J.Natn.Sci.Foundation Sri Lanka 2016 44 (2): 155-165
DOI: http://dx.doi.org/10.4038/jnsfsr.v44i2.7996

RESEARCH ARTICLE

Multi-label classification of computer science documents using fuzzy logic

Naseer Ahmed Sajid*, Muhammad Tanvir Afzal and Muhammad Abdul Qadir
Centre for Distributed and Semantic Computing, Mohammad Ali Jinnah University, Islamabad, Pakistan.

Revised: 24 August 2015; Accepted: 28 August 2015

Abstract: Classification has been already used for the prediction of predefined topics in many diversified domains including research paper classification task. A research paper may belong to one or more than one topic (classes). The state-of-the-art techniques in this area have the following limitations such as: (1) most of the techniques classify documents to at most one principal topic and do not identify all of the topic associations for research papers, (2) considers the classification problem of research documents in discrete domain and the accuracy of these techniques remain low when considering multiple classes for a single document. These limitations led us to explore the fuzzy domain for the classification of Computer Science documents because we are not sure whether the documents belong to one category or more than one category. Furthermore, fuzzy classification will help to identify the degree to which papers belong to different topics. To validate the findings of our research, we need a comprehensive dataset. Such a dataset has been made available by the scientific community for Computer Science domain. Therefore, in this paper, we restrict our focus to the Computer Science domain. Key features are extracted from the Title and Keywords of the research paper. We used term frequency (TF) as the weight scoring methodology. As a paper may belong to more than one category, we used fuzzy classifier, which automatically identifies all possible categories. Subsequently based on a threshold, the final one or more than one topic is assigned. We propose a generic framework and two algorithms for category (ies) identification. Our rules have been evolved (updated) by rules updater after the classification has been done by the fuzzy classifier. Performance of the technique with respect to accuracy has been compared with different classification techniques. The proposed approach has outperformed the state-of-the-art approaches.

Keywords: Category identification, document classification, fuzzy rules, research paper classification.

INTRODUCTION

Researchers produce a large number of scientific documents and scientific information is doubling every five years (Davis *et al*., 1995). Most of these documents are searchable over the internet using search engines, digital libraries and citation indexes. However, most of these documents are unstructured. Due to this unstructured nature of documents, search systems provide too many generic hits (Koller & Sahami, 1997). For example, a user query 'Ontology Engineering' in a Google search engine retrieves 2.6 million resources. This is due to the fact that the documents are not classified according to a proper hierarchy or structure. Moreover, the search systems retrieve all the documents or links, which contain one or more keywords of the user query. Therefore, recall of the user query is too high because in such a huge number of returned results, the probability of retrieving relevant documents becomes high; and the precision at the top remains low. This points to the necessity to develop an automatic technique to properly classify documents. A proper classification will help search engines, citation indexes, and digital libraries to: 1) classify research documents, 2) index them according to a structure, and 3) to answer user queries more effectively.

Classification has been already used for the prediction of predefined classes (Shin *et al*., 2008) and helps to assign documents to one or more categories. The state-of-the-art systems of document classification employ a number of approaches such as: decision tree (Gerstl *et al*., 2001), naïve Bayes classifier (Kononenko, 1990), term frequency and inverse document frequency (Salton, 1990), support vector machines (Cortes & Vapnik, 1995),

* Corresponding author (nasajid2005@gmail.com)

soft set based classifier, rough set based classifier, artificial neural network and approaches based on natural language processing (Salton & McGill, 1983; Sebastiani, 2002).

However in research paper classification, there are some limitations such as: 1) most of the above techniques categorise documents into one category from multiple categories (Gerstl *et al*., 2001), Computer Science documents belong to more than one category; 2) the state-of-the-art techniques consider the classification problem of computer science documents in discrete domain and the accuracy remains very low for the few systems, which consider multiple topics for classification (Salton *et al*., 1981; Salton, 1990; Gerstl *et al*., 2001; Kok-Chin & Choo-Yee, 2006). These issues (limitations) led us to explore the fuzzy domain for a proper classification of research documents. To evaluate the system we need a benchmark or user judgments and making a benchmark, which contains a comprehensive set of documents is a challenging task. Such a comprehensive benchmark dataset is available in the domain of Computer Science; therefore, the proposed technique has been validated on the comprehensive dataset of the Computer Science domain.

Fuzzy logic or domain was introduced by the mathematician Lotfi A. Zadeh (Zadeh, 1965). Zadeh is not only the founder of this but also the founder of fuzzy sets and fuzzy based systems. Fuzzy logic and sets are used to solve a variety of problems like pattern recognition, decision support, medicine, law, information retrieval, taxonomy and topology etc. (Perry, 1995). Fuzziness is about uncertainty and it indicates the probability that something is true. It has been used in information retrieval to account for data itself (Gershon, 1992), for result visualisation (Deller *et al*., 2007), for ontology to support in matching (Zhai *et al*., 2008) and for methods of matching (Ji & Yao, 2007).

In this paper, we propose a fuzzy classifier for the classification of Computer Science papers. We used research papers or articles (documents) from the dataset of the Journal of Universal Computer Science (J.UCS). This dataset contains research papers of different domains of Computer Science. The reason for the selection of this dataset is twofold: 1) the J.UCS covers all areas of computer science topics; 2) the authors belong to diversified domains, which gives a fair chance to the proposed technique to evaluate the system. Both of these helped us in the comprehensive evaluation of the proposed approach. We extracted key feature terms from the Title and Keywords of the papers. We selected the Title and Keywords of scientific publications because usually they contain the theme of the work and are also easily available online, which does not require extensive effort to acquire this metadata. Set of documents are represented below:

$$x_1, x_2, x_3, \ldots, x_n \in D$$

$$\{x_1, x_2, x_3, \ldots, x_n \mid C_1, C_2, C_3, \ldots, C_n\}$$

where $x_1, x_2, x_3, \ldots, x_n$ are key features (terms) from the given set of documents D and $C_1, C_2, C_3, \ldots, C_n$ are categories of these set of documents D. On the basis of above representation of documents in the form of key features, set of rules are represented as follows:

$$\{x_{i1}, x_{i2}, x_{i3}, \ldots, x_{in} \rightarrow C_{i1}, C_{i2}, C_{i3}, \ldots, C_{in}\}$$

where x_{ij} are terms, i represents the document and j (1,2,3,..) represents the terms of that document i, as the document may belong to more than one category. In C_{ij}, i represents the document and j (1,2,3,..) represents the categories of that document i. These terms are extracted from the Title and Keywords of the research papers and help the prediction of categories.

According to our knowledge, there are some existing techniques, which focus on the Computer Science documents. Kok-Chin and Choo-Yee (2006) used Keywords and four categories for document classification. Zhang *et al*. (2004) used the title and abstract for text classification. Senthamarai and Ramaraj (2008) used small vocabulary for text document classification approach.

In this study, we initially extracted the Title and Keywords from research papers and evaluated our approach using the documents (research papers) in the J.UCS dataset. First we trained our framework on 80 % of papers from the dataset and then 20 % of papers were used for the testing purpose. Fuzzy approach was used because we were not sure whether the documents belonged to one category or more. Therefore, we had to assign one or more categories to those documents. In training, initially we generated each rule for each paper of the category. Then, rules belonging to the same categories were merged by fuzzy based rule merger algorithm. We assigned weights to each merge-rule for deciding or predicting the category. The test document's rule weight was then compared and by using fuzzy classifier algorithm, the category for test document was predicted. Details of our algorithms and framework have been explained in the proposed

framework section (Figure 4). We also assigned some weights to those terms, which appeared repeatedly in the Title or Keywords of the document (research paper). For this purpose we used the term frequency technique for calculating the weights of each term in the rule. Our rules were also evolved and updated regularly whenever the new test document appeared for automatic classification. Rules were evolved to improve the performance of our approach for document classification (Computer Science documents). Our results for category prediction were better than the existing techniques.

PROPOSED FRAMEWORK

Document classification of Computer Science papers has been done using a number of techniques and datasets. The datasets used are normally the content and metadata of the papers. The content gives better precision due to rich number of features (Dendek *et al*., 2014), however, the content of scientific documents is not always available openly. Therefore, some authors have tried to classify papers based on metadata. Metadata is often defined as data about data or description about the actual data. In the domain of research papers it describes the creation, context or content of the actual documents. By using metadata, inconsistency or redundancy can be identified easily because the dataset of metadata is not too large. Metadata of a scientific document are the title, authors, keywords etc. However, metadata provide limited number of features, which does not give very accurate classification. The objective of this research is to use freely available metadata and test which metadata features are better suited for classification using a number of innovative approaches.

This research has proposed, developed, and tested a technique on metadata and have reported the results achieved so far. Another important finding from literature was that most of the works only focus on single classification of research papers. This means a paper is categorised to be associated with only one topic. However, research papers belong to more than one topic. This phenomenon (multi-label classification) has also been focused in this research.

The reason to select fuzzy classification is that research papers do not belong to only one category. There is a great possibility that a paper is partially associated with one topic and partially related to other topics. For example a paper on 'Network Routing Algorithm' has two associations: one with the network topic and the second with the algorithm topic. To identify such overlaps, fuzzy based systems have great accuracy and flexibility (Dehzangi *et al*., 2007; Yaguinuma *et al*., 2014).

To solve these types of problems, we used fuzzy logic. Fuzzy logic has been used to deal with the improbability and ambiguity of real world problems (Gershon, 1992). We proposed a framework for the categorisation of papers into one or more than one categories. First, we applied some preprocessing techniques to enable our dataset for the input of the framework. Then, we proposed an algorithm 'fuzzy based rules merger (FBRM)', to merge the rules generated. Next, we proposed second algorithm 'fuzzy classifier' to classify the papers into one or more than one categories. Finally, to increase the performance of our approach, rule updater is used to enrich our knowledge base (training set) for document classification. The details of the proposed framework is described below.

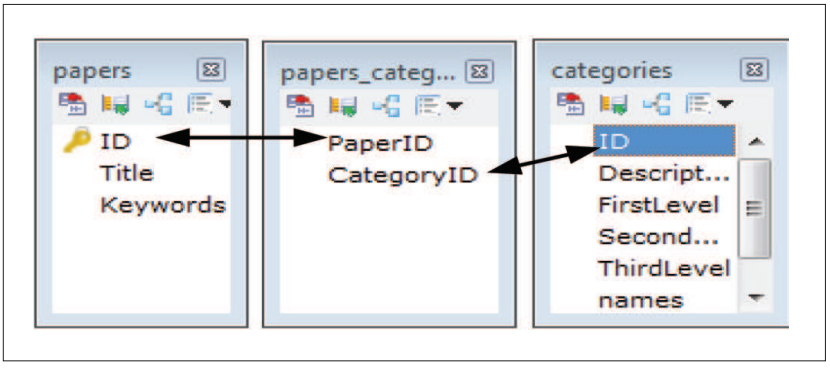

Figure 1: Tables selected from J.UCS dataset

Preprocessing

Features selection is an important part for document classification. Document classification's performance may be affected by the increase of features. So some preprocessing steps are necessary. For this purpose we take three tables (papers, papers_category and categories) from the J.UCS dataset, which is shown in Figure 1. From these tables, we generate the training dataset for our approach. The sample of training dataset is shown in Figure 2. Each row in Figure 2 represents the rule. The number of rules is equal to the number of rows in the dataset, which are *R1, R2, R3….,Rn*. In Figure 2, we can see that some papers belong to more than one category. That is why we used fuzzy approach in this paper. After that, we combined (merged) those papers, which belonged to the same category. There is a chance that some papers may belong to two or more categories. For this purpose, we apply the fuzzy logic to identify the most relevant category of the paper. The relevance of documents with relation to categories can be represented by means of linguistic terms. In addition, the importance of the document categories *via* linguistic variables allow the generation of fuzzy rules that can be used for identifying the most relevant category for that particular paper (Senthamarai & Ramaraj, 2008).

PaperID	Title	Keywords	category
1	Integration of Communities into Process-Oriented S...	cooperative knowledge generation,knowledge commun...	H
3	Small Groups Learning Synchronously Online at the ...	professional training,workplace learning,computer-...	H
3	Small Groups Learning Synchronously Online at the ...	professional training,workplace learning,computer-...	J
4	Using Weblogs for Knowledge Sharing and Learning i...	Experience-based Information System,wiki,weblog,pe...	A
4	Using Weblogs for Knowledge Sharing and Learning i...	Experience-based Information System,wiki,weblog,pe...	D
4	Using Weblogs for Knowledge Sharing and Learning i...	Experience-based Information System,wiki,weblog,pe...	H
4	Using Weblogs for Knowledge Sharing and Learning i...	Experience-based Information System,wiki,weblog,pe...	J
4	Using Weblogs for Knowledge Sharing and Learning i...	Experience-based Information System,wiki,weblog,pe...	K
5	Modelling and Implementing Pre-built Information S...	modelling method,introduction method,context-aware...	H
5	Modelling and Implementing Pre-built Information S...	modelling method,introduction method,context-aware...	I
5	Modelling and Implementing Pre-built Information S...	modelling method,introduction method,context-aware...	J
6	Tube Map Visualization: Evaluation of a Novel Know...	knowledge visualization,information visualization...	H
7	Reconciling Knowledge Management and Workflow Mana...	workflow,knowledge management	H
8	A Methodology and a Toolkit that Integrate Technol...	knowledge management,knowledge networks,inter-orga...	C
8	A Methodology and a Toolkit that Integrate Technol...	knowledge management,knowledge networks,inter-orga...	I
9	KMDL - Capturing, Analysing and Improving Knowledg...	Process-oriented Knowledge Management,knowledge-in...	D
9	KMDL - Capturing, Analysing and Improving Knowledg...	Process-oriented Knowledge Management,knowledge-in...	H
9	KMDL - Capturing, Analysing and Improving Knowledg...	Process-oriented Knowledge Management,knowledge-in...	I
10	The Role of Knowledge Management Solutions in Ente...	knowledge management,business process,enterprises,...	A
10	The Role of Knowledge Management Solutions in Ente...	knowledge management,business process,enterprises,...	H

Figure 2: Retrieval of required data for training from selected tables (sample)

Rules		Paper	Title	Keywords	Category
R3	R35	172	Knowledge,Integration,Source,Competitive,Adv	management,strategic,human, resource, managem	A
R5		173	Systmatic,Approach,Knowledge,Audit,Analysis	map,social,network,flow	A
R7		266	Benefits,knowledge,Management,Results,Germa	balanced,scorecard	A
R9		267	post-Nonaka, Knowledge, Management	generation,productivity,worker,scientific,Cynefi	A
R12		307	Strong, Effects, Soft, Factors,Knowledge,Manag	corporate,culture,leadership	A
R14		311	Effective,Integration,Knowledge,Management,Bu	Business,diagnostics,measurements,cost,benefit,c	A
R18		354	Knowledge,Attention,Gap,Underestimate,Proble	document,explosion,intelligent,agents,positive,ign	A
R29	R36	665	Estimation,Metrics,Courseware,Maintenance,Effort		D
R31		764	Formal,Analysis,Kerberos,Authentication,Syste	Methods,Security,Protocol,specification,Refinem	D
R23	R37	410	Error,Correction,Finite,Delay,Decodability	channel,decoding,detection,regular,language,trans	F
R25		491	Codifiable,Languages,Parikh,Matrix,Mapping	injectivity	F

Figure 3: Selection of rules belonging to same category (sample)

As some papers (documents) may belong to two or more categories, we have to find the most relevant categories for those papers. For this purpose, we developed a formula, and for calculation we first find the membership (here we find term frequency weights) of those papers with respect to their categories and then applied an alpha-cut "φ" (threshold) on that membership to identify the most relevant categories to those papers. Formal representation of identifying those categories is as follows:

$$\forall_i \mu_{ci}(P) \geq \Phi \rightarrow P : C_i$$

where P is the paper (document), C_i is the set of categories, Φ is an alpha-cut (threshold), which can be assigned to any value determined by domain experts; $\mu_{ci}(P)$ is the membership (term frequency weight) of P in category C_i and $P{:}C_i$ represents that paper P belongs to category set C_i.

In Figure 3, rules such as *R3, R5, R7, R9, R12, R14* and *R18* represent the paper's ID belonging to the same category. Papers belonging to the same category are then merged into a single rule such as *R35* for papers of category *A*. Similarly all the rules, which represent

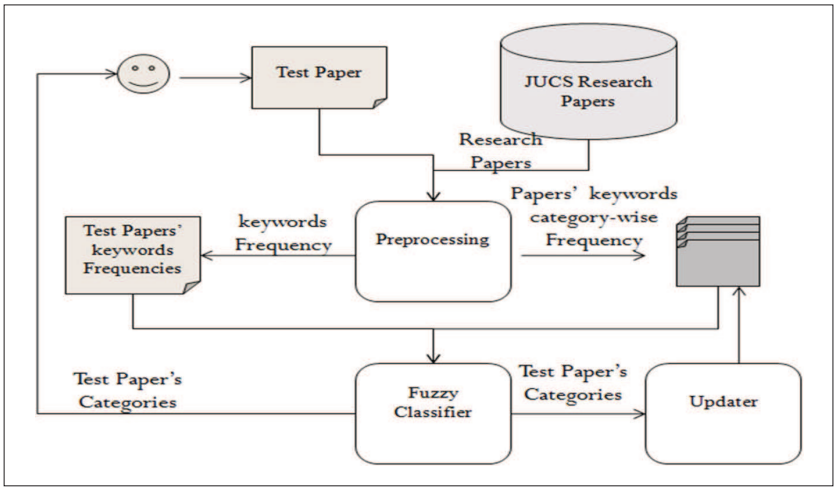

Figure 4: Proposed framework for classification

Input: Rules belong to the same category of Dataset $\mathfrak{D}$.
Output: Term Frequency of each category.

1. $\mathcal{K}$ ← Array of *KEYWORDS* of category c.
2. $\mathcal{T}$ ← Array of *TITLE* of category c.
3. $\mathfrak{R}$ ← Resultant array of $\mathcal{K}$ and $\mathcal{T}$ of category c.
4. **For each** C_i in c // c is a set of categories
5. **For each** d_j in $\mathfrak{D}$ // $\mathfrak{D}$ is a set of documents in a C_i
6. Locate *KEYWORDS* or *TITLE* in d_j
7. **If** found **then**
8. $\mathcal{K}$ = Concatenate ($\mathcal{K}$, *KEYWORDS* in d_j)
9. $\mathcal{T}$ = Concatenate ($\mathcal{T}$, *TITLE* in d_j)
10. **End if**
11. **End For loop**
12. $\mathfrak{R}$ = Concatenate ($\mathcal{K}$,$\mathcal{T}$) // for each category
13. Count_Terms_Frequency ($\mathfrak{R}$)
14. **End For Loop**

Figure 5: Fuzzy-based rules merger algorithm

the same categories are merged into a single rule. At the end, each category has only one rule. All this is done by our fuzzy based rules merger (FBRM) algorithm. When all the training papers (documents) are assigned to their respective categories as shown in Figure 3, to remove the unrelated, unnecessary and not meaningful words from the Keywords and Title, we used an approach to remove the stop words and stemming algorithm (Porter, 1997) to break the compound words into single words. After that we applied our FBRM algorithm to calculate the term frequencies of Keywords, Title and Keywords + Title against each category. Our proposed framework is shown in Figure 4. It has two main components. One is FBRM algorithm and the other is fuzzy classifier.

Input: Term Frequency of Test Document ($\mathcal{TD}$),
Term Frequency of each category c,
α -cut (threshold)
Output: Set of categories

1. $w \leftarrow$ Array of category's weights.
2. **For each** y_j in c // terms in each category.
3. **For each** term x_i in $\mathcal{TD}$ // terms in test document
4. **If** x_i is found in y_j **then**
5. //Calculate the term weight of test document in each category.
6. $w_k = \sum_{k=0}^{n}[\text{TermWeight}(xi) * \text{TermWeight}(yj)]$
7. **End if**
8. **End For loop**
9. $WeightSum = \sum_{k=0}^{n} w_k$ // total terms weights of all category against test document
10. **End For Loop**
11. **For each** weight w_k in w
12. $\mu_{yj}(\mathcal{D}) = (w_k / WeightSum)$ //membership of test document in each category
13. **If** $\mu_{yj}(\mathcal{D}) \geq \varphi$ then //applying α -cut(threshold)
14. $c_i \leftarrow \mathcal{D}$ // assign document to that category
15. Update_Category (c_i)
16. **End if**
17. **End For loop**

Figure 6: Fuzzy classifier algorithm

Manual Output

Category Description	Names	No. Of rules	Metadata	Features Selected									
				Knowledge	*Integration*	*Learning*	*Model*	*Management*	*Information*	*Retrieval*	*System*	*User*	.....
Information System	H	257	Title + Keywords	97	9	56	76	71	95	34	95	44	
			Title	42	4	23	31	26	28	13	42	19	
			Keywords	55	5	33	31	45	67	21	53	25	
Software	D	263	Title + Keywords	12	5	10	81	14	11	3	65	8	
			Title	5	2	4	38	2	5	1	29	3	
			Keywords	7	3	6	43	12	6	2	36	5	
General Literature	A	36	Title + Keywords	44	8	8	2	31	16	3	7	0	
			Title	20	4	4	0	13	7	0	2	0	
			Keywords	24	4	4	2	18	9	3	5	0	
Computer Applications	J	16	Title + Keywords	8	4	6	2	2	5	0	1	1	
			Title	4	2	2	1	1	1	0	1	0	
			Keywords	4	2	4	1	1	4	0	0	1	

System Generated Output

Paper ID=172-------Origional Categroy =A,H

---------System Genereated Results----------
Category A=0.255
Category D=0.098
Category H=0.599
Category J=0.048
Final category(ies) of test paper=A,H,

Test Paper (172), Category (H)	
Features	(TF)
Knowledge	3
Integration	2
Source	1
Competitive	2
Advantage	2
Management	3
Resourse	1

Category	Metadata	Weights	Membership(μ)
H	Title + Keywords	(3*97)+(2*9)+(1*0)+(2*0)+(2*0)+(3*71)+(1*0)=522	522/877=**0.60**
D		(3*12)+(2*5)+(1*0)+(2*0)+(2*0)+(3*14)+(1*0) =88	88/877=0.1
A		(3*44)+(2*5)+(1*0)+(2*0)+(2*0)+(3*31)+(1*0) =235	235/877=**0.27**
J		(3*8)+(2*1)+(1*0)+(2*0)+(2*0)+(3*2)+(1*0) =32	32/877=0.04
H	Title	(3*42)+(2*4)+(1*0)+(2*0)+(2*0)+(3*26)+(1*0)=212	212/357=**0.59**
D		(3*5)+(2*2)+(1*0)+(2*0)+(2*0)+(3*2)+(1*0)=25	25/357=0.07
A		(3*20)+(2*3)+(1*0)+(2*0)+(2*0)+(3*13)+(1*0)=105	105/357=**0.29**
J		(3*4)+(2*0)+(1*0)+(2*0)+(2*0)+(3*1)+(1*0)=15	15/357=0.04
H	Keywords	(3*55)+(2*5)+(1*0)+(2*0)+(2*0)+(3*45)+(1*0)=310	310/520=**0.60**
D		(3*7)+(2*3)+(1*0)+(2*0)+(2*0)+(3*12)+(1*0)=63	63/520=0.12
A		(3*24)+(2*2)+(1*0)+(2*0)+(2*0)+(3*18)+(1*0)=130	130/520=**0.25**
J		(3*4)+(2*1)+(1*0)+(2*0)+(2*0)+(3*1)+(1*0)=17	17/520=0.03

Figure 7: Comparison of manual *vs* system generated output of fuzzy classifier algorithm

			Threshold=0.1	Threshold=0.15	Threshold=0.2
S #	PID	Origional_Category	Predicted_Category	Predicted_Category	Predicted_Category
1	1	C,	C,D,H,I,	D,H,	D,H,
2	2	H,K,	D,H,I,K,	D,H,I,K,	D,H,
3	3	C,	C,D,H,I,K,	C,D,H,K,	C,D,
4	9	C,	C,D,H,I,K,	D,	D,
5	12	K,C,	C,D,H,I,	C,D,H,	C,D,
6	13	C,	C,D,H,I,K,	C,D,H,	C,D,
7	14	H,K,	B,C,D,H,I,K,	C,D,	D,
8	15	C,	C,D,H,I,K,	C,D,H,K,	D,
9	17	C,G,	C,D,H,I,K,	C,D,H,	D,
10	18	D,	C,D,H,I,	C,D,H,	C,D,
11	20	H,K,	C,D,H,K,	D,H,	D,H,
12	21	C,	C,D,H,I,	C,D,H,	C,D,
13	22	C,	C,D,H,I,	C,D,	D,
14	24	C,	C,D,H,K,	C,D,H,	D,
15	25	D,	C,D,H,I,K,	D,	D,
16	26	B,	B,C,D,H,K,	D,H,	D,
17	27	D,	C,D,H,	C,D,	D,
18	28	H,	B,C,D,H,I,	D,H,	D,
19	29	D,B,E,F,	C,D,H,I,K,	D,H,	D,
20	30	B,	B,C,D,H,I,	D,H,	D,

Figure 8: Sample of system generated output of fuzzy classifier algorithm

Fuzzy based rules merger (FBRM) algorithm

FBRM algorithm merges rules, which belong to the same category. Initially in preprocessing, we assigned a rule for each document (eg: *R3, R5, R7, R9, R12, R14, R18*) and then combined those rules (eg: *R35*), which belonged to the same category. This algorithm extracts Keywords and Title from research papers and concatenates them against each category.

We separately concatenate the Keywords and Title against each category and also concatenate both Keywords and Title together against each category. In addition, we have calculated the term frequency (TF) against the resultant Keywords string, resultant Title string and resultant of both Keywords and Title together. The FBRM algorithm is shown in Figure 5.

Fuzzy classifier

When a user submits a test document for classification, preprocessing steps are performed as discussed above and term frequency weights of the test document are computed. After comparing test document terms weights with the rules weights of each category, we got some results against each category. For that particular test document, fuzzy based classifier predicted the most relevant category or categories on the basis of membership of each category by applying the "φ" α-cut (threshold). The algorithm of fuzzy classifier is shown in Figure 6. In the selected dataset, manual selection of topics by the authors of the papers is available. The proposed system was evaluated against those predefined topics. The comparisons have been shown in Figure 7. The sample output of fuzzy classifier algorithm is presented in Figure 8.

Rules updater

When the fuzzy classifier assigned a category or categories for a test document, we have to update rule weights for that particular category or categories where the classifier assigned the test document. In this way, we enrich our knowledge base (training set) for document classification. By doing this, the performance of our classification approach will increase due to increase of our training rules weights.

The working of our framework is explained in the following steps:

1. Extract Keywords and Title of the research papers.
2. Tokenize it and remove stop words and apply stemming.
3. Calculate the term frequency of the test document.
4. Compare the term frequency with the term frequency of merge rules.
5. Calculate the weights against each category.
6. Apply fuzzy similarity measures to get the most relevant category (or categories) for the test document.

7. Assign test document to those relevant categories.
8. Update the term frequency of the relevant categories (rules) according to the term frequency of the test document.

RESULTS AND DISCUSSION

To evaluate the proposed scheme we calculate precision and accuracy on the Journal of Universal Computer Science (J.UCS) dataset. Related features of the J.UCS dataset and the number of research papers used for training and testing the dataset are also provided in Tables 1 and 2. Figure 9 shows the categories-wise papers of the J.UCS dataset.

In Table 3, 'YES' and 'NO' represent a crisp decision given for document classification where document d_i assigns to category(ies) C_i. Prediction of each document's category entry in the table indicates the number of documents specified against each type (YES or NO).

The description of each type of contingency table is as follows: In *True Positive (TP)*, system predicts the numbers of true positive documents which actually belong to category C_i; in *False Positive (FP)*, system predicts the numbers of false positive documents which actually do not belong to category C_i; in *False Negative (FN)*, system predicts the numbers of false negative documents which actually belongs to category C_i and in *True Negative (TN)*, system predicts the number of true negative documents which actually do not belong to category C_i.

Based on the above parameters, the standard performance measures for evaluation are computed such as: precision and recall. Precision is the percentage of True Positive as correct and recall is the percentage of True Positive as predicted.

Table 1: Related features of J.UCS dataset

Total number of research papers	1460
Average number of research papers in each main category	112
Average number of multi-class research papers	234
Number of research papers for training	1010
Number of research papers for testing	450

Table 2: Categories-wise papers of J.UCS dataset

Categories	Papers
A	41
B	77
C	172
D	585
E	62
F	445
G	125
H	760
I	372
J	105
K	236
L	33
M	26

Table 3: Contingency table for categories

	YES (T)	NO (F)
Category predict	True +ve (TP)	False +ve (FP)
Category predict	True –ve (TN)	False –ve (FN)

Table 4: Contingency table for the proposed system

	YES (T)	NO (F)
Category predict (P)	82 %	6 %
Category predict (N)	9 %	3 %

Table 5: Comparison of proposed algorithm with other techniques

Sr#	Approaches	Accuracy (%)
1	Fuzzy classifier (proposed)	91.00
2	Similarity based technique (Senthamarai & Ramaraj, 2008)	90.00
3	PSO based (Ali *et al*., 2013)	84.71
4	Bayesian (Kok-Chin & Choo-Yee, 2006)	83.75
5	Naïve Bayesian network (Kok-Chin & Choo-Yee, 2006)	82.50
6	Bayesian network learned from training documents (Kok-Chin & Choo-Yee, 2006)	76.25
7	Reference based (Sajid *et al*., 2011)	70.00
8	Majority GP (Zhang *et al*., 2004)	60.81
9	Support vector machine (Zhang *et al*., 2004)	57.74
10	Majority based evidence (Zhang *et al*., 2004)	53.60

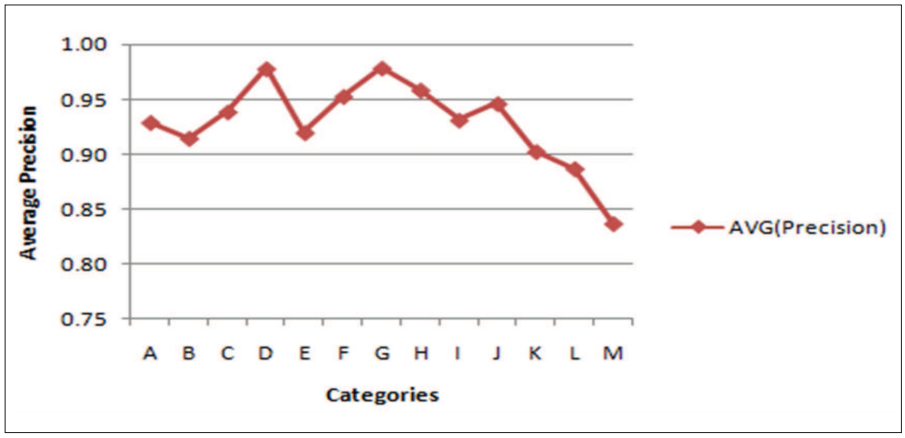

Figure 10: Category-wise average precision (fuzzy)

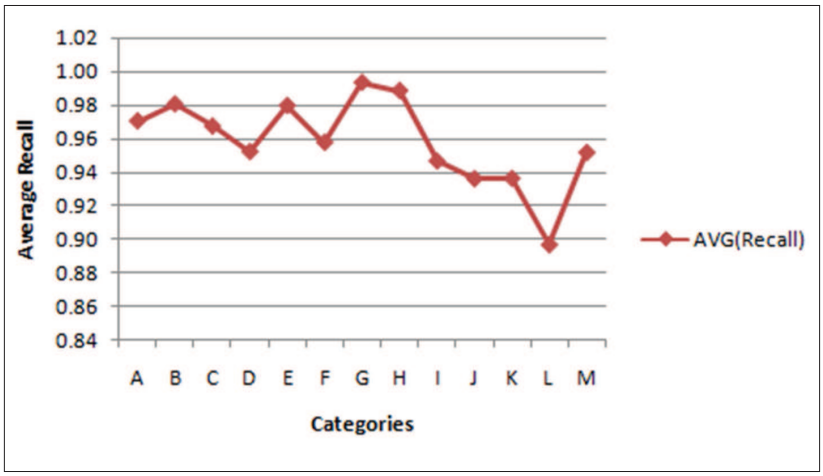

Figure 11: Category-wise average recall (fuzzy)

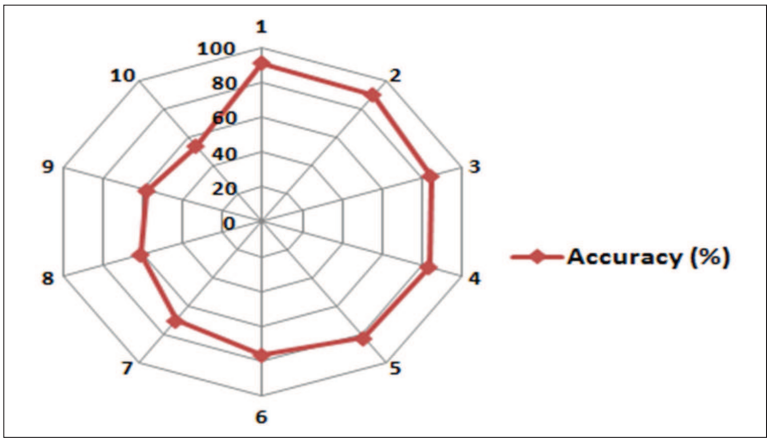

Figure 12: Performance measures graph

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error} = \frac{FP + FN}{TP + TN + FP + FN}$$

where $TP+FP > 0$ and $TP+TN+FP+FN > 0$.

After detailed analysis of our results, precision and recall of our approach are 93 % and 96 %, respectively. We calculated the precision and recall for each paper with respect to each category, counted the papers in each category and added their precision and recall percentage. After that we determined average precision and recall for each category, which are shown in Figures 10 and 11.

We have compared our approach with different document classification approaches, which are techniques for text document classification on the basis of similarity (Senthamarai & Ramaraj, 2008), PSO based (Ali *et al*., 2013), Bayesian (Kok-Chin & Choo-Yee, 2006), Naïve Bayesian (Kok-Chin & Choo-Yee, 2006), Bayesian network learned from data (Kok-Chin & Choo-Yee, 2006), reference based (Sajid *et al*., 2011), majority GP (Zhang *et al*., 2014), support vector machine (Zhang *et al*., 2014) and majority best evidence (Zhang *et al*., 2014). In Table 5, comparison has been done on the basis of accuracy. We have concluded that our approach performs better than other mentioned document classification approaches. The performance measure graph (accuracy) of all mentioned approaches is shown in Figure 12.

CONCLUSION

This paper proposed, implemented and evaluated a framework for fuzzy based classification of Computer Science documents. Both algorithms, fuzzy based rules merger and fuzzy classifier worked well for Computer Science document classification. Rules updating mechanism increased the performance of our approach for Computer Science document classification. In this paper, we tested the proposed framework on the comprehensive dataset of J.UCS against ACM categorisation hierarchy. According to the comparison with state-of-the-art classification systems, the accuracy of the proposed approach proved to be better.

REFERENCES

1. Ali T., Sajid N.A., Asghar S. & Ahmed M. (2013). Classification of scientific publications using swarm intelligence. *Proceedings of the Pakistan Academy of Sciences* **50**(2): 115 − 126.
2. Cortes C. & Vapnik V. (1995). Support vector networks. *Machine Learning* **20**: 273 − 297. DOI: http://dx.doi.org/10.1007/BF00994018
3. Davis J., Weeks R. & Revett M. (1995). Jasper: communicating information agents for WWW. *Proceedings of the Fourth International World Web Conference*, issue 1, Boston, Massachusetts, USA, pp. 11 – 14.
4. Dehzangi O., Zolghadri M.J., Taheri M.J. & Fakhrahmad S.M. (2007). Efficient fuzzy rule generation: a new approach using data mining principles and rule weighting, *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, volume 2, Haikou, China, 24 - 27 August, pp. 134 − 139. DOI: http://dx.doi.org/10.1109/FSKD.2007.267
5. Deller M., Ebert A., Bender M., Agne S. & Barthel H. (2007). Preattentive visualization of information relevance. *Proceedings of the International Workshop on Human-centered Multimedia HCM'07*, Augsburg, Bavaria, Germany, 23 – 28 September, pp. 47 – 56. DOI: http://dx.doi.org/10.1145/1290128.1290137
6. Dendek P.J., Czeczko A., Fedoryszak M., Kawa A., Wendykier P. & Bolikowski L. (2014). Content analysis of scientific articles in apache hadoop ecosystem. *Intelligent Tools for Building a Scientific Information Platform: From Research to Implementation* (eds. R. Bembenik, L. Skonieczny, H. Rybiński, M. Kryszkiewicz & M. Niezgódka), pp. 157 − 172. Springer International Publishing, Switzerland. DOI: http://dx.doi.org/10.1007/978-3-319-04714-0_10
7. Gershon N.D. (1992). Visualization of fuzzy data using generalized animation. *Proceedings of the IEEE Conference on Visualization*, Boston, MA, 19 – 23 October, pp. 268 – 273. DOI: http://dx.doi.org/10.1109/visual.1992.235199
8. Gerstl P., Hertweck M. & Kuhn B. (2001). Text mining: grundlagen, verfahren and anwendungen. *Praxis der Wirtschaftsinformatik - Business Intelligence* **39**: 22238 − 22248.
9. Ji R. & Yao H. (2007). Visual and textual fusion for region retrieval: from both fuzzy matching and Bayesian reasoning aspects. *Proceedings of the 9th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Augsburg, Bavaria, Germany, 24 – 29 September, pp. 159 – 168. DOI: http://dx.doi.org/10.1145/1290082.1290106
10. Journal of Universal Computer Science (J.UCS). Available at *http://www.jucs.org/*, Accessed 20 November 2011.
11. Kok-Chin K. & Choo-Yee T. (2006). A Bayesian approach to classify conference papers. *MICAI 2006: Advances in Artificial Intelligence* (eds. A. Gelbukh & C.A. Reyes-Garcia), pp. 1027 − 1036. Springer Berlin Heidelberg, Germany.

12. Koller D. & Sahami M. (1997). Hierarchically classifying documents using very few words, *Proceedings of the 14th International Conference on Machine Learning (ICML-97)*, Nashville, USA, pp. 170 − 178.
13. Kononenko I. (1990). Comparison of inductive and naïve bayesian learning approaches to automatic knowledge acquisition. *Current Trends in Knowledge Acquisition*. IOS Press, Amsterdam, The Netherlands.
14. Perry T.S. (1995). Lofti A. Zadeh (the inventor of fuzzy logic). *IEEE Spectrum* **32**(6): 32 – 35. DOI: http://dx.doi.org/10.1109/6.387136
15. Porter M.F. (1997). An algorithm for suffix stripping. *Readings in Information Retrieval*, pp. 313 − 316. Morgan Kaufmann Publishers Inc., San Francisco, USA.
16. Sajid N.A., Ali T., Afzal M.T., Qadir M.A. & Ahmed M. (2011). Exploiting reference section to classify paper's topics. *Proceedings of the International Conference on Management of Emergent Digital EcoSystems* (*MEDES'2011*), San Francisco, California, USA, 21 – 23 November, pp. 220 – 225. DOI: http://dx.doi.org/10.1145/2077489.2077531
17. Salton G. (1990). Developments in automatic text retrieval. *Science* **253**: 974 − 980. DOI: http://dx.doi.org/10.1126/science.253.5023.974
18. Salton G. & McGill M.J. (1983). *Introduction to Modern Retrieval*. McGraw-Hill Book Company, New York, USA.
19. Salton G., Wu H. & Yu C.T. (1981). The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science* **32**(3): 175 – 186. DOI: http://dx.doi.org/10.1002/asi.4630320304
20. Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* **34**(1): 1 − 47. DOI: http://dx.doi.org/10.1145/505282.505283
21. Senthamarai K. & Ramaraj N. (2008). Similarity based technique for text document classification. *International Journal of Soft Computing* **3**(1): 58 − 62.
22. Shin K., Abraham A. & Han S., (2008). Enhanced centroid-based classification technique by filtering outliers. *Text, Speech and Dialogue*, pp.159 − 163. Springer Verlag, Berlin Heidelberg, Germany.
23. Yaguinuma C.A., Santos M.T.P., Camargo H.A., Nicoletti M.C. & Nogueira T.M. (2014). A meta-ontology for modeling fuzzy ontologies and its use in classification tasks based on fuzzy rules. *International Journal of Computer Information Systems and Industrial Management Applications* **6**: 89 − 101.
24. Zadeh L.A. (1965). Fuzzy sets. *Information and Control* **8**(3): 338 − 353.
25. Zhai J., Chen Y., Wang Q. & Lv M. (2008). Fuzzy ontology models using intuitionistic fuzzy set for knowledge sharing on the semantic web. *Proceedings of the 12th International Conference on Computer Supported Cooperative Work in Design*, Xi'an, China, 16 - 18 April, pp. 465 – 469.
26. Zhang B., Gonçalves M.A., Fan W., Chen Y., Fox E.A., Calado P. & Cristo M. (2004). Combining structural and citation-based evidence for text classification. *Proceedings of the 13th ACM International Conference on Information and Knowledge Management*, pp. 162 − 163. DOI: http://dx.doi.org/10.1145/1031171.1031204