Research on Model of Web Pages Clustering

Kuanjiu Zhou [1] Yifei Qu [1] Tiyun Huang [2]

[1] Systems engineering institute, Dalian University of Technology, PRC,116023

zhoukj@dlut.edu.cn

[2] School of management, Harbin Institute of Technology, PRC,150001

tyhuang@hope.hit.edu.cn

ABSTRACT

Web pages clustering is to divide different web pages into different classes according to traveling information of users so as to recommend relative pages to a user when the user is searching a website for his expectant information and to push relative pages to a user when the user is browsing one page. A similarity matrix, which shows similarity degree for any two pages in a website, is constructed according to users traveling paths to implement web pages clustering. A directional graph is created according to the users traveling paths at first and then a usage similarity matrix is constructed according to the directional graph. Structure disintegration model is adopted to analyze the similarity matrix and to transform the matrix into a catercorner matrix or a nether triangle matrix, which shows web pages clustering classes and relations among these classes.

Keywords: Web Pages Clustering, Similarity Matrix, Structure Disintegration Model, Web Usage Mining

1 INTRODUCTION

Clustering is a fundamental activity for human being to explore and understand the world. Sometimes, clustering should be made for human being to explore and study some objective matters easily so as to master the inside running laws [7]. Here, clustering means to divide a group of objects into some classes according to a certain similarity measurement model. There are two kinds of clustering in web usage mining research field, namely web users clustering and web pages clustering. Web pages clustering is to classify all web pages into different groups, in which the pages are relative in content, to recommend relative pages to a user when he is searching a website for his expectant information and to push relative pages from web server to a user when he is browsing one web page[9]. A similarity matrix, which shows similarity degree for any two pages in a website, is constructed according to users traveling paths to implement web pages clustering. At last section of this paper, the structure disintegration model is adopted to analyze the similarity matrix and to transform the similarity matrix into a catercorner matrix or a underside triangle matrix after a series of transformations on row elements and column elements of the matrix. The new matrix shows that web pages clustering classes and relations among these classes and more detailed web pages clustering process is shown as Fig.1.

2. CONTRUCTION OF FUZZY SIMILARITY MATRIX

The vital work of Web pages clustering is to construct similarity matrix of web pages, which shows similarity degree of any two pages in some attributes. A model of web pages similarity calculation is issued to construct web pages similarity matrix. If one web page refers to another web page, or the two web pages cite each other or the two web pages have some similar content, well then we can draw a conclusion that the two web pages

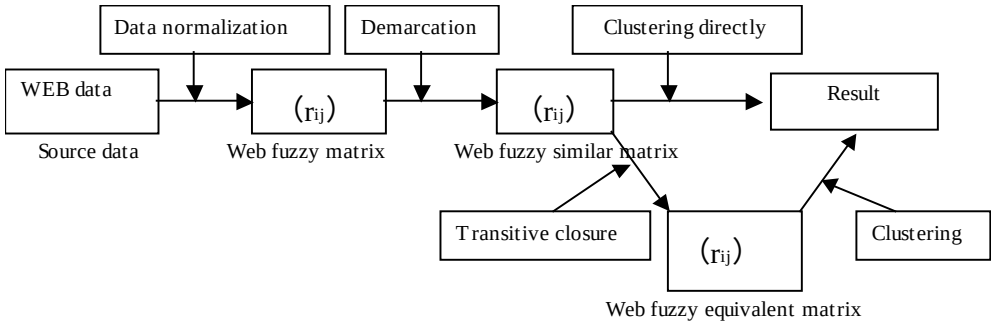

Fig.1 Web pages clustering process

have some similarity. Here, a similarity degree of the two web pages can be calculated by the following model. Thereinto, content-based similarity and reference-based similarity should be included. The calculation model is as formula(2-1):

$$sim(p_i, p_j) = \begin{cases} 1 & if \quad i = j \\ \alpha_1 \times sim_{cont}(p_i, p_j) + \alpha_2 \times sim_{usage}(p_i, p_j) & if \quad i \neq j \end{cases} \tag{2-1}$$

$\alpha_1 + \alpha_2 = 1, and \alpha_1 \geq 0, \alpha_2 \geq 0$, their weight values can be geared according to the importance degree of the content similarity degree and the reference similarity degree. If only the reference similarity is taken into consideration, the two parameters can be set as $\alpha_1 = 0, \alpha_2 = 1$. If they are of the same importance, the two parameters can be set as $\alpha_1 = 0.5, \alpha_2 = 0.5$.

2.1 Model of content similarity-based calculation

Here, let's discuss how to calculate content similarity degree for a character of web pages. At first, we suppose SET_{key} as a key words vector of a website W, and FV_i、FV_j as two character vectors of web page p_i and p_j respectively, and then to check whether or not every key word SET_{key} [t] in key words vector SET_{key} appears in web page p_i or p_j .if it does, FV_i[t] or FV_j[t] will be set 1 ,otherwise the two variables will be set 0. Consequently, web page content similarity matrix is constructed according to the following formula(2-2).

$$sim_{cont}(FV_i, FV_i) = \frac{\sum_{t=1}^{|SETkey|} FV_i[t] * FV_j[t]}{\sqrt{\sum_{t=1}^{|SETkey|} FV_i[t]^2 * \sum_{t=1}^{|SETkey|} FV_j[t]^2}} \tag{2-2}$$

Here, we suppose that every key word in SET_{key} has the same importance. If one keyword is different from other keywords for their importance, every keyword in SET_{key} can be assigned different weight value [8].

2.2 Model of reference similarity-based calculation

In order to describe and interpret the reference similarity matrix algorithm clearly, here, we discuss reference directional graph at first. A reference directional graph is a directional graph , which is transformed from user traveling paths. A reference directional graph is formalized as G =<N,A>. Thereinto, N is a set of nodes in the directional graph G as $N = \{i \mid p_i \in \mu\}$, and here μ is a set of web pages. A is a set of edges: for any edge $a_{ij} \in A$, which connects node p_i and node p_j if and only if there exists a traveling sequence S, $p_i \in s, p_j \in s$。We suppose ten users have visited one website as an example to clearly narrate reference similarity matrix algorithm, and the following are all traveling paths of these users:

S_1: <A, B, C, B, D> S_2: <E, F>

S_3: <A, B, D> S_4: <A, B>

S_5: <A, E , F> S_6: < F, E >

S_7: <A, B, C> S_8: <A, B, D, B, C>

S_9: <A, B, D> S_{10}: <A, B, A, E>

The above users traveling paths can be transformed into the following directional graph as Fig.2:

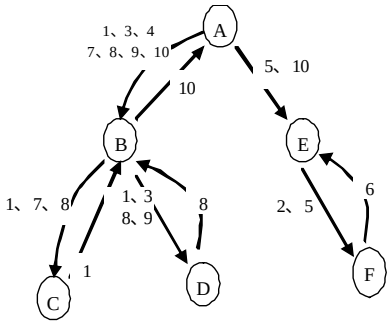

Fig.2 Directional graph of users traveling paths

We can create reference similarity matrix according to the above Fig.2 directional graph. The most length among the above ten user traveling paths is 2. Hence, the similarity matrix for web pages can be calculated by this formula $M = \sum_{1}^{\infty} w^l M^l$,here M^l is a similarity matrix of all web pages with distance *l* between two web pages. w^l is weight for this matrix and usually assigned $\frac{1}{2^l}$. Accordingly, shorter distance of two pages appearing in path means greater weight assigned to the similarity matrix and vice versa. Algorithm of calculating M^l as follows:

Input: user traveling paths set S = { S_1,S_2,S_3,…,S_N } and *L*, *L* is the most length for all path S. Fig.2 shows that the longest path is 2, so L should be 2 .

Output: reference similarity matrix *M*.

Begin

Initialize M, for every i,j $m_{ij} \leftarrow 0$;

For l = 1 to L do

{

Initialize M^l , for every i,j $m_{ij}^{l} \leftarrow 0$;

For every $s \in S$, if $p_i \in S, p_j \in S$, and there are *l* sections between p_i and p_j in path S, then

$m_{ij}^l = m_{ij}^l + \frac{1}{N}$;

M = M + $\frac{1}{2^l}$ $* M^l$

}

Re-compute the similarity matrix M, for every i,j,

re-compute $m_{ij} = \frac{m_{ij} + m_{ji}}{1 + | m_{ij} - m_{ji} |}$

Return M.

End

According to the above algorithm, two similarity matrixes M1,M2 can be calculated from the above directional graph and the calculating results as follows:

$$M^1 = \begin{bmatrix} 0 & 0.7 & 0 & 0 & 0.2 & 0 \\ 0.1 & 0 & 0.3 & 0.4 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.2 \\ 0 & 0 & 0 & 0 & 0.1 & 0 \end{bmatrix}$$

$$M^2 = \begin{bmatrix} 0 & 0 & 0.3 & 0.4 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Web pages reference similarity matrix M can be calculated according to the formula

$$sim_{usage}(p_i, p_j) = \sum_{l=1}^{\infty} w^l M_{ij}^{(l)}.$$

$$M=\begin{bmatrix} 0 & 0.35 & 0.075 & 0.1 & 0.1 & 0.025 \\ 0.05 & 0 & 0.15 & 0.2 & 0.025 & 0 \\ 0 & 0.05 & 0 & 0.025 & 0 & 0 \\ 0 & 0.05 & 0.025 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0.05 & 0 \end{bmatrix}$$

According to the formula $m_{ij}=\dfrac{m_{ij}+m_{ji}}{1+|m_{ij}-m_{ji}|}$, the matrix M is geared into the last similarity matrix A.

$$A=\begin{bmatrix} 0 & 0.3 & 0.07 & 0.09 & 0.091 & 0.024 \\ 0.3 & 0 & 0.18 & 0.22 & 0.024 & 0 \\ 0.07 & 0.18 & 0 & 0.05 & 0 & 0 \\ 0.09 & 0.22 & 0.05 & 0 & 0 & 0 \\ 0.091 & 0.024 & 0 & 0 & 0 & 0.143 \\ 0.024 & 0 & 0 & 0 & 0.143 & 0 \end{bmatrix}$$

3. STRUCTURE MODELING

Fuzzy clustering algorithms based on fuzzy similarity matrix include system clustering, transitive closure[10-11], most support tree PRIM algorithm, KRUSKAL algorithm , dynamic direct clustering algorithm, FCMBP、fuzzy C-means, fuzzy ISODATA and artificial neural network algorithm etc[1-4], a clustering algorithm based on structure disintegration model is issued in this paper to analyze the above similarity matrix A. The difference between structure model and traditional transitive closure is that structure model can classify all elements in a system into some sub-classes and find the relations among these sub-classes.

The matrix A can be mapped into a directional graph, every element in the matrix A shows how and how much two pages for the element contact or associate directly. The transference closure matrix R of the adjacent matrix A shows how and how much two pages for this element contact or associate directly or indirectly. So, the transitive closure matrix R should be figured out to analyze the graph deeply. The transitive closure matrix R in which every element values are between 0 and 1, is figured out after a series of matrix A power calculation.

We set a threshold α to transform every elements in the matrix R into 0 or 1.

$$x_{ij}=\begin{cases}1, & (x_{ij}\geq\alpha),\\ 0, & (x_{ij}<\alpha),\end{cases}$$

Here, K-mean clustering algorithm is used to cluster all elements x_{ij} in the matrix R into two classes , and two center values α_1、α_2 of the two classes are calculated , then α value can be figured out :

$$\alpha=\frac{\alpha_1+\alpha_2}{2} \qquad (3-1)$$

So, Two Boolean matrix A and R are acquired from the above matrix A and R after all elements in the matrix A and R are substituted with 0 or 1 and if there are loops in direct graph of the matrix A, then there must be full sub-matrixes in the matrix R. To give prominence to the loops, let's calculate $R \cap R^T$ at first.

So, there is only loop part left in this new matrix besides self relations. If the matrix R is a full matrix (all elements are 1), then the graph for the matrix A is strong connective otherwise there exists an integer $v(v\leq n)$ to make $A^k=0, K\geq v$ true. Consequently ,from the view point of reachable matrix, the formula $R \cap R^T = I$ is true.

Here, we should explain that similarity relation meets self property, symmetry property except transfer property. R is supposed as similarity relation, if x R y ^ y R z is satisfied , x R z can't be drawn. But for fuzzy similarity relation, $u_R(x,y)=\lambda_1 \wedge u_R(y,z)=\lambda_2 \Rightarrow u_R(x,z)=\min(\lambda_1, \lambda_2)$. so transfer property of the matrix R can be assured.

4. DISINTEGRATION OF STRUCTURE MODEL

For a directional graph G without any loops completely, we suppose that serial numbers of rows whose elements are 0 in an adjacent matrix A as $i_1^1, i_2^1, \Lambda, i_{n_1}^1$, the set including these serial numbers is

$$S_1 = \left\{ i_1^1, i_2^1, \Lambda, i_{n_1}^1 \right\}.$$

After all $i_1^1, i_2^1, \Lambda, i_{n_1}^1$ rows and $i_1^1, i_2^1, \Lambda, i_{n_1}^1$ columns in matrix A are deleted, the rest rows and columns still keep their original serial numbers. To do so means getting ride of all output nodes and all branches that come from these nodes from the directional graph G. The new graph still has new output nodes, and there exist rows with all elements as 0. We still suppose that serial numbers for these rows as $i_1^1, i_2^1, \Lambda, i_{n_1}^1$, the set including these serial numbers is

$$S_2 = \left\{ i_1^2, i_2^2, \Lambda, i_{n_2}^2 \right\}$$

the process doesn't stop until all rows and columns have been classified. And at last , we have got the following classes:

$$S_1 \, Y \, S_2 \, Y \Lambda \, Y \, S_p = \{1,2,\Lambda, n\}$$

Next step, we will do transformations on the matrix A as the following sequence:

$$\begin{array}{ccccccc} i_1^1, & i_2^1, \Lambda, & i_{n_1}^1 & i_1^2, \Lambda, & i_{n_p}^p \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 2 & n_1 & n_1+1 & n \end{array}$$

Consequently, $\hat{A} = P^T A P$ can be transformed into a catercorner matrix or a underside triangle matrix.

But for directional graphs with loops, we suppose that it isn't connective intensively at first, it means that there exist elements 0 in the reachable matrix R. We suppose that these elements on certain column (for example, j_1 th column) and $i_1^1, i_2^1, \Lambda, i_{n_1}^1$ th rows are 1 in matrix $R \, I \, R^T$, and all rest elements are 0. Let's set

$$S_1 = \left\{ i_1^1, i_2^1, \Lambda, i_{n_1}^1 \right\}$$

Because elements on catercorner line in matrix $R \, I \, R^T$ must be 1, j_1 belongs to S_1 and all nodes in S_1 are strong connective. S_1 will not be strong connective if S_1 is added other nodes which don't belong to S_1.Next, we will select j_2 th column out of S_1, and suppose that only $i_1^2, i_2^2, \Lambda, i_{n2}^2$ elements are 1, and let

$$S_2 = \left\{ i_1^2, i_2^2, \Lambda, i_{n_2}^2 \right\}$$

then j_2 belong to S_2. This process will not stop until $S_1 \, Y \, S_2 \, Y \Lambda \, Y \, S_p = \{1,2,\Lambda, n\}$ becomes true.

Apparently,

$$S_i \, I \, S_j = \Phi, j \neq i$$

Adjacent matrix A is counterchanged as the following transformation

$$\begin{array}{ccccccc} i_1^1, \Lambda, & i_{n_1}^1 & i_1^2, \Lambda & \Lambda, & i_{n2}^2 & \Lambda, & i_{n_p}^p \\ \downarrow & \downarrow & \downarrow & & \downarrow & & \downarrow \\ 1 & \Lambda & n_1 & n_1+1 & n_1+n_2 & & n \end{array}$$

into a block matrix like the following matrix:

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{21} & \Lambda & \hat{A}_{p1} \\ \hat{A}_{12} & \hat{A}_{22} & \Lambda & \hat{A}_{92} \\ M & M & \Lambda & \Lambda \\ \hat{A}_{1p} & \hat{A}_{2p} & \Lambda & \hat{A}_{pp} \end{bmatrix} \begin{matrix} S_1 \\ S_2 \\ \\ S_p \end{matrix},$$

$$S_1 \qquad S_2 \quad \ldots\ldots\ldots\ldots$$

Connect relation among node groups S_1, S_2, Λ, S_p is usually expressed with $p \times p$ Boolean matrix B and elements in matrix are defined as follows:

$$b_{ii} = 0 \qquad i = 1, \Lambda, p$$

$$b_{ij} = \begin{cases} 1 & A_{ij} = 0 \\ & , \quad i, j = 1,2, \Lambda, p, \quad i \neq j \\ 0 & A_{ij} = 0 \end{cases}$$

The graph based on matrix B has no loops. So the matrix B will be counterchanged into nether triangle matrix according to the following transformation:

$$\begin{matrix} j_1, j_2, \Lambda, j_p \\ \downarrow \ \downarrow \quad \downarrow \\ 1 \quad 2 \quad p \end{matrix}$$

The above serial numbers $j_1, j_2, \ldots . j_p$ .are substituted with its original serial number groups as follows:

$$\begin{matrix} i^{j}_{1^{1}}, \Lambda, i^{j}_{n^{1}j_1} & i^{j}_{1^{2}}, \Lambda \Lambda, i^{2}_{n^{2}j_2} & \Lambda, i^{j}_{n^{p}j_n} \\ \downarrow \quad \downarrow & \downarrow \quad \downarrow & \downarrow \\ 1 \ \Lambda \ n_{j_1} & n_{j_1} + 1 \quad n_{j_1} + n_{j_2} & n \end{matrix}$$

So, the adjacent matrix A has been counterchanged into nether triangle matrix after the above serial transformations. All web pages in the same block should belong to the same class.

5. AN EXAMPLE

Next, We will illustrate web page clustering algorithm with an example. The example is the same as the above example. So we still take a similarity matrix as example.

$$A = \begin{bmatrix} 0 & 0.3 & 0.07 & 0.09 & 0.091 & 0.024 \\ 0.3 & 0 & 0.18 & 0.22 & 0.024 & 0 \\ 0.07 & 0.18 & 0 & 0.05 & 0 & 0 \\ 0.09 & 0.22 & 0.05 & 0 & 0 & 0 \\ 0.091 & 0.024 & 0 & 0 & 0 & 0.143 \\ 0.024 & 0 & 0 & 0 & 0.143 & 0 \end{bmatrix}$$

We can figure out a threshold $\alpha = 0.134$ with formula（3－1）, then a structure model matrix can be constructed as the following matrix A:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

A catercorner matrix R is figured out according to the algorithm in section 4:

$$R = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The upper matrix shows that these web pages can be classified into two groups: one group is composed of four web pages A,B,C,D and another group is composed the rest two web pages E and F. Anymore, because the matrix is a catercorner matrix but not a nether triangle matrix , so there is no any relation between the two groups.

6. CONCLUSION

Web page similarity model issued in this paper is composed of content similarity model and usage similarity model, which has solved the demerit of traditional model that only takes content similarity into account. Structure disintegration model is used to

analyze the similarity matrix constructed by web page similarity model. An example is taken to prove that Web pages clustering has been improved with the new model. Our web page clustering algorithm based on user traveling paths is proved feasible by our prototype running on a PC. The user traveling paths was mined from a Web Log on Qingdao Hisense group E-business website server. The total user traveling paths are over 8k for one day. The clustering result of Web Log on the website is helpful to improve the website architecture.

REFERENCES

[1] Shi Zhongzhi, knowledge discovery, Tsinghua university press, Jan. 2002

[2] Zhong Maosheng, Fuzzy clustering of web pages, East china jiaotong university press, Oct. 2004

[3] Liu Qi, Application of most tree algorithm of fuzzy clustering in Web pages classification, Research on computer application, Nov. 2004

[4] Hao Xianchen, Application of fuzzy clustering mining in E-business, North-east University press, 2001.8

[5] Wang shi, Path clustering, Knowledge discovery in Websites, Development and research of computer, Apr. 2001

[6] Wang Zhongtuo, Systems Engineering, Dalian university of technology press, 1990

[7] Jiawei Han, Micheline Kamher . Data mining: Concepts and Techniques[M]. Academic Press, 2000.

[8] Alexios Chouchoulas, Qiang Shen. Rough Set—Aided Keyword Reduction For Text Categorization[J]. Application Artificial intelligence, 2001, 8: 857- 861.

[9] Ragab M Z, Emam E G. On the Min-max Composition of Fuzzy Matrices[J] Fuzzy Sets and Systems , 1995 (75) 83-92

[10] Mohamed K S. New Algorithm for Solving the fuzzy C-means Clustering Problem[J] Pattern Recognition. 1994. 27: 421-428

[11] Bezdek. Pattern Recognition with Fuzzy Objective function Algorithm[J] Plenum Press, NewYork,1 997