

Automatic generation of document semantics for the e-science Knowledge Grid [☆]

Hai Zhuge ^{*}, Xiangfeng Luo

*China Knowledge Grid Research Group, Key Lab of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, P.O. Box 2704, Beijing 100080, PR China*

Received 24 November 2004; received in revised form 21 August 2005; accepted 26 August 2005
Available online 7 October 2005

Abstract

This paper proposes an approach to automatically generate semantics for scientific e-documents, and presents its applications in e-document understanding, question answering and question refinement. The approach uses not only keywords and their relations in e-documents, but also the implied meaning of co-occurred keywords that is hard to be exploited, represented and derived by previous semantic representation approaches. The proposed approach facilitates automatic construction, composition, decomposition and derivation of semantics at different granularity levels, which lay the basis for realizing intelligent services of the e-science Knowledge Grid. © 2005 Elsevier Inc. All rights reserved.

Keywords: Fuzzy cognitive map; Document understanding; Semantics; Knowledge Grid; Semantic Web

1. Introduction

The Knowledge Grid is an intelligent interconnection environment that enables people or roles to effectively capture, publish, share and manage knowledge resources. It provides on-demand services to support innovation, cooperative teamwork, problem-solving and decision-making by adopting the technologies developed during work toward the future interconnection environment (Zhuge, 2004b). The e-science Knowledge Grid is an application where scientific documents need to be efficiently processed based on the understanding of content to effectively support scientific activities. The intelligent services of the Knowledge Grid require semantics to be automatically constructed, composed, decomposed and derived at different granularity levels.

Current Semantic Web approaches (Berners-Lee et al., 2001) can help but are not enough to meet these requirements. The XML-based RDF (Resource Description Framework) describes Web resources by using the object-attribute-value model. RDFS (RDF schema) expresses the metadata of Web resources by defining vocabulary, class-based structure, and constraints (Heflin and Hendler, 2001; W3C). SHOE (frame-based Simple HTML Ontology Extensions) supports the Horn clause axioms. OIL (Ontology Inference Layer) supports the description logics and frame. OWL (Web Ontology Language, Smith et al., 2003) describes classes and their relations.

Fuzzy Cognitive Map (FCM) uses adjacency matrix to represent relational knowledge. The reasoning of FCM is realized by matrix operations (Liu and Satur, 1999; Noha and Lee, 2000; Kosko, 1997; Leea and Lee, 2003).

An active document framework (ADF) is a self-representable, self-explainable, and self-executable document mechanism (Zhuge, 2003). It represents document content in four aspects: granularity hierarchy, template hierarchy, background knowledge, and semantic links between fragments.

[☆] Research work is supported by the National Basic Research Program of China (973 project no. 2003CB317000) and the National Science Foundation of China (grants 60273020, 60402016 and 70271007).

^{*} Corresponding author. Tel.: +86 1062565533; fax: +86 1062567724.
E-mail address: zhuge@ict.ac.cn (H. Zhuge).

Based on the FCM and the idea of ADF, we propose an approach that generates document semantics by considering not only keywords and their relations, but also the implied meaning of co-occurred keywords in documents. Co-occurred keywords imply certain meaning. For example, if the keywords “terrorist”, “casualty”, “explode” and “panic” co-occur in the same section or paragraph, then the topic about “terror event” is likely to be discussed. The topic determined by multiple co-occurred keywords usually implies rich semantics, further, the meaning of keywords will be specified within the determined topic.

2. The generation of document semantics based on fuzzy cognitive maps

2.1. Fuzzy cognitive maps

The Fuzzy Cognitive Map (FCM) is a graphical model for causal knowledge representation. It can represent not only the causal relations between keywords or phrases but also the knowledge of different granularity levels. An FCM comprises concepts (nodes) and the relations between concepts (arrows). The mathematic model of FCM is as follows (Kosko, 1997):

$$V_{c_j}(t + 1) = f \left(\sum_{\substack{i=1 \\ i \neq j}}^N V_{c_i}(t)w_{ij} \right) \tag{1}$$

where V_{c_i} and V_{c_j} are the state values of the cause concept and the effect concept respectively; w_{ij} is the weight of the causal relation from C_i to C_j . $f(x)$ is the threshold function of concept C_j (this paper uses $\tanh(x)$ function).

Fig. 1(a) illustrates a simple FCM, where C_i is a concept with a state value. The state value can be a fuzzy value within $[0, 1]$ that represents the existent degree of a concept, or a bivalent logic in $\{0,1\}$ that represents a concept’s open/close state. The weight w_{ij} of an arrow indicates the influence degree from the cause concept C_i to effect concept C_j , which can be a fuzzy value within $[-1, 1]$ or a trivalent logic within $\{-1,0,1\}$. If the weight is positive, then the increase/decrease of the value of concept C_i leads to

the increase/decrease of the state value of concept C_j . If the weight is negative, the increase/decrease of the state value of concept C_i leads to the decrease/increase of the state value of concept C_j . The adjacency matrix corresponding to the simple FCM is shown in Fig. 1(b).

FCMs have been used in many fields such as Geographic Information Systems (Liu and Satur, 1999), tacit knowledge management (Noha and Lee, 2000), and virtual world (Kosko, 1997).

Documents include not only the cause–effect relations, but also many other relations among concepts. Previous FCMs only support the reasoning of cause–effect relations. We extend the reasoning mechanism of FCM by first deriving out other relations (like sequential or probabilistic causal relations) according to the rules (Zhuge, 2003, 2004a) and the computing model (Zhuge, 2004b) before carrying out the reasoning of FCMs.

2.2. Document semantic templates based on fuzzy cognitive maps

Scientific documents generally include title, abstract, keywords, introduction, body, conclusions and reference. The body part is composed of sections, which further consist of paragraphs or subsections.

Definition 1. *Knowledge point* refers to the basic questions, topics, viewpoints and methods of a domain. A section may have several knowledge points.

For example, “FCM-based document representation”, “FCMs’ automatic construction” and “FCM-based document understanding” imply three knowledge points.

Definition 2. *Theme concept* (denoted as C_i^0) is a special concept of FCM, which reflects the implied meaning of the other co-occurred concepts in the FCM.

For example, C_0^0 is an FCM’s theme concept in Fig. 2, which can reflect the implied meaning of these co-occurred keywords (concepts) “FCM”, “semantic”, “reasoning”, “causal”, “keywords”, “representation”, “template” and “document”.

Definition 3. *Element semantic template* (denoted as E-FCM) is an FCM that can represent the content of a knowledge point.

According to the definitions of E-FCM and the theme concept, a causal relation exists and directs from theme concept C_j^0 to C_i in reverse while there is a causal relation from the cause concept C_i to the theme concept C_j^0 in E-FCM. For example, E-FCM1 (shown in Fig. 2, E1 is its adjacency matrix) can be represented as Fig. 3 according to the “reverse” characteristic of E-FCM.

Definition 4. *Section semantic template* (denoted as S-FCM) is an augmented E-FCM belonging to a section or several paragraphs, which constructs the semantics of a section or several paragraphs.

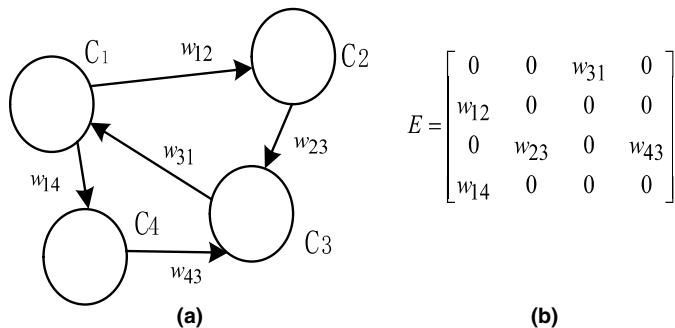
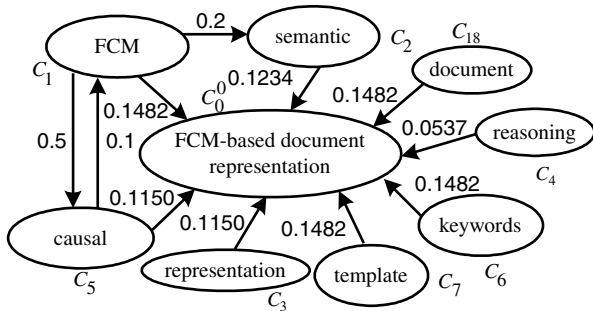


Fig. 1. A graphical FCM and its matrix representation.



$$E1 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1482 & 0 & 0.2 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0.1234 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1150 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0537 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1150 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1482 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1482 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1482 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Fig. 2. The element semantic template about the knowledge point “FCM-based e-document representation” (denoted as E-FCM1).

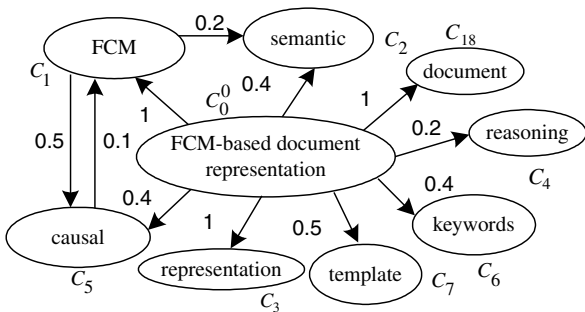


Fig. 3. The FCM formed by reversing the causal relations in E-FCM1.

Fig. 12 is an example of S-FCM.

Definition 5. Document semantic template (denoted as D-FCM) is an augmented S-FCM belonging to a document or its body, which generates the semantics of a document or its body.

D-FCM is an integrated pattern for a document. Fig. 14 is an example of D-FCM.

According to the addible characteristic of FCMs, D-FCM can be automatically constructed by combining S-FCMs, and S-FCM can be automatically constructed by combining E-FCMs. So the problem of automatic generation of document semantics is transformed to the problem of automatic construction of E-FCMs.

2.3. Generation of document semantics

The automatic generation process of document semantics is described in Fig. 4. This paper focuses on the gray squares. The candidate document passes the domain keywords repository, if the document belongs to the domain, generates its E-FCMs, S-FCMs and D-FCM. If the document does not belong to the domain, then re-selects the candidate document from the candidate document repository.

The domain keyword repository is constructed as follows:

- Step 1. Select documents from the domain and then find their keywords;
- Step 2. Recommend these keywords to experts; and,
- Step 3. The experts modify these keywords or add some keywords.

2.3.1. Automatic generation of document’s element semantic templates

The semantics and content of knowledge points represented by E-FCMs are machine-understandable. E-FCM consists of a theme concept and a set of co-occurred concepts (keywords) that can reflect the content and semantics of the theme concept. The construction criterions of E-FCM are as follows:

Criterion 1. Theme concept in E-FCM corresponds to the title of the training section or paragraphs.

Criterion 2. Concepts in E-FCM correspond to the keywords in a training section or paragraphs, or a set of basic keywords and phrases or keyword repository of this domain.

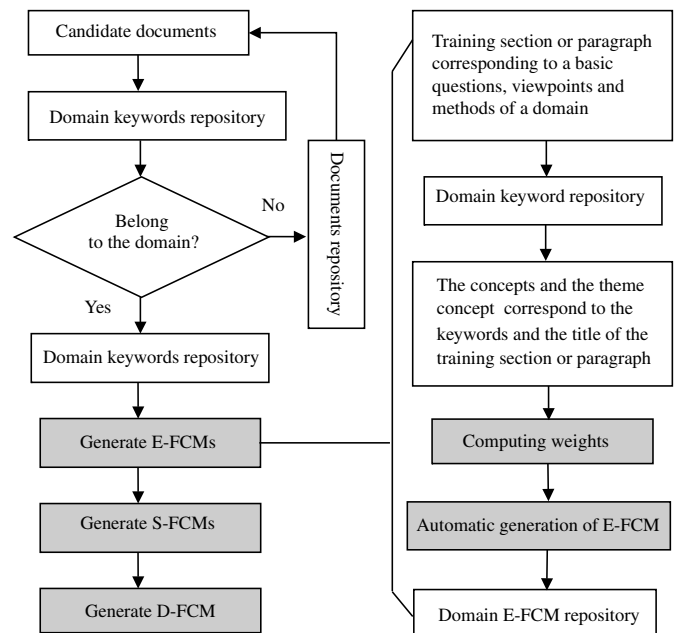


Fig. 4. The process for generating document semantics.

Table 2
E-FCM2's weights are determined automatically

	C_1	C_4	C_6	C_7	C_{22}^0	C_{24}	C_{25}	C_{26}	C_{29}	C_{30}
C_1	0	0.5	0	0.5	0.1331	0	0	0	0	0
C_4	0	0	0	0	0.0490	0	0	0	0	0
C_6	0	0	0	0	0.1331	0	0	0	0	0
C_7	0	0	0	0	0.1331	0.7	0	0	0	0
C_{22}^0	0	0	0	0	0	0	0	0	0	0
C_{24}	0	0	0	0.7	0.1331	0	0.6	0	0	0
C_{25}	0	0	0	0	0.1056	0.6	0	0	0	0
C_{26}	0	0	0	0	0.0490	0	0	0	0	0
C_{29}	0	0	0	0	0.1308	0	0	0	0	0
C_{30}	0	0	0	0	0.1331	0	0	0	0	0

2.3.1.3. *Experiments on the automatic construction of element semantic template.* The experiment is to construct the E-FCMs reflecting the knowledge points: “FCM-based complex question automatic answering”, “FCM-based e-document representation”, “FCMs’ automatic construction”, “fuzzy cognitive map”, “FCM-based scientific e-document understanding” and “FCM-based vague question refining”. Here Sections 3.2, 2.2, 2.3, 2.1, 3.1 and 3.3 of this paper are used as training sections to construct the E-FCMs. From the training sections and the keyword repository of the document semantic representation and understanding (denoted as *Domain1*), the E-FCMs weights can be obtained. The weights of “FCM-based complex question automatic answering” are shown in Table 2, and its E-FCM is shown in Fig. 4. Other E-FCMs are shown in Figs. 2 and 5–9 respectively.

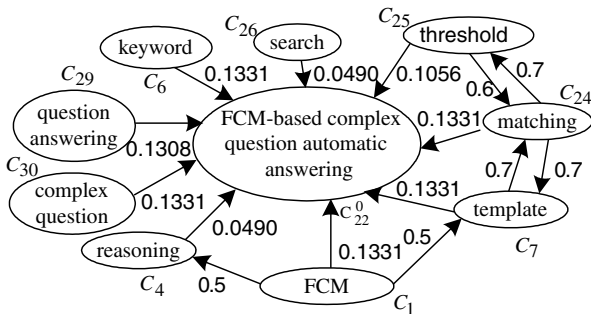


Fig. 5. The element semantic template about the knowledge point “FCM-based complex question automatic answering” (denoted as E-FCM2).

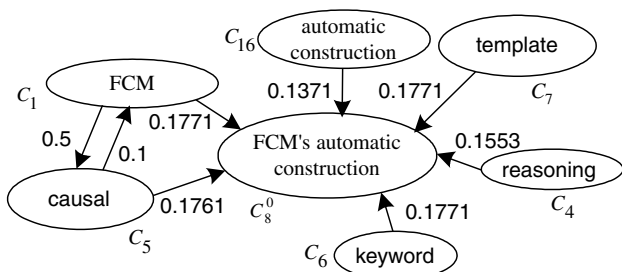


Fig. 6. The element semantic template about the knowledge point “FCMs’ automatic construction” (denoted as E-FCM3).

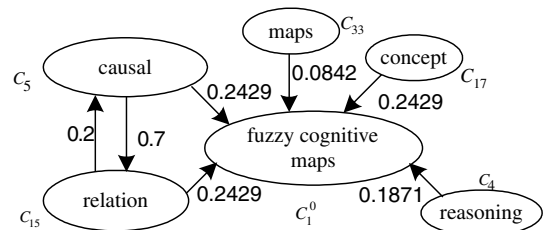


Fig. 7. The element semantic template about the knowledge point “fuzzy cognitive map” (denoted as E-FCM4).

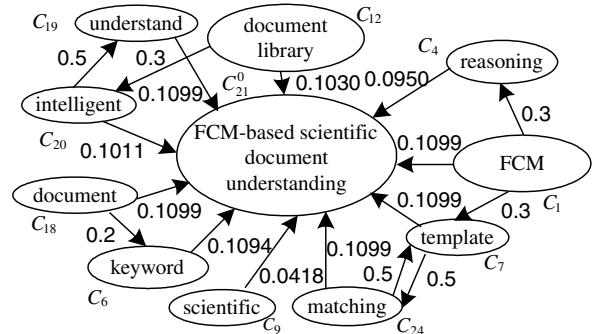


Fig. 8. The element semantic template about the knowledge point “FCM-based scientific document understanding” (denoted as E-FCM5).

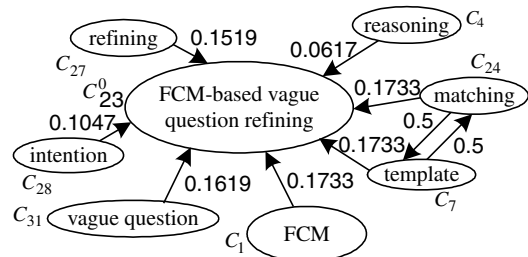


Fig. 9. The element semantic template about the knowledge point “FCM-based vague question refining” (denoted as E-FCM6).

2.3.1.4. *Validity of the constructed element semantic templates.* Here Sections 1, 2.1–2.4 and 3.1–3.3 are used to verify the validity of E-FCM2 and E-FCM5. The results are shown in Tables 3 and 4. If all the thresholds are 0.85,

Table 3
Reasoning results when using Sections 1, 2.1–2.4 and 3.1–3.3 to verify the validity of E-FCM2

Section	1	2.1	2.2	2.3	2.4	3.1	3.2	3.3
V_{C22}^0	0.5831	0.2118	0.5423	0.6293	0.2283	0.7200	0.9954	0.5662

Table 4
Reasoning results when using Sections 1, 2.1–2.4 and 3.1–3.3 to verify the validity of E-FCM5

Section	1	2.1	2.2	2.3	2.4	3.1	3.2	3.3
V_{C21}^0	0.6613	0.1849	0.5874	0.5754	0.2118	0.9929	0.7438	0.4816

then Sections 3.2 and 3.1 can be the answer of E-FCM2 and E-FCM5 respectively.

2.3.2. Automatic generation of section semantic template

The section semantic template (S-FCM) is constructed by combining E-FCMs as follows:

- Step 1. Combine E-FCMs belonging to a section to form an augmented E-FCM;
- Step 2. Delete the concepts that have no direct relations with the other E-FCMs' theme concepts in the augmented E-FCM to form an S-FCM.

Figs. 10 and 11 shows the augmented E-FCM of E-FCM1 (Fig. 2), E-FCM3 (Fig. 6) and E-FCM4 (Fig. 7). S-FCM1 (Fig. 12) is obtained by deleting the redundant concepts, those having no direct relations with the other E-FCMs' theme concepts in the augmented E-FCM. S-FCM1 is machine understandable and expresses the semantics of knowledge points “FCM-based e-document representation”, “FCMs automatic construction” and “fuzzy cognitive map”, and their relations in Section 2 of this paper. The existence degrees of these knowledge points can be reflected by theme concepts' state values and weights of S-FCM1. The importance of concepts can be determined by

$$V_{ck} = \begin{cases} 1 & \text{if } V'_{ck}(1 + \alpha u) \geq 1 \\ V'_{ck}(1 + \alpha u) & \text{otherwise} \end{cases} \quad (8)$$

where V'_{ck} is the maximal state value of the overlapped concept C_k , u is the times of overlapping, and α is the coefficient measuring overlapping between matrices.

2.3.3. Automatic generation of document semantic templates

The construction of the document semantic template (D-FCM) needs to know the relations and their weights between theme concepts.

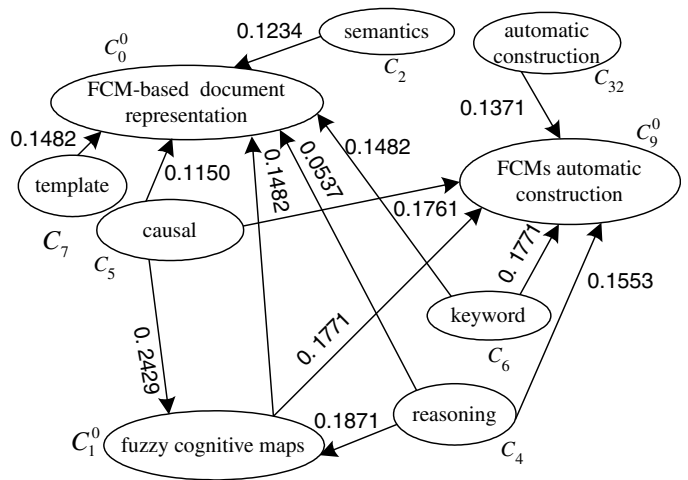


Fig. 11. Augmented E-FCM generated from element semantic template 1 (Fig. 2), element semantic template 3 (Fig. 5) and element semantic template 4 (Fig. 6).

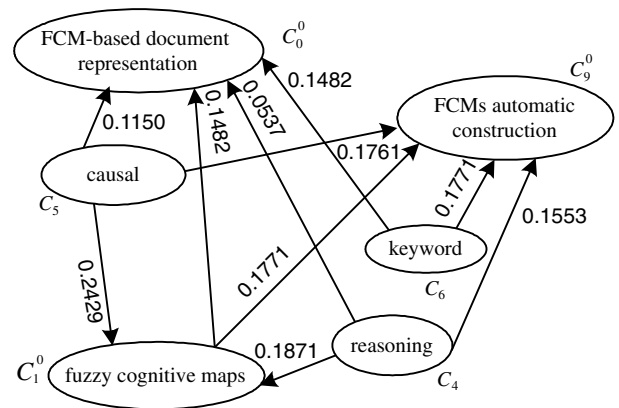


Fig. 12. Section semantic template of Section 2 (S-FCM1).

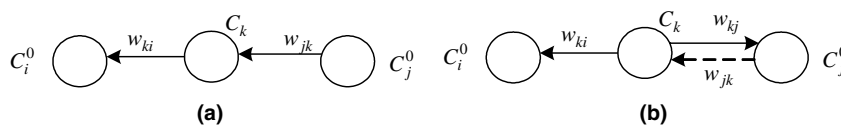


Fig. 10. Relation between theme concepts.

2.3.3.1. *Determination of the relations and weights between theme concepts.* We use an example to illustrate the basic principle for the determination of relations between theme concepts. There is a concept C_k and two theme concepts C_j^0 and C_i^0 shown in Fig. 10(a), if C_j^0 is a cause of C_k , and C_k is a cause of C_i^0 , then there exist causal relations between C_j^0 and C_k as well as between C_k and C_i^0 . So there is a causal relation between C_j^0 and C_i^0 . On the other hand, if C_k is C_j^0 's cause concept shown in Fig. 10(b), then knowing from the “reverse” characteristic of the E-FCM, there also exists a causal relation from C_j^0 to C_i^0 , since there is a causal relation from C_j^0 to C_k . Therefore, relation between theme concepts can be determined by reversing the relation between the fine-granularity concepts.

The weight (T_{ji}) between C_j^0 and C_i^0 can be computed as follows:

$$T_{ji} = \tanh \left(2 * \sum_{k=1}^{N1} \beta_k V_{ck} w_{jk} w_{ki} \right) \quad (9)$$

where $N1$ is the number of concepts that have relations with C_j^0 and C_i^0 ; β_k is a reversing coefficient of C_k , V_{ck} is the state value of concept C_k , and w_{kj} is the weight between concept C_k and theme concept C_j^0 .

2.3.3.2. *Algorithm for constructing document semantic template.* The construction of D-FCM consists of the following steps:

- Step 1. Combine E-FCMs belonging to a section to form S-FCM.
- Step 2. Combine S-FCMs belonging to a document to form augmented S-FCM.
- Step 3. Calculate weights (T_{ji}) from C_j^0 to C_i^0 in the augmented S-FCM.
- Step 4. Delete the concepts that have no direct connection with each theme concept in the augmented S-FCM to form D-FCM.

We take this paper as an example to illustrate the automatic construction process of D-FCM. There are seven subsections in the body part. After each subsection’s keywords being matched with E-FCMs in *EFCM-R1*, an E-FCM library that is under the control of *Domain1.E-FCM1* (Fig. 2), E-FCM3 (Fig. 6), and E-FCM4 (Fig. 7) belong to Section 2, while E-FCM2 (Fig. 5), E-FCM5 (Fig. 8), and E-FCM6 (Fig. 9) belong to Section 3. In steps 1 and 2, S-FCM1 (Fig. 12) and S-FCM2 (Fig. 13) are constructed by combining E-FCM1, E-FCM3 and E-FCM4, and by combining E-FCM2, E-FCM5, and E-FCM6 respectively. D-FCM (Fig. 14) is generated in steps 3 and 4.

D-FCM shows that this paper mainly discusses: “FCM-based document representation”, “FCMs’ automatic construction”, “fuzzy cognitive maps”, “FCM-based document understanding”, “FCM-based complex question automatic answering”, “FCM-based vague question refining”, and all

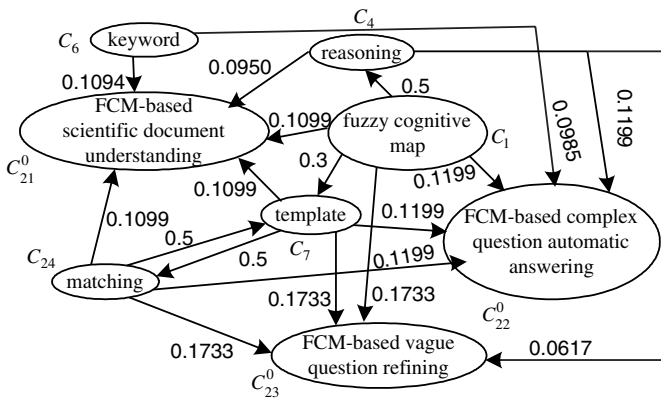


Fig. 13. Section semantic template of Section 3 (S-FCM2).

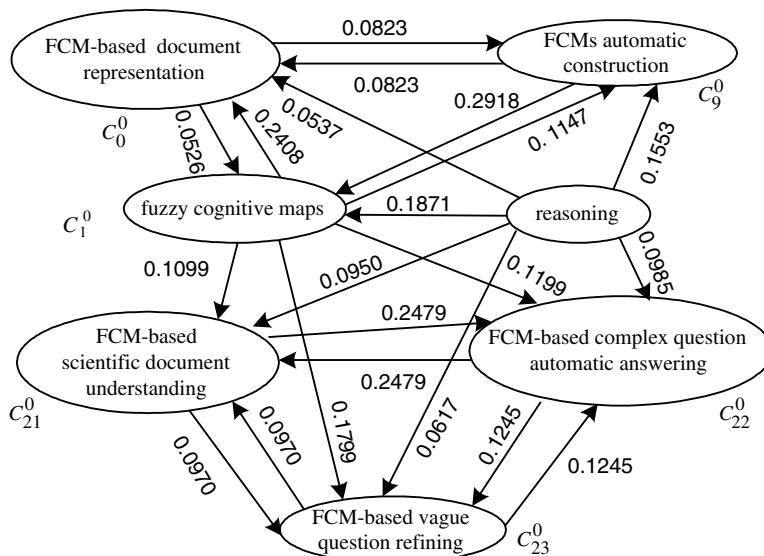


Fig. 14. Document semantic template of this paper.

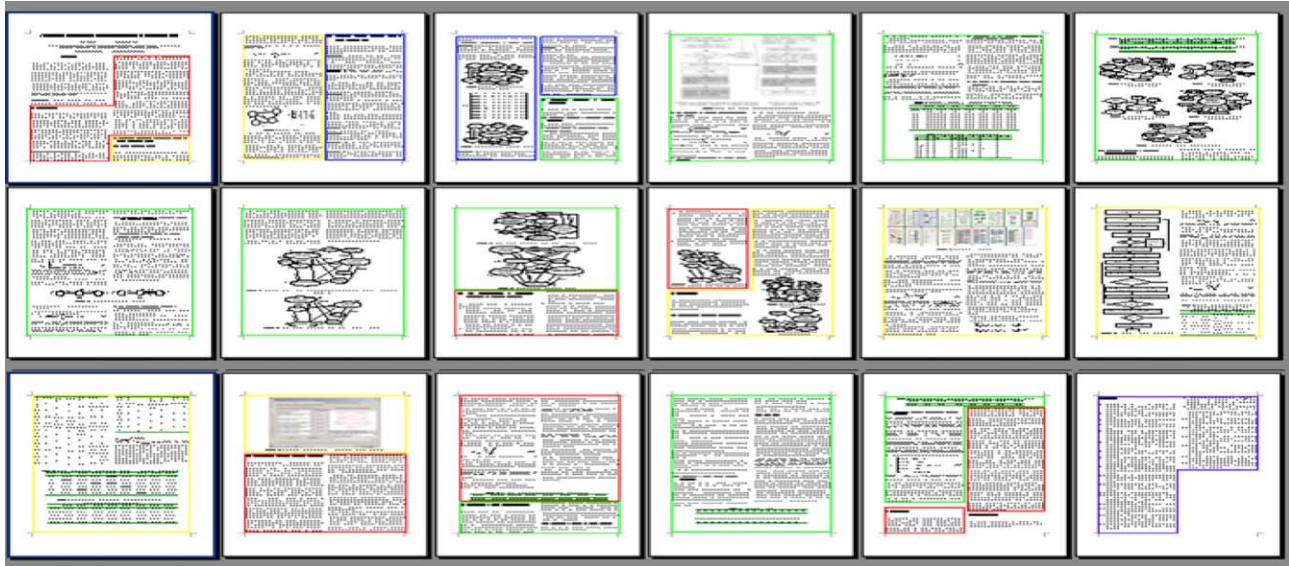


Fig. 18. Original structure of this paper.

- Step 1. Get a candidate document from the document repository.
- Step 2. Choose an E-FCM repository, for example, EFCM-R1.
- Step 3. Get K_1 and the keyword repository of *Domain* corresponding to EFCM-R1.
- Step 4. Get all the concepts (denoted as TE-keywords) of EFCM-R1.
- Step 5. Calculate the degree of document l belonging to domain i as follows:

$$md_{BG} = \sum_{k=1}^R \alpha_k f(x_k) / N \quad (10)$$

where x_k is the frequency of the keyword n_k ; $f(x_k)$ is a frequency function of n_k , $f(x_k) = \tanh(x_k/3)$; α_k is the weight of n_k that reflects its important degree in a domain; R and N are the number of the keywords of D_l and K_i respectively. K_i is a set of the keywords that can reflect domain i 's background knowledge; D_l is a set of keywords that are the results of matching K_i in document l 's title, abstract, keywords, introduction, section heading, conclusion and reference.

In the matching process, it is necessary to match synonymic keywords with the WordNet, and define a compound keywords set to match compound and abbreviation keywords in a domain.

- Step 6. If $md_{BG} \geq 0.75$, then go to step 7, otherwise, go to step 2. Here $md_{BG} = 0.86$. If $md_{BG} < 0.75$, a new E-FCM repository needs to be selected.
- Step 7. Parse the whole paper; sections and subsections are gained.
- Step 8. Match TE-keywords in each section. According to Eqs. (2) and (3), the matching degrees between TE-keywords and this paper's subsections are

obtained (denoted as $Index_i$, $i \in (1, 2, 3, \dots, 9)$). For example, Table 5 shows that $Index_2$ – $Index_4$ and $Index_6$ – $Index_8$ correspond to Sections 2.1–2.3, 3.1–3.3.

- Step 9. Assign $Index_i$ ($i \in (1, 2, 3, \dots, 9)$) to the E-FCMs in the EFCM-R1. For example, in Section 2.2, state values of concepts in E-FCM1 and E-FCM2 are $\{V_{c1}, V_{c2}, V_{c3}, V_{c4}, V_{c5}, V_{c6}, V_{c7}, V_{c18}\} = \{1.0000, 0.7616, 0.7616, 0.3215, 0.9640, 1.0000, 0.9925, 0.9997\}$ and $\{V_{c1}, V_{c4}, V_{c6}, V_{c7}, V_{c24}, V_{c25}, V_{c26}, V_{c29}, V_{c30}\} = \{1.0000, 0.3215, 0.9925, 0.5828, 0, 0, 0, 0\}$ respectively. In Section 3.2, the concepts' state values of E-FCM1 and E-FCM2 are $\{1.0000, 0, 0, 0, 0.3215, 0.9997, 1.0000, 0.9904\}$ and $\{1.0000, 0.3215, 0.9997, 1.0000, 0.9311, 0.5828, 0.3215, 0.9814, 0.9814\}$ respectively.

- Step 10. Match E-FCMs in this paper's sections. Such as, the steps of matching E-FCM1 in Section 2.2 are as follows:

- (a) Reason the probabilistic causal relation between C_1 and C_2 , C_1 and C_5 respectively;
- (b) Reason the concepts' state values of E-FCM1 as follows:

$$\begin{aligned} & \{V'_{c0}, V'_{c1}, V'_{c2}, V'_{c3}, V'_{c4}, \dots, V'_{c18}\} \\ & = \{V_{c0}, V_{c1}, V_{c2}, V_{c3}, V_{c4}, \dots, V_{c18}\} \times E1; \end{aligned}$$

- (c) According to Eq. (11), calculate the theme concept state value in EFCM-R1.

$$md_{1TE-S} = V_{c0} \quad (11)$$

- (d) Assign $Index_i$ ($i \in (1, 2, 3, \dots, 9)$) to the concepts of E-FCMs in EFCM-R1. E-FCM1-8's matching degrees in sections of this paper are been reasoned. Table 6 shows the reasoning results.

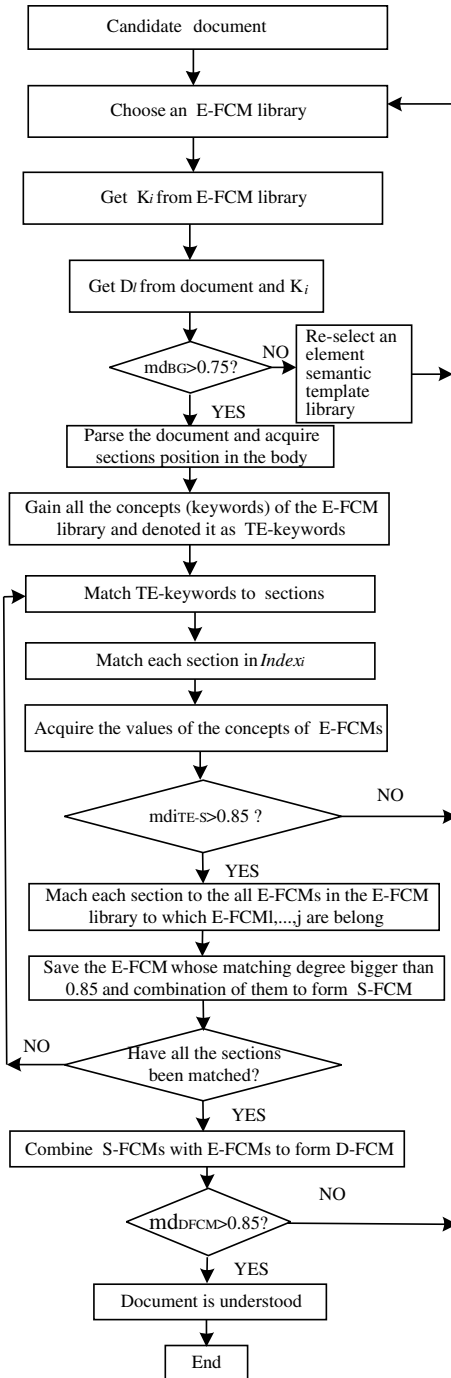


Fig. 19. Flow chart of document understanding.

Table 5
Degrees of matching (Index₂–Index₄ and Index₆–Index₈) between TE-keywords and Sections 2.1–2.3 and 3.1–3.3 of this paper

	Index ₂	Index ₃	Index ₄	Index ₆	Index ₇	Index ₈
V _{c1}	0.7616	1.0000	1.0000	1.0000	1.0000	1.0000
V _{c2}	0	0.7616	0	0	0	0
V _{c3}	0	0.7616	0.3215	0	0	0
V _{c4}	0.7616	0.3215	0.8704	0.7616	0.3215	0.3215
V _{c5}	0.9640	0.9640	0.9640	0.3215	0	0
V _{c6}	0	0.9997	1.0000	0.9951	0.9997	0
V _{c7}	0	0.9925	1.0000	1.0000	1.0000	0
V _{c8}	0	0	0	0	0	0
V _{c9}	0.3215	0.5828	0.5828	0.3215	0.9814	0
V _{c10}	0	0	0	0	0	0
V _{c11}	0	0	0	0	0	0
V _{c12}	0	0	0	0.9311	0.3215	0.3215
V _{c13}	0	0	0	0	0	0
V _{c14}	0	0	0	0	0	0
V _{c15}	0.9904	0.7616	1.0000	0.3215	0.3215	0
V _{c16}	0	0	0.9975	0	0	0
V _{c17}	1.0000	0.9975	1.0000	0.9951	0.7616	0.9993
V _{c18}	0.3215	0.9997	0.9311	1.0000	0.9904	0.9640
V _{c19}	0	0	0	0.9997	0.5828	0.5828
V _{c20}	0	0	0	0.7616	0	0
V _{c21}	0	0	0	0	0	0
V _{c22}	0	0	0	0	0	0
V _{c23}	0	0	0	0	0	0
V _{c24}	0	0.5828	0	0.9987	0.9311	0.9904
V _{c25}	0.3215	0	0.5828	0.3215	0.5828	0.5828
V _{c26}	0	0	0	0	0.3215	0.3215
V _{c27}	0	0	0	0	0	0.8701
V _{c28}	0	0	0	0	0	0.5828
V _{c29}	0	0	0	0.5828	0.9814	0
V _{c30}	0	0	0	0	0.9814	0.3215
V _{c31}	0	0	0.3215	0	0	0.9311
V _{c32}	0.9951	1.0000	1.0000	1.0000	1.0000	1.0000
V _{c33}	0	0	0.3215	0	0	0
V _{c34}	0	0	0	0	0	0
V _{c35}	0	0	0	0	0	0
V _{c36}	0	0	0	0	0	0
V _{c37}	0	0	0	0	0	0
V _{c38}	0	0	0	0	0	0
V _{c39}	0	0	0	0	0	0
V _{c40}	0	0	0	0	0	0

Step 11. If $mdi_{TE-S} > 0.85$, keep this matching degree and the E-FCM; otherwise, delete the E-FCM.

Finally, Sections 2.1–2.3 correspond to E-FCM1, E-FCM 3, E-FCM 4, and Sections 3.1–3.3 correspond to E-FCM2, E-FCM 5, and E-FCM 6 respectively.

If the reference (Zhuge, 2003) (its body includes five sections) is chosen from the document reposit-

Table 6
Degrees of matching between element semantic template 1–8 and Sections 2.1–2.3 and 3.1–3.3

Section	E-FCM	E-FCM	E-FCM	E-FCM	E-FCM	E-FCM	E-FCM	E-FCM
2.1	0.3392	0.2118	0.4393	0.9911	0.1849	0.1799	0.3984	0.2762
2.2	0.9958	0.5423	0.7646	0.7288	0.5874	0.4960	0.4273	0.0973
2.3	0.7924	0.6293	0.9903	0.8994	0.5754	0.4810	0.4531	0.3591
3.1	0.6678	0.7200	0.7388	0.5665	0.9929	0.5700	0.3709	0.0973
3.2	0.5978	0.9954	0.6132	0.3427	0.7438	0.5426	0.2957	0.0973
3.3	0.4484	0.5662	0.4345	0.3056	0.4816	0.9945	0.3236	0.0973

Table 7
Degrees of matching between E-FCM1-8 and Section 2–6 of Zhuge (2003)

	E-FCM1	E-FCM2	E-FCM3	E-FCM4	E-FCM5	E-FCM6	E-FCM7	E-FCM8
Section 2	0.1751	0.1137	0.1950	0.2105	0.2164	0.1446	0.1257	0.2032
Section 3	0.1887	0.1149	0.2034	0.1149	0.1299	0.0000	0.0000	0.2030
Section 4	0.1800	0.1128	0.1891	0.1514	0.1498	0.1096	0.1257	0.1951
Section 5	0.1849	0.1129	0.1891	0.1473	0.1257	0.0000	0.0000	0.2084
Section 6	0.1865	0.1137	0.2468	0.2398	0.2423	0.1128	0.0000	0.1800

tory, then matching degrees between the body’s Sections 2–6 and E-FCM1-8 are shown in Table 7. It also shows that the understanding system needs to select a new E-FCM repository or document.

- Step 12. If all the E-FCMs have been matched, go to step 13; otherwise, go to step 10.
- Step 13. Apply E-FCMs’ adjacency matrixes to construct S-FCMs.
- Step 14. If all the sections have been scanned, go to step 15; otherwise, go to step 13.
- Step 15. Apply adjacency matrixes of S-FCMs to construct D-FCM (Fig. 13).
- Step 16. Reason on D-FCM, and calculate the matching degree between D-FCM and this paper according to Eq. (12):

$$md_{D-FCM} = \sum_{i=1}^k V_{c_i^0} / k \tag{12}$$

where k is the number of the theme concepts in D-FCM; $V_{c_i^0}$ is the theme concept’s state value after doing one time reasoning on D-FCM. This paper’s md_{D-FCM} is 0.9959.

Step 17. If $md_{D-FCM} > 0.85$, then the D-FCM corresponding to the document and its concepts as well as the state values are recommended to users; otherwise, go to step 2. Fig. 20 is one interface representing the final understanding of this paper. It is suitable for machine understanding so can support intelligent services.

3.2. Question answering

Previous question answering (QA) system needs to identify the question keywords that determine the expected answer type (Webber, 1987; Pasca and Harabagiu, 2001). Recent question answering systems (Voorhees, 2001)

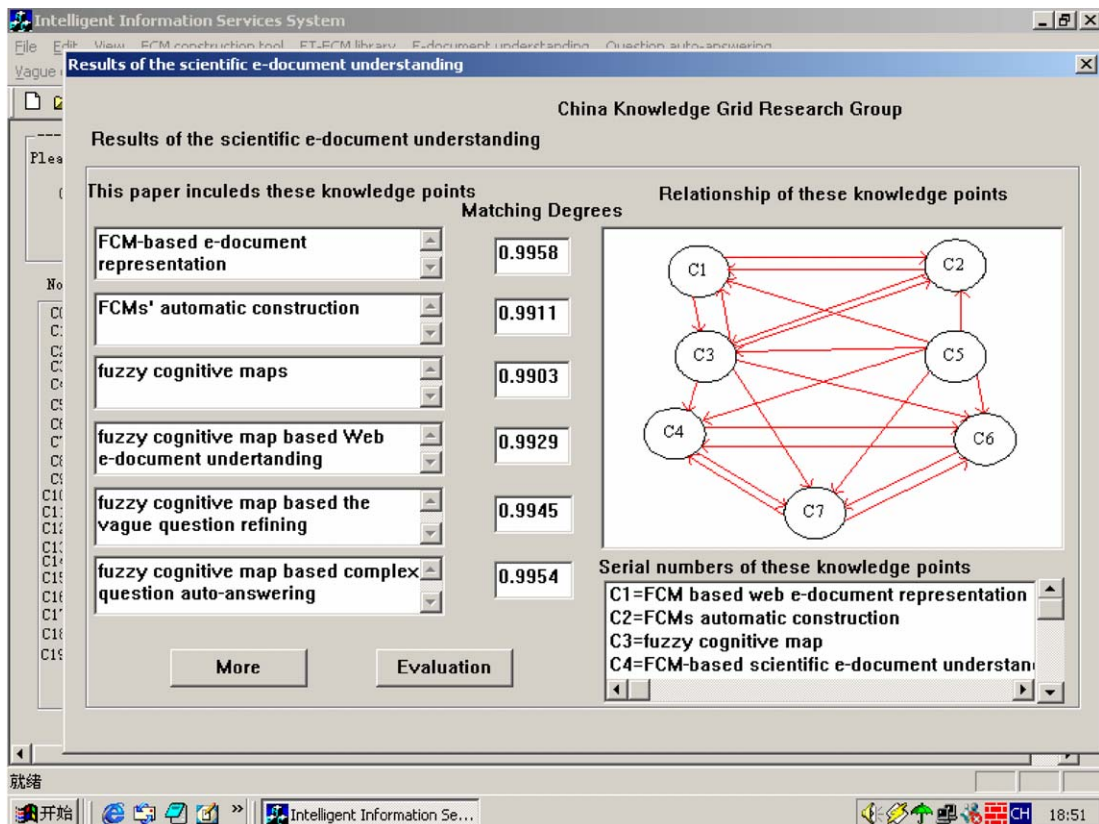


Fig. 20. Interface for displaying document understanding.

requires *sophisticated linguistic knowledge* or tools, such as named-entity recognizer, ontology and WordNet. In TREC10 QA track, the best performing system used many textual patterns (Soubotin and Soubotin, 2001). The power of textual patterns for question answering looks quite amazing. But the patterns were *compiled manually based on human analysis of a large set of questions* (Sasaki et al., 2004). Most of the question answering taxonomies and textual patterns in these QA systems (Voorhees, 2001) are based on fine-granularity level, such as keywords and phrases, and the question answering taxonomies and textual patterns are constructed *manually*.

Unlike other passage-retrieval algorithms, we use FCM semantic templates to select documents, sections or paragraphs that may contain the answer within a corpus and can generate and represent the question focus and the question keywords. A complex scientific question generally involves several knowledge points. Given a scientific question, the sentence of the question is firstly parsed to keywords, which are then matched in the theme concepts' keywords in different E-FCM libraries. Following that, the selected E-FCMs are used to find a closely related section or paragraph from the document. Finally, the S-FCMs corresponding to the closely related sections or paragraphs of the document and its D-FCM are recommended to users for the detailed explanation. Of course, we can reorganize the selected sections or paragraphs coming from different documents as the detailed answer. The process of question answering is illustrated by an example as follows:

Step 1. Input question “How to represent and understand document with FCMs?”;

Step 2. Get the question's keyword set $Q_k = \{\text{“represent”, “understand”, “scientific”, “document”, “FCM”}\}$ from the question;

Step 3. Compute the matching degree between Q_k and the theme concepts in different E-FCM repositories by:

- (a) matching Q_k with the theme concepts in different E-FCM repositories, e.g., the results of matching Q_k in EFCM-R1 are shown in Table 8, where $md_i = p/k$, k is the number of the theme concept keywords, and p is the number of theme concept keywords in Q_k ;
- (b) computing the matching degree between Q_k and EFCM-R1 as follows:

$$md_{Q_k} = \sum_{i=1}^g q_i/g \quad (13)$$

where q_i is the md_i that is bigger than 0.5, and g is the number of md_i that is bigger than 0.5.

In the matching process, it is necessary to use WordNet to match the synonyms, and define a keywords repository to match the compound and abbreviation keywords in a domain.

Step 4. Choose the E-FCM repository corresponding to the maximum md_{Q_k} . In this case, EFCM-R1 is selected;

Step 5. Get K_1 and the keyword repository of *Domain1*, which correspond to EFCM-R1;

Step 6. Get the TE-keywords of EFCM-R1;

Step 7. Get a document (e.g., this paper) from document repository;

Step 8. Select the E-FCMs whose md_i is bigger than 0.50 in EFCM-R1. Here E-FCM1 and E-FCM 5 are selected;

Step 9. Match TE-keywords in each subsection, then $Index_i$ ($i \in (1, 2, 3, \dots, 9)$) are obtained;

Step 10. The matching degrees between $Index_2$ – $Index_4$, $Index_5$ – $Index_8$ and Sections 2.1–2.3 and 3.1–3.3 are shown in Table 6.

Table 6 shows that E-FCM1 corresponds to Section 2.2; E-FCM5 corresponds to Section 3.1. Therefore, this paper can be recommended to the user who asks. If the matching degree is less than the threshold, another document will be selected from the document repository and then go to step 7;

Step 11. Assign $Index_i$ ($i \in (1, 2, 3, \dots, 9)$) to the E-FCMs in EFCM-R1 and select the matched E-FCMs, that is, searching for knowledge points related to E-FCM1 and 5. Here, E-FCM 1-6 are selected.

Step 12. Using E-FCM1-6 to understand the selected paper; S-FCM1 (Fig. 12), S-FCM2 (Fig. 13) and D-FCM (Fig. 14) are acquired;

Step 13. If $md_{D-FCM} \geq 0.85$, display the sections or paragraphs corresponding to the S-FCMs to the user as a succinct answer for the question; recommend D-FCM and this paper to the user as the detailed answer to the question; otherwise, go to step 7.

The above algorithm can handle erroneous and avoid sophisticated linguistic issues.

3.3. Vague question refining

Sometimes users are unable to exactly express their intentions with words or questions. The FCM-template-based vague question refinement system works this way to help users: recommend some candidate documents first for review, then a user can select one or more documents, sections or paragraphs from the candidates to express

Table 8
The results of match Q_k in the EFCM-R1's theme concepts

E-FCM1	E-FCM2	E-FCM3	E-FCM4	E-FCM5	E-FCM6	E-FCM7	E-FCM8
$md_1 = 75\%$	$md_2 = 33\%$	$md_3 = 20\%$	$md_4 = 33\%$	$md_5 = 100\%$	$md_6 = 33\%$	$md_7 = 33\%$	$md_8 = 50\%$

Table 10
The degrees of matching between selected element semantic templates and NCs

E-FCM1	E-FCM2	E-FCM3	E-FCM4	E-FCM5	E-FCM6	E-FCM7	E-FCM8
100%	56%	100%	60%	67%	50%	30%	40%

tively represents the interests of users. We can further reorganize these selected sections or paragraphs that come from different documents according to the user's interest.

3.3.4. Adaptive threshold determination of theme concepts

The determination of C_i^0 's threshold θ_i is a key issue. If $V_{c_i} > \theta_i$, the knowledge point represented by the semantic template belongs to this section or paragraph. Relations exist between θ_i and the number of the recommended documents as well as their quality in vague question refining or document recommendation. If θ_i is lower/higher, the number of recommended documents are increased/decreased, but the qualities become bad/better consequently. In order to automatically adjust θ_i , the adaptation of the theme concept's threshold is defined as follows:

$$\theta_i = \begin{cases} \theta_i - \Delta & I - N = 0 \\ \theta_i + 2\Delta & f \leq -I \\ \theta_i + \Delta & f > -I \text{ and } f \leq \delta \\ \theta_i - \Delta & f > \delta \\ \theta_i & \text{otherwise} \end{cases} \quad (14)$$

where $f = 3 * I - N$, Δ is a constant, I is the number of documents or sections relevant to the user's interests, N is the number of documents or sections irrelevant to the user's interests. N and I can also be regarded as the number of keywords (or knowledge points) that represent the user's interest or non-interest. Eq. (14) shows that the document selection is accomplished in the process of automatically adjusting the thresholds of E-FCMs.

4. Conclusions

By defining semantic templates for e-documents of different granularities, this paper proposes an approach to automatically generate semantics for scientific e-documents. The approach uses document's keywords, their relations, and the implied meaning of co-occurred keywords that are hard to be exploited and reasoned by previous semantic representation approaches. The semantics can be constructed, reasoned, composed and decomposed at different granularity levels according to requirement. So it plays an important role in implementing intelligent services based on document understanding in the e-science Knowledge Grid environment.

Given training passages and keyword repository, the understanding can be automatically carried out. The semantic templates enable the question answering mechanism to return a relatively precise and overall answer by

avoiding sophisticated linguistic approaches and many manual works for constructing answer taxonomy. The templates also support users to express and refine their intentions by using element semantic templates.

The proposed approach can be extended to support richer semantic relationships and reasoning by combining the semantic link network (Zhuge, 2003, 2004a) with FCM. The reasoning mechanism for the combination model needs to use the semantic link reasoning rules (Zhuge, 2004b). We are exploring relevant mathematic principles.

References

- Altamura, O., Esposito, F., et al., 2000. Transforming paper documents into XML format with WISDOM++. International Journal of Document Analysis and Recognition 3 (2), 175–198.
- Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic Web. Scientific American 284 (5), 34–43.
- Heflin, J., Hendler, J., 2001. A portrait of the semantic Web in action. IEEE Intelligent Systems 16 (2), 54–59.
- Hobbs, J.P., Stickel, M., et al., 1988. Interpretation as abduction. In: Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics, Buffalo, NY, June, pp. 95–103. Available from: <<http://acl.ldc.upenn.edu/P/P88/P88-1012.pdf>>.
- Kosko, B., 1997. Fuzzy Engineering. Prentice-Hall.
- Leea, K.C., Lee, S., 2003. A cognitive maps simulation approach to adjusting the design factors of the electronic commerce Web sites. Expert Systems with Applications 24, 1–11.
- Liu, Z.Q., Satur, R., 1999. Contextual fuzzy cognitive maps for decision support in geographic information systems. IEEE Transactions on Fuzzy Systems 7 (10), 495–502.
- Pasca, M.A., Harabagiu, S.M., 2001. High performance question/answering. In: Processing of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 366–374.
- Sasaki, Y., Isozaki, H., et al., 2004. NTT's QA systems for NTCIR QAC-1. In: Processing of the Third NTCIR Workshop. Available from: <<http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-QAC-SasakiY.pdf>>.
- Smith, M.K., Welty, C., et al., 2003. OWL Web ontology language guid. Available from: <<http://www.w3.org/tr/2003/WD-owl-guide-20030331/>>.
- Soubbotin, M.M., Soubbotin, S.M., 2001. Patterns of potential answer expressions as clues to the right answer. In: Proceedings of the 10th Text Retrieval Conference (TREC10), NIST, Gaithersburg, MD, pp. 175–182.
- Noha, J.B., Lee, K.C., 2000. A case-based reasoning approach to cognitive maps-driven tacit knowledge management. Expert Systems with Applications 19, 249–259.
- Voorhees, E., 2001. Overview of the TREC 2001 question answering track. In: Proceedings of the 10th Text Retrieval Conference (TREC10), NIST, Gaithersburg, MD, pp. 157–165.
- W3C. RDFCore Working Group, W3Cpublication. URL Available from: <<http://www.w3.org/2001/sw/RDFCore/>>.
- Webber, B., 1987. Question answering. In: Shapiro, S.,C. (Ed.), Encyclopedia of Artificial Intelligence, vol. 2. Wiley, pp. 814–822.

Zhugue, H., 2003. Active E-document framework ADF: model and tool. *Information and Management* 41, 87–97.

Zhugue, H., 2004a. Retrieve images by understanding semantic links and clustering image fragments. *Journal of Systems and Software* 73 (3), 455–466.

Zhugue, H., 2004b. *The Knowledge Grid*. World Scientific, Singapore.

Hai Zhuge is the chief scientist of the China Semantic Grid project funded by the National Basic Research Program of China. He is a professor and the director of the Key Lab of Intelligent Information Processing at the Institute of Computing Technology in Chinese Academy of Sciences, and the founder of the China Knowledge Grid Research Group (<http://kg.ict.ac.cn>), which owns over 30 young researchers. He presented over 10 keynotes at international conferences. He was the co-chair of the 2nd International Workshop on Knowledge Grid and Grid Intelligence, the program co-chair of the 4th International Conference on Grid and Cooperative Computing, and the co-chair of the 1st International Conference on Semantics, Knowledge and Grid. He organized several journal special issues on Knowledge Grid and Semantic Grid. He is serving as the Area Editor of the *Journal of Systems and Software*, the Associate Editor of *Future Generation Computer Systems*, the area editor of the *Journal of Computer Science and Technology*, and the editorial member of the

Information and Management and the *Electronic Commerce Research and Applications*. His major research interest is the model, theory and methodology on the future interconnection environment. His monograph *The Knowledge Grid* is the first book in the area, and receives 2005's Top Award of SONY Excellent Research. He is the author of over ninety papers appeared mainly in leading international journals such as *Communications of the ACM*; *IEEE Computer*; *IEEE Transactions on Knowledge and Data Engineering*; *IEEE Intelligent Systems*; *IEEE Computing in Science and Engineering*; and *IEEE Transactions on Systems, Man, and Cybernetics*. One of them was among the Top 1% highly cited papers in the area according to ISI Essential Science Indicator. He is a senior member of the IEEE and a member of the ACM. He was among the Top scholars in software engineering and systems area (1999–2003) according to the assessment report published in the *Journal of Systems and Software*.

Xiangfeng Luo is a postdoctor of China Knowledge Grid Research Group at Institute of Computing Technology in Chinese Academy of Sciences. His research fields include knowledge capturing, Semantic and Knowledge Grid, artificial intelligence and pattern recognition. He is in charge of a research project supported by National Science Foundation of China.