

Sistema Evolutivo Generador De Bases De Conocimiento

Área de conocimiento: Sistemas Basados en Conocimiento

Guadalupe Gaxiola Castro ¹, Olaf Yadir Cazarez Sarabia ², Jesús Manuel Olivares Ceja ³

¹ Universidad Autónoma de Sinaloa, Calle Puerto Zihuatanejo #2219, Colonia Francisco Villa, CP: 80110, Culiacán, Sinaloa, México.

Teléfono Casa (01667) 7-17-42-05

Teléfono Celular 016671-00-59-65.

guadalupegaxiola_castro@hotmail.com

² Universidad Autónoma de Sinaloa.

oycs_10@hotmail.com

³ Centro de Investigación en Computación del Instituto Politécnico Nacional.

jesusoc@yahoo.com

Resumen. Actualmente uno de los problemas que se enfrentan los Ingenieros de Conocimiento es la obtención de base de conocimiento a partir de información no estructurada. El problema ha aumentado debido al auge de Internet donde se tiene una gran cantidad de información no estructurada.

En este trabajo se propone una alternativa basada en el enfoque de Sistemas Evolutivos para construir de forma semi automática bases de conocimiento utilizando el formalismo de Redes Semánticas.

Palabras clave: Sistemas Evolutivos, Bases de Conocimiento, Lenguaje Natural, Internet

Introducción

En este trabajo se presenta una herramienta para ayudar a solucionar la problemática de la extracción de conocimiento a partir de información no estructurada mediante un enfoque de Sistemas Evolutivos [1] [3].

Este trabajo está apoyado en Redes Semánticas (RS) [4] y es una extensión al trabajo presentado en [2] en el sentido de considerar además de la Distribución Lingüística el problema de la anáfora [5] para construir la red semántica.

El objetivo de la herramienta es generar una Base de Conocimientos que refleje el contenido de varios documentos que el usuario le especifica. Se pretende es que este sistema pueda ser utilizado por cualquier persona aunque no tenga conocimientos de Ingeniería de Software u operaciones lingüísticas.

1 Arquitectura del Sistema

El sistema está desarrollado en Java y consta de tres módulos principales. En el análisis léxico se identifican los verbos, proposiciones y aquellas otras palabras. Los verbos y proposiciones forman las relaciones de la red semántica, en forma similar a como se hace en [3]. Los sustantivos acompañados con sus adjetivos forman los nodos de la red semántica.

El Analizador Léxico obtiene una oración canónica que se procesa mediante el modulo de operaciones gramaticales que incluyen un parser que permite obtener oraciones del tipo $N \ r \ N$, donde N es un nodo y r es una relación.

Los elementos N están formados por secuencias de sustantivos o sustantivos y adjetivos. Por ejemplo, Benito Juárez, Buenos Aires, México, Gato Blanco.

Los elementos r están formados por verbos, grupos de verbos y proposiciones. Por ejemplo, vive en, alimenta de, corre.

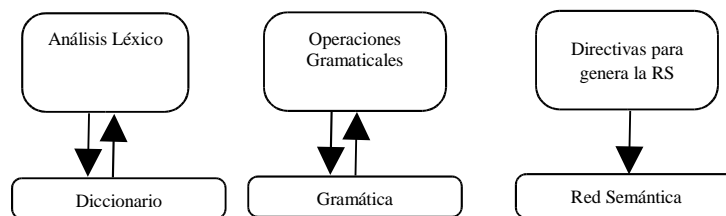


Fig 1. Arquitectura del Sistema

1.1 Análisis Léxico

Esta parte del Sistema es en el que se almacena cada una de las palabras sin repetirse de cada texto, así como también su tipo (verbo, preposición, nodo (es a quien me refiero como un sujeto), ignora (palabras que se pueden omitir), relaciones (se conforman por un verbo o un verbo y una preposición), los tipo punto (separadores punto), signos (separadores como lo es la coma)) y por último contiene una definición de cada palabra.

En el ejemplo siguiente se presenta la asignación de tipos de unidades léxicas con ayuda del diccionario.

Benito_Juárez nació en San_Pablo_Guelatao, Oaxaca, en 1806. De extracción_indígena, habló solamente zapoteco durante gran parte de su_niñez. En la_ciudad_de_Oaxaca vivió con su hermana Josefa, quien servía en la casa_de_don_Antonio_Maza.

Benito_Juárez	n1	.	tp
nació	v1	De	p3
En	p1	extracción_indígena	n5
San_Pablo_Guelatao	n2	,	c
,	c	habló	v2
Oaxaca	n3	solamente	i
,	c	zapoteco	n6
en	p2	durante	i
1806	n4	gran	i

parte	i	Josefa	n9
de	p4	,	c
su_niñez	n7	quien	n9
.	tp	servía	v4
En	p5	en	p7
la_ciudad_de_Oaxaca	n8	la	i
vivió	v3	casa_de_don_Antonio_Ma	n10
con	p6	za	
su	i	.	tp
hermana	i		

En donde:

v = verbo

n = nodo

p = preposición

tp = tipo punto

r = relación

c = coma

i = ignora

1.2 Tratamiento de la Gramática

Se da tratamiento a la gramática (cada uno de los elementos del texto y sus combinaciones) que se encuentra almacenada en un archivo, lo cual se hace de la siguiente manera: se toma la gramática del archivo, almacenándose en tablas en memoria para hacer el procedimiento más eficiente, es decir, que al buscar una producción sea más rápido.

En este modulo se generan directivas, mediante distribución lingüística [2], de la forma:

$$N1 \ r \ N2$$

En el siguiente ejemplo podemos observar que estamos manejando Benito Juárez sin el guión bajo, esto es porque cuando los datos se van a introducir al diccionario entran con guiones para identificar cada una de las partes pero al momento de volver a utilizar la misma palabra ya no es necesario utilizar el guión bajo por que lo va a reconocer automáticamente.

A partir de las oraciones canónicas se obtiene la gramática siguiente:

$$S \rightarrow D \ v \ p \ D$$

$$D \rightarrow n \ D$$

$$D \rightarrow n$$

Donde:

S y D son axiomas

D v p D es una producción

v p son terminales porque ya no tienen una secuencia

D es no terminal porque hay un axioma que se llama D el cual contiene más elementos

Si se reconoce que la oración tiene estructura $D \vee p D$ entonces se invoca a un programa que busca en la red semántica la respuesta a la consulta.

La gramática con atributos resultante es la siguiente:

$S \rightarrow D \vee p D$

$D \rightarrow n D$

$D \rightarrow n$

Trabajamos la regla de producción $S \rightarrow D \vee p D$ aplicada a la oración ¿Dónde nació Benito Juárez? de donde resulta la oración canónica $n n \vee p n n n$

De lo anterior tenemos como entrada la gramática

$S \rightarrow D \vee p D$

$D \rightarrow n D$

$D \rightarrow n$

y la oración

Benito_Juárez	Juárez	nació	en	San	Pablo	Guelatao
z						
n	N	v	p	n	n	n

Lo que queremos tener como resultado de esta oración es la consulta a la Red Semántica. El manejador de la gramática toma una producción y sino está entonces no puede procesar la oración canónica. Si está entonces verifica que la oración cumpla con la estructura indicada en la gramática.

Refiriéndonos al ejemplo de 1.1, las oraciones canónicas que se obtienen del ejemplo son:

$n_1 v_1 p_1 n_2$	$r_1 = v_1 p_1$	n_1
$r_1 n_2$		
$n_1 v_1 p_1 n_3$	$r_2 = v_1 p_1$	$n_1 r_2 n_3$
$n_1 v_1 p_2 n_4$	$r_3 = v_1 p_2$	$n_1 r_3 n_4$
$n_1 v_1 p_3 n_5$	$r_4 = v_1 p_3$	$n_1 r_4 n_5$
$n_1 v_2 n_6$	$r_5 = v_2$	$n_1 r_5 n_6$
$n_1 v_2 p_4 n_7$	$r_6 = v_2 p_4$	$n_1 r_6 n_7$
$n_1 v_2 p_5 n_8$	$r_7 = v_2 p_5$	$n_1 r_7 n_8$
$n_1 v_3 p_6 n_9$	$r_8 = v_3 p_6$	$n_1 r_8 n_9$
$n_9 v_4 p_7 n_{10}$	$r_9 = v_4 p_7$	$n_9 r_9 n_{10}$

En este modulo se resuelve la anáfora directa sustituyendo las palabras como él, ella, ellos, quien por su correspondiente nodo. También se cambian las anáforas indirectas por su nodo correspondiente como en el caso de aquella, ese, aquello.

1.3 Obtención de la Red Semántica

La Red Semántica se construye con las directivas generadas en el modulo explicado en 1.2, en este caso, si un nodo no se encuentra, se da de alta; pero si el nodo ya está entonces únicamente se liga con el nodo que aparece a continuación.

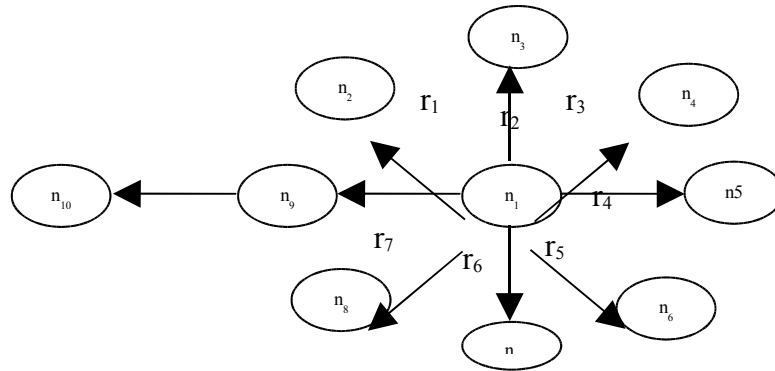


Fig 2. Red semántica

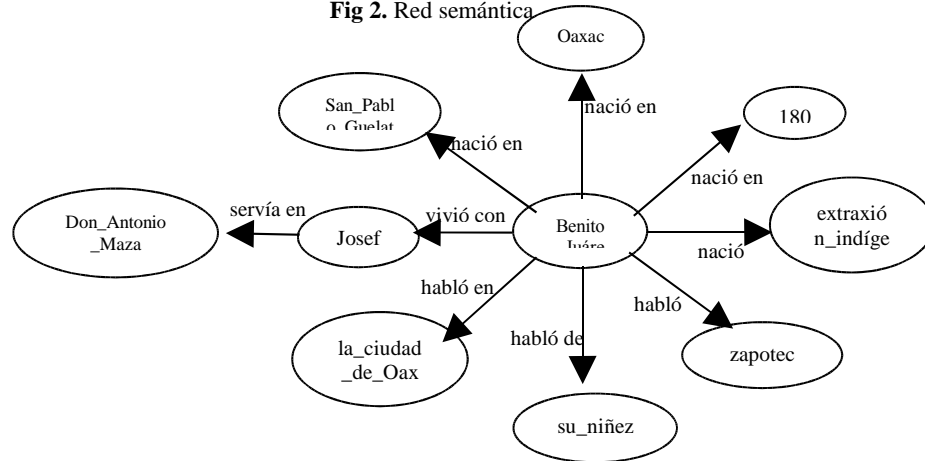


Fig 3. Red semántica equivalente con las palabras

Conclusiones

Se ha presentado un sistema basado en el enfoque de Sistemas Evolutivos para hacer posible la generación automática de bases de conocimiento, por ahora el usuario participa en el establecimiento de los elementos léxicos de tipo nodo. Los resultados muestran que es posible automatizar el proceso de generación de base de conocimiento con una mínima intervención del usuario empleando técnicas de procesamiento del lenguaje natural [6].

Trabajo a futuro

Este Sistema Evolutivo proporciona las estructuras de conocimiento para que sea posible su explotación mediante Sistemas Expertos o sistemas de búsqueda de información, por ejemplo: Si un usuario desea saber algo acerca del nacimiento de Benito Juárez le dará la respuesta accediendo a la Base de Conocimientos construida.

Referencias

- [1] Fernando Galindo Soria, Sistemas Evolutivos: Nuevo Paradigma de la Informática, 2.- Congreso Iberoamericano de Inteligencia Artificial, IBERAMIA 90, Morelia, Mich., Ed. Limusa Noriega, Julio de 1990
- [2] Jesús Manuel Olivares Ceja, Sistema Evolutivo para la Representación del Conocimiento, (tesis de licenciatura), Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas (UPIICSA), abril de 1991
- [3] Elsa Barruecos Rodríguez, Sistema Evolutivo Generador de Esquemas Lógicos de Bases de Datos, trabajo de seminario de titulación en la Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas (UPIICSA) del Instituto Politécnico Nacional (IPN), 5 septiembre de 1988.
- [4] Ross Quillian, Semantic Memory en Semantic Information Processing (Editor Marvin Minsky), MIT Press, 1968
- [5] Maximiliano Saiz Noeda, Influencia y aplicación de papeles sintácticos e información semántica en la resolución de la anáfora pronominal en español, trabajo realizado en la Universidad de Alicante, departamento de Lenguajes y Sistemas Informáticos, junio de 2002.
- [6] [Alexander Gelbukh](#), Tendencias recientes en el procesamiento de lenguaje natural. In: SICOM-2002, Simposium Nacional de Computación. Villahermosa, Tabasco, Mexico, May 29-31, 2002. CD edition: ISBN 970-18-7980-5, 14 pp.

