Journal of Science and Technology (JST)
Volume 2, Issue 4, July - August 2017, PP 42-50
www.jst.org.in *ISSN: 2456 - 5660*

Domain Extraction From Research Papers

Dr. R. Jayanthi[1], S. Sheela[2]

[1] *Dr. Mrs. R. Jayanthi, Assistant Professor, PG and Research Department of Computer Science, Quaid E – Millath college for Women, Chennai, India..*

[2]*Ms. S. Sheela, M.Phil Research Scholar, Computer Science, Quaid E – Millath college for Women, Chennai, India.*

Abstract: *Automatically finding domain specific key terms from a given set of research paper is a challenging task and research papers to a particular area of research is a concern for many people including students, professors and researchers. A domain classification of papers facilitates that search process. That is, having a list of domains in a research field, we try to find out to which domain(s) a given paper is more related. Besides, processing the whole paper to read take a long time. In this paper, using domain knowledge requires much human effort, e.g., manually composing a set of labeling a large corpus. In particular, we use the abstract and keyword in research paper as the seeing terms to identify similar terms from a domain corpus which are then filtered by checking their appearance in the research papers. Experiments show the TF –IDF measure and the classification step make this method more precisely to domains. The results show that our approach can extract the terms effectively, while being domain independent.*

Keywords - *Text Mining, Information Extraction, Domain keyword extraction, Term Frequency and Inverse Document Frequency (TF – IDF)*

I. INTRODUCTION

Domain – specific terms are term that have significant meaning (s) in a specific domain [1]. We extract the term from the research papers. Here a "term" refers to a word or compound words representing a concept of a specific domain, e.g., in chemistry, "alcohol" is a term that refers to an organic compound in which the hydroxyl functional group is bound to a saturated carbon atom [2]. Terminology extraction are using rule based techniques, supervised learning techniques or a combination of these two types of techniques all of which rely on some domain knowledge. Acquiring such domain knowledge requires much human effort (e.g., manually labeling a large corpus) [3].

To overcome this approach uses semantic extraction by building knowledge from a large corpus. The semantic extraction refers to range of processing techniques that identify and extract entities, facts, attributes, concepts, and events to populate meta- data fields. The purpose of this is to enable the analysis of semi-structured or unstructured content. Semantic extraction is usually based on three approaches,

*1. Rule based***:** Matching similar to entity extraction, this approach requires the support of one or more vocabularies.

2. Machine learning: A statistical analysis of the content, that potentially compute intensive application that can benefit within the document corpus.

3. Hybrid solution: Statistically driven, but enhanced by a vocabulary. This is typically the best approach if the content set is focused on a specific subject area.

The extracted information also should be machine understandable as well as human understandable in term of research paper from set of domain corpus [4]. A statistical method is proposed in this paper and it is based on, 3 steps.

First, extract the abstract and keyword in the collection of research paper using pdfifilter. Second, terms which are similar to a certain domain occur frequently in research paper. Third, count word introduced learning to Predict from Text (Weighted Scoring Method). The highly count value is the domain name of particular research paper. The TF–IDF is introduced, the weighting method of measure the value of domain corpus. User can easily find out the domain name from research paper as well as to separate folder for each domain paper to save on your computer. Later users don't waste the time for searching the research paper. We can save the time for searching the research paper with domain name.

II. LITERATURE REVIEW

For text mining of domain extraction techniques we have studied few related papers. In this section we describe the different techniques with different authors which are related to the domain extraction.

In this paper [2] existing terminology extraction approaches are mostly domain dependent. They use domain specific linguistic rules, supervised machine learning techniques. In particular, we use the title words and the keywords in research papers as the seeing terms and word2vec to identify similar terms from an open-domain corpus as the candidate terms, which are the filtered by checking their occurrence in research papers.

Rakhi Chakraborty, explains [5] it is extremely time consuming and difficult task to extract keyword or feature manually. So an automated process that extracts keywords or features needs to be established. This paper proposes a new domain keyword extraction technique that includes a new weighting method on the base of the conventional TF - IDF. Term frequency-Inverse document frequency is widely used to express the documents feature weight, which can't reflect the division of terms in the document, and then can't reflect the significance degree and the difference between categories.

Hospice Houngbo, Robert E. Merer, Method Mention Extraction from Scientific Research Paper, explains [6] scientific publications contain many references to method terminologies used during scientific experiments. In this study we report our attempt to automatically extract such method terminologies from scientific research papers, using rule-based and machine learning techniques. We first used some linguistic features to extract fine-grained method sentences from a large biomedical corpus and then applied well established methodologies to extract the method terminologies.

The author of [7] the computational linguistics community and its sub-fields have changed over the years with respect to their foci, methods used, and domain problems. We extract these characteristics by matching semantic extraction patterns, learned using bootstrapping, to the dependency trees of sentences in an article's abstract.

In this paper [8] the dynamics of a research community can be studied by extracting information from its publications. Such information cannot be extracted using approaches that assume words are independent of each other in a document. We use dependency trees, which give rich information about structure of a sentence, and extract relevant information from them by matching semantic patterns.

III. DOMAIN KEYWORD EXTRACTION TECHNIQUES

Domain Corpus

In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. This paper, uses seven Domain name with set of large corpus.

SONY.domainExtraction - dbo.corpus

	domainname	keywords
	Cryptography	Key Management, Encryption, secret key encryption, public key encryption, RSA, DES, AE...
	Network Security	Wireless communication, Wireless Sensor Networks, packets, Time Synchronization, Multi-...
	Data Mining	Automated, semi-automated techniques, data discovery, DiDrip, Data mining, K-Mean clu...
	Image Processing	Image processing, Diabetic Retinopathy Retinal images, Biomedical image Processing, exu...
	Software Engine...	Test Case Generation, UML Activity Diagram, Software Testing, Test Cases, Test Automat...
	Big Data	Big Data, Process, Analysis, Hadoop, Storage, Map Reduce, Cloud Storage, Disaster Reco...
	Text Mining	Content-based filtering, Clusters, Classification, kNN, Demographic Information, precision ...
▶*	NULL	NULL

Figure 1: Domain Corpus in Research Paper

Figure1 shows the list for each domain some number of keywords is manually setting such as Text Mining 51, Data Mining 66, Cryptography 56, Software Engineering 53, Big Data 57, Network Security 78, and Image Processing 73.

PDF IFILTER

An IFilter is a plugin that allows Microsoft's search engines to index various file formats (as documents, email attachments, database records, audio metadata etc.) So that they become searchable. Without an appropriate IFilter, contents of a file cannot be parsed and indexed by the search engine. An IFilter acts as a plug-in for extracting full-text and metadata for search engines.

A search engine usually works in two steps: The search engine goes through a designated place, e.g. a file folder or a database, and indexes all documents or newly modified documents, including the various types documents, in the background and creates internal data to store indexing result. A user specifies some keywords he would like to search and the search engine answers the query immediately by looking up the indexing result and responds to the user with that contains the keywords. During Step 1, the search engine itself doesn't understand format of a document.

Therefore, it looks on Windows registry for an appropriate IFilter to extract the data from the document format, filtering out embedded formatting and any other non-textual data. Figure 2 explain, one full research paper (paper format pdf) read after using this tool to extract abstract and keywords in research paper.

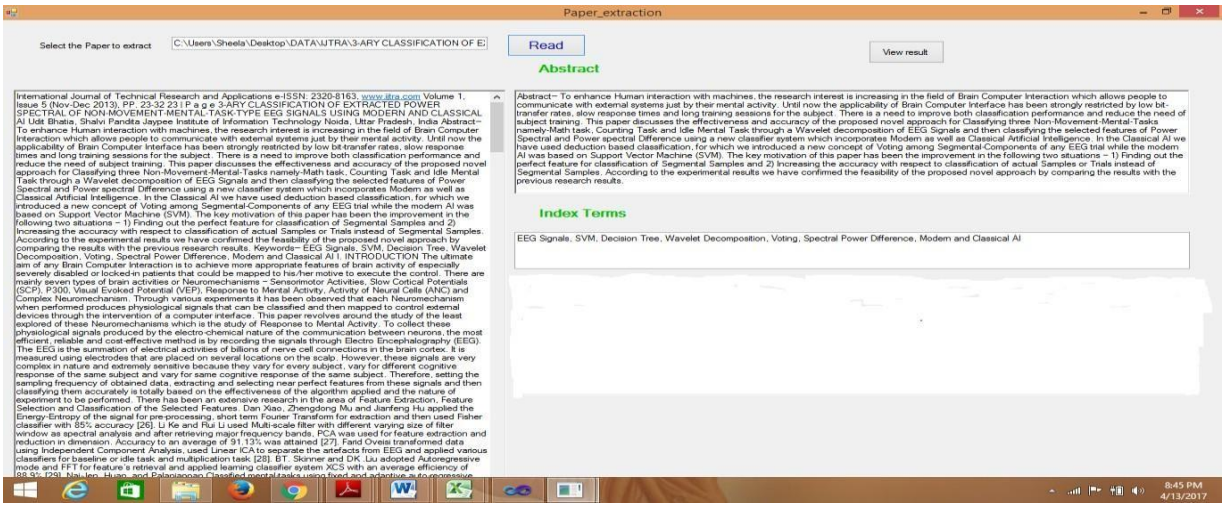


Figure 2: Extract abstract and Index Terms from research paper TEXT

Similarity Mining

Text Vector must be generated before text similarity calculation between domain corpus and extraction of research paper. A Chinese corpus converting economy, military, education, culture, and other fields, contains approximately 100,000 documents [9]. Generating word figure3 sets from research paper such as domain corpus. The process is relatively simple feature words were extracted from research paper and put in the domain corpus.

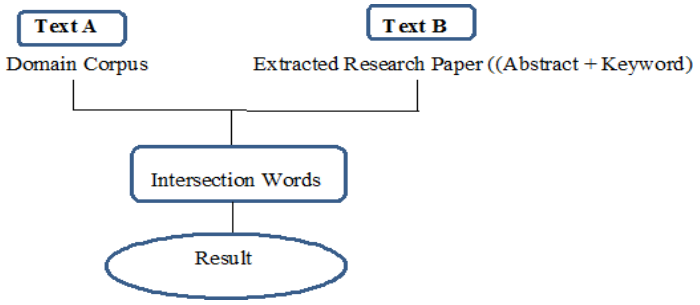


Figure 3: Intersection of Domain Corpus and Extracted Research Paper

Figure 4, getting the result to calculate the total word for each domain. Total result of particular domain name in given research paper, t is the maximum number appearance of the word.

t = count (Maximum Number of words appearance in domain corpus for each domain)

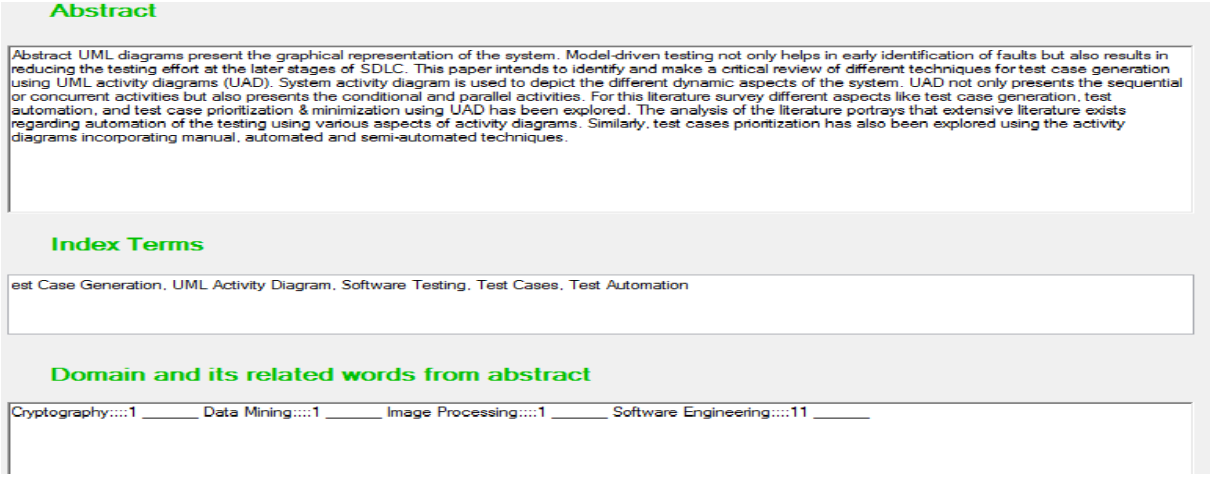


Figure 4: Count the words appearance in domain corpus for each domain

Finally, figure 5 explain the maximum number of count in the domain result of one research paper. For each paper for calculating and get the result for 200 research paper.

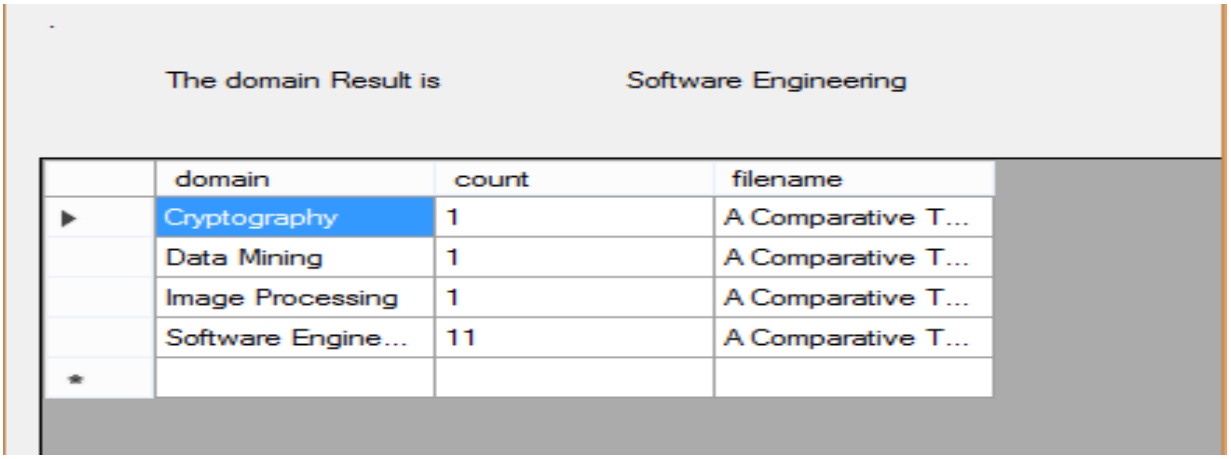


Figure 5: Domain Result

EVALUATION MEASURE

Precision

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

Precision = (relevant items retrieved) / (retrieved items) = P (relevant|retrieved)

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision. For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results. Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system. Note that the meaning and usage of "precision" in the field of information retrieval differs from the definition of accuracyand precision within other branches of science and technology.

Recall

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

Recall = (relevant items retrieved) / (relevant items) = P(retrieved|relevant)

For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by computing the precision.

F –Measure

F1 measure is a derived effectiveness measurement. The resultant value is interpreted as a weighted average of the precision and recall. The best value is 1 and the worst is 0.

F –Measure = 2((precision*recall) / (precision+recall))

TF – IDF

TF: Term Frequency, which measures how frequently a term occurs in a document. Since Table1, every document is different in length it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (the total number of terms in the document) as a way of classification [10].

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)

The domain result is Data mining then value is 1 otherwise 0.

Total number of paper: 200

TF (t) = 40/56 = 0.7142857142

Where, text mining paper total count is 40 and domain corpus count is 56. The TF calculation number of times t appear in the document value dividing the total number of terms in the document. Table 1, explain the number of paper appears in the particular domain and the number of times appears, then the total calculation for each domain.

Table 1: TF calculation

Domain name	**Paper1**	**Paper 2**	**Paper 3**	**Paper 4**	**Paper ….**	**Paper 200**	**Total**
Data mining	1	0	0	1	…	1	20
Text mining	0	1	1	1	…	1	40
Network security	1	0	1	1	…	0	20
Cryptography	1	1	0	0	…	1	35
Software engineering	1	0	1	0	…	0	25
Image processing	1	0	1	1	...	0	30
Big data	1	1	0	1	…	1	30

IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following [10]:

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

IDF(t) = log_e (200 / 20)

Total no of paper =200

No of paper occur = 10

Finally,

tf-idf = tf *idf

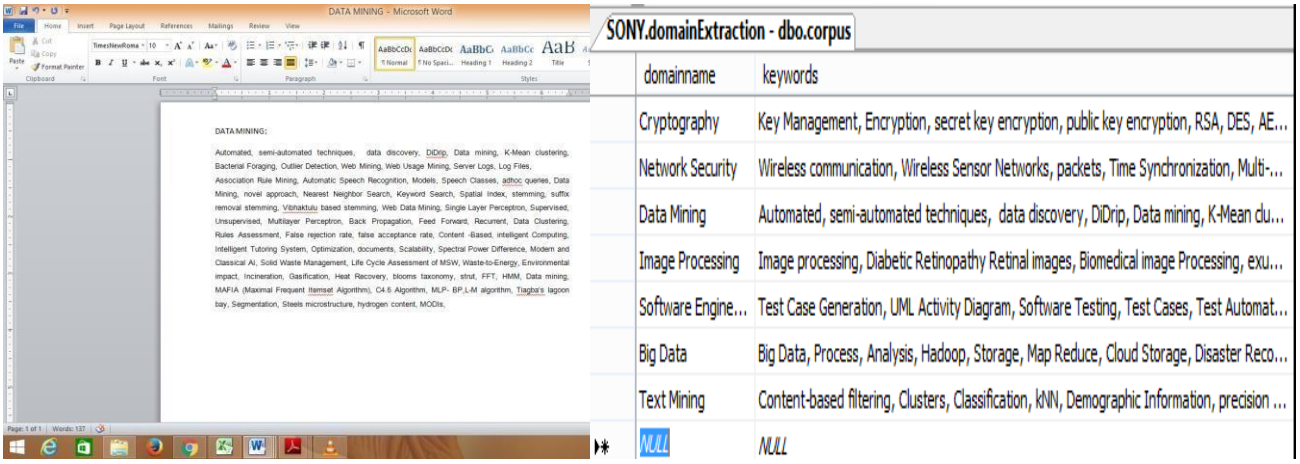
= 0.7142857142 * 10

tf-idf = 7.142857142

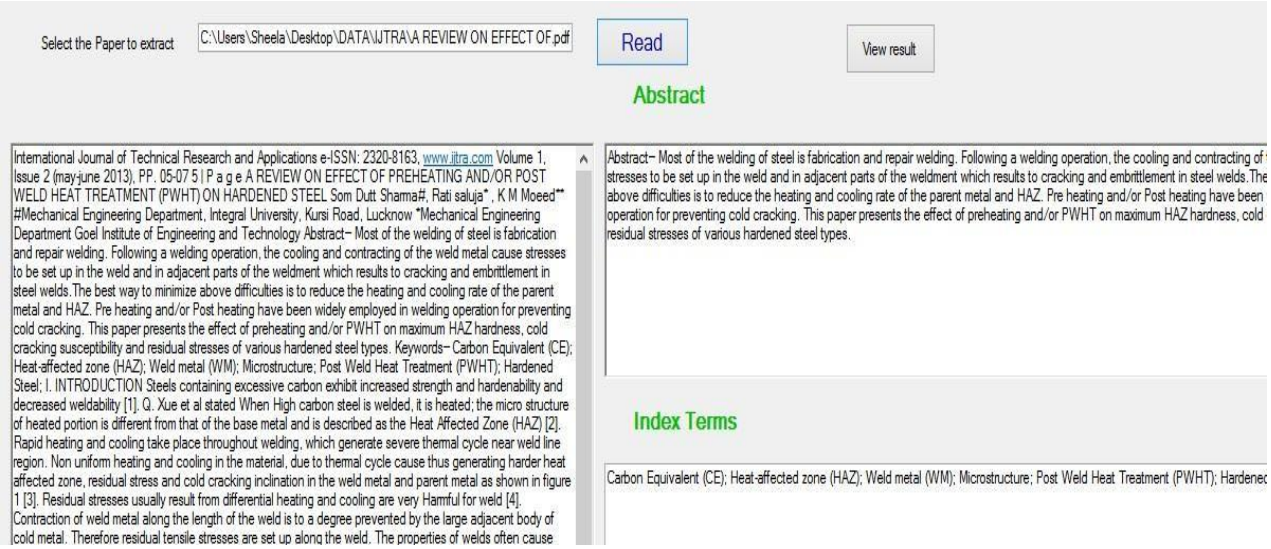
Example, Consider a document containing 100 words wherein the word cat appears 3 times. The term frequency (i.e., tf) for cat is then (3 / 100) = 0.03. Now, assume we have 10 million documents and the word cat appears in one thousand of these. Then, the inverse document frequency (i.e., idf) is calculated as log(10,000,000 / 1,000) = *4.* Thus, the Tf-idf weight is the product of these quantities: 0.03 * 4 = 0.12.

IV. EXPERIMENTAL RESULT IN DOMAIN EXTRACTION

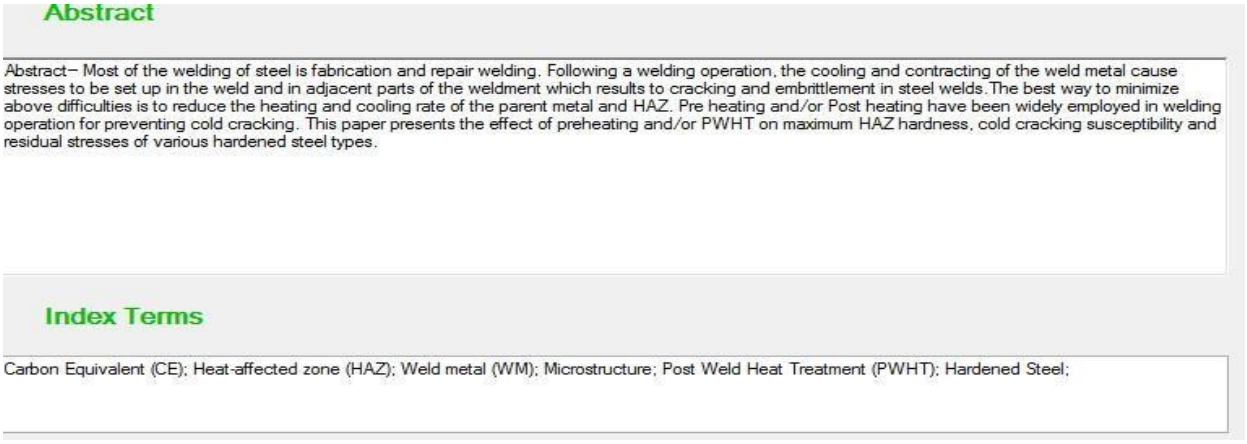
Step 1: Create Domain Corpus in Word document or notepad



Step 2: Read Research paper(pdf format) using Pdf IFilter Tool

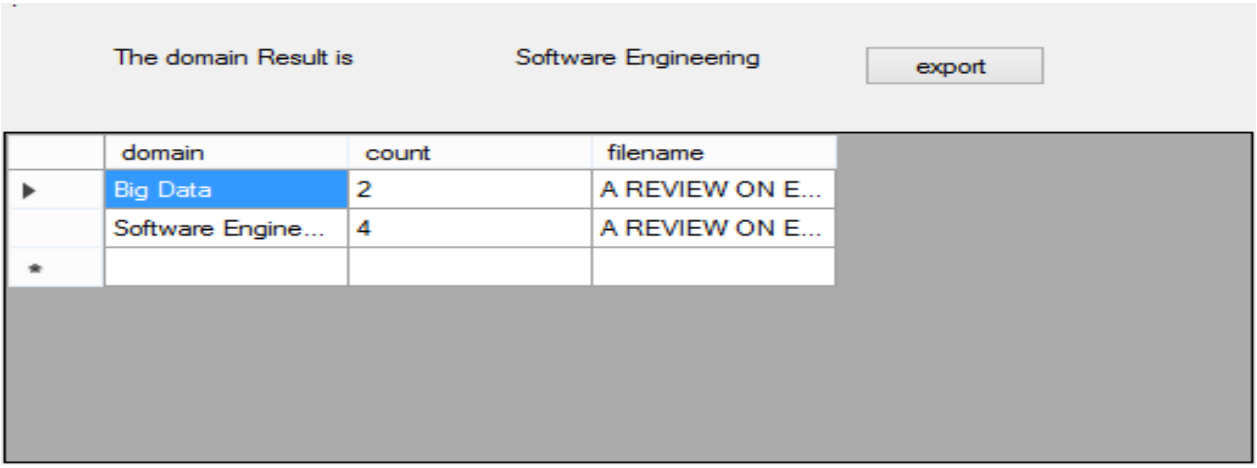


Step 3: Extract Abstract and Index Terms from Research Paper



Step 4: Then count the term frequency from the collected term and rank them according to the high term frequency (Using Term Frequency Similarity)

Step 5: Domain Extraction Result from Research Paper



Step 6: Precision, Recall and F – Measure calculation from Research Paper

id	precision	recall	fmess	filename	domainname
2823	210	10	18	07-NTCIR8-PAT...	Data Mining
2824	118	9	17	07-NTCIR8-PAT...	Text Mining
2825	650	10	19	2014 research p...	Data Mining
2826	630	10	19	2014 research p...	Text Mining
2827	203	10	18	2D Conditional R...	Text Mining
2828	155	9	18	3-ARY CLASSIFI...	Data Mining
2829	790	10	20	3-ARY CLASSIFI...	Image Processing
2830	185	9	18	3-ARY CLASSIFI...	Network Security
2831	150	9	18	3-ARY CLASSIFI...	Text Mining
2832	203	10	18	4415ijnlc05.pdf	Text Mining
2833	118	9	17	A - research pap...	Text Mining
2834	203	10	18	A -Based Paper....	Text Mining
2835	910	10	20	A CLINICAL STU...	Big Data

Step 7: TF – IDF calculation

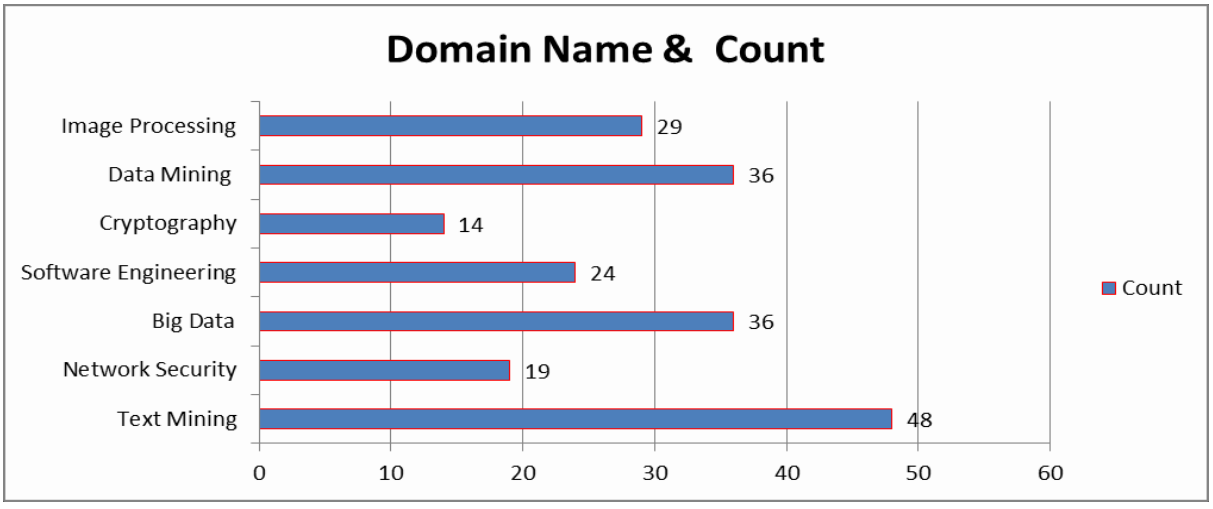
Domain name	Papers	TF	IDF	TF * IDF
Text Mining	48	0.941176471	4.166666667	3.921568627
Network Security	19	0.243589744	10.52631579	2.564102563
Big Data	36	0.631578947	5.555555556	3.508771929
Software Engineering	24	0.452830189	8.333333333	3.773584905
Cryptography	14	0.25	14.28571429	2.04
Data Mining	36	0.545454545	5.555555556	3.03030303
Image Processing	29	0.50877193	6.896551724	3.50877193

Step 8: Domain Extraction from Research paper for each paper domain name for the following image

Results Messages

	id	domain	filename
1	184	Text Mining	07-NTCIR8-PATMN-CaiY.pdf
2	162	Text Mining	2D Conditional Random Fields for Web Information Extraction....
3	203	Text Mining	4415ijnlc05.pdf
4	150	Text Mining	A - research paper.pdf
5	163	Text Mining	A -Based Paper.pdf
6	62	Big Data	A CLINICAL STUDY TO EVALUATE THE EFFICACY OF TRI...
7	1	Software Engineering	A Comparative Testing from UML Design using.pdf
8	2	Network Security	A Constructive Analysis of Time Synchronization.pdf
9	63	Image Processing	A FAST OPTIMIZED PARALLEL GRAPH.pdf
10	3	Software Engineering	A Hybrid Approach Using GA and ACO for Risk.pdf
11	4	Text Mining	A Hybrid Online Genre-based Recommender System.pdf

Step 9: Graph Representation from Research paper for each Domain



V. CONCLUSION

The abstract and index terms extracting can be used for extracting domain name from specified research paper and applied to document classification. This paper uses PDF IFilter is one of the best-known and most commonly used research paper extractions currently in use. Term similarity mining approach is the term frequency value counting from research paper. We use different performance measurements, including precision, recall and F – measures with respect to individual human annotations and a weighted measure. Finally TF –IDF measurement is calculated for each domain from overall research paper. This paper experimented with the identification of individual research paper with domain name.

REFERENCES

[1] Su Nam Kim and Lawrence Cavedon, Classifying Domain-Specific Terms Using a Dictionary, In Proceedings of Australasian Language Technology Association Workshop, 2011, 57−65.
[2] Birong Jiang, Edong Xun and Jianzhong Qi, A Domain Independent approach from Extracting Terms from Research Papers, Springer International Publishing Switzerland 2015, 155- 166.
[3] P. Velardi, M. Missikoff, R. Basili, Identification of Relevant Terms to Support the Construction of Domain Ontologies, ACL workshop on Human Language Technology and Knowledge Management, 2001, 5:1 – 5:8.
[4] Rishabh Upadhyay, Akihiro Fujii, Semantic Knowledge Extraction from Research Documents, Computer Science and Information System (Fed CSIS) 2016.
[5] Rakhi Chakraborty, Domain Keyword Extraction Technique: A new weighting method based on frequency analysis, National Conference on Advancement of Computing in Engineering Research, ACER 2013.
[6] Hospice Houngbo, Robert E. Merer, Method Mention Extraction from Scientific Research Paper, Proceeding of COLING 2012, 1211 – 1222.
[7] Sonal Gupta, Christopher D.Manning, Analyzing the dynamics of Research by Extracting key Aspects of Scientific Papers, International Joint Conference on Natural Language Processing (IJCNLP), 2011.
[8] Sonal Gupta, Christopher D.Manning, Identifying Focus, Techniques and Domain of Scientific Papers, NIPS workshop on Computational Social Science and the Wisdom of Crowds (NIPS-CSS), 2010.
[9] Gang Chen, Feng Liu, Mohammad Shojafer, Fuzzy System and Data Mining proceeding of FSDM 2015, Frontiers in Artificial Intelligence and Applications 281, IOS Press 2016, ISBN 978-1-61499-618-7.
[10] H. Wu and R. Luk and K. Wong and K. Kwok, Interpreting TF-IDF term weights as making relevance decisions, ACM Transactions on Information Systems, 2008.

BIBLIOGRAPHY OF AUTHORS

	Dr. R. Jayanthi, MCA, M.Phil., Ph.D., working as an Assistant Professor in PG & Research Department of Computer Science at Quaid-E-Millath Govt. College for Women (Autonomous), Chennai. Her areas of interests are Data Mining, Text Mining, Natural Language Processing, Information Extraction and Business Intelligence. She has published articles in more than 10 International Journals.
	S. Sheela received her M.sc. Computer science in 2015 from Queen Mary's College for Women. She is pursuing her M.Phil. Computer Science under the supervision of Dr. Mrs. R. Jayanthi, MCA., M.Phil., Ph.D., Assistant Professor in PG & Research Department of Computer Science at Quaid-EMillath government College for Women, Affiliated to university of Madras. She has presented papers in International conferences and published papers in International Journals. Her area of interest is Text Mining, Cryptography and Network Security.