

# Using Fuzzy Cognitive Map to Effectively Classify E-Documents and Application<sup>\*</sup>

Jianzeng Wang<sup>1,3</sup>, Yunpeng Xing<sup>1,3</sup>, Peng Shi<sup>1,3</sup>, Fei Guo<sup>1,3</sup>, Zhen Wang<sup>1,3</sup>,  
Erlin Yao<sup>1,2,3</sup>, Kehua Yuan<sup>1,2,3</sup>, and Junsheng Zhang<sup>1,2,3</sup>

<sup>1</sup> Key Lab of Intelligent Information Processing, Institute of Computing Technology,  
Chinese Academy of Sciences, Beijing, 100080, China

{wjz, ypxing, pengshi, guofei, wangzhen,  
alin.yao, kehua, junsheng}@kg.ict.ac.cn

<sup>2</sup> Hunan Knowledge Grid Lab, Hunan University of Science and Technology,  
Hunan, China

<sup>3</sup> Graduate School of the Chinese Academy of Sciences,  
Beijing, China

**Abstract.** In the current Web, e-document has been the most common vehicle for delivering and exchanging information. As the amount of e-documents has grown enormously, effective classification facilities are urgently needed to classify and query e-documents users want. In this paper, we propose a method to classify e-documents into a set of predefined categories based on Fuzzy Cognitive Map (FCM). The e-documents are collected from Internet by a meta-search engine. FCM has been employed to capture the semantic relationships between keywords of e-documents. Experiments with a set of local e-documents have proved that this approach has high performance and can help users getting the e-documents efficiently and effectively. The proposed method has been implemented and integrated into the Dunhuang Feitian System to manage and classify e-documents.

## 1 Introduction

The World Wide Web provides us with a large-scale and universal information space. E-document has been the most common vehicle for delivering and exchanging information in the current Web. However, the enormous amount of e-documents makes it difficult to search, access, present, and maintain the information. Searching for specific information or discovering useful information from the large amount of information in the Web has been becoming a difficult, time-consuming and challenging task. So effective classification and search facilities are urgently needed to efficiently classify and query e-documents users want in the open and dynamic World Wide Web environment.

Document classification, as originally used to improve the precision, especially top-level precision, of information retrieval systems, is the process of assigning a

---

<sup>\*</sup> This work was supported by the National Basic Research Program of China (973 project no.2003CB317000) and the National Science Foundation of China (Grants 70271007, 60273020, 60402016).

document to one of the predefined categories based on the document content [1]. In the classification process, traditional document classification methods such as *vector space model*, only consider the appearance frequency of keywords. But it has been well acknowledged in such fields as traditional database management system and information retrieval that the more semantics about data are understood and considered by a system, the more precise queries and searches can be achieved [2, 6].

In this paper, a method to effectively classify e-documents into a set of predefined categories based on FCM is proposed. FCMs capture the relationships between the keywords to improve the classification accuracy. The proposed method has been implemented and integrated into the Dunhuang Feitian System to manage and classify e-documents. One of the important aspects about such classification is that it could provide semantics (similarity at the conceptual level between documents) without requiring the classification process to use any extra semantics. And the classification process is adaptive.

## 2 Effective E-Document Classification Based on FCM

### 2.1 Fuzzy Cognitive Map

Fuzzy Cognitive Map is a directed and weighted graph of concepts and relationships between the concepts [3]. FCM is derived from Cognitive Map. Based on the CM structure, FCM was proposed by Kosko [4, 5]. As a great improvement compared to CM, FCM introduces fuzzy quantitative relationship between concepts to describe the weight of the causal relationship. FCM has iterative characteristic. In FCM, the arcs are not only directed to show the direction of causal relations, but also accompanied by a quantitative weight within the interval [0, 1].

The formulation for calculating the state value of concepts is proposed as follows:

$$A_i^t = f(k_1 \sum_{\substack{j=1 \\ j \neq i}}^n A_j^{t-1} W_{ji} + k_2 A_i^{t-1}) \quad (1)$$

$A_i^t$  is the state value of  $C_i$  at step  $t$ ;  $A_j^{t-1}$  is the state value of the interrelated causal concept  $C_j$  at step  $t-1$ ;  $A_i^{t-1}$  is the state value of  $C_i$  at step  $t-1$ ;  $W_{ji}$  is the interrelation's weight from concept  $C_j$  to  $C_i$  and  $f$  is a threshold function, here  $f(x) = \tanh(x)$ . In this paper, it is assumed that  $k_1 = k_2 = 1$ . The coefficient  $k_2$  represents the contributive proportion of the previous value to the new state value.

### 2.2 Construction of Basic Category FCM (BC-FCM)

In order to effectively classify documents into a set of predefined categories, it is necessary to construct a BC-FCM to represent a category first. BC-FCM plays an important part in the process of classification. It may greatly affect the performance of document classification systems, because its broadness and granularity influence the coverage and specificity of classification categories.

Based on the representative documents whose categories have been predetermined, the BC-FCM for each category is constructed as the following steps:

First, keywords should be extracted as BC-FCMs’ concepts from the representative documents. Using the statistical characteristic of the keywords, we pick out those keywords that can represent each category best. Intuitively, the best way to represent each a category is to select only the exclusive keywords [1]. But in BC-FCMs, the quantitative weight of each causal relation in different BC-FCMs may be different, so the proposed method also considers those overlapped keywords.

Then the concept state value should be determined. In the paper, the concept state value  $V_{ci}$  is calculated by Eq. (2):

$$V_{ci} = \frac{1}{1 + e^{-cy}} \tag{2}$$

where  $c$  is a constant, herein  $c = 0.6$  and it can be manually adjusted by user according to the experiment result.  $y = f(location, frequency)$ ,  $V_{ci} \in [0, 1]$ . In the proposed classification system, we only consider the appearance frequency of keywords.

### 2.3 Effective E-Documents Classification Based on FCM

In this subsection, the classification method based on FCM is described in detail. The proposed method has been implemented and integrated into the Dunhuang Feitian System, and it is named as SRCC component, representing Search Results Clustering and Classification. The architecture of the SRCC component is given in fig. 1.

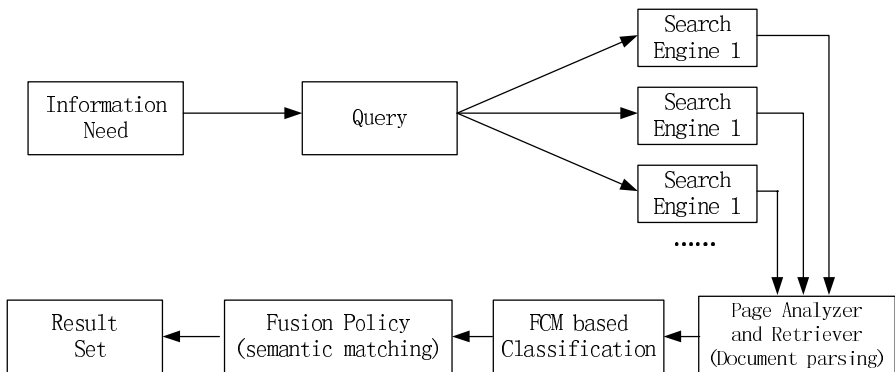


Fig. 1. The architecture of the SRCC component

As a whole, collecting e-documents from Internet by the meta-search engine firstly, and the search results are stored in a database. Then the FCM based classification method is used to classify the search results into predefined subclasses for helping users to access the web pages effectively and efficiently.

After choosing out keywords from an e-document, we match these keywords with concepts included in each BC-FCM; then we get the matching keywords’ state values, and put them into Eq. (3) to calculate the matching degree ( $md$ ). The matching degrees are used to determine which category the document belongs to.

$$md = \frac{1}{1 + e^{-c \times \sum_1^n V_{ci}}} \tag{3}$$

where  $c$  is a constant, herein  $c = 0.6$  and it can be manually adjusted;  $V_{ci}$  is calculated by Eq. (2) and  $n$  is the total number of concepts in a BC-FCM.

A document  $D_i$  is classified into the category  $C_j$  where the matching degree ( $md$ ) is the maximum among all the categories and also it is bigger than the threshold. The threshold is an adjustable variable in the system, and usually it is 0.75.

### 3 Implementation in the Dunhuang Feitian System

The proposed method is implemented and integrated into the Dunhuang Feitian System, a part of Dunhuang Culture Knowledge Grid [7-14]. Feitian is a typical representative of the ancient Dunhuang art. Because of the cultural relic conservation and geographical restrictions, she hasn't been familiar to most people in the world. Under such circumstances, the Dunhuang Feitian System is undertaken to help people enjoying this humankind's ancient culture more conveniently. The SRCC component is integrated into the Dunhuang Feitian System for providing useful knowledge of Dunhuang Feitian art to visitors and researchers.

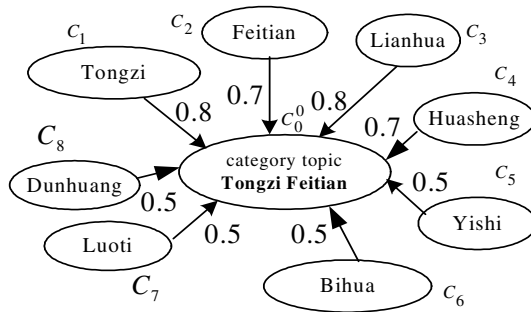


Fig. 2. BC-FCM for “Tongzi Feitian”

In the implementing process, it is also necessary to construct BC-FCMs for each subclass of Dunhuang Feitian. Under the domain background knowledge of Dunhuang Culture as well as referring to some professional dictionaries, the BC-FCMs for subclasses of “Dunhuang Feitian”, such as “Tongzi Feitian”, are constructed. Fig. 2 is the BC-FCM for “Tongzi Feitian”.

#### 3.1 Major User Interface and Results Analysis

We use the software *Macromedia Flash* to organize the pictures and generate the whole animation of the Feitian System. The SRCC component is used to search and access the corresponding on-line e-documents.

Users can input query keyword(s) in the search box, and then click the “Go” button. Query keyword(s) is (are) transmitted to Web search engines. Search results are returned by the meta-search engine and classified by the proposed classification method.

The scenario of the search results corresponding to subclass “Tongzi Feitian” is shown in fig. 3. When user selects a subclass, the corresponding e-document list is shown to the user, with query words highlighted. This e-document list could be in the original order, or be re-ranked according to some ranking algorithms.

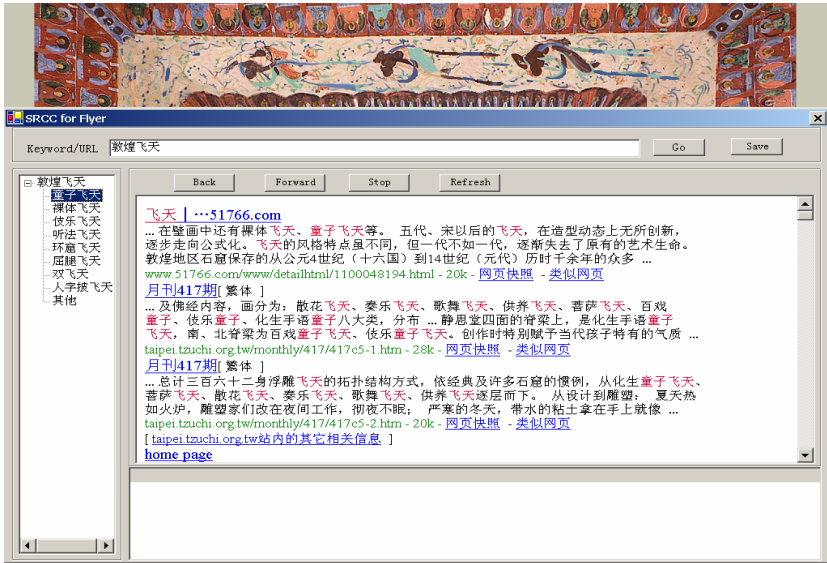


Fig. 3. The scenario of the search results corresponding to “Tongzi Feitian”

In the meta-search engine, if a user inputs “Dunhuang Feitian” as the query keywords, there are nearly 30,000 search results returned by included search engines. Among these search results, using the proposed classification method, there are about 6430 search results belonging to “Tongzi Feitian”; 8160 search results belonging to “Luoti Feitian”; 826 search results belonging to “Jiyue Feitian”; 454 search result belonging to “Shuang Feitian” and 709 search results belonging to “Qutui Feitian” etc.

## 4 Conclusions

An approach of using FCM to effectively classify e-documents returned by a meta-search engine into a set of predefined categories is proposed. FCM has been employed to capture the semantic relationships between keywords extracted from an e-document. The classification process is adaptive: if some documents don't belong to any of the predefined categories, they will not be wrongly assigned to any existing categories. Experiments with a set of local e-documents have proved that this approach has high performance and can help users getting the e-documents. This

method has been implemented to help users know culture effectively, and can help implement the future interconnection environment [15]. Zhuge's Resource Space Model [10-12] is used to store and effectively manage documents. Ongoing work is to incorporate the FCM with the SLN [7-9] to form more powerful semantic representation approach for the Semantic Grid [7, 14].

## References

1. Haruechaiyasak, C., Shyu, M. and Chen, S.: Web Document Classification Based on Fuzzy Association. In: Proceedings of the 26th Annual International Computer Software and Applications Conference (COMPSAC'02), Oxford, England, 2002.
2. Lee, J., Lee, K. and Kim, W.: Preparations for Semantics-Based XML Mining. In: Proceedings of the 2001 IEEE International Conference on Data Mining, California, USA, 2001.
3. Tsadiras, A. and Margaritis, K.: Cognitive Mapping and Certainty Neuron Fuzzy Cognitive Maps. *Information Sciences*, 101 (1997) 109-130.
4. Hart, J.: Comparative Cognition: Politics of International Control of Oceans. In: Axelrod, R. (ed.): *Structure of Decision*. Princeton: Princeton University Press, (1976) 180-217.
5. Tsadiras, A., Margaritis, K. and Mertzios, B.: Strategic Planning Using Extended Fuzzy Cognitive Maps. *Studies in Informatics and Control*, 4(3) (1995) 237-245.
6. Berners-Lee, T., Hendler, J. and Lassila, O.: The Semantic Web. *Scientific American*, 284(5) (2001) 34-43.
7. Zhuge, H.: Clustering Soft-device in the Semantic Grid. *IEEE Computing in Science and Engineering*, 4(6) (2002) 60-62.
8. Zhuge, H.: Active e-Document Framework ADF: Model and Platform. *Information and Management*. 41(1) (2003) 87-97.
9. Zhuge, H. and Jia, R.: Semantic Link Network Builder and Intelligent Browser. *Concurrency and Computation: Practice and Experience*, 16(14) (2004) 1453-1476.
10. Zhuge, H.: *The Knowledge Grid*. World Scientific, 2004.
11. Zhuge, H.: Resource Space Grid: Model, Method and Platform. *Concurrency and Computation: Practice and Experience*, 16(14) (2004) 1385-1413.
12. Zhuge, H.: Resource Space Model, Its Design Method and Applications. *Journal of Systems and Software*, 72(1) (2004) 71-81.
13. Zhuge, H.: China's E-Science Knowledge Grid Environment. *IEEE Intelligent Systems*, 19(1) (2004) 13-17.
14. Zhuge, H.: Semantic Grid: Scientific Issues, Infrastructure, and Methodology. *Communications of the ACM*. 48(4) (2005) 117-119.
15. Zhuge, H.: The Future Interconnection Environment. *IEEE Computer*, 38(4) (2005) 27-33.