DYNAMIC WEB DOCUMENT CLASSIFICATION IN E-CRM USING NEURO-FUZZY APPROACH

Iraj Mahdavi[a], Babak Shirazi[a], Namjae Cho[b], Navid Sahebjamnia[a] and Meysam Aminzadeh[a]
[a] *Mazandaran University of Science & Technology, Babol, Iran*
[b] *The School of Business, Hanyang University, Seoul, Korea*

Keywords: e-CRM; data mining; Web document clustering; neuro-fuzzy approach.

Abstract: Internet technology enables companies to capture new customers, track their performances and online behavior, and customize communications, products, services, and price. The analysis of customers and customer interactions for electronic customer relationship management (e-CRM) can be performed by data-mining (DM), optimization methods, or combined approaches. Some of web mining techniques include analysis of user access patterns, web document clustering and classification. Most existing methods of classification are based on a model that assumes a fixed-size collection of keywords or key terms with predefined set of categories. We propose a new approach to obtain category-keyword sets with unknown number of categories. On the basis of the training set of Web documents, the approach is used to classify test documents into a set of initial categories. Finally evolutionary rules are applied to these new sets of keywords and training documents to update the category-keyword sets to realize dynamic document classification.

1 INTRODUCTION

The purpose of CRM is to identify, acquire, serve, and retain profitable customers by interacting with them in an integrated way across a range of communication channels. Swift (2002) describes analytical e-CRM as a four-step iterative process consisting of (1) collecting and integrating online customer data, (2) analyzing this data, (3) building interactions with customers, and (4) measuring the effectiveness of these interactions. A typical performance metric used in many CRM applications deals with finding the optimal life time value (LTV) of a customer.

Customer analysis includes two major procedures under the context of e-CRM (Padmanabhan and Tuzhilin 2003): (1) preprocessing data, and (2) building customer profiles from this and other data. Data preprocessing is one critical step of the knowledge discovery process in e-CRM, and the success of most DM methods, to a large extent, depends on this step (Zheng et al. 2003). Most current literature considers different variations of heuristic methods for analyzing click- stream data gathered from websites.

Shamim Khan and Khor (2004) describe a method developed for the automatic clustering documents of World Wide Web documents, according to their relevance to the user's information needs, by using a hybrid neural network. The objective is to reduce the time and effort the user has to spend to find the information sought after. Yager (1992) indicates that Neuro-fuzzy approach allows for classification of fuzzy systems based on training set.

In this paper, we illustrate how neuro-fuzzy approach can facilitate improved data preprocessing to achieve rich and accurate profiles of customers through dynamic Web document classification.

2 WEB DOCUMENT CLASSIFICATION

The neuro-fuzzy approach offers a data preprocessing module to help users to prepare his/her data for high-quality mining results .The model is described in the following section.

Mahdavi I., Shirazia B., Chob N., Sahebjamniaa N. and Aminzadeha M. (2007).
DYNAMIC WEB DOCUMENT CLASSIFICATION IN E-CRM USING NEURO-FUZZY APPROACH.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 378-381

Notation:

m : Number of keywords
n : Number of documents
nc : Number of Category-keyword sets
k : Index for category size ($k = 1,2,\cdots, nc$)

w_i : i[th] keyword
d_j : j[th] document
f_{ij} : Number of occurrence of keyword w_i in the document d_j
f_{ijk} : Number of occurrence of keyword w_i in the document d_j to the category k
α_{ij} : Fuzzy membership value of keyword w_i and document d_j
$M = [f_{ij}]_{m\times n}$: Keyword document incidence matrix (*KDIM*)
$MF = [\alpha_{ij}]_{m\times n}$: Fuzzy *KDIM* matrix
$MB = \lfloor \beta_{ij} \rfloor_{m\times n}$: Binary matrix
$TD_j = [f_{ijk}]_{1\times m}$: Test document matrix
$TDB_j = [\beta_{\mathrm{ijk}}]_{1\times \mathrm{m}}$: Binary test document matrix

2.1 Neuro-fuzzy Approach

Given an initial training set of Web documents denoted as $D_0 = \{d_1, d_2, \cdots d_n\}$, corresponding set of keywords exist as $W_0 = \{w_1, w_2, \cdots w_m\}$ at time T_0. Each document is transformed into a vector that contains maximum of m keywords (existing keywords). For each keyword we compute its number of occurrences in the document (denoted by "frequency"). We use keyword-document incidence matrix *KDIM* as the basic input data to obtain the category-keyword sets *CK* as an unsupervised process of clustering. *KDIM* matrix is represented as $M = [f_{ij}]_{m\times n}$, where m is the number of keywords, n is the number of documents and f_{ij} is the number of occurrence of keyword w_i in the document d_j .We define matrix $MF = [\alpha_{ij}]_{m\times n}$, a fuzzy *KDIM* matrix where α_{ij} is a corresponding keyword-document fuzzy membership as defined in equation (1).

$$\alpha_{ij} = \frac{f_{ij}}{\sum_{i=1}^{m} f_{ij}} \times \frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}} = \frac{(f_{ij})^2}{\sum_{i=1}^{m} f_{ij} \sum_{j=1}^{n} f_{ij}} \quad (1)$$

The ratio of $\frac{f_{ij}}{\sum_{i=1}^{m} f_{ij}}$ can be viewed as a reliability value for keyword w_i in document d_j and the ratio of $\frac{f_{ij}}{\sum_{j=1}^{n} f_{ij}}$ as a reliability value for document d_j in keyword w_i.

The value of α_{ij} is considered as a reliability value for the incidence of keyword w_i – document d_j in the whole *KDIM* matrix. This reliability shows the membership value of f_{ij} and it is the probability of a document d_j possesses keyword w_i .

If this value is greater than or equal to the threshold value θ as given in formula (2), then we can interpret that document d_j possesses keyword w_i reasonably and strongly, otherwise due to weak value of α_{ij}, we ignore that document d_j possesses keyword w_i.

$$\theta = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\alpha_{ij} - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{ij}}{nm})^2}{nm} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\frac{(f_{ij})^2}{\sum_{i=1}^{m} f_{ij} \sum_{j=1}^{n} f_{ij}} - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(f_{ij})^2}{\sum_{i=1}^{m} f_{ij} \sum_{j=1}^{n} f_{ij}}}{nm})^2}{nm} \quad (2)$$

The Dynamic Web Document Clustering (*DWDC*) algorithm has been introduced to achieve the profile of customers in e-CRM.

2.2 DWDC Algorithm

Step 0. Set *i = 0*; (i is considered as time period), *C = 0* (*C* is a counter for test documents that we can not classify to existing categories).

Step 1. Considering a threshold value θ as a degree of membership value for qualification, where the value of θ would be in the range of (*min* non-zero α_{ij} , *max* α_{ij}). We convert the *MF* matrix into a binary matrix *MB* such that; $MB = \lfloor \beta_{ij} \rfloor_{m\times n}$ and $\beta_{ij} = 1$ if $\alpha_{ij} \geq \theta$ otherwise $\beta_{ij} = 0$

Step 2. Use Graph-Neural Algorithm (GNA) to obtain unsupervised clustering of keywords as given in next section.

Step 3. From time T_i to T_{i+1}, set $j=1$ (classify all test documents)

Step 4. Use the threshold vector $\theta_c=[\theta_1,\theta_2,...,\theta_{nc}]^T$ as a vector to determine membership value using equation (2) for each cluster, and then convert the vector TD_j to the fuzzy test document using equation (3)

as $FTD_j=[\alpha_{\mathrm{ijk}}]_{1\times \mathrm{m}}$. Then create a binary vector $TDB_j=[\beta_{ijk}]_{1\times m}$ such that;

$\beta_{ijk}=1$ if $\alpha_{ijk}\geq\theta_k$, otherwise $\beta_{ijk}=0$.

$$\alpha_{ij}=\frac{f_{ij}}{\sum_{i=1}^{m} f_{ij}} \quad , \quad \forall j \qquad (3)$$

Step 5. Select the max $\{\sum_i \beta_{ijk}\}$

for $k\in\{1,2,\cdots,nc\}$, (Break ties arbitrarily), then assign each test document j to a category *k* which shows the maximum obtained value.

Step6. If $\sum_i \beta_{ijk}=0$ for $i\in\{1,2...,m\}$, then document *j* is not appeared in the classification, set $C=C+1$.

Step7. If $C\geq c$ (*c* is the maximum value of unspecified test document), go to step 8, otherwise $j=j+1$ then go to step4.

Step8. Set *i=i+1*; (i can be considered as a time stamp).

Step9. Apply the evolutionary rule for document clustering to planning time period T_i ($i=1,2,\cdots,p$) to revise the category-keyword sets.

Step 10. Update the *KDIM* matrix, and then obtain *MF* matrix.

Step 11. Go to step1.

Keywords have life. If a specific keyword is not used over certain number of (say"*c*") testdocuments, then it is disqualified to remain in the network. The clustering approach is updated whenever changes appear in the structure of keywords. After the evolutionary rule is applied at time T_i for restructuring the keyword sets, new training document sets are created and used for clustering.

2.3 Graph-neural Network Approach

The neuro-fuzzy diagram has been shown in figure 1.

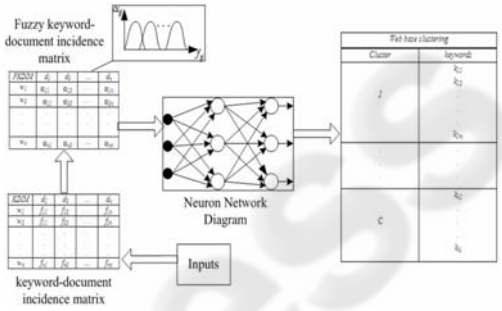

Figure 1: Neuro-fuzzy diagram.

We consider $MG=\lfloor g_{ij}\rfloor_{m\times m}$ as a multigraph matrix that we obtain from *MB* matrix such that; $g_{ij}=g_{ji}=\sum_{k=1}^{n} x_{ik}.x_{jk}$, where $x_{ik}=1$,if in matrix *MB* the value of β_{ik} had been 1, otherwise $x_{ik}=0$. (*i* and *j* are considered as keywords and *k* is considered as document in *MB* matrix). The algorithm includes the following steps.

Graph-Neural Algorithm (GNA):

Step 1. Convert the *MB* matrix into the matrix of a multigraph, *MG*.

Step 2. Create one category for the output layer, i.e. *k*=1.

Step 3. Calculate the sum of entries by rows in the matrix of the multigraph.

SUM (i) = $\sum_{j=1}^{m} g_{ij}$ for i =1, 2,…, *m*.

Step 4. Select the row in the matrix with the highest sum (Break ties arbitrarily).

SUM-MAX = *max* {SUM (i)} for i=1, 2,…, *m*.

Step 5. With the highest sum, consider next non-zero entries of that row in decreasing order.

Step 6. If the first entry is g_{ij} (corresponding to the i^{th} row and j^{th} column), then the keywords w_i and w_j are assigned to the current category.

Step 7. If the second non-zero entry of the i^{th} row is g_{ir} (corresponding to the i^{th} row and r^{th} column) and if keyword w_r has identical number(s) with keyword w_j then keyword w_r is also assigned to the current category, otherwise it cannot be assigned to this category. Continue this process for the remaining non-zero entries of the i^{th} row.

For assigning a new keyword to the current category, it should be confirmed that the candidate keyword must have identical number(s) to all the assigned keywords in that category.

Step 8. If all the keywords have been assigned, stop. Consider the number of category keyword sets as (*nc*) and go to step3 in ***DWDC*** **algorithm.**

Step 9. Construct a new matrix of multigraph for the remaining keywords and create a new category.
Step 10. k=k+1 and go to step3.

7 CONCLUSIONS

Most firms today recognize the importance of building and maintaining strong relationship with their customers. As firms increasingly rely on their online presence to interact with customers, e-CRM will continue to grow in importance. In this paper, an approach to automatically classify the web documents into categories was suggested using neuro-fuzzy approach. A method for identifying categories in an evolutionary scale-free keyword network and clustering test documents is proposed to facilitate preprocessing of click-stream data in e-CRM that incorporates dynamic changes in web document.
This paper provides a novel approach on Web document clustering as there is no predefined category or fixed number of keywords assumed in the model. And such dynamic formulation is highly realistic in the context of World Wide Web by in the sense that it allows one to dynamically change and update the category keyword sets for web document classification. The practical and dynamic keyword clustering process identified by the method suggested in this research will help to create ideal patterns of Web document for effective and efficient management of Web contents.
Moreover, it provides interesting opportunities for DM to help develop better solutions to e-CRM problems, as many e-CRM applications require concise profiles that contain the most important set of information about customers.
The prototype of system has been designed to show the computerized results of web document clustering.

REFERENCES

Padmanabhan, B. and Tuzhilin, A., 2003, On the use of optimization for data mining: Theoretical interactions and eCRM Opportunities, Management Science, 49(10), 1327-1343.

Shamim Khan,M. and Khor,S.W., "Web document clustering using a hybrid neural network" Applied Soft Computing, Volume 4, Issue 4, September 2004, Pages423-432

Swift, R. ,2002, Analytical CRM powers profitable relationships: Creating success by letting customers guide you, *DM Rev.* (February).

Yager, R., "Implementing fuzzy logic controllers using a neural network framework", Fuzzy Sets and Systems 48 (1992) p. 53 - 64.

Zheng, Z., Padmanabhan, B., and Kimbrough, S., 2003, On data preprocessing biases in web usage mining, *INFORMS J. Comput*. 15(2), 148-170.