Semantic Concept Operations for Fuzzy Document Ordering System and Their Application to Knowledge Discovery

Tadashi Ohashi, Hajime Nobuhara, Kaoru Hirota
Tokyo Institute of Technology
G3-49,4259 Nagatsuta, Midori-ku, Yokohama 226-8502, JAPAN
{ohashi,nobuhara,hirota}@hrt.dis.titech.ac.jp

Abstract：A shrink operation of FOCUS (Fuzzy dOCUment ordering System) is proposed as a tool for changing user's viewpoint in knowledge structures of documents. The FOCUS produces the user's preference vector space and its concept hierarchical structure based on CSD (Concept System Dictionary) of EDR (Japan Electronic Dictionary Research Institute). By using the hierarchical structure, the shrink operation, i.e., shrinkage of user's preference vector space, avails to do fuzzy document marching in user's laugh point of view. Through document ordering experiments by 4 examinees using 20 given documents and 40 unknown ones extracted from IMDB (Internet Movie Data Base), the effectiveness of shrink operations is confirmed.

１. Introduction

To extract user's preference documents from a deluge of information, FISHVIEW[1][2] and FOCUS (Fuzzy dOCUment ordering System)[3] that is based on fuzzy set theory[4], have been proposed. The FOCUS can be extended to perform knowledge discovery using semantic concept structure. A constructive method of semantic concept structure by using CSD (Concept System Dictionary) that includes 400,000 concepts in EDR (Japan Electronic Dictionary Research Institute [5]) and the shrink operation that perform fuzzy document marching in user's laugh point of view, is proposed.

In total, 20 given documents and 40 unknown documents extracted from Internet Movie DataBase (IMDB[6]) are used for documents ordering experiments by 4 examinees from different countries and the effectiveness of the proposed shrink operation is confirmed.

２. Fuzzy dOCUment ordering System (FOCUS)

The FOCUS orders web-based unknown documents in accordance with user's preference and performs knowledge discovery by concept structure operations as data mining. The FOCUS can learn user's preference by which user selects given documents to be suitable for user's preference in advance. Architecture of the FOCUS is given by the following five units (Fig. 1).

・**Basic Feature Vector Generator (BFVG)**

Given (training) documents are translated into the Basic Feature Vectors (BFV) using the TFIDF (Term Frequency and Inverse Document Frequency) [7].

・**User's Preference Configuration Unit (UPCU)**

In the process of document selection, a user assigns Membership Value (MV) expressing his/her preference degree to the given web-based documents. UPCU generates User's Preference Vector (UPV) based on the assigned MV and BFV.

・**Fuzzy Document Matching Unit (FDMU)**

Web-based unknown documents are translated into Unknown Document Vector (UDV) using Term Frequency (TF). User's Preference Degree (UPD) of each web-based document is calculated based on UPV, generated in ***2.3***.

・**Semantic Concept Operation Unit （SCOU）**

By using CSD (Concept System Dictionary (CSD) by Japan Electronic Dictionary Research Institute, Ltd. (EDR)), Semantic Concept [8] [9] [10] Operation Unit offers the shrink operation for concept structure to perform data mining and knowledge discovery.

2.1 Basic Feature Vector Generator (BFVG)

The given (training) web-based documents are expressed as a set D_G ,

$$D_G = \{d_1^{(G)}, d_2^{(G)}, \dots, d_n^{(G)}, \dots, d_N^{(G)}\}. \qquad (1)$$

The word set W is defined as

$$W = \{W_1, W_2, .., W_m, .., W_M\}, \qquad (2)$$

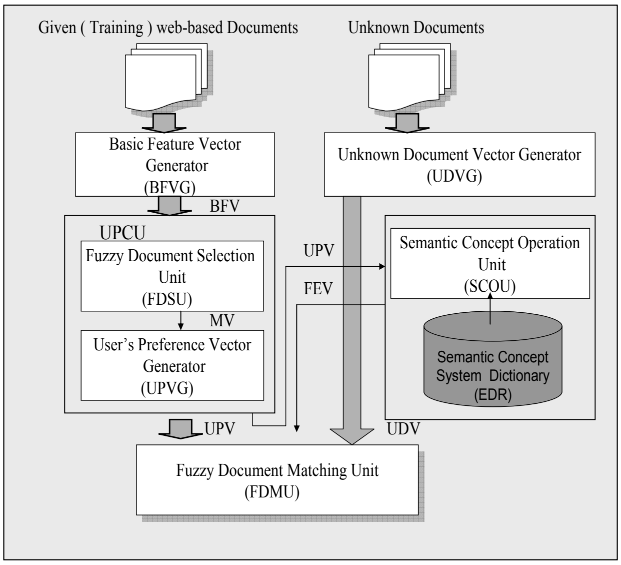
Given (Training) web-based Documents
Unknown Documents
Basic Feature Vector Generator (BFVG)
BFV
Unknown Document Vector Generator (UDVG)
UPCU
Fuzzy Document Selection Unit (FDSU)
MV
User's Preference Vector Generator (UPVG)
UPV
FEV
Semantic Concept Operation Unit (SCOU)
Semantic Concept System Dictionary (EDR)
UPV
UDV
Fuzzy Document Matching Unit (FDMU)

Fig. 1 The data flow of FOCUS

where W_m ($m = 1, 2, \dots\dots, M$) . In the vector space, the document $d_j^{(G)}$ is represented by the basic feature vector $O_j^{(G)}$,

$$O_j^{(G)} = (w_j^{(G)}(W_1), w_j^{(G)}(W_2), \dots, w_j^{(G)}(W_m), .., w_j^{(G)}(W_M)), \qquad (3)$$

in which $w_j^{(G)}(W_m)$ is obtained as TFIDF defined as

$$w_j^{(G)}(W_m) = tf_j^{m(G)} \cdot \log \frac{N}{df_m} \quad (\in \mathbb{R}), \qquad (4)$$

where tf_j^m (≥ 0) is occurrence frequency of term W_m, df_m (≥ 1) is the number of documents including term W_m, and N denotes the number of all documents.

2.2 User's Preference Configuration Unit (UPCU)

For given (training) web-based documents D_G, the user determines whether they belong to the preferable document set D_P or to the non-preferable document set D_N. Users determine the membership value of $d_j^{(G)} \in D_G$ as

$$\mu_P(d_j^{(G)}), \mu_N(d_j^{(G)}) \in [0,1]. \qquad (5)$$

Each weight of each axis of the vector space is represented as

$$u = ((u(W_1), u(W_2), \dots, u(W_m), \dots, u(W_M)), \qquad (6)$$

where

$$u(W_i) = \sum_{j=1}^{N} (\alpha\mu_P(d_j^{(G)}) - \mu_N(d_j^{(G)}))w_j^{(G)}(W_i), \qquad (7)$$

$$\alpha \in [0,1]\text{:constant}. \qquad (8)$$

As in Eq. (5), a user can assign the membership values $\mu_P(d_j^{(G)})$ and $\mu_N(d_j^{(G)})$ appropriately to a document $d_j^{(G)}$. By using fuzzy document selection, the FOCUS can generate user's preference vector spaces based on the fuzzy evaluation.

2.3 Fuzzy Document Matching Unit (FDMU)

Unknown documents $d_x^{(U)}(x=1,2,\dots,N)$ are translated into the unknown document vectors $F_x^{(U)}(x = 1, 2, \dots, N')$ defined as

$$F_x^{(U)} = (tf_x^{1(U)}, tf_x^{2(U)}, \dots, tf_x^{m(U)}, \dots, tf_x^{M(U)}), \qquad (9)$$

where $tf_j^{m(U)}$ denotes term W_m occurrence frequency in unknown documents $d_x^{(U)}$.

User's Preference Degree $UPD(u, d_x^{(U)})$ of document $d_x^{(U)}$ in the vector space with user's preference is obtained by

$$UPD(u, d_x^{(U)}) = \sum_{k=1}^{M} u(W_K) \cdot tf_x^{K(U)} \quad (\in \mathbb{R}). \qquad (10)$$

The FDMU arranges unknown documents in ascending order by using the obtained user's preference degree as described in (10).

3. Rough Documents Matching by Shrink Operations

By using concept system dictionary (CSD) which contains 400,000 concepts and is one of the dictionaries of Japan Electronic Dictionary Research Institute Ltd. (EDR [5]), semantic concept structures are formatted from user's preference vector space (Eq. (6)). Figure 2 shows an example of semantic concept structures. Given/unknown documents are expressed by the semantic concept structures [8] [9] [10].

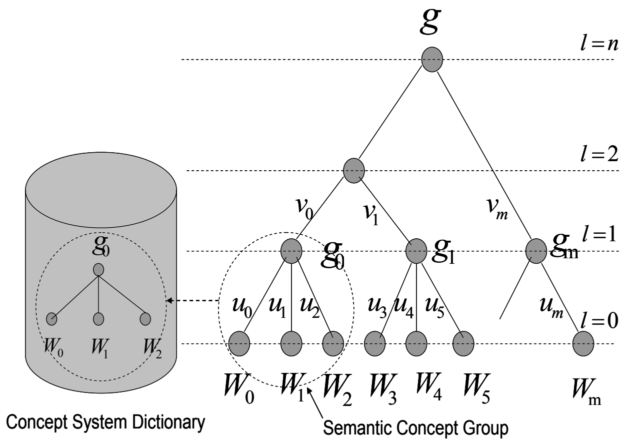
g
$l=n$
$l=2$
$l=1$
$l=0$
v_0
v_1
v_m
g_0
g_1
g_m
u_0 u_1 u_2 u_3 u_4 u_5 u_m
W_0 W_1 W_2 W_3 W_4 W_5 W_m
Concept System Dictionary
Semantic Concept Group

Fig. 2 Semantic Concept Structure

In Fig. 2, l denotes layer number of hierarchical structure, g denotes semantic concept group and v denotes weight of semantic concept group, respectively.

3.1 Shrink Operation

Shrinks operations are done by the following steps.

・If $l=0$

$$G^{(1)}=S(W)=\{g_i\in C|\exists w_j\in W, W_j\to g_i\}, \qquad (11)$$

where $G=\{g_1,\ldots\ldots,g_m\in C\}$ stands for semantic concept group with layer $l=0$, C denotes all concepts defined in CSD, $W_j\to g_i$ denotes that a concept relation exists in CSD. Figure 3 shows an example of a shrink operation from W to $C^{(1)}$.

Each weight of concept g_i summarizes the user's preference degree relating word set, and is computed by

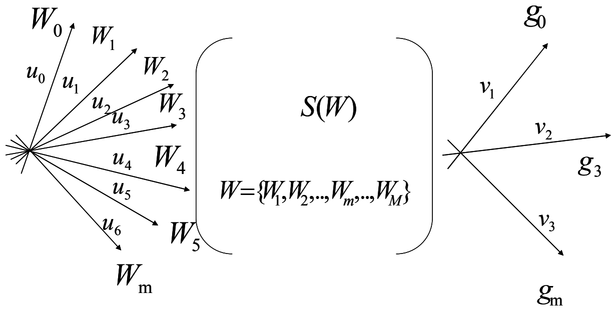

Fig.3 Shrink operation of Wlist.

$$v(g_i)=\sum_{W_k\in W(g_i)}u(W_k), \qquad (12)$$

where $W(g_i)=\{W_j\in W\,|\,W_j\to g_i\}$.

・If $l\neq 0$

$$G^{(L+1)}=S(G^{(L)})=\{g_i\in C|\exists g_j\in G^{(L)}, g_j\to g_i\}. \qquad (13)$$

Each weight of concept g_i is computed by

$$v(g_i)=\sum_{g_k\in G^{(L)}(g_i)}v(g_k), \qquad (14)$$

where $G^{(L)}(g_i)=\{g_j\in C\,|\,g_j\to g_i\}$.

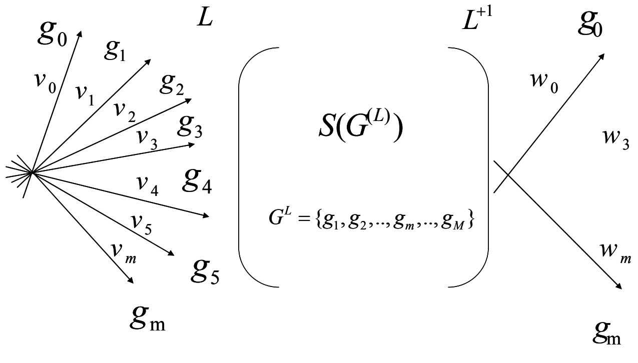

Fig.4 Shrink Operation Slist $(l\neq 0)$

3.2 Fuzzy Document Matching for Shrink Operation

After applying the shrink operation, User's Preference Degree $UPDs(u,d_x^{(U)})$ of document $d_x^{(U)}$ in the vector space with user's preference is obtained by

・If $l=0$

$$UPDs(u,d_x^{(U)})=\sum_{k=1}^{M}\{(\sum_{i=0}^{i}u(S_K))\cdot(\sum_{j=0}^{j}tf_x^{K(U)}(S_K))\}(\in\mathbb{R}), \qquad (15)$$

・If $l\neq 0$

$$UPDs(g,d_x^{(U)})=\sum_{k=1}^{M}\{(\sum_{i=0}^{i}g(S_K))\cdot(\sum_{j=0}^{j}tf_x^{K(U)}(S_K))\}(\in\mathbb{R}). \qquad (16)$$

The fuzzy document matching using (15) and (16) is illustrated in Fig.5.

Given documents and unknown documents are converted into Fisheye vectors which are organized by semantic concept structures and changed by shrink operations. Similarity between both given documents and unknown documents are calculated in the fisheye vector spaces.

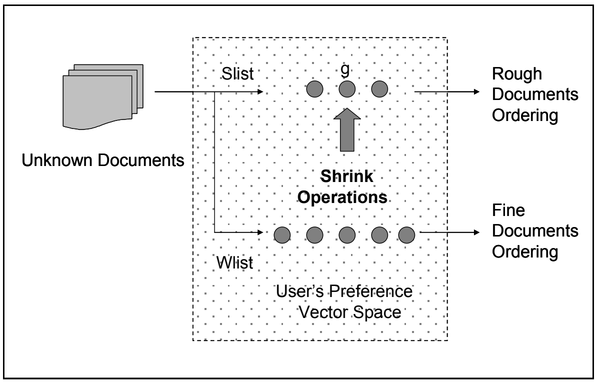

Fig.5 Semantic Concept Operations applied to Fuzzy Document Matching

4. Document Ordering Experiments by Shrink Operations

Document ordering experiments are done by 4 examinees to evaluate the effectiveness of shrink operations of the FOCUS. The documents of 60 movies are selected from the internet movie database (IMDB) [6] , where each document contains 300 words synopses and these documents are composed into 4 genres i.e., (1) action (2) comedy (3) horror (4) music. For each genre, 5/10 movies are selected as given/unknown documents (total 20/40 documents). The procedure for experiments are as follows. First, 20 documents from 4 genres are presented at random to each examinee. First and second evaluation sheets is given to each examinee. An examinee must select one of the genres as desired one and declare what he wants to see on the evaluation sheet. 40 unknown documents are

given to an examinee. One must judge whether he wants to see each movie or not by assigning desirability and non-desirability membership degrees (Eq.5) to each movie. The FOCUS reads an evaluation sheet data values. Then, FOCUS executes unknown document ordering in accordance with user's preference degree from 40 documents in the 4 genres. Ordered documents and evaluation sheet are given to each examinee. An examinee calculates the document precision ratio β defined as (17) to confirm that the FOCUS orders documents to meet user's preference properly.

$$\beta=\frac{\text{Number of desired documents}}{\text{Number of total documents}} \qquad (17)$$

An example of ordered unknown documents generated by the FOCUS is shown in Table 3 and 4.

	QUESTIONNAIRE	EVALUATION POINTS
1	Ordered documents are desired？	Very Satisfied So so Very Unsatisfied ▲5 ▲4 ▲3 ▲2 ▲1
2	Ordered documents are desired？	Very Satisfied So so Very Unsatisfied ▲5 ▲4 ▲3 ▲2 ▲1
3	Experiments are easy to understand？	Very Satisfied So so Very Unsatisfied ▲5 ▲4 ▲3 ▲2 ▲1
4	Noise （Desirded／Undesired） is suppressed？	Very Satisfied So so Very Unsatisfied ▲5 ▲4 ▲3 ▲2 ▲1
5	Please Write Your Knowledge Discovery From Results	
6	Precision Rate	

Table.1 Questionnaire for 1st Evaluation Sheet

	QUESTIONNAIRE	EVALUATION POINTS
1	Is second experiment better orderThan first experiment？	
2	What is your own knowledge discovery？	
3	What is Your total impressions？	

Table.2 Questionnaire for 2nd Evaluation Sheet

In Table.3, ordered documents are mentioned in details.The examinee has selected music genre. There are 4 action movies in top 10, where Wlist is described as non-duplicated words (excluding stop words).

In Table.4, UD10 which corresponds to the movie "Singin'in in the rain" is on the top after shrink operations. The movie is a comedy but action-related words like “dance”, “time” and “caused” are involved.

Order	Doc. No.	Genre	Name of movie	Wlist [words]
1	UD06	Action	“General”	257
2	UD28	Horror	“Others”	340
3	UD19	Commedy	”Annie Hall”	271
4	UD03	Action	“North by Northwest”	82
5	UD08	Action	“Saving Private”	144
6	UD04	Action	“Laurence”	320
7	UD29	Horror	“Rosemary”	148
8	**UD38**	**Music**	“Un Coeur en Hiver”	168
9	UD00	Action	“Buono”	257
10	**UD33**	**Music**	“Hard days”	83

Table.3 Detailed Ordered Documents (Before Shrink Operation)

Table4 shows the result of shrink operation.

Order	Doc. No.	Genre	Name of movie	Wlist[words] /Slist[words]
1	UD10	Comedy	”Singin'in In The Rain”	127/5
2	UD19	Comedy	“Annie Hall”	205/2
3	UD28	Horror	“Othes”	293/5
4	UD00	Action	“Buono”	257/4
5	UD33	**Music**	“Hard Days”	68/2
6	UD27	Horror	“Invasion of The Body”	30/1
7	UD29	Horror	“Rosemary”	124/6
8	UD21	Horror	“Bridge of Frankenstein”	120/2
9	UD31	**Music**	“Yankee Double”	55/4
10	UD06	Action	“General”	219/1

Table.4 Detailed Ordered Documents (After Shrink Operation)

The result of average first evaluation sheet value of 4 examinees is as follows.

Qustionnaire	Averege Evaluation
Ordered documents are desired?	4
Contents are desired?	3. 5
Experiments are easy to understand?	3. 5
Noise (desired/undesired) supressed?	3
Precision Rate β	0.45

Table.5 Result of 1st Evaluation Sheet

Fig.6 shows user’s preference vector (uWi) whose length is 1821 words and its shrink user’s preference vector (sWi) whose length is 21 words.

As results of 2nd evaluation sheets, knowledge discoveries are, e.g.,

(1) “I felt that second result as shrink operation is better than first result without shrink operation”

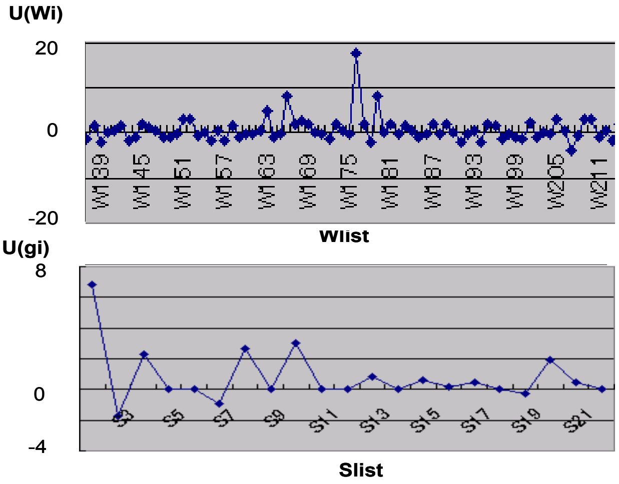

Fig.6 User’s Preference Vectors after Shrink Operations.

(2) “Comparing with first experiments, different order of the documents can give person a transition of contents".

(3) “From the comfortable topics, it is easy to accept, and the feeling about these documents is not changed strongly”.

5. Conclusions

A shrink operation of FOCUS is proposed as a tool of data-mining for knowledge discovery. The effectiveness of the shrink operation of the FOCUS is confirmed by 4 examinees from different countries.

The results of the experiments are summarized as:

(1) The shrink operation makes users easy to select documents in accordance with users preferable vector spaces.

(2) The fuzzy document matching with shrink operations is effective to find rough document in accordance with users preference.

(3) Processing time of shrink operation’s document matching is reduced by a factor of 10 from that of initial fuzzy document matching.

(4) The shrink operations of FOCUS is programmed by Java Language (Visual J++6.0) to realize the

performance as quickly as possible.

The shrink operation of FOCUS is applied to only web documents but its use is expected to cover more general human oriented semantic web concepts [8]:

(1)The FOCUS's semantic concept operation technology aims the future semantic web age to human-centered daily activities like ubiquitous life and broad band age.

(2) The multimedia information retrieval is also possible by applying the FOCUS technology.

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] T. Ohashi, H. Nobuhara, K. Hirota., *A document Ordering Support System Employing Concept Structure based on Fuzzy Fish View Extraction*, ISIS2003, pp.98-101 (2003)

[4] L. A. Zadeh, Fuzzy sets, Information Control, 8, pp.338-353 (1965)

[5] http://www.iijnet.or.jp/edr/

[6] http://www.imdb.com/

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

[8]Berners-Lee, Tim, Handler, James and Lassila, Ora: *The Semantic web*, Scientific America, Vol. 284, No.5, pp.34-43 (May, 2001)

[9]B. Richardson and L.J.Mazlack, *Approximate Ontology Merging For The Semantic,* 23th NAFIPS International Conference (NAFIPS2004).

[10] V. Cross, *Fuzzy Semantic Distance Measures Between Ontological Concepts,* 23th NAFIPS International Conference (NAFIPS2004).
