Adquisición de Conocimiento Mediante Estrategias de Navegación Web con Gramáticas Hipertextuales

Acquisition of knowledge through Web navigation strategies hypertext grammars

Augusto Cortez Vásquez [1], **Zoraida Mamani Rodriguez** [2]

Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, Perú

Resumen— Internet se ha convertido en el mayor medio de comunicación e interacción entre seres humanos así como entre estos y las máquinas, La adquisición de conocimiento a partir de sesiones web es una de las áreas que reviste mayor interés. En esta investigación se pretende ofrecer en primer lugar una presentación de los fundamentos para la adquisición de conocimiento identificando los sitios web más relevantes o visitados. En segundo lugar se representa las sesiones web mediante grafos y gramáticas libres de contexto probabilístico de tal forma que las sesiones que tengan mayor probabilidad son consideradas como las más visitadas o más preferidas, por tanto las más relevantes en relación a un tópico determinado.

Abstract— Internet has become a major means of communication and interaction between human beings and between them and the machines, Knowledge acquisition from web sessions is one of the areas of greatest interest . This research aims to provide first a presentation of the rationale for the acquisition of knowledge by identifying the most relevant or visited websites. Second web sessions is represented by graphs and probabilistic context -free grammars so that sessions that are most likely are considered the most popular or most preferred , therefore the most relevant in relation to a particular topic.

Palabras clave— **Adquisición de conocimiento, patrones de navegación, aprendizaje de patrones, gramática probabilística, hipertexto, recuperación de información.**

Keys words— **words of knowledge acquisition , navigation patterns, learning patterns, probabilistic grammar hypertext, hypertext information retrieval.**

I. INTRODUCCIÓN

El desarrollo de los sistemas hipermedia desempeñan un papel capital dentro del proceso de presentación de contenidos al aprendizaje del usuario, para ello, generalmente se utilizan evaluaciones *on-line* de adquisición del conocimiento con las que evitan los problemas asociados a las evaluaciones post navegación[14]. Una de las principales limitaciones de estas medidas es que proporcionan un valor global de la adquisición del conocimiento, sin distinguir entre el aprendizaje del texto y de la estructura del mismo [15]. Kintsch propone que el usuario aprende el contenido de un hipertexto a partir de la adquisición del texto base (la mera secuencia de palabras) y su combinación con el modelo de la estructura (el modelo sobre las relaciones entre las diferentes ideas expresadas en el texto).. El usuario puede aprender tan sólo el texto base del contenido, sin adquirir una comprensión profunda del mismo (modelo de la estructura), y viceversa. Existen interesantes opiniones al respecto, sobre todo la relevancia del aprendizaje a través de sistemas hipertexto, donde la estructura de enlaces entre los distintos textos ayuda a la integración del contenido del sistema. La noción de hipertexto se usado en diversos sentidos y contextos, siendo la cuestión fundamental que los sistemas hipertexto facilitan la adquisición de un mejor modelo de la estructura que los sistemas tradicionales en papel, aunque no se encuentren diferencias en la adquisición del texto base [16]. De lo anterior se deriva el hecho de que, las medidas de adquisición del conocimiento existentes son insuficientes, por una u otra razón, para evaluar la evolución del aprendizaje de textos a través de la navegación en un hipertexto. Una medida alternativa debe ser capaz de superar estos inconvenientes y medir los cambios en el aprendizaje de los contenidos a medida que el usuario navega por el sistema. Por otro

[1] Augusto Cortez Vásquez, E-mail: acortezv@unmsm.edu.pe
[2] Zoraida Mamani Rodríguez, E-mail: zmamanir@unmsm.edu.pe
Recibido: Noviembre 2015 / Aceptado: Diciembre 2015

lado, debe proporcionar información tanto del aprendizaje del texto base, como de la estructura del texto. [14].

Cuando un usuario visita la web y quiere recuperar páginas en relación a un concepto, debe evitar muchas páginas irrelevantes, el objetivo es pues recuperar las páginas significativas, es decir aquellas que sean autoridad en el tópico. Hay dos conceptos relacionados: páginas más visitadas y páginas más relevantes. Por ello se parte de la premisa de que las páginas más relevantes son aquellas que son más visitadas. La presente investigación captura a partir de la información contenida en los *logs* del servidor, las actividades de los usuarios durante su conexión en la web y extrae patrones de comportamiento que permitirán ayudar a comprender las preferencias de navegación de los usuarios, permitiendo así adaptar las interfaces de futuras páginas a los usuarios individuales. Para conseguir el propósito se utilizó un modelo simple de hipertextos representados mediante grafos, y se utiliza una representación de las sesiones de navegación de los usuarios inferidas de los archivos log como una gramática probabilística de hipertexto

II. OBJETIVOS

A. *Objetivo general*

Obtener un instrumento para identificar las preferencias de los usuarios en la web.

B. *Objetivos específicos*

- Representar las sesiones web mediante grafos dirigidos.
- Representar las sesiones web mediante gramáticas libre de contexto probabilísticas de hipertexto.

III. MARCO CONCEPTUAL

A. *Recuperación de información*

La recuperación de información (*IR Information Retrieval*) es un término utilizado en un sentido muy amplio, que requiere precisión, a menudo vagamente definido y en este contexto se refiere solamente a los sistemas automáticos de recuperación de información. Contreras señala en su tesis [1]: "Lancaster proporciona una definición: *Un sistema de recuperación de información no informa (es decir cambia el conocimiento) al usuario del propósito de su pregunta. Este informa simplemente de la existencia (o la no existencia) y paradero de documentos referentes a su petición*".

Los sistemas automáticos a los que se refiere requieren velocidad, consistencia, precisión, y facilidad de uso en la recuperación de textos relevantes para satisfacer las consultas de los usuarios.

B. *Minería web*

Cada vez se acentúa la necesidad de conocer como los usuarios interactúan con los sitios web. La minería web (MW) se refiere esencialmente al descubrimiento y análisis de información de los usuarios en la web, con el objetivo de descubrir patrones de comportamiento. Alcívar se refiere al término MW, como la tecnología usada para descubrir información no obvia a partir de fuentes de datos que incluyen los logs del servidor [9].

C. *Lenguaje formal*

Un lenguaje natural, se rige por reglas gramaticales, que aunque están ya definidas, pueden ser modificadas posteriormente (ver Fig.1). Esto constituye una ventaja para el lenguaje natural, pues lo enriquece, sin embargo al mismo tiempo dificulta su procesamiento computacional dado que puede ser ambiguo e impreciso. Un lenguaje formal, por el contrario es formal y libre de ambigüedades, es un lenguaje desarrollado por el hombre para expresar las situaciones que se dan en específico en cada área del conocimiento científico. Los lenguajes formales pueden ser utilizados para modelar una teoría de la mecánica, física, matemática, ingeniería eléctrica, o de otra naturaleza, con la ventaja de que en estos toda ambigüedad es eliminada. Revisten especial importancia los lenguajes de programación de computadoras, y estas se definen considerando un conjunto de componentes léxicos, reglas gramaticales y una delimitación semántica [2], [4].

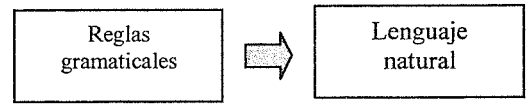

Fig. 1. Gramática y lenguaje.

Se define un *alfabeto A* como un conjunto finito de símbolos. Los elementos de un alfabeto constituyen las unidades básicas o *primitivas* de un lenguaje. Estos, a su vez, se agrupan en cadenas [5], [12].

Se denomina *cadena* o *palabra* sobre un alfabeto A, a una secuencia finita de elementos de A. [7],

D. *Gramática*

Una gramática G es un modelo lingüístico-matemático que describe el orden sintáctico que deben cumplir las frases bien formadas de un lenguaje [6], [13]. Una gramática se define formalmente como en (1),

$$G = (V_T, V_N, P, S) \quad (1)$$

donde:

V_T: conjunto finito de símbolos terminales del lenguaje.

V_N: conjunto finito de símbolos no terminales.

P: conjunto finito de reglas de producción.

$S \in V_N$: Símbolo distinguido o axioma inicial.

A partir del axioma S se reconocerán las secuencias de L aplicando sucesivamente las reglas de producción.

E. Gramática libre de contexto probabilístico

Chomsky clasificó las gramáticas de acuerdo a la forma de sus reglas de producción, así una gramática libre de contexto tiene sus reglas de la siguiente forma:

$$P: A \rightarrow \alpha$$

donde: $A \in V_N \;\; Y \; \alpha \in (V_N \cup V_T)$

El lado izquierdo consta de solo un no terminal, mientras que el lado derecho consta de una secuencia de terminales y no terminales [2], [6].

Una gramática libre de contexto probabilístico (**GLCP**) es una gramática libre de contexto en la cual cada regla tiene asignada una probabilidad. La probabilidad de un análisis sintáctico es el producto de las probabilidades de cada una de las reglas usadas en éste. De esta manera existen análisis que son más consistentes que otros. Nótese que las GLCP extienden las gramáticas libre de contextos incluyéndoles una función de probabilidad [8], [9].

Una GLCP se define entonces como una quíntupla $G = (V_T, V_N, P, S, £]$ donde $£$ es una función para asignar probabilidades a cada regla en P. La función $£$ expresa la probabilidad de que un no-terminal dado será expandido a la secuencia β. Una gramática probabilística tiene para cada regla P una probabilidad condicional:

$$A \rightarrow \beta \quad [p]$$

Luego de definir la gramática asignamos una probabilidad a cada regla de producción (ver Fig. 2). Consideremos el ejemplo siguiente tomado de [2]:

```
P: {
S          ⟶ NOMBRE VERBO          [1.0]
NOMBRE     ⟶ ADJ NOMBRE            [0.4]
NOMBRE     ⟶ ADJ NOMB-SING         [0.6]
VERBO      ⟶ VERB-SING ADVERBIO    [1.0]
ADJ        ⟶ El                    [0.25]
ADJ        ⟶ La                    [0.25]
ADJ        ⟶ Los                   [0.15]
ADJ        ⟶ Las                   [0.15]
ADJ        ⟶ Esos                  [0.10]
ADJ        ⟶ Pequeño/traviesa      [0.10]
NOMB-SING  ⟶ niño                  [0.50]
NOMB-SING  ⟶ niña                  [0.50]
VERB-SING  ⟶ estudia               [0.27]
VERB-SING  ⟶ corre                 [0.16]
VERB-SING  ⟶ juega                 [0.34]
VERB-SING  ⟶ salta                 [0.23]
ADVERVIO   ⟶ rápidamente           [0.45]
ADVERVIO   ⟶ despacio              [0.28]
ADVERVIO   ⟶ mucho                 [0.27]
}
```

Fig. 2 Gramática con probabilidades [2].

El término Hipertexto se refiere al Sistema de organización y presentación de datos basados en la vinculación de fragmentos textuales o gráficos a otros fragmentos, lo cual permite al usuario acceder a la información no necesariamente de forma secuencial sino desde cualquiera de los distintos ítems relacionados, como se muestra en la Fig. 3.

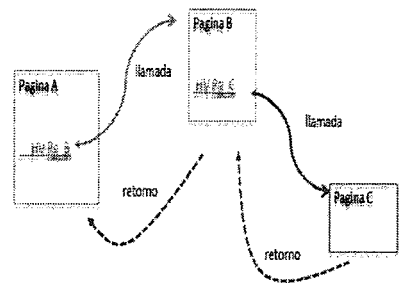

Fig. 3. Hipertexto.

F. Navegación hipertextual

Para comprender más claramente la naturaleza de la navegación por el hiperespacio de la información es preciso descomponer el problema como lo han intentado varios autores. Se distinguen en ese sentido, la clasificación que hacen Wright y Lickorish, a la que hace referencia [9], [10], navegación interna, esto es, la que forma parte del hipertexto, y externa, aquella posibilitada por las herramientas de navegación genérica, independiente del hipertexto. La navegación hipertextual, se refiere al proceso de transitar varias páginas cuando se visita la web. La cuestión fundamental aquí, sin embargo, es que La amplitud de las zonas de navegación libre indica el control que tiene el usuario, en este sentido intervienen las capacidades cognitivas del usuario, la estructuración del propio dominio y la destreza con la interfaz para la navegación que posea el sistema [Fischer & Mandl, 1990]. Cuanto mayor sea destreza con la interfaz (o más simple sea ésta) mayores podrán ser las zonas de navegación libre que pueda controlar un alumno dado [16]

G. Gramática probabilista de hipertexto

Una Gramática probabilista de hipertexto (GPH) se define como $G = (V_T, V_N, P, S, £)$ gramática regular (definida por una expresión regular) con una relación uno-a-uno entre V_N y V_T.

Hernández señala [8] que las sesiones de navegación de los usuarios inferidas desde los archivos log pueden representarse como una gramática probabilista de hipertexto. Cada símbolo no terminal de G corresponde a una página visitada y cada regla de derivación corresponde a un enlace entre las paginas, Así la regla A a B, significa la transición desde la página A hacia la página B. Se advierte en ese sentido, que este método consiste en que las cadenas generadas por la gramática con mayor probabilidad corresponden a los caminos preferidos por los usuarios [3],14].

La probabilidad de una cadena de la gramática es el producto de las probabilidades de las producciones usadas en su derivación [10].

H. Log de Servidor web

Esencialmente, los log de servidor constan de uno o más archivos de texto que son creados automáticamente y administrados por un servidor, en donde se almacena toda actividad que se hace sobre este. Cada servidor, dependiendo de su implementación y/o configuración podrá o no crear determinados logs. Uno de los logs más típicos son los logs de acceso de un servidor web, en donde se almacenan datos con la dirección IP, navegador, fecha y hora, etc., de cada acceso al mismo, lo que permite crear las estadísticas de un sitio web[3][11].

IV. METODOLOGÍA

La investigación se realizó con una muestra de archivos logs del servidor del laboratorio de computadoras de la Facultad de Ingeniería de Sistemas. A partir de estos archivos se construyó la gramática de hipertexto (GH), para ello se determinó la cantidad de veces que se aplicó una determinada regla gramatical y se realizaron cálculos estadísticos calculando la frecuencia en que aparecen las páginas en las sesiones de navegación. Para tal efecto cada símbolo no terminal de GH corresponde a una página y cada regla de derivación una transición de una página a otra, luego se le asignó las probabilidades a cada una de las reglas de producción. Para modelar las sesiones de navegación se construyó un grafo. Se desarrolló un programa en Java usando la plataforma Netbeans Ide 7.3.

Definición de la gramática: Se definió la gramática G identificando los símbolos terminales, no terminales y las reglas de derivación. A cada página identificada se le asignó un símbolo no terminal.

Definición de la gramática GPH: Se calcula la probabilidad de cada regla de producción asociada a la gramática.

Definición de sesiones de navegación: A partir de los log del servidor se construyó un conjunto P que contiene a las sesiones de navegación objeto de estudio.

Construcción de grafo de sesiones. Se modelo las sesiones mediante una estructura de grafo G.

Implementación: se construyó un prototipo para identificar las páginas más relevantes.

V. RESULTADOS

A. Definición de Gramática probabilística de hipertexto

A partir del conjunto P de sesiones de navegación obtenidas desde los archivos logs del servidor, se identificó las páginas involucradas, las que se representaron mediante símbolos no terminales de G.

$$V_T: \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$$

$$V_N: \{A_1, A_2, A_3, A_4, A_5, A_6, A_7\}$$

Las reglas de producción se visualizan en el grafo de la Fig. 4, en donde las aristas están etiquetadas con la probabilidad P_{ij} de derivar A_i a A_j.

El siguiente paso consistió en realizar el cálculo estadístico para asignar probabilidades (ver Tabla I). Luego de determinar la cantidad de veces que se enlazan las páginas se calculó todas las probabilidades medias y condicionadas y el número de veces que se aplicó una regla gramatical.

Luego se amplío la gramática G a una gramática GPH. Se distinguen las producciones en dos tipos:

Producciones de inicio, aquellas que comienzan con el axioma (S) y corresponde al inicio de una sesión.

Producciones transitivas aquellas que inician con un no terminal distinto a S, y corresponden a los enlaces entre páginas.

La Tabla II muestra la gramática con sus probabilidades.

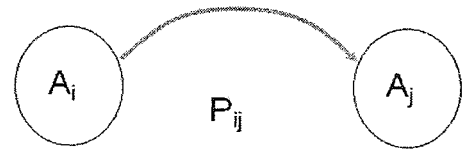

Fig. 4. Diagrama de transiciones.

B. Grafo de sesiones

Las reglas de producción se visualizan en el siguiente grafo (ver Fig. 5), en donde las aristas están etiquetadas con la probabilidad P_{ij} de derivar A_i a A_j

TABLA I.
DETERMINACIÓN DE LAS PROBABILIDADES

Regla	Ocurrencia de α	Ocurrencia de α → β	Probabilidad
S → a_1A_1	100	12	0.12
S → a_2A_2	100	3	0.03
S → a_3A_3	100	8	0.08
S → a_4A_4	100	9	0.09
S → a_5A_5	100	25	0.25
S → a_6A_6	100	33	0.33
S → a_7A_7	100	10	0.10
.....			
A_6 → a_2A_7	50	16	0.32
A_6 → a_2A_7	50	34	0.68
A_7 → F	15	15	1.00

TABLA II.
GRAMÁTICA CON PROBABILIDADES

1) S →	a_1A_1	(0.12)	14) A_2 →	a_5A_7	(0.32)
2) S →	a_2A_2	(0.03)	15) A_4 →	a_5A_5	(0.26)
3) S →	a_3A_3	(0.08)	16) A_3 →	a_2A_4	(0.63)
4) S →	a_4A_4	(0.09)	17) A_3 →	a_5A_6	(0.37)
5) S →	a_5A_5	(0.25)	18) A_5 →	a_3A_6	(0.23)
6) S →	a_6A_6	(0.33)	19) A_5 →	a_2A_1	(0.30)
7) S →	a_7A_7	(0.10)	20) A_6 →	a_2A_7	(0.32)
8) A_1 →	a_2A_3	(0.35)	20) A_1 →	F	(0.30)
9) A_1 →	a_4A_4	(0.12)	21) A_4 →	F	(0.57)
10) A_1 →	a_3A_7	(0.23)	22) A_5 →	F	(0.47)
11) A_4 →	a_2A_6	(0.17)	23) A_6	F	(0.68)
12) A_2 →	a_2A_3	(0.23)	24) A_7 →	F	(0.10)
13) A_6 →	a_4A_2	(0.45)			

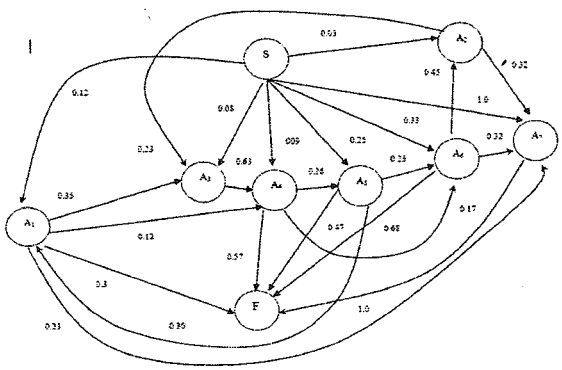

Fig. 5. Grafo de sesiones.

C. Determinación de probabilidades de sesiones

Se distinguieron las producciones en dos tipos: *producciones de inicio*, aquellas que comienzan con el axioma (S) y corresponde al inicio de una sesión, mientras que el resto, las que inician con un no terminal distinto a S, se denominan *producciones transitivas* y corresponden a los enlaces entre páginas[8].

A partir de la gramática se las cadenas que representan a las sesiones de navegación de los usuarios (ver Tabla III), se realizó un cálculo estadístico sobre una colección de sesiones de navegación que permitió obtener el número de veces que una página aparece como página inicial, el número de veces que aparece como página final y el número de veces que no es ni página inicial ni final. A partir de esta estadística se obtendrá un patrón.

TABLA III. SESIONES DENAVEGACION

ID	Sesión
1	$A_1 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow A_6 \rightarrow A_7$
2	$A_1 \rightarrow A_7$
3	$A_1 \rightarrow A_4 \rightarrow A_5 \rightarrow A_6$
4	$A_3 \rightarrow A_4 \rightarrow A_6$
5	$A_1 \rightarrow A_5 \rightarrow A_6 \rightarrow A_2 \rightarrow A_7$
6	$A_1 \rightarrow A_4 \rightarrow A_5 \rightarrow A_1$
7	$A_5 \rightarrow A_6 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5$
8	$A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow A_1 \rightarrow A_4 \rightarrow A_6$

Sea:

S_i : una sesión del conjunto P

A_i : Página involucrada en una sesión S_i

r_i : número de veces que una página A_i fue requerida en las sesiones de P

p_i : número de veces que una página A_i fue el primer estado en una sesión S_i de P.

u_i : número de veces que una página i fue el último estado en una sesión S_i de P.

t_{ij} : número de veces que una subsecuencia de dos páginas aparece en la sesión , o lo que es lo mismo, el número de veces que el enlace fue atravesado de P.

$\alpha > 0$: se pueden generar cadenas desde cualquier estado.

$\alpha = 0$: solo los estados que fueron primeros en las sesiones actuales tienen probabilidad mayor que cero de ser una producción de inicio.

$\alpha > 0$: se pueden generar cadenas desde cualquier estado.

$\alpha = 1$: la probabilidad de una producción de inicio es proporcional al número de veces que el correspondiente estado fue visitado. El nodo destino de una producción con más alta probabilidad corresponde al estado que fue visitado más a menudo.

N: $N \geq 1$ Determina la memoria del usuario cuando se navega por la red, es decir el número de URL

anteriores que puede influir en la elección del siguiente URL.

Si N=1, el resultado será la que se conoce formalmente como una cadena de Markov, que es un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende del evento inmediatamente anterior. Esta característica de falta de memoria recibe el nombre de propiedad de Markov así como se muestra en (2):

Si N=1 y $\alpha = 0$.

$$P(S \rightarrow \alpha_1 A_1) = \frac{\alpha * N\text{-}V\text{-}A_1}{N\text{-}T\_V} + \frac{\alpha * N\text{-}I\text{-}A_1}{N\text{-}T\text{-}I} \qquad (2)$$

donde:

N-V-A_1 : número de visitas a $A1 = 6$
N-I-A_1 : número de inicios de $A1 = 4$
N-T-V : número total de visitas = 36
N-T-I : número total de inicios = 8

A partir del axioma S se pueden elegir los símbolos entre A_1 y A_7, aplicando la formula se obtiene que la página A_1 tiene mayor probabilidad de ser seleccionada, siguiéndole en orden A_3, A_4, A_5 y A_6, A_2 y A_7 tienen la misma probabilidad (Tabla IV) lo cual se visualiza en la Fig. 6.

TABLA IV
ESTADÍSTICA DE ELECCIÓN DE PRODUCCIÓN A PARTIR DEL AXIOMA S

Produccion p	α	NVA_1	NTA_1	NIA_1	NTI		P(p)
S--> a1A1		0.5	6	36	4	8	0.33333333
S--> a_2A2		0.5	2	36	0	8	0.02777778
S--> a3A3		0.5	4	36	2	8	0.18055556
S--> a4A4		0.5	7	36	1	8	0.15972222
S--> a5A5		0.5	6	36	1	8	0.14583333
S--> a6A6		0.5	6	36	0	8	0.08333333
S--> a7A7		0.5	2	36	0	8	0.02777778

$$P(S \rightarrow \alpha_1 A_1) = \frac{0.5 * 6}{36} + \frac{0.5 * 4}{8} = 0.33 \qquad (3)$$

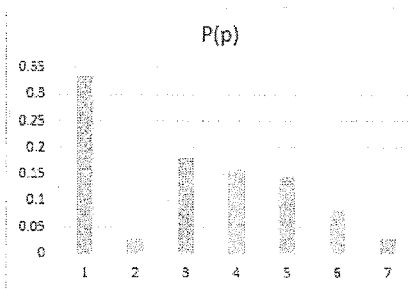

Fig. 6. Cuadro comparativo de probabilidad de página seleccionada a partir del axioma S.

A. Implementación

1. Ingreso y almacenamiento de los archivos logs del servidor(ver Tabla V). A partir de los archivos log del servidor se crea la gramática probabilística de hipertexto.
2. Limpieza de los datos almacenados. Se depuran los datos irrelevantes que no transfieren contenido.
3. Identificación de usuarios.
4. Identificación de sesiones y reconocimiento de páginas consideradas como peticiones.

TABLA V.
FORMATO DE LOG

Id sesión	Identificador de la sesión
Id Usuario	Identificador del usuario que inicia la sesión
IP	IP del usuario que inicia la sesión
Hora Inicio	Fecha y hora que inicia la sesión
Hora fin	Fecha y hora que finaliza la sesión
NPV	Número de páginas accedidas en el sitio web
NS	Número total de peticiones realizadas durante la sesión
BD	Total de Bytes transferidos durante la sesión

VI CONCLUSIONES

En el presente trabajo se describen los distintos pasos que se han llevado a cabo para obtener una medida *on-line* de adquisición del conocimiento en sistemas hipertexto. Para ello se parte de un análisis de las estrategias de navegación de los usuarios, y de un modelo para identifica r a partir de las sesiones de navegación las páginas más relevantes.

Muchas de las actuales medidas de aprendizaje se toman después de que el usuario haya navegado por el sistema hipertexto. Sin embargo, a la luz de la investigación desarrollada, se concluye que el aprendizaje de los contenidos de un sistema hipertexto es un proceso que se desarrolla permanentemente durante la navegación. Durante el el procesamiento de la información de un determinado texto del sistema, se debe integrar esos datos en la estructura de conocimiento adquirido con anterioridad acerca del resto de contenidos del sistema, de tal forma que el conocimiento de un usuario cambia tanto cuantitativa como cualitativamente.

Se pone de relieve la importancia de las gramáticas libres de contexto muy utilizadas en teoría de lenguajes, como instrumento para detectar las preferencias de los usuarios por las páginas web. Este instrumento permite que las empresas comerciales puedan perfeccionar sus sitios electrónicos para maximizar el impacto comercial en función de la conducta dinámica de sus visitantes.

El método permitió inferir a partir de los archivos log las sesiones de navegación de los usuarios representándolas mediantes gramática probabilística de hipertexto, de tal forma que las secuencias generadas o reconocidas por la gramática corresponden a las sesiones o caminos preferidos por los usuarios.

La principal dificultad de construir la gramática libre de contexto probabilística fue primero construir la gramática y luego asignar las probabilidades a cada regla de producción

El modelo desarrollado puede servir para calcular la probabilidad de alcanzar una página si el usuario está en una página dada

REFERENCIAS

[1] H. Contreras. "*Procesamiento del Lenguaje Natural basado en una gramática de estilos para el idioma español*", Universidad de los Andes, 2001.

[2] A. Cortez. "*Lenguajes y Traductores*", 1st ed. Lima: UCSS, 2013, pp. 34–36.

[3] A. Cortez Vásquez. "*Sistema de Aprendizaje de Patrones de Navegación Web Mediante Gramáticas Probabilísticas de Hipertexto*", *INGE CUC*, vol. 11, no. 1, pp. xx-xx,2015.

[4] J. E. Hopcroft."*Introduccion a la Teoria de Automatas, Lenguajes y Computacion*", 3rd ed. Madrid: Pearson, 2005, pp. 3–8.

[5] S. Russell and P. Norvig. "*Inteligencia Artificial, Un enfoque moderno*", 2nd ed. Mexico: Pearson, 2004.

[6] A. Aho, R. Sethi, and J. Ullman. "*Compiladores, principios, técnicas y herramientas*", 1st ed. México: Adison Wesley Longman, 1998.

[7] J. G. Brookshear. "*Teoría de la computación: lenguajes formales, autómatas y complejidad*", 1st ed. México: Pearson, 1993.

[8] A. Cortez, H. Vega, and J. Pariona. "*Procesamiento de lenguaje natural*", Rev. Investig. Sist. e Informática, vol. 6, no. 2, pp. 45–54, 2009.

[9] J. Hernández., M. Ramírez, and C. Ferri. "*Introducción a la minería de datos*", 2nd ed. España: Pearson, 2008.

[10] F. Iriarte. "*Patrones de navegación hipertextual en usuarios inexpertos de sexto grado*", Zo. próxima Rev. del Inst. Estud. Super. en Educ., vol. 1, no. 6, pp. 116–129, 2005.

[11] P. Alcivar Zambrano, F. Idrovo Chiriboga, and V. Macas Pizarro. "*Sistema de análisis de patrones de navegación usando minería web*", Escuela Superior Politécnica del Litoral, 2007.

[12] A. Cortez. "*Gramáticas probabilistas*", Revista Algorithmic Vol 4 Nª 1 2013, Pg 9-16 ISSN 2220-3982 Lima Perú.

[13] T.Pratt. "*Lenguajes de programación, Diseño e implementación*", Prentice Hall Hispanoamericana, 1988

[14] J. Sánchez. "*Estimación de gramáticas incontextuales probabilísticas y su aplicación en modelización del lenguaje*"; Universidad Politécnica de Valencia, Tesis para optar al grado de Doctor en Informática Valencia, 1999.

[15] L. Salmeron. "*Análisis de la adquisición del conocimiento en sistemas hipertexto a partir de las estrategias de navegación del usuario*", Dept de Psicología Experimental,, Universidad de Granada

[16] T.Perez. "*Estrategias pedagógicas con hipermedia: Limitar el acceso al hiperespacio con fines educativos*", Universidad del País Vasco (UPV-EHU), Depto. de Lenguajes y Sistemas Informáticos, Facultad de Informática, apdo. de correos 649, 20080 San Sebastián, Gipuzkoa.