y herramientas de la RI.

Se implementó el modelo del espacio vectorial y la lematización para el desarrollo de las consultas a realizar. Se puede deducir que este nuevo programa tiene una precisión del 80%.

Referencias

[1] Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27, 379–423, 623–656.

[2] Baeza Yates, R. R. (1999). Modern infomation Retrieval. ACM. New York: Addison Wesley .

[3] Kobayashi, M. a. (2000). Information Retrieval on the Web. ACM Computing Survey (CSUR) , 144-173.

[4] R, B.-Y. (2004). Computational linguistics and intelligent text processing. Challenges in the interaction of information retrieval and natural language processing , 2945, 445-456.

[5] R. Kranft, C. c. (2006). Searching whit Context. International Word Wide web conference Commitee, (págs. 23-26). USA.

[6] BERNERS-LEE, T; HENDLER, J.y LASSILA, O (2001). The semantic Web. Scientific America, vol. 284, no. 5, p. 34-43.

[7] Blair, D.C. (1990). Language and representation in information retrieval. Amsterdam [etc.]: Elsevier Science Publishers.

[8] Chang, G. et al. (2001). Mining the World Wide Web: an information search approach". Norwell, Massachusetts: Kluwer Acad. Publishers.

[9] GRUBER, T.R. (1993). Toward principles for the design of ontologies used forknowledge sharing. En: Formal Ontology in Conceptual Analysis andKnowledge Representation. The Netherlands: Kluwer Academic Publishers.

[10] Padrò, Muntsa y Lluìs Padr´o. 2004. Comparing methods for language identification. ProceAnalizadores multilingues en FreeLing ¨ Linguamatica ´ – 19 samiento del Lenguaje Natural, (33):155–162, September.

[11] Larry Wall (1991). Programming Perl, O'Reilly and associates.

[12] Aho, A.V., Sethi, R., Ullman, J.D. (1990), Compiladores: principios, técnicas y herramientas, capitulo 1, páginas: 1-25, 743-747.

[13] Juan Venegas. 1985 . «El entorno lingüístico». *Documentos Lingüísticos y Literarios*11: 29-38.

[14] Maria Moliner (2009). Diccionario del uso del Español. Tercera Edición, Santillana USA Publishing Company, Gredos S. A.

[15] Feliu Arquiola, E. (2009). Palabras con estructura interna", en E. de Miguel (ed.): Panorama de la Lexicologìa, Barcelona, ariel pp. 51-82.

[16] Matthews, P.H. (1974). Morfologìa. Introduccion a la teorìa de la estructura de la palabra, Madrid, Paraninfo.

[17] Guzmán Luna, Jaime A.; Torres Prado, Durley; Ovelle, Demetrio A. "SABIOS: una aplicación de la Web semántica para la gestión de documentos digitales,". *Revista Internacional de bibliotecología*, Vol. 30, núm. 1, pp. 51-71, Universidad de Antioquia Medellín, Colombia, Enero-Junio 2007.

[18] Gòmez Dìaz, Raquel (2005). La lematizacion en Español: Una aplicaciòn para la recuperaciòn de informaciòn., Trea, pag. 175 y 176.

[19] Martìnez Mèndez, Francisco J. (2015). Recuperacion de Informaciòn: Modelos,Sistemas y Evaluaciòn.El Kiosko JMC. Capitulo 1, paginas: 09-12.

[20] Corripio, Fernando (1985). Diccionario de Ideas Afines. Herder A. A., Barcelonna. Uso del diccionarion. Pàginas: 9 y 10.