

Noise Control in Document Classification Based On Fuzzy Formal Concept Analysis

Sheng-Tun Li

Institute of Information Management
National Cheng Kung University
Tainan, Taiwan
stli@mail.ncku.edu.tw

Fu-Ching Tsai

Institute of Information Management
National Cheng Kung University
Tainan, Taiwan
r7895104@mail.ncku.edu.tw

Abstract—Document classification is critical due to explosive increasing of text in modern world. However, most of existing document classification algorithms are easily affected by noise data. Therefore, in document classification tasks, the ability of noise control is as important as the ability to classify exactly. In this paper, we propose a novel classification framework based on fuzzy formal concept analysis to moderate the impact from noise. In addition, the well-organized concepts also provide inherent relations, which support knowledge codification and distribution effectively. Experimental results using Reuters 21578 dataset demonstrates significant noise control benefit and superior classification accuracy.

Keywords—text classification; fuzzy formal concept analysis; noise control; information retrieval

I. INTRODUCTION

The explosive increasing of text in modern world contributes the increasing need to provide automatic document classification and categorization. The well-organized categories of document are efficient in browser, retrieval and dissemination. Most existing document classification algorithms operate in a high dimensional vocabulary space while computing document similarities and relevance. However, the document classification procedure are easily affected by term-based inner product, cosine coefficient operations and noise data, such as synonym, antonym or words with multiple meanings [1]. Thus, the ability of handling noise is one of the important indicators for a text analysis algorithm and the reason why it can produces more useful results than others [2]. Moreover, for effectively controlling noise without eliminating relevant information, the parameter setting is another delicate topic. For example, when excessive parameter is applied for filtering irrelevant data, it is possible to remove some significant messages as well. Thus, the tradeoff parameter settings may vary regarding to different sources of data and needs to be careful examined especially in practical implementations.

One way to solve the noise problem in document classification is to decrease the impact from terms in document vectors. Since a vector is composed of a document and key terms, we consider that the technique, which can transforms the inherent relations between document and terms to a more abstract form, may appropriate for noise control. The more general elements are capable of resisting specific noise-sensitive variations. Formal concept analysis (FCA) formulates

concepts in terms of documents and their properties, terms, and provides a way of combining and organizing individual concepts of a given context into a hierarchical ordered conceptual structure [3]. In the last few years, several researches have shown the capacity of FCA in information retrieval and management. In FCA lattice, the document collection can be organized into a reasonable consensus with good efficiency in retrieval performance, its structural comprehension is valuable to extract implication rules and classify documents in a suitable category [4], [5]. The path, location and hierarchical relations of concepts in FCA are successful to provide significant relations of keywords and documents [6]. While the weight of each term in a document is not simply belongs or does not belongs, we further utilize fuzzy formal concept analysis (FFCA) to take uncertain information into consideration.

The objective of this research is to develop a document classification scheme which can tolerate noise data. The structure of the paper is organized as follows. Section 2 reviews underlying technologies on formal concept analysis and Section 3 describes the specifics of the research methodology and system development process. Section 4 illustrates the results and discusses performance evaluation. Finally, Section 5 draws conclusions on the paper and provides insights on future work.

II. FUZZY FORMAL CONCEPT ANALYSIS

A. Formal Concept Analysis

Formal concept analysis (FCA) is a mathematical approach proposed by Wille[7] used for structuring, analyzing and visualizing data, based on a notion of duality called Galois connection[8]. Data consisting of a set of entities and their features are structured into formal abstractions called formal concepts, which together form a concept lattice ordered by a partial order relation. Concept lattices are constructed by identifying the objects and their corresponding attributes for a specific domain, called conceptual structures, and then displaying the relationships between them.

In FCA, a context is represented as a triple (O, A, R) where O is a set of objects, A is a set of elements and R is a binary relation between O and A . If for an object o and an attribute a , oRa holds, when we say that “object o has attribute a ”, or “attribute a applies to object o ”. Table 1 shows an context of an animal environment, where each row is an object representing

an animal, each column is an attribute, and each X represents the existence of the relationship.

TABLE I. THE CONTEXT OF A ANIMAL ENVIRONMENT

	Hunt	Fly	Bird	Mammal
Lion	×			
Warbler		×	×	
Eagle	×	×	×	
Rabbit				×
Ostrich			×	

Two sets can be derived from a given formal context (O, A, R) , one from the perspective of objects and another from the viewpoint of attributes. The former is a subset of objects $I' \subseteq O$ such that I' is defined by the attributes that apply to all objects belonging to I and the later is a subset of attributes $J' \subseteq A$ such that J' is defined by the objects that have all attributes belonging to J . Correspondingly,

$$I' = \{a \subseteq A | oRa \ \forall o \in I\},$$

$$J' = \{o \subseteq O | oRa \ \forall a \in J\},$$

thus the pair (I, J) , where $I \subseteq O$ and $J \subseteq A$, is a formal concept of the context (O, A, R) if $I' = J$ and $J' = I$. The formal concept (I, J) is a pair where I consist of those objects that have all attributes from the set J , and equally, J consists of those attributes that apply to all objects from the set I . additionally, the set of objects is known as the extent of the concept and the set of attributes as the intent.

B. Fuzzy Formal Concept Analysis

Fuzzy formal concept analysis (FFCA) was introduced by Tho et al.[9] by incorporating fuzzy logic into FCA in order to represent vague information. Traditional FCA uses a binary approach indicating the presence or absence of the relation between an object and an attribute and cannot take uncertain information into consideration, while FFCA extends the relation into a value in the interval of real numbers between 0 and 1 called membership value. In this way, the relation is not simply belongs or does not belong anymore, but can take a set of values indicating a degree of belonging. This membership value of an attribute towards an object is proportional to its significance and can be used to calculate a similarity value between two formal concepts. These notions are represented in the following definitions.

Definition 1 (Fuzzy formal context) A fuzzy formal context is a triple $(G, M, F = \varphi(G \times M))$. Where G is a set of objects, M is a set of attributes, and F is a fuzzy relation between G and M . Each relation $(g, m) \in F$ has a membership value $\mu(g, m)$ in $[0, 1]$.

Definition 2 (Fuzzy formal concept). A fuzzy formal concept of a fuzzy formal context (G, M, F) is a pair $(\varphi(Y), Z)$ for $Y \subseteq G$ and $Z \subseteq M$, where each object $g \in \varphi(Y)$ has a membership value $\mu_Y(g) = \min_{m \in Z} \mu(g, m)$ with $\mu(g, m)$ being the membership value of $(g, m) \in F$.

Definition 3 (Fuzzy formal concept cardinality). The fuzziness of a fuzzy formal concept is represented by membership values of objects of the concept, hence the cardinality of a fuzzy formal concept $(\varphi(Y), Z)$ is defined as $|\varphi(Y)|$, where $|\varphi(Y)|$ is the cardinality of the fuzzy set $\varphi(Y)$.

Definition 4 (Fuzzy formal concept similarity). The similarity of a fuzzy formal concept $(\varphi(Y_1), Z_1)$ and its subconcept $(\varphi(Y_1), Z_1)$ is defined as (1)

$$E((\varphi(Y_1), Z_1), (\varphi(Y_1), Z_1)) = E(\varphi(Y_1), \varphi(Y_2)) = \frac{|\varphi(Y_1) \cap \varphi(Y_2)|}{|\varphi(Y_1) \cup \varphi(Y_2)|} \quad (1)$$

where $\cap$ and $\cup$ are the intersection and union operations of fuzzy sets $\varphi(Y_1)$ and $\varphi(Y_2)$, respectively.

III. TEXT CLASSIFICATION WITH FFCA

This research uses FFCA to generalize documents to concepts in order to decrease the impact from noise term features. The documents are preprocessed by natural language processing to identify significant terms. Subsequently, FFCA is employed to provide the more generalized perspective, concept, by the way of analyzing inherent relationships between documents and its term features. Every formal concept on lattice diagram is used to recommend the precise category for new documents. An overview of the proposed FFCA document classification system is shown in Fig. 1. Five main phases, namely natural language processing, lattice construction, attribute-concept transformation, document-concept mapping and collaborative recommendation, are discussed in the following subsections.

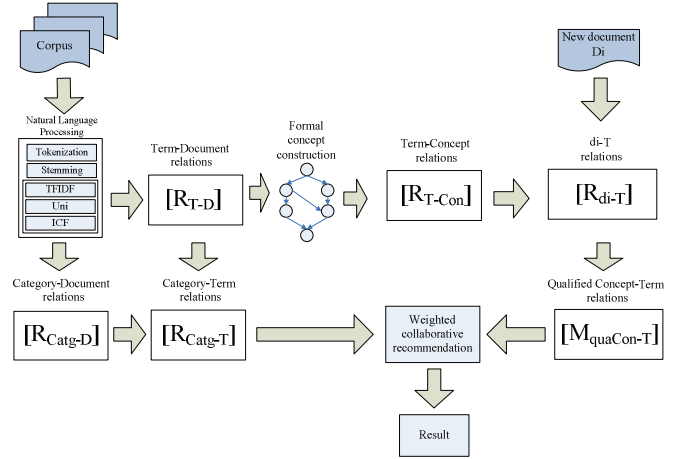


Figure 1. The FFCA classification framework

A. Natural Language Processing

Documents containing data in natural language format are useful sources of information, but due to the several variations in human language, it is problematical for automated systems to understand its meaning accurately. Natural language using computers, but has faced numerous problems and limitations since in order to really understand its connotation, extensive

knowledge about the outside world is required. The general process starts by separating textual input into tokens, followed by stemming to map related words to the same root and reduce the amount of tokens, stop-word elimination, part-of-speech tagging and finally syntax analysis.

In order to retrieve significant features from the corpus, information retrieval technique is applied to estimate the importance of retrieved tokens. Three indexes, term frequency and inverse document frequency (TF-IDF), Uni and ICF provide effective criteria to identify important keywords, which are $TF\text{-}IDF > 26$, $Uni < 0.2$ and $ICF > \log 2$. Terms, which meet these three criteria, are treated as significant features of a specific document[10].

B. Lattice Construction

The source of lattice is composed of the relationships between key terms and documents. For more precisely demonstrate the degree of significance of the key terms, the traditional binary relations are improved to fuzzy numbers. The TF-IDF values that previously calculated for terms are used as membership degrees to build up fuzzy formal context. A α -cut acting as a membership degree threshold is set to eliminate relations that have low membership and in this way, reduce a accordingly the complexity of the analysis process. All membership values below the α -cut are considered equivalent to zero. Fig. 2(b) shows the fuzzy formal context in Fig. 2(a) with $\alpha=0.5$, where the null values signify the absence of the topic or its exclusion due to low significance. In each row, d, represents a document, and each column, t, represents a key term.

	t1	t2	t3
d1	0.8	0.12	0.61
d2	0.9	0.85	0.13
d3	0.1	0.14	0.87

(a)

	t1	t2	t3
d1	0.8	-	0.61
d2	0.9	0.85	-
d3	-	-	0.87

(b)

Figure 2. The formal context with α -cut=0.5

The lattice organizes the documents and key terms via partial order which reveals the inherent hierarchical relationships of knowledge. The formal concepts which produced by FFCA further provide more general perspective than documents to represent the content of corpus. Therefore, this research focuses on operating concepts instead of documents in classification problems. The relationships between concepts and key terms are transformed from lattice structure with matrix shows in Fig. 3.

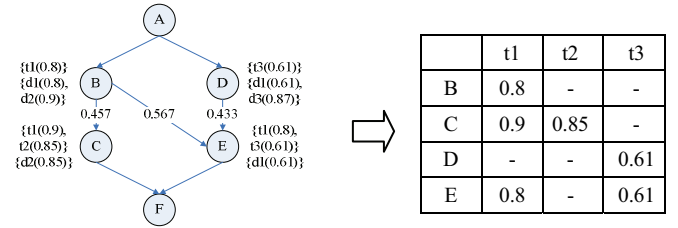


Figure 3. The transformation from lattice structure to concept-term relations

C. Term-category Transformation

The objective of this research is to provide concepts, a more general perspective than document, to facilitate noise control in document classification issue. However, the category information is mainly tagged on documents. Neither terms nor concepts can be the basis to decide the category of a new document. Therefore, the term-category transformation is necessary for a concept-based classification approach. The transformation mechanism is based on fuzzy compositions[11]. The standard composition of two fuzzy relations $P(X, Y)$ and $Q(Y, Z)$ is represented by $P(X, Y) \circ Q(Y, Z)$ which generate $R(X, Z)$ on $X \times Z$. The $R(X, Z)$ is calculated by (2).

$$R(x, z) = [P \circ Q](x, z) = \max_{y \in Y} \min [P(x, y), Q(y, z)] \quad (2)$$

where all $x \in X$ and all $z \in Z$. Therefore, the relations between terms and categories, M_{T-Catg} , can be derived by composing term-document relations, and document-category relations, as shown in (3).

$$R_{T-Catg} = R_{T-D} \circ R_{D-Catg} \quad (3)$$

where R_{T-D} is term-document relations and M_{D-Catg} are document-category relations. The values in M_{T-D} are the previously calculated TF-IDF value. In traditional approaches, each row represents a document vector which is used for similarity calculation of documents. In M_{D-Catg} , the binary data, 1 or 0, is used to represent document and category relations. The value 1 indicates the document belongs to the category. Otherwise, the value will be 0.

D. Concept-category Transformation

Since the a more abstract perception, concepts, are proposed to replace documents in text classification task, the category information for each concept is essential to conclude the appropriate category for a new incoming document. As mentioned above, the category information from documents has been transformed to terms in M_{T-Catg} , the next challenge is how to reveal the concept-category relations, $M_{Con-Catg}$. Similarly, the standard composition of concept-key term and key terms-category relations can be composited by $M_{Con-T} \circ M_{T-Catg}$. Thus, the category information, which previously based on documents, has been successively transformed to concepts.

E. Unclassified document mapping

The main objective of FFCA classification approach is to classify new documents based on well-organized concepts in the lattice space. As $M_{Con-catg}$ has provided concept-category relations, it is capable of mapping unclassified documents to a set of relevant concepts and assigning the most appropriate category by collaborative recommending of its significant neighbor concepts. In order to search relevant concepts for a new document, we represent the new document as fuzzy formal concepts with TF-IDF values. Further, the fuzzy similarity operation, (1), is applied to calculate the similarity measure between new document and fuzzy formal concepts. The parameter δ is designed to determine the similarity criterion that the qualified fuzzy formal concepts should include in classification procedure. A more detailed pseudo code is presented in Fig. 4.

```

Procedure Build_MSC( $d_i, \delta, R_{Con-T}$ )
 $d_i$  : new document
 $\delta$  : most similar concept threshold,  $0 < \delta < 1$ 
 $R_{Con-T}$  : concept-term relations
 $MSC$  : the most similar concept set

 $MSC = \emptyset$ 
FOR each  $concept_i$  exists in  $R_{Con-T}$ 
  IF  $fuzzysimilarity(d_i, concept_i) > \delta$ 
     $MSC \leftarrow concept_i$ 
  END IF
END FOR

return  $MSC$ 

```

Figure 4. Algorithm to find relevant concepts of new document

F. Classification

After retrieving the qualified concepts which highly correlated with a new document, the classification procedure are based on weighted collaborative recommendation of retrieved concepts as (4).

$$\begin{aligned}
 & \text{weighted_sum}(C_i) \\
 &= \sum_{k=1}^j fuzzysimilarity(d, concept_k) \\
 & \times membership(concept_k, C_i)
 \end{aligned} \quad (4)$$

where $fuzzysimilarity(d, concept_k)$ is the fuzzy similarity between document, d , and $concept_k$. As $membership(concept_k, C_i)$ is the part of $M_{Con-Catg}$ which provide the category information of qualified concepts. A more detailed pseudo code is presented in Fig. 5.

```

Procedure Categorize( $d_i, MSC$ )
 $d_i$  : new document
 $MSC$  : the most similar concept set
 $C$  : assigned category

```

```

FOR each  $C_i$ 
  FOR each  $concept_j \in MSC$ 
     $weighteds\_sum(C_i) = fuzzysimilarity(d_i, concept_j) \times membership(concept_j, C_i)$ 
  END FOR
END FOR
 $C = \arg \max_{C_i} weighteds\_sum(C_i)$ 
return  $C$ 

```

Figure 5. Algorithm to categorize new document

IV. EXPERIMENTS AND ANALYSIS

We conduct several experiments to evaluate the effectiveness of proposed FFCA classification. In the first experiment, we demonstrate the performance in terms of f-measure, the weighted harmonic mean of recall and precision, based on ten-fold cross-validation. In the second experiment, we investigate the impact of different parameter settings on classification performance. Further, the information which contributed from various lattice construction perspectives are analyzed with regards to their structure to affect the classification results. In the following section, we first describe the experimental dataset and then analyze the evaluation results.

A. Data Collection

In this research, we choose a public corpus, Reuters 21578, to evaluate the performance of classification. Reuters 21578 is the main benchmark for text classification evaluation. Each article is well tagged with several properties. The tag, TOPIC, indicates what categories each article belongs to. Articles with multiple classes are allowed in Reuters 21578, which means there are multiple values in TOPIC tag. However, our research is mainly focus on the single classification problem. Thus, articles belongs to multiple classes are ignored in this experiment. In order to avoid the effects of small sample, the classes which have less than one hundred articles are not included in this experiment as well. The qualified five classes containing 5550 articles are implemented using ten-fold cross-validation experiment. The statistics of experimental data are demonstrated in table 2.

TABLE II. THE STATISTICS OF EXPERIMENTAL DATASET

Category	Testing data	Training data	Total
acq	151	1359	1510
crude	26	234	260
earn	355	3195	3550
ship	10	90	100
trade	13	117	130

B. Evaluation of accuracy

Since the FFCA classification approach is affected by two parameters, α and δ , we investigate their effect and correlation by conducting different parameter settings. α represents the tfidf threshold of terms in new documents based on the training data. Only terms with tfidf value higher than α are included in

lattice construction. Higher α denotes less keywords are qualified as the features of new document and vice versa. δ depicts the similarity criterion between new document and concepts. Since k concepts which are similar enough to new document are chosen as neighbors to predict the category, δ provides the information to manipulate the number of concepts participating in classification. The range of δ is between 0 and 1. When $\delta=0$, all concepts are involving in classification procedure. On the contrary, only the concept which is exact the

same with new document is involving the classification procedure when $\delta=1$. The results of different combinations of α and δ are shown in Fig 6.

Fig. 6 depicts that the highest classification f-measure is when $\alpha=0.1$ and $\delta=0.1$. $\alpha=0.1$ demonstrates that the terms of new document should be kept as factors to provide information in classification. The same situation, $\delta=0.1$, also shown in deciding δ , the results also support that when more the concepts involving in classification the accuracy is higher.

		α								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
δ	0.1	96.03%	95.86%	95.80%	95.58%	94.53%	93.80%	88.39%	88.46%	87.14%
	0.2	93.56%	93.49%	95.42%	95.34%	92.47%	89.32%	86.87%	86.65%	85.39%
	0.3	90.43%	91.17%	91.57%	90.18%	81.85%	85.36%	85.29%	85.39%	84.38%
	0.4	85.82%	87.14%	83.33%	81.78%	65.79%	72.10%	85.00%	85.13%	84.29%
	0.5	74.74%	84.10%	75.29%	71.45%	63.13%	67.68%	85.07%	84.95%	84.17%
	0.6	66.60%	73.55%	63.79%	62.57%	62.65%	67.49%	85.07%	84.95%	84.17%
	0.7	63.95%	68.85%	58.96%	61.53%	62.65%	67.49%	85.07%	84.95%	84.17%
	0.8	63.77%	67.36%	58.70%	61.52%	62.65%	67.49%	85.07%	84.95%	84.17%
	0.9	63.77%	67.36%	58.70%	61.52%	62.65%	67.49%	85.07%	84.95%	84.17%
	1	63.77%	67.36%	58.70%	61.52%	62.65%	67.49%	85.07%	84.95%	84.17%

Figure 6. Prediction accuracy for different parameter combinations

C. Implications for KNN-based approaches

For KNN-based approaches, how to decide the most appropriate K neighbors remains a great challenge. Information may lose when choosing small K , but choosing big K may also produce poor results due to too many noise data. In practical KNN applications, K is always suggested to be neither too big nor too small, but lack of an efficient way to determine the most appropriate number. In this research, the parameters, α and δ , are used to determine the number of terms and concepts for classification, which is comparable to K in KNN approaches. Therefore, we further investigate the results from different levels of parameters and analyze the impacts of different parameter setting strategies.

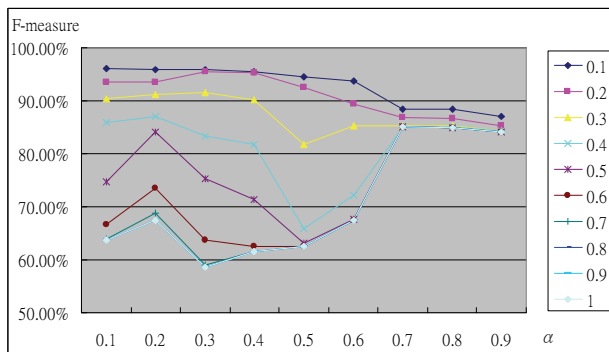


Figure 7. F-measure curves under different level of α

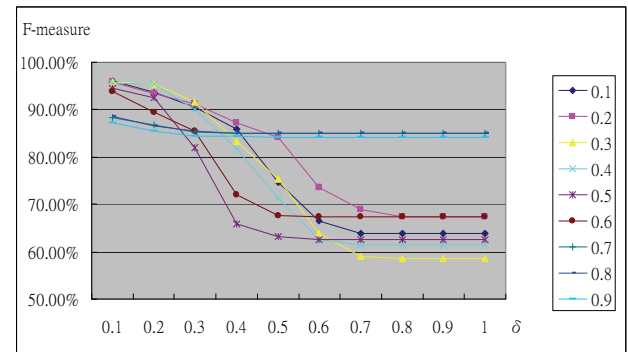


Figure 8. F-measure curves under different level of δ

Fig. 7 demonstrates the curves of δ under different levels of α . For all δ F-measure curves, their rank from high to low under different level of α are the same sequence as its value from 0.1 to 1. This fact indicates that under different thresholds of TF-IDF, more concepts are capable of providing information for document classification. As for the perspective of each single δ curve, we observe that drop and up patterns obviously appears except when $\delta=0.1$. Since α represents TF-IDF which controls the amount of key terms involving in classification task, the drop F-measure tendency between α equals 0.2 and 0.5 reflects that the valuable key terms are eliminated by rising α and contribute decline in performance. However, tendencies reverse when $\alpha>0.5$ indicate that the rising α further eliminate moderate terms and strictly keep significant ones facilitate the performance of F-measure. The F-measure for all levels of δ are stable when α is larger than 0.7. It is worth noting that, the drop and up pattern is not obvious when δ is 0.1 and 0.2. The

reason is that when δ is low indicates most of the concepts are engaging in classification task, thus, the impact of reducing key terms is negligible in F-measure.

As for Fig. 8, the tendencies of different α are divided into two different typical patterns. The first pattern, when α is between 0.1 and 0.6, demonstrate that the F-measure is declined when δ is rising. This pattern reveals that the decreasing concept caused by rising δ not only neglect relevant information but also impede the F-measure significantly. However, the disproportional relations between α and δ are limited when $\delta < 0.7$. While only important concepts are kept for document classification, $\delta > 0.7$, the impact from key term selection also becomes not significant. The second pattern appeals when $\alpha > 0.6$ which are stable confronting the variation of δ . The second pattern reveals that the number of concept involving in document classification is slightly affected F-measure when only high TF-IDF value key terms are selected. It is worth noting that there is no drop and up tendency appears in Fig. 8. As mentioned above, the drop and up tendency is caused by eliminating noise data, however, each tendencies of α in Fig. 8 are monotonically decreasing, which means every concept is capable of providing valuable information and none of them should be eliminated in the process of document classification. In traditional k-means methodology, how to decide K is very critical and may vary from different parameter settings. Thus, finding appropriate K remains a great challenge. In this experiment, the monotonically descending reveals that all concepts are suitable for providing valuable information. The highest F-measure of all α are consistently located when $\alpha = 0.1$. This fact indicates that the transformation from terms to concepts can successively retrieve the abstractness of terms and avoid the noise terms, which is inevitable in feature selection phase.

V. CONCLUSION

Nowadays, with the fast increasing of text, document classification is important to knowledge management and distribution. However, noise data is inevitable in practical document classification tasks. How to eliminate the impact from noise data is as important as achieving high classification accuracy. For this purpose, we develop a novel classification approach to generalize document to concept and successively reduce the impact from noise. Furthermore, according to accuracy assessment results, FFCA method demonstrates the impressive ability by achieving 96.17% in Reuters 21578 dataset. By further investigate the result tendencies of different parameter settings, FFCA method also provide benefits in determining parameter settings. Future research directions may include exploring more experimental data for evaluation and retrieving IF-THEN rules from lattice to support decision-making issues.

REFERENCES

- [1] O.-W. Kwon and J.-H. Lee, "Text categorization based on k-nearest neighbor approach for Web site classification," in *Information Processing and Management*, vol. 39 Elsevier Science, 2003, pp. 25-44.
- [2] P. J. Årgensen, "Incorporating context in text analysis by interactive activation with competition artificial neural networks," in *Information*

- Processing and Management*, vol. 41: Elsevier Science, 2005, pp. 1081-1099.
- [3] R. K. Rajapakse and M. Denham, "Text Retrieval with More Realistic Concept Matching and Reinforcement Learning," in *Information Processing and Management*, vol. 42, 2006, pp. 1260-1275.
- [4] M. Kim and P. Compton, "Evolutionary document management and retrieval for specialized domains on the web," *International Journal of Human-Computer Studies*, vol. 60, pp. 201-241, 2004.
- [5] L. E. Zárate, S. Mariano Dias, and M. A. Junho Song, "FCANN: A new approach for extraction and representation of knowledge from ANN trained via Formal Concept Analysis," in *Neurocomputing*, vol. 71: Elsevier Science, 2008, pp. 2670-2684.
- [6] T.-C. Hsieh and T.-I. Wang, "A mining-based approach on discovering courses pattern for constructing suitable learning path," *Expert Systems with Applications*, vol. 37, pp. 4156-4167, 2010.
- [7] R. Wille, "Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies," *Lecture Notes in Computer Science*, vol. 3626, pp. 1-33, 2005.
- [8] U. Priss, "Formal concept analysis in information science," *Annual Review of Information Science and Technology*, vol. 40, pp. 521-543, 2006.
- [9] Q. T. Tho, "A fuzzy FCA approach for citation-based document retrieval," in *Proceedings of the 2004 IEEE conference on cybernetics and intelligent systems*, Singapore, 2004, pp. 578-583.
- [10] T. Y. Wang and H. M. Chiang, "Fuzzy support vector machine for multi-class text categorization," *Information Processing and Management*, vol. 43, pp. 914-929, 2007.
- [11] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: theory and applications*. NJ: Prentice Hall, 1995.