Learning System of Web Navigation Patterns through Hypertext Probabilistic Grammars*

Sistema de Aprendizaje de Patrones de Navegación Web Mediante Gramáticas Probabilísticas de Hipertexto

DOI:http://dx.doi.org/10.17981/ingecuc.11.1.2015.07

Research Article - Reception Date September 23, 2014 - Acceptance Date: December 15, 2014

Augusto Cortez Vasquez
Master of Computer and Information Sciences, Universidad Nacional Mayor de San Marcos. Lima (Perú). acortezv@unmsm.edu.pe

To reference this paper:
A. Cortez Vasquez, "Learning System of Web Navigation Patterns through Hypertext Probabilistic Grammars" *INGE CUC*, vol. 11, no. 1, pp. 72-78, 2015. DOI: http://dx.doi.org/10.17981/ingecuc.11.1.2015.07

Abstract--- One issue of real interest in the area of web data mining is to capture users' activities during connection and extract behavior patterns that help define their preferences in order to improve the design of future pages adapting websites interfaces to individual users. This research is intended to provide, first of all, a presentation of the methodological foundations of the use of probabilistic languages to identify relevant or most visited websites. Secondly, the web sessions are represented by graphs and probabilistic context-free grammars so that the sessions that have the highest probabilities are considered the most visited and most preferred, therefore, the most important in relation to a particular topic. It aims to develop a tool for processing web sessions obtained from a log server represented by probabilistic context-free grammars.

Keywords-- Probabilistic Grammars, Navigation Patterns, Pattern Learning Hypertext Probabilistic Grammar, Hypertext, Information Retrieval.

Resumen-- Uno de los problemas que reviste real interés en el área de minería de uso de la web es capturar las actividades de los usuarios durante su conexión y extraer patrones de comportamiento que permitan definir sus preferencias con el fin de mejorar el diseño de futuras páginas adaptando las interfaces de los sitios web a los usuarios individuales., En esta investigación se pretende ofrecer en primer lugar una presentación de los fundamentos metodológicos del uso de lenguajes probabilísticos para identificar sitios web más relevantes o visitados. En segundo lugar se representa las sesiones web mediante grafos y gramáticas libres de contexto probabilísticas de tal forma que las sesiones que tengan mayor probabilidad son consideradas las más visitadas o más preferidas, por tanto las más relevantes en relación a un tópico determinado. Se pretende desarrollar una herramienta para procesamiento de sesiones web obtenidas a partir de log de servidor representado mediante gramáticas probabilísticas libres de contexto.

Palabras claves-- Gramáticas probabilísticas, patrones de navegación, aprendizaje de patrones, gramática probabilística de hipertexto, hipertexto, recuperación de información.

* Research paper deriving from the research Project "Categorización de textos mediante máquinas de soporte vectorial.". Funded by el Consejo Superior de Investigaciones UNMSM Lima – Peru. Starting date: January 2012. Ending date: December 2012.

INGE CUC vol. 11 no. 1, pp. 72-78. January - June, 2015 Barranquilla. ISSN 0122-6517 Printed, ISSN 2382-4700 Online
DOI: http://dx.doi.org/10.17981/ingecuc.11.1.2015.07

I. Introduction

What science and technology have achieved so far has been truly spectacular. We just have to look around to witness what the extraordinary power of our understanding of nature has helped us achieve. In the early eighties the first text mining endeavors were made with the inconvenience of needing a lot of human effort, but technological advances have allowed this area surprisingly progress in the last decade. Text mining is a multidisciplinary area based on information retrieval, data mining, machine learning, statistical and computational linguistics. Like most of the information (over 80%) is currently stored as text, it is believed that text mining has great commercial value. When users browse the Web and want to retrieve pages in relation to a particular concept, they should avoid many irrelevant pages; the objective is therefore to recover significant pages, that is, those that are authority on the subject.

There are two related concepts: most relevant and most visited pages. Therefore, we start from the premise that the most relevant pages are those that are most visited. This research captures, from the information contained in the server logs, the users' activities during their connection to the web and extracts behavioral patterns that will help understand the preferences of users' browsing, allowing adapting the interfaces of future pages to individual users. To achieve this purpose, a simple model of hypertext represented by graphs was used; that is, a representation of the users' navigation sessions which were inferred from the log files as a hypertext probabilistic grammar.

II. Objectives

A. General Objective: To obtain a tool to identify the preferences of users on the Web.

B. Specific Objectives:

1. To represent the web session by directed graphs.
2. To represent web sessions using hypertext probabilistic context-free grammars.

III. Conceptual Framework

A. Information Retrieval

Information retrieval *(IR)* is a term used in a very broad sense that requires precision; it is often vaguely defined, and in this context refers only to automated information retrieval systems. Contreras points out in her thesis [1] that:

> "Lancaster provides a definition: "*An information retrieval system does not inform (i.e., change the knowledge of) the users on the subject of their inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to their* request."

The referred automated systems require speed, consistency, accuracy and ease of use in the retrieval of relevant texts to satisfy users' queries.

B. Web Mining

There is a growing need to know how users interact with websites. W*eb mining* (WM) essentially concerns with the discovery and analysis of users' information on the web in order to uncover behavior patterns. Alcívar refers to the term WM as technology used to discover non-obvious information from data sources that include server logs [2].

C. Formal Language

Although a natural language is governed by grammatical rules that are already defined, they can be modified later (see Fig.1). This is an advantage for natural language, because this possibility enriches language, yet at the same time, it hinders its computer processing since it can be ambiguous and imprecise. On the contrary, a formal language is unambiguous and exact; it is a language developed by man to express situations that occur specifically in each area of scientific knowledge. Formal languages can be used to model a theory of mechanics, physics, mathematics, electrical engineering, or otherwise, with the advantage that in these languages all ambiguities are eliminated. Of particular importance are computer programming languages which are defined considering a set of lexical components, grammatical rules and semantic delimitation [3], [4].

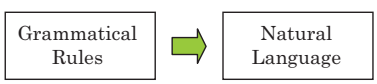

Fig. 1. Grammar and language
Source: Author

1. *Definition of Alphabet A:* An *alphabet A is defined* as a finite set of symbols. The elements of an alphabet constitute the basic units or *primitives* of a language. These, in turn, are grouped into strings [5], [6]
2. *Definition of Word:* It is called *string* or *word* on an alphabet A, to a finite sequence of elements of A [7]

D. Grammar

A grammar *G* is a linguistic and mathematical model that describes the syntactic order to be met by well-formed sentences of a language [8], [9]. A grammar is formally defined as in (1):

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

Where:

$\mathbf{V_{T:}}$ finite set of terminal symbols of language
$\mathbf{V_{N:}}$ finite set of non-terminal symbols
P: finite set of production rules
$\mathbf{S \in V_{N:}}$ distinguished symbol or initial axiom

From axiom *S,* sequences *L* are recognized by applying successively the rules on production grammar.

E. Probabilistic Context-Free Grammar

Chomsky classified grammars according to the form of its production rules, thus a context-free grammar has its rules as follows:

$$P: A \rightarrow \alpha$$

Where:

$$A \in V_N \quad Y\, \alpha \in (V_N \cup V_T)$$

The left side contains only a *non-terminal,* while the right side consists of a sequence of *terminals and non-terminals* [3], [8].

A probabilistic context-free grammar (PCFG) is a context-free grammar in which each rule is assigned a probability. The probability of a parsing is the product of the probabilities of each of the rules used in it. Thus there are analyses that are more consistent than others. Note that the PCFG extend the contexts-free grammars incorporating a probability function [2], [10].

A PCFG is then defined as fivefold $\mathbf{G = (V_T, V_N, P, S, £)}$ *where* £ is a function to assign probabilities to each rule in *P.* Function £ expresses the probability that a non-terminal given will be expanded to sequence ß. A probabilistic grammar has for each rule *P* a conditional probability.

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

1. *Assign Probabilities to Every Production Rule:* After defining the grammar, a probability is assigned in each production rule (see Fig. 2)

Consider the following example taken from [3]

P: {

S → NAME VERB	[1.0]
NAME → ADJ NAME	[0.4]
NAME → ADJ NAME - SING	[0.6]
VERBOVER → B → SING ADVERB	[1.0]
ADJ → he	[0.25]
ADJ → she	[0.25]
ADJ → the	[0.15]
ADJ → the	[0.15]
ADJ → those	[0.10]
ADJ → small/naughty	[0.10]
VERB → SING → boy	[0.50]
VERB → SING → girl	[0.50]
VERB → SING → estudies	[0.27]
VERB → SING → runs	[0.16]
VERB → SING → plays	[0.34]
VERB → SING → jumps	[0.23]
ADVERB → fast	[0.45]
ADVERB → slow	[0.28]
ADVERB → much	[0.27]

}

Fig. 2. Grammar with probabilities
Source: [3]

The term *hypertext* refers to the organization system and presentation of data based on the linking of text fragments or graphics to other fragments, allowing the user to access information not necessarily sequentially but from any of several related items, as shown in the Fig. 3.

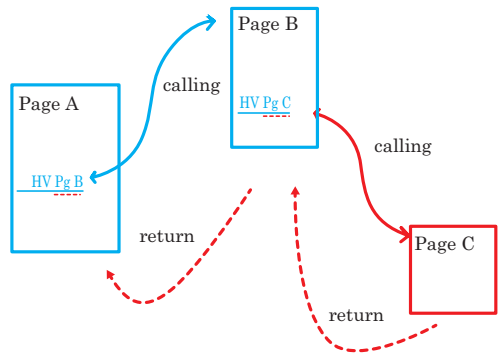

Fig. 3. Hypertext
Source: Author

F. Hypertext Navigation

To understand more clearly the nature of navigation through the information hyperspace, it is necessary to decompose the problem as several authors have tried. In this sense, there is a discrepancy in the classification made by Wright and Lickorish, with the references [2] [11]; internal navigation, is what is part of the hypertext; and external, the one made possible by generic navigation tools, independent of hypertext. Hypertext navigation refers to the process of moving through multiple pages when you visit the Web.

G. Hypertext Probabilistic Grammar

A hypertext probabilistic grammar (HPG) is defined as $\mathbf{G = (V_T, V_N, P, S}$, £) and a regular grammar, defined by a regular expression, has a one-to-one relationship between $\mathbf{V_N}$ and $\mathbf{V_T}$.

Hernandez noted [2] that the sessions of users' navigation inferred from the log files can be represented as a hypertext probabilistic grammar. Each non-terminal symbol belonging to *G* corresponds to a visited page each derivation rule corresponds to a link between pages. Thus, the rule *A* to *B* means the transition from page *A* to page *B.* In this regard, it is important to note that this method consists of the fact that the strings generated by the grammar with the highest probability correspond to the users' preferred paths [12].

The probability of a grammar string is the product of the probabilities of the productions used in its derivation [11].

H. Web Server Logs

Essentially server logs consist of one or more text files that are automatically created and managed by a server, where all activity that is done on it is stored. Each server, depending on its implementation and / or configuration may or may not create a particular log. One of the most typical logs is the access log of a web server that stores in each access and at the same time data such as an IP address, browser, date and time, etc., allowing the creation of the website statistics [2] [13].

IV. Methodology

The research was conducted with a sample of the server log files from the computer lab of the Systems Engineering Faculty. Using these files, a hypertext grammar (HG) was built; for this purpose, It was determined the number of times a particular grammatical rule was applied and statistical calculations were done by estimating the frequency in which the pages appear in the navigation session. For this purpose, each non-terminal symbol of HG corresponds to a page and each derivation rule to a transition from one page to another; then the probabilities of each of the production rules were assigned. To model the navigation sessions, a graph was constructed; and finally, a Java program was developed using the platform NetBeans IDE 7.3.

A. Grammar Definition

Grammar *G* was defined identifying the terminals, non-terminals symbols and derivation rules. A non-terminal symbol was assigned to each identified page.

B. Definition of Grammar H PG

The probability of each production rule associated to grammar is calculated.

C. Definition of Navigation Sessions

Using the server logs, a *set P* containing the navigation sessions was constructed.

D. Session Graph Construction

Sessions were modeled by a graph structure *G*.

E. Implementation

A prototype was constructed to identify the most relevant pages.

V. Results

A. Definition of Hypertext Probabilistic Grammar

Using the navigation session set *P* obtained from the server log files, the identified pages were represented by non-terminals symbols of *G*.

Production rules are displayed in Fig. 4, where the line is labeled with the probability P_{ij} resulting from derivation A_i to A_j

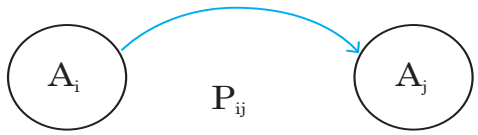

Fig. 4. Transitions Diagram
Source: Author

The next step was to perform statistical calculations to assign probabilities (see Table I). After determining the number of times pages were linked, it was calculated all middle and conditional probabilities and the number of times that a grammar rule has been applied.

Table I. Determination of Probabilities

Rule	Ocurrence of α	Ocurrence of α→ β	Probability
S→A1A1	100	12	0.12
S→A2A2	100	3	0.03
S→A3A3	100	8	0.08
S→A4A4	100	9	0.09
S→A5A5	100	25	0.25
S→A6A6	100	33	0.33
S→A7A7	100	10	0.10
.....			
A6→A2A7	50	16	0.32
A6→A2A7	50	34	0.68
A7→F	15	15	1.00

Source: Author

Then grammar *G* was expanded to a grammar *HPG*. The productions are distinguished into two types:

1. *Start Productions*: those that begin with axiom (S) and represent the start of a session.
2. *Transitive Productions:* Those that start with a non-terminal different from *S* and correspond to the links between pages [2].

Table II shows the grammar with its probabilities:

Table II. Grammar with Probabilities

1) S → a1A1(0.12)	14) A2 → a5A7 (0.32)
2) S→ a2A2 (0.03)	15) A4→a5A5 (0.26)
3) S → a3A3 (0.08)	16) A3→a2A4 (0.63)
4) S → a4A4 (0.09)	17) A3 → a5A6 (0.37)
5) S→ a5A5 (0.25)	18) A5→a3A6 (0.23)
6) S→ a6A6 (0.33)	19) A5→ a2A1 (0.30)
7) S→ a7A7 (0.10)	20) A6 → a2A7 (0.32)
8) A1 →a2A3 (0.35)	20) A1→ F(0.30)
9) A1→ a4A4 (0.12)	21) A4 → F(0.57)
10) A1→a3A7 (0.23)	22) A5 → F(0.47)
11) A4 →a2A6 (0.17)	23) A6→F(0.68)
12) A2→ a2A3 (0.23)	24) A7→ F(0.10)
13) A6 →a4A2 (0.45)	

Source: Author

B. Sessions Graph

Production rules are shown in the following graph (see Fig 5), where the lines are labeled with the probability P_{ij} resulting from derivation A_i to A_j

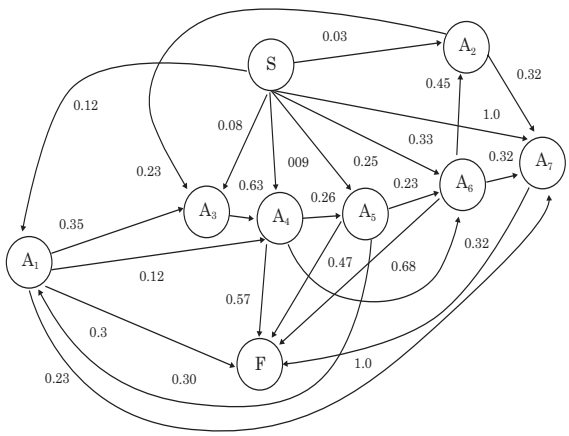

Fig. 5. Session Graph
Source: Authors

C. Determination of Sessions Probability

As already established, the productions were distinguished into two types: *production start* and *transitive productions.*

Using grammar strings, representing users' navigation sessions (see Table III), a statistical calculation was made over a collection of navigation sessions that yielded the number of times a page appears as initial page, the number of times it appears as the final page, and the number of times that is not initial or final page. From this statistics, a pattern is obtained.

Table III. Sessions of Navigation

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	$A_4 \rightarrow A_5 \rightarrow A_6 \rightarrow A_2 \rightarrow A_7$
6	$A_1 \rightarrow A_4 \rightarrow A_5 \rightarrow A_1$
7	$A_5 \rightarrow A_6 \rightarrow A_2 \rightarrow A_3 \rightarrow A_4 \rightarrow A_5$
9	$A_3 \rightarrow A_4 \rightarrow A_5 \rightarrow A_1 \rightarrow A_4 \rightarrow A_6$

Source: Author

Where:

S_i a session of set *P*

A_i a page involved in a session S_i

r_i the number of times a page A_i was requested in the sessions *P*

p_i the number of times a page A_i was the first state in a session S_i *of* P.

u_i the number of times a page *i* was the last state in a session S_i of P

t_{ij} the number of times a subsequence of two pages appears on the session, or what is the same, the number of times the link was crossed of *P*

$\alpha > 0$ strings can be generated from any state

$\alpha = 0$ only states that took the top places in the current sessions have a probability higher than zero to be start production

$\alpha = 1$ the probability of a start production is proportional to the number of times the corresponding state was visited. The destiny node of a production with higher probability corresponds to the state that was visited more often

N: $N \geq 1$ determines the user's memory when navigating the Web, that is, the number of previous URLs that may influence the choice of the following URL

If *N = 1,* the result will be what is formally known as a Markov string, which is a special type of discrete stochastic process in which the probability of an event occurring depends on the immediately preceding event. This lack of memory feature is called Markov property as shown in (2) and solved in (3):

Si N=1 y α **= 0**

$$P(S \rightarrow a_1 A_1) = \frac{\alpha * N - V - A_1}{N - T\_V} + \frac{\alpha * N - I - A_1}{N - T - I} \quad (2)$$

Where:

N-V-A_1: number of visits to A1 = 6

N-S-A_1: number of starts from A1= 4

T-N_V: total number of visits = 36

T-N-S: total number of starts = 8

$$P(S \rightarrow a_1 A_1) = \frac{0.5 * 6}{36} + \frac{0.5 * 4}{8} = 0.33 \quad (3)$$

Using axiom *S,* symbols between A_1 and A_7 can be chosen. Applying the formula, it yields that page A_1 has higher probabilities to be selected, followed by A_3, A_4, A_5 and A_6; A_2 and A_7 are equally probable (Table IV).

Table IV. Production Choice Statistics From Axiom S

Producción 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

Source: Author

This probability is shown in Fig 6

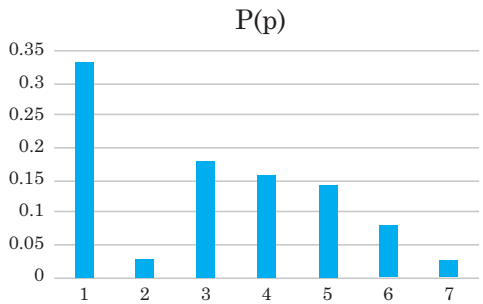

Fig. 6. Comparative table of selected page probability from axiom S.
Source: Author

D. Implementation

1. Entry and storage of log files on the server (see Table V). Using the server log files, a hypertext probabilistic grammar is created.
2. Cleaning of the stored data. Irrelevant data that do not transfer content is debugged.
3. Users' identification.
4. Identification of sessions and recognition of pages considered as petitions.

Table V. Log Format

ID session	Session identifier
ID User	Identifier of user who logs in
IP	IP of user who logs in
Start time	date and time of user's log in
End Time	Date and time of user's logout
NPV	number of accessed pages in the website
NS	total number of requests made during the session
BD	Total transferred bytes during the session

Source: Author

VI. Conclusions

This research emphasizes the importance of context-free grammars (widely used in language theory) as a tool to detect the preferences of website users. This instrument allows commercial companies to improve their websites to maximize the business impact in terms of the dynamic behavior of its visitors.

The method allowed inferring, from the log files, users' navigation sessions representing them through hypertext probabilistic grammar, so that the sequences generated or recognized by the grammar correspond to preferred users' sessions or paths.

The main difficulties of building probabilistic context free grammar were, first, to build the grammar, and then assign the probabilities in each production rule.

The developed model can be used to calculate the probability of reaching a page if the user is on a given page.

There are many tools for websites analysis and statistics that together with web servers provide really good data views and summaries to generate reports and graphs, but do not allow other activities like drawing patterns on user behavior or explore the relevance and ranking of pages. Our analysis of web sessions modeled by context-free grammars is equated with the ability to extract and use information from sessions to learn users' behavior patterns. The patterns obtained from past uses can determine web customizing, meaning by customization any action that adapts the Web to suit the user.

Computational linguistics is not only a method but a paradigm with a computational scheme of language processing that has led to a wide variety of applications, in this case, to the learning of navigation patterns.

References

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

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

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

[4] J. E. Hopcroft, *Introducción a la Teoría de Autómatas, Lenguajes y Computación*, 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. Cortez, "Gramáticas probabilistas", *Revista Algorithmic* Vol 4 No. 1, 2013, Pg 9-16. ISSN 2220-3982. Lima, Perú.

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

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

[9] T.Pratt, *Lenguajes de programación: Diseño e implementación*; Prentice Hall Hispanoamericana, 1988.

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

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

[12] 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.

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