Winter 12-5-2004

Markov Chain-based Clustering Analysis of Customers and WebPages

Changshou Deng

Pie Zheng

Yanling Yang

Bingyan Zhao

Markov Chain-based Clustering Analysis of Customers and WebPages

Changshou Deng [1,2], Pie Zheng[1], Yanling Yang[2], Bingyan Zhao[2]

[1]Institute of Systems Engineering, Tianjin University, Tianjin 300072, China
[2]Jiujiang University, Jiangxi 332005, China
dengtju@eyou.com

ABSTRACT

This paper focuses on users' behavior towards an EC website. A novel Markov Chain-based way combining the web log file information and the topology of an EC website is presented to rank a user's interest in a WebPage. Then a URL-USERID relevant matrix is set up, with URL taken as a row and USERID as column, and each element's value is the probability of a user to access a WebPage when time goes infinitely. The similarity of each column vector can be used to cluster customers, and relevant web pages can be found from the similarity of each row vector. The knowledge discovered by this dynamic model can be fairly helpful to the design and maintenance of a website, to provide personalized service, and can be used in an effective recommending system of an EC website etc.

Keywords: Markov chain, electronic commerce, customers clustering, WebPages clustering

1. INTRODUCTION

With the high development of Internet, Web-based applications have been warmly welcomed. For instance, Remote Education, Network Bank, and in particular, the Electronical Commerce(EC) based on Internet have been increasing quickly. When a customer is browsing an EC website, his behavior information will be collected by the Web Server automatically, and stored in a weblog file. On the other hand, the topology of website's WebPages indicates some relationship between WebPages. If knowledge can be discovered from the weblog and the structure of an EC website, it can do good to the design and maintenance of a website, and provide personalized service, and Customer Relationship Management (CRM) can be conducted effectively. In a word, it can lead an EC website to succeeding in today's competitive business market. Customers clustering and WebPages Clustering are one of the most important issues in discovering knowledge from the Web. They are the preconditions for most web-based applications.

Intuitively, when time goes infinitely, users' ultimate probability of browsing each WebPage is a constant, and it should reflect the worthiness of each page to the users. This hypothesis is proved in this paper, and this has potential use in web-based data mining. Firstly, a user's behavior is abstracted as a homogeneous Markov Chain dynamically, which then is proved to be ergodic. Secondly with the help of the ergodic Markov Chain, a novel way to evaluate a user's interest in a WebPage combining the weblog and the structure of an EC website is presented. Lastly, a URL-USERID relevant matrix is set up, where URL is taken as a row and USERID is taken as column, and each element' s value is the probability of a user to access a WebPage when time goes infinitely. Relevant WebPages and similar customers can be clustered efficiently by using the URL-USERID matrix.

This paper is organized as follows: Section 2 presents the necessary mathematical knowledge. The model of a user's behavior is presented in Section 3.
Section 4 presents the way to cluster the WebPages and customers. Section 5 concludes this paper and discusses the direction of the future research.

2. MATHEMATICAL BACKGROUND

Definition 2.1 Suppose the state space $\{S=i_0,i_1,i_2,i_3...\}$ of the random variables $\{x(t),t=0,1,2,..,\}$ is discrete; for only time t and any state space $i_0,i_1,i_2\ldots,i_{i-1},i,j$, the equation:
$P\{x(t+1)=j,x(t)=i,x(t-1)=i_{t-1},\ldots,x(0)=i_0\}=P\{x(t+1)=j|x(t)=i\}$ is always true, then $\{x(t),t=0,1,2,..,\}$ is named a Markov chain.$P\{x(t+1)=j|x(t)=i\}$(abbreviated to $p_{ij}(t)$) represents the probability of transition from the state i to j at the time t. If the transition probability $p_{ij}(t)$ is independent of t, i.e., for any $i,j\in S$ and any different time t_1,t_2, $p_{ij}(t_1)=p_{ij}(t_2)$, then this Markov chain is called homogeneous.

Definition 2.2 $P=(p_{ij})_{n\times n}$ is one step transition probability matrix of this homogeneous Markov chain, where

(1) $p_{ij}\in(0,1)$ $\quad 1\leqslant i,j\leqslant n$

(2) $\sum_{i=1}^{n} p_{ij}=1 \quad j=1,2,3,...,n$

Definition 2.3 Suppose $\{x(t),t=0,1,2,\ldots,\}$ is a homogeneous Markov Chain on the state space S,the probability $p_{ij}(k)=P\{x(t+k)=j \mid x(t)=i\},i,j\in S$ represents the system will transit from state i to state j after Kth step which is called k-step probability. And the matrix $P=(p^{(k)}_{ij})_{n\times n}$ is named k-step transition matrix, where

(1) $p^{(k)}_{ij}\in(0,1)$ $\quad 1\leqslant i,j\leqslant n$

(2) $\sum_{i=1}^{n} p_{ij}(k)=1 \quad j=1,2,3,\ldots,n$

3. MODELING A USER'S BROWSING BEHAVIOR

For a virtual user A, $R=\{r_1,r_2,...,r_n\}$ denotes the set of URLS of an EC website, and is considered as the space state. At time t,the user A is browsing the URL r_i in the probability $p_i(t)$,and will leave for the URL r_j in probability p_{ij}. The user's surfing action on the web can be abstracted as a homogeneous Markov Chain, for $p_{ij}(t)$ is independent of time t.

Theorem 1 Suppose $\{x(t),t=0,1,2..\}$ is a homogeneous Markov Chain on the state space S,for $m\geq 1$, $n\geq 1$ and $k=m+n$, the following is true.

$$p_{ij}(k)=\sum_{r\in S} p_{ir}(m)p_{rj}(n) \qquad i,j\in S \quad (3-1)$$

Eq.(3-1) is the famous Champaman-Kolmogorov equation. This equation denotes that if a user wants to move from space i to space j by $(m+n)$ steps, he must move to any space r by m steps and then move to space j by n steps.

The matrix form of this theorem is $P_{m+n}=P_m \cdot P_n$.Generally $P_n=P_1^n$ can be got. From the equation, the n-step transition matrix of homogeneous Markov Chain can be calculated by one step transition Matrix.

Theorem 2 At time t, a Markov Chain's distribution vector $p(t)$ is totally determined by the initial distributed vector $p(0)$ and one step transition matrix, that is $p(t)=p(0)p^t$.

3.1 Analysis of user's interest in WebPages

How to evaluate the importance of a WebPage and a user's interest in the WebPage is crucial to web data mining. As the user's interest in a WebPage is concerned, it should at least include the following metrics:

Relevance: It's an intuitive metric, denoting how much the user cares about the WebPage. It has been used as a useful metric in today's most Web Search Engine.

Authority: The authority metric means how many WebPages refer to the WebPage r. Moreover, the Web resources referred by higher quality resources should be assigned with higher authority.

Hub: The hub metric means how many WebPages are pointed by the Web resource r. Moreover, the Web resources pointing to higher quality resources should be assigned with higher hub.

Novelty: It denotes that how much a WebPage differs from other WebPages.

3.2 User's Browsing Behavior

Suppose that a user is browsing the WebPage r_i of an EC website, his behavior is as follows. At time t, he is browsing the WebPage r_i, and at next time $t+1$, he may have one of the following options:

(1) Continue browsing the WebPage r_i;

(2)Click one hyperlink in r_i to jump to a new URL r_j;

(3)Press the "BACK" button in the browser so as to return to the last browsing WebPage;

(4)To enter a new URL of the website.

Towards each options above, the probability of a user's tendency is measured as $\alpha, \beta, \gamma, \delta$ respectively, which satisfies the condition $0<\alpha,\beta,\gamma,\delta<1$, $\alpha+\beta+\gamma+\delta=1$

Definition 3.1 Web Graph

The linkage structure graph of the WebPages, $G=(V,E)$, is a directed graph, where V is the set of nodes ($|V|=n$), a node corresponds to a WebPage r_i ,for instance,in R; and E is the set of edges, $E=\{(v_i,v_j)|v_i,v_j\in V$ and WebPage r_i points to r_j through hyperlink$\}$.

Definition 3.2 Tendency Matrix

The Tendency Matrix of resources R is $U=(u_{ij})_{n\times n}$,and

$$u_{ij}=\begin{cases}\alpha & i=j\\ \beta & (v_i,v_j)\in E\\ \gamma & (v_j,v_i)\in E\\ \delta & otherwise\end{cases} \qquad (3-2)$$

It is noted that the Tendency Matrix here has synthesized the four metrics(α , β , γ , δ) mentioned above (Relevance,Authority,Hub and Novelty).With the tendency matrix being normalized, one-step transition probability matrix in the Markov Chain for a user's surfing behavior on the EC Website can be got.

Definition 3.3 One Step Transition Matrix

One step transition matrix of a virtual user A on the resource R can be defined as $P=(p_{ij})_{n\times n}$,and each element of P is as follows:

$$p_{ij}=\frac{u_{ij}}{\sum_{j=1}^{n} u_{ij}} \qquad i=1,2,...,n \qquad (3-3)$$

The distribution vector p(t) can be calculated by theorem 2 and definition 3.3.It's obvious that α , β , γ , δ actually reflect the relative importance, in the user's opinion, of relevance,authority,hub and novelty metrics respectively. And thus α is called relevance parameter, β authority parameter, γ hub parameter and δ novelty parameter.

3.3 Stationary distribution of a Markov Chain

Definition 3.4 Suppose that S is the state space of one homogeneous Markov Chain {x(t),t=0,1,2…},if for any i ,j $\in$ S,there exists a constant p_j,which is independent of i, making the following equation always true .then the Markov Chain is ergodic.

$$\lim_{t \to \infty} p_{ij}(t) = p_j$$

Theorem 3
A homogeneous Markov chain, on the state space $S=\{s_1,s_2,..s_n\}$, is ergodic, if for any $s_i,s_j \in S$,there exists a positive integer *k*, the state can transit from s_i to s_j via the edge between s_i and s_j in a positive probability within *k* steps, that is

$$\lim_{k \to \infty} p_{ij}(k) = p_j$$

where $p_j(j=1,2,\ldots n)$ is the unique solution to the following equation:

$$p_j = \sum_{i=1}^{n} p_i p_{ij} \qquad (j = 1,2,3,...,n) \qquad (3-4)$$

which satisfies the restricted (3-5) conditions.

$$p_j > 0 \quad (j = 1,2,..., n)$$

$$\sum_{j=1}^{n} p_j = 1 \qquad (3-5)$$

Theorem 3 denotes that when a homogeneous Markov Chain is ergodic, the distribution vector *p(t)* is only determined by one step transition matrix, independent of time t and of it's initial vector *p(0)*.

Definition 3.5 Positive Stochastic Matrix
Suppose $P=(p_{ij})_{n\times n}$ is an n-dimensional non-negative matrix, if it satisfies the following two conditions:①the sum of all elements in each row is 1 ②if there exists positive integer *k*,making $P^k>0$,then the matrix P is called positive stochastic matrix.
It is true without proof that each element in one-step matrix is positive and the sum of each row is 1, respectively. So one step matrix is positive stochastic matrix.

Theorem 4 A Markov Chain on the finite State Space $S=\{s_1,s_2,s_3...,s_n\}$ is ergodic if and only if it's transition matrix is positive stochastic matrix.

In a word, with the help of formulas (3-4) and (3-5) ,the ultimate probability can be got easily. Suppose the user A is rational and experienced enough in an EC website, the probability of browsing the WebPage r_i should be proportional to its worthiness. In real life, user usually spends more time with the WebPages he cares more. Firstly,since the user has no knowledge about the value of each WebPage, he browses WebPages randomly. The user becomes more and more experienced while more and more WebPages he browsed. So his judgment on the value of WebPages becomes more and more accurate. When time goes infinitely, the ultimate probability of browsing each WebPage should reflect the worthiness of each WebPage to the user accurately. From Theorem 3,the ultimate distribution vector *p(t)* of a ergodic Markov chain is independent of the initial distribution vector *p(0)*, but totally determined by the one step transition probability matrix. So given an EC Website WebPages R, the transition probability matrix can be set up through definition 3.3,and the ultimate distribution vector can be calculated based on Theorem 3. The ultimate distribution vector is user's interest in WebPages. The group of equations in Theorem 3 can be solved using "Gauss method" .The ultimate constant probability of user's browsing the web can be used to discover the similar customers and relevant WebPages.

4. CUSTOMERS CLUSTERING AND WEBPAGES CLUSTERING

4.1 Data Preparation

An EC website and the weblog file will be used inknowledge discovering. The weblog file complies with W3C standard. It at least includes the following items:
①user's IP address ;②Access time;③Request method; ④the URL of the accessed WebPage;⑤the transferred protocol;⑥return code;⑦the bytes transferred. Before discovering knowledge from weblog,data preparation is needed. It has two stages: data cleaning and transaction identification. Suppose L is a weblog,and l belongs to L.The items of l are,respectively,l.ip,l.uid and l.url.The algorithm to find a transaction is as fellows:
Step1: to partition weblog by the user's IP address,that is to determine each visitor's browsing set;
Step2:to calculate the times of each customer's browsing each WebPage in every set;
Step3: to sort the browsing set by IP address, and then the users' browsing set can be got.

4.1.1 Topology of an EC website

A website's topology is a directed graph, and customer's browsing pattern is its subgraph. The customers who have similar visiting subgraphs are similar. Suppose that an EC website's topology is as the following directed graph: *G*=(N, Np, E),where N is the set of nodes,N_p($N_p=\{Node \in N,\{userid,p\}^n\}$ n>=1)stores the customers' USERID and the ultimate probability to browse each Webpage, which can be calculated by theorem 3.N_p is the attribute of node; E is the set of edges. From the set N of G, all the urls of the EC website can be got. And from the N_p, each node's userid and his ultimate probability to visit this WebPage can be found. According to this, A URL-USERID relevant matrix $M_{m\times n}$ can be set up as follows:

$$M_{m\times n} = \begin{bmatrix} p_{11}, p_{12}, \cdots, p_{1n} \\ \vdots, \ \vdots, \ \cdots, \ \vdots \\ p_{m1}, p_{m2}, \cdots, p_{mn} \end{bmatrix}$$

Analysis of the matrix can be made as follows.
(1) Each element p_{ij} of this matrix presents the probability of the jth customer's to visit the ith URL when time goes infinitely.
(2) The *ith* row vector represents that the probabilities

of the customers to visit the ith URL when time goes infinitely.

(3) The *jth* column vector represents the probabilities of the jth user to visit each URL when time goes infinitely.And the other form of matrix $M_{m\times n}$ can be written as $M_{m\times n}=(p_1,p_2,...,p_n)$.

At the same time, URL-USERID relevant matrix can be represented by row vectors, and each row vector is of a constant. So a conclusion can be drawn that the row vectors represent the structure of the EC website and also imply the common browsing pattern; the column vectors represent the types of customers and the personalization browsing pattern. The similarity of each row vector represents the relevant WebPage and the similarity of each column vector represents the similar customers. The similarity can be judged by the definition4.1.The larger value of the angle of between vectors a and b,the more similar they will be.In practical use,a threshold can be set to determine the similarity.

Definition 4.1 Suppose vectors $a=(a_1,a_2...,a_n)$,and $b=(b_1,b_2,...,b_n)$,then, the cosine of the angle between vectors a and b is

$$\cos(a,b)=\frac{a\cdot b}{|a|\cdot|b|}=\frac{(a_1\cdot b_1+a_2\cdot b_2+...+a_n\cdot b_n)}{\sqrt{{a_1}^2+{a_2}^2+...+{a_n}^2}\cdot\sqrt{{b_1}^2+{b_2}^2+...+{b_n}^2}}$$

4.1.2 One-Step Matrix and Ultimate Probability

Four metrics described in Section 3 can be used to evaluate the interest of a user in a WebPage during given time t. Times of the user's visiting the WebPage represent the relevance of a user to a WebPage.The inverse of a url's Indegree(the number of hyperlinks pointed to the URL) represents the authority; the Outdegree(the number of hyperlinks pointing to the other URL) of the url represents the hub of a url; the novelty can be judged by the inverse of the Indegree of a url approximately. The detailed method is as following.

Firstly, the trend matrix $U=(u_{ij})_{n\times n}$(at time t) is formed by the following

$$u_{ij}=\begin{cases}\alpha & i=j\\ \beta & (v_i,v_j)\in E\\ \gamma & (v_j,v_i)\in E\\ \delta & otherwise\end{cases}$$

Secondly, the following is defined

.

$$a=p_1\times\frac{l_i^t.time}{\sum_{i=1}^{m}l_i^t.time}\qquad b=p_2\times\frac{outgree(v_j)}{|E|}$$

$$c=p_3\times\frac{in\deg ree(v_i)}{|E|}\qquad d=p_4\times\frac{1}{in\deg ree(v_i)}$$

Third, normalizations of a,b,c,d is as follows,

$$a=\frac{a}{\sqrt{a^2+b^2+c^2+d^2}}\quad b=\frac{b}{\sqrt{a^2+b^2+c^2+d^2}}$$

$$c=\frac{c}{\sqrt{a^2+b^2+c^2+d^2}}\quad d=\frac{d}{\sqrt{a^2+b^2+c^2+d^2}}$$

respectively.

let

$$\alpha=a^2\qquad\beta=b^2\qquad\gamma=c^2\qquad\delta=d^2$$

It can be noticed that

(1) p_1,p_2,p_3 and $p_4(p_1,p_2,p_3,p_4>0)$ are coefficients.

(2) If the outdegree or indegree of a node is zero, then let the outdegree or indegree is 1.

(3) It is obvious that a,b,c,d can be adapted by the coefficients p_1,p_2,p_3,p_4.

4.2 Customer clustering and WebPages clustering

With the help of the above URL-USERID matrix, clustering can be done in an EC website.

4.2.1 Customers Clustering

As described above, the column vector $(p_1,p_2,...,p_n)$ of relevant matrix $M_{m\times n}$ is the personalized browsing pattern. The customers who have similar browsing patterns can be clustered as similar customers. The cosine of the angle between the two column vectors can be calculated by definition 4.1 to cluster similar customers. In order to cluster the similar customers, a matrix Q, of which each element is the cosine of the angles between two column vectors, can be formed. It's obvious that the matrix Q is symmetrical, i.e. $Q=(q_{ij})_{n\times n}$,$q_{ij}=q_{ji}$,$q_{ii}=0$.The matrix Q is as fellows:

$$Q=\begin{bmatrix}0 & q_{12} & ... & q_{1n}\\ q_{21} & 0 & ... & q_{2n}\\ ... & ... & 0 & ...\\ q_{n1} & q_{n2} & ... & 0\end{bmatrix}$$

A threshold is given by formula(4-1)

$$\theta=\frac{2\times\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}q_{ij}}{n\times(n-1)}\qquad(4-1)$$

The threshold is the average of all elements in matrix Q,and it can be adapted to satisfy the requirements in practical usage.For any $q_{ij}\in Q$ ($1\leqslant i\leqslant n, i<j\leqslant n$) ,if $q_{ij}>\theta$,then the ith customer and the jth customer can be clustered into the same group. The more the threshold is, the more accurate the clustering can be.

All in all, any two customers can be clustered if the similarity of their ultimate probability to each WebPage vector satisfies the given threshold .In our daily life, as

time goes by, the similar customers will come into being after their long browsing and transaction .The method proposed in this paper can track this procedure and grasp the trend of this procedure.

4.2.2 WebPages clustering

As has been said that the row vectors represents what all the customers have browsed each WebPage of an EC website. If all the customers have the similar experience in some WebPages, then these WebPages will be thought as relevant WebPages. A matrix Q' is introduced, elements are the cosines of the angles of between row vectors. For ease usage, let q_{ij} be 0 (if i=j). The matrix Q' is as follows:

$$Q' = \begin{bmatrix} 0 & q'_{12} & \cdots & q'_{1m} \\ q'_{21} & 0 & \cdots & q'_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ q'_{m1} & q'_{m2} & \cdots & 0 \end{bmatrix}$$

The threshold can also be calculated by formula (4-1) and adapted according to the requirements. For any $q'_{ij} \in Q'$ ($1 \leqslant i \leqslant m, i < j \leqslant m$),if $q'_{ij} > \theta$,then the ith URL and the jth URL($j \in (i,m]$)can be considered as one group.

In a word, if all the customers browse a given WebPage, their ultimate probability of the WebPage is a constant. The row vectors of URL-USERID are also constants. For two different WebPages, if all the customers' probabilities of browsing them are the same or similar, it is reasonable that they are relevant WebPages.

4.3 Algorithm analysis

The method proposed in this paper to cluster relevant WebPages and similar customers is based on Markov Chain ,which combines the weblog file and website topology. The way to evaluate a user's interest in a WebPage has integrated four metrics: Relevance, authority, hub and novelty. The linear equations can be used to calculate user's interest in WebPages. When WebPages are clustered, we only need to evaluate the similarity of the two n-dimension vectors. Both space complexity and time complexity are $o(n^2)$. When similar customers are to be discovered, it is only needed to evaluate the similarity of the two n-dimension vectors. Both space complexity and time complexity are $o(n^2)$.

5. CONCLUSION

A Markov Chain-based dynamic model is put forward to investigate the customers' browsing behavior towards an EC website. As is well known that the similar customers will come into being after their long browsing and transacting with a EC website as time goes by. It is assumed that the ultimate probability of browsing each WebPage reflects the worthiness of each WebPage when time goes infinitely,and,thus ,this probability is a constant. A URL-USERID relevant matrix is set up to cluster customers and WebPages. This novel method can be intergrated in an EC recommending system and in personalized service system. That will be our further research interest.

REFERENCES

[1] S. Brin, L. Page,"The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems, Vol.30, No. 7: pp107-117,1998.

[2] MS.Chen, JS.Park,"Data Mining for Path Traversal Patterns in A Web Environment",proceeding of the 16th Int'l Conf on Distributed Computing Systems,pp385~392, HongKong,May,1996.

[3] O.R.Zaiane,M.Xin,J.Han,"DiscoveringWeb Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs",proceeding of Advances in Digital Libraries Conf, pp19~29， Santa Barbara,CA ,1998.

[4] QinBaoSong,Junyi Sheng,"An Efficient and Multi-Purpose Algorithm for Mining Weblog" , Computer Research and Development,Vol.38, No.3,pp328~332,2001.

[5] Rong xing Wang,StochasticProcess,Xi'an Jiao Tong Univer sity press.1987.

[6] Dell Zhang, Yisheng Dong,"An Efficient Algorithm to Rank Web Resources",proceeding of the 9th www conference on THE WEB:THE NEXT GENERATION.Amsterdam,May 15,2000.