2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

A Document Ordering Support System Employing Concept Structure based on Fuzzy Fish View Extraction

Tadashi Ohashi, Hajime Nobuhara, Kaoru Hirota
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama 226-8502, JAPAN
{ohashi,nobhara,hirota}@hrt.dis.titech.ac.jp

Abstract **- To classify desired and undesired documents on the web according to each user's view, FOCUS (Fuzzy dOCUment ordering System) is developed based on fuzzy concept extraction, fuzzy fish eye matching, and fuzzy selection. Experiments are done using the concept-system-dictionary by EDR (Electronic Dictionary Research Institute) including 140,000 words and web-based documents related to movie.**

I . Introduction

FISH VIEW system[1][2] with visual support functions is one of important tools to produce effective support functions for ordering the vast electronic web-based documents. It has been widely used in the area of medical applications and movie documents, and is effective to pick up un-noticed documents by the knowledge discovery technology embedded in the system. Recently huge databases appear on web and many users utilize the data based on so called semantic web technology, where human factor is an important issue for data retrieving and concept extraction/ construction.

Fuzzy sets have been used to represent human factors successfully in the field of control and expert systems. Concept extraction by FISH VIEW with fuzzy sets, proposed as Fuzzy FISH VIEW extraction, is a key technology in semantic web applications. It is applied to a document ordering system, called FOCUS (Fuzzy dOCUment ordering System), on computers connected with internet. In the FOCUS, users put the web-based documents into the system, and concept groups are constructed based on EDR concept-system-dictionary composed of about 140,000 English words. A concept structure could be extended in terms of the fuzzy clustering that is able to deal with more human-centered concept structure suitable for extraction of user's viewpoint. Furthermore, the proposed FOCUS constructs the concept structure based on a fuzzy clustering and produces user's desired documents by using fuzzy selection, fuzzy concept extraction, and fuzzy fisheye matching. An experimental result of ordered documents produced by the FOCUS is shown.

The FOCUS is proposed in II. In III , the experiment using movie materials, the effectiveness of the proposed FOCUS is confirmed.

II. Fuzzy dOCUment ordering System (FOCUS)

To put in order disordered documents on the web, the FISH VIEW system[1][2] has been used to choose the web documents in the user's point of view. In the FISH VIEW, input data is web-based documents as vast disordered documents (e.g., movie materials), and output data is ordered documents in user's point of view, respectively. The system generates Fish vectors from the input data and EDR (Electronic Dictionary Research Institute) concept-system-dictionary. Huge databases appeared on web as semantic web need the function to accept human factors for data retrieving and concept extraction/ construction. By applying fuzzy sets to the FISH VIEW system, FOCUS (Fuzzy dOCUment ordering System) is proposed taking into account fuzzy document selection, fuzzy concept extraction, and fuzzy fisheye

A. Overview of FOCUS

Before going into practical operations (operation mode) of the FOCUS, a user is requested to instruct his/her favorable documents to the FOCUS in order to adjust the fuzzy fisheye vectors to the user. In this instruction mode, the input of the FOCUS is desired movie materials and the output is ordered movie

materials in user's view (cf. Fig.1). User's desired fuzzy fisheye vectors are generated from the desired movie materials.

After finishing the instruction mode above, the FOCUS moves to the operation mode, where the input and the output are vast movie materials and ordered movie materials, respectively. Fuzzy fish view vectors to be matched by user's desires are generated by the vast movie materials.

Fig.1 shows the data flow of the FOCUS in both instruction and operation modes, where the two types of the input data are processed as follows;

Step.1: TFIDF

The TFIDF (Term Frequency and Inverse Document Frequency) assigns basic feature vectors to the documents of the input data based on fuzzy concept structure.

Step.2: Fuzzy Documents Selection

Users select documents as desired ones or desired ones by this step, in order to construct human-centered viewpoints.

Step.3: Fuzzy Fish View Extraction

Fuzzy semantic group sets are extracted by this step and the user's viewpoint is obtained in the step 2. By using the basic feature vectors, fuzzy fisheye vectors are extracted based on the fuzzy semantic group sets.

Step.4: Fuzzy Fisheye Matching

By using the fisheye vectors obtained in steps.1 -3, the fuzzy fisheye matching can be performed in terms of user's view.

The detailed descriptions are mentioned in the followings.

B. TFIDF

The input document dj is decomposed into words by eliminating stop words, e.g., space, period, comma. By using TFIDF (Term Frequency and Inverse Document Frequency) [3] and the decomposed words, the basic feature vector defined as

$$O_j = (w_j(W_1), \ldots, w_j(W_n)), \qquad (1)$$

is assigned to the document dj, whereas the weight $w_j(W_i)$**is the TFIDF of the word** W_i **coresponding to the document** d_j.

C. Fuzzy Document Selection

The FISH VIEW system [1][2] has D_P (user's "desired" document sets) and D_N (user's "undesired" document sets). A user is supposed to input either "desired" or "undesired" answer for the presented document dj through GUI. The system adds the dj to D_P or D_N according to the user's input "desired" or "undesired", respectively.

In the FOCUS, a user is able to make a fuzzy answer to the presented document dj through GUI of the system, i.e., "desired" membership degree and "undesired" membership degree independently. The system adds the fuzzy singletons dj with "desired" and "undesired" degrees to the fuzzy sets D_P and D_N, respectively.

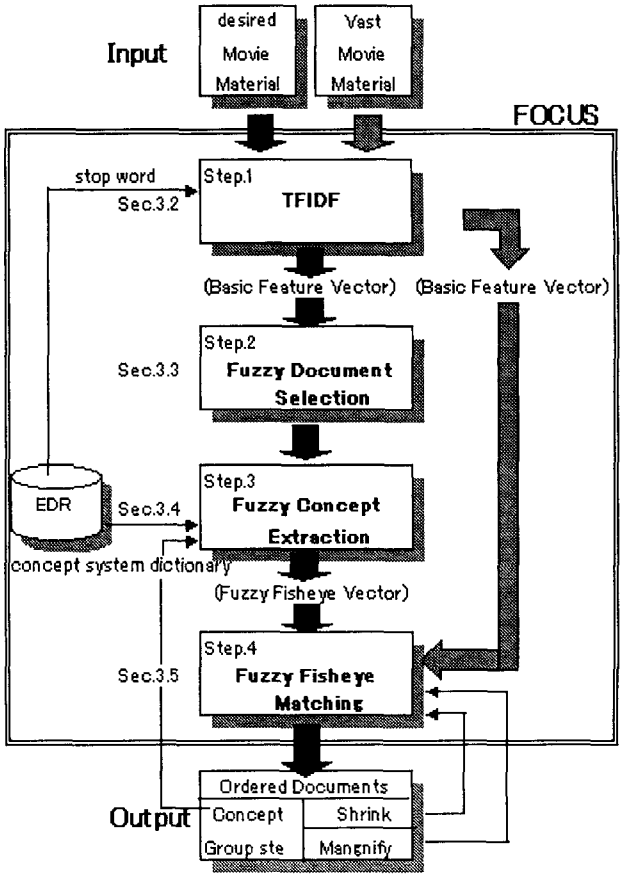

Fig.1 Flowchart of FOCUS

D. Fuzzy Fish View Extraction

The FOCUS organizes fuzzy semantic groups on

words appeared in all input documents by using EDR concept-system-dictionary as shown in Fig.2. The FOCUS reads EDR concept-system-dictionary according to input document dj to form fuzzy semantic groups. The fuzzy semantic group sets are generated by remaining links with bigger assigned weight defined by

$$f(W_i)=\alpha\frac{1}{|D_p|}\sum_{d_j\in D_p}w_j(W_i)-\frac{1}{|D_N|}\sum_{d_k\in D_N}w_k(W_i), \quad (2)$$

where $w_j(W_i)$ and $w_k(W_i)$ denote the W_i element of the basic feature vector assigned to document dj and dk, respectively. Using fuzzy semantic group and fuzzy semantic group set, new concept structure is organized to produce fuzzy fish eye vectors where fuzzy fish eye (vector) is a vector whose component is a pair of word W_i and its weight $f(W_i)$ in (2).

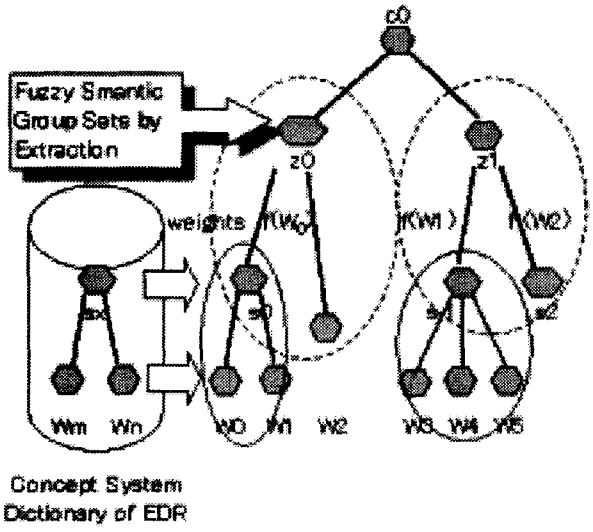

Fig.2 Fuzzy semantic groups and group sets

E. Fuzzy Fisheye Matching

The fuzzy fisheye matching between a fisheye vector and a basic feature vector of vast movie materials is done based on a similarity between two vectors and ordered documents are produced from the vast movie materials. The FOCUS supports the following functions.

(1) Shrink function

The fuzzy semantic group sets can be observed in terms of upper layer in Fig.2 to shrink the number of sets.

(2) Magnify function

The fuzzy semantic group sets can be observed in terms of lower layer in Fig.2, to magnify the number of sets.

III. Document Ordering Experiments by FOCUS

A screen shot of the proposed FOCUS is shown in Fig.3. A user operates FOCUS using mouse click button. On the upper left hand in Fig.3 EDR concept-system-dictionary, semantic groups, and semantic fuzzy group sets are displayed. The related fuzzy sets graphs are also displayed in the same place. On the lower left hand. desired or vast movie materials are displayed.

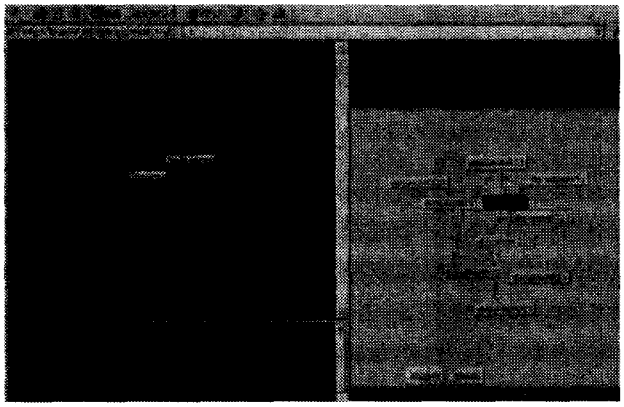

Fig.3 Screen shot of the FOCUS

A result of document ordering is displayed on the right hand. Size of document number box shows how interest it is, width of a bar between document number boxes shows how strong the relation is.

The shrink function and the magnify function are activated by click button at the bottom of right hand.

An experiment of ordered documents by the proposed FOCUS is shown in Fig.4 through some web-based documents related to movie topics. At this stage, it is unable to select document by clicking document number box. Instead of selecting document box from the lower left hand of screen, it is possible to select as shown in Fig.4.

A result of ordered documents produced by the FOCUS is illustrated in Fig.5.

In Fig.4,Document number 0413 has the title of "Film Threat Online" with hyper link. A user can

easily select this document by clicking this item as shown in Fig.5

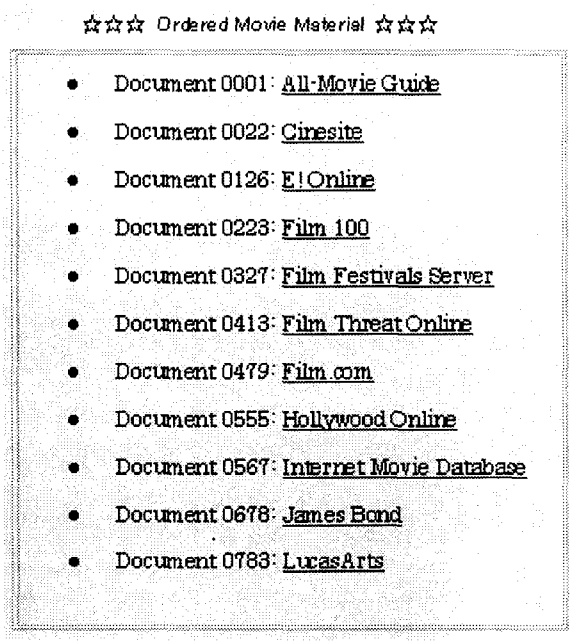

Fig.4 An example of ordered documents list

APE CANYON" TOURS NORTH AMERICA

(2003-08-14)

If you haven't seen Ape Canyon, the funniest damn Sasquatch movie out there, then here's your chance.

Beginning in October, filmmaker Jon Olsen will be taking Ape Canyon on a four to six month road trip across the United States and Canada.If anyone has a venue they would like to host an Ape Canyon screening at, go ahead and contact Jon at jon@normalpeoplelikeyou.com. *For more info, check out the Ape Canyon website.*

Fig.5 An example of selected movie (Document0413)

IV. Conclusions

The FOCUS (Fuzzy dOCUment ordering System) is proposed by a fuzzy extension of the conventional FISH VIEW system to deal with human factors in the semantic web, where fuzzy selection, fuzzy concept extraction, and fuzzy fish eye matching are applied to;

(1) Fuzzy Documents Selection: User's desired/undesired weights are expressed by fuzzy values..

(2) Fuzzy Concept Extraction; The semantic fuzzy group set is extracted as fuzzy value.

(3) Fuzzy Fisheye Matching; This fuzzy fisheye matching corresponds to a measurement of documents based on a fuzzy similarity

The experiment using movie materials shows that the ordered documents produced by FOCUS are harmonize with human-oriented document selections by users.

The FOCUS loads documents to computers connected with internet. In the semantic web age, however, web itself will be a vast data base consisting of XML based technologies instead of a closed data base in a computer. The FOCUS constructed by XML technology will open new semantic web applications.

References

[1]Y. Takama, M. Ishizuka, FISH VIEW System :A Document ordering Support System Employing Concept-structure-based Viewpoint Extraction (in Japanese), Journal of Information Society of Japan, Vol. 41, No.7, pp.1976-1986

[2]Y. Takama, M. Ishizuka, FISH-Eye Matching :A Document organizing Function Based on the Extraction of User's Viewpoint Using Concept Structure (in Japanese), Journal of Japanese Society for Artificial Intelligence, Vol. 14, No. 1, pp.93-101

[3] G. Salton, C. Buckley, Term Weighting Approachesin Automatic Text Retrieval, Information Processing and Management, Vol. 24, No.5, pp.513-523, 1998.

[4] Y. Takama, M. Kawabe, K. Hirota, Kansei-keyword Extraction from Japanese Film Scenario Using Sensitivity Information. Joint 9th IFSA World Congress and 20th NAFIPS International Conference (IFSA/NAFIPS2001), pp. 2900-2905, 2001 (Canada)