A Model for Auto-Tagging of Research Papers based on Keyphrase Extraction Methods

MG Thushara, Krishnapriya MS, Sangeetha S Nair
Dept. of Computer Science and Applications
Amrita School of Engineering, Amritapuri
Amrita Vishwa Vidyapeetham
Amrita University
India
thusharamg@am.amrita.edu, krishnapriya.ms009@gmail.com, sangeethasnair01@gmail.com

***Abstract*— Tagging provides a convenient means to assign tokens of identification to research papers which facilitate recommendation, search and disposition process of research papers. This paper contributes a document centered approach for auto-tagging of research papers. The auto-tagging method mainly comprises of two processes: -classification and tag selection. The classification process involves automatic keyword extraction using Rapid Automatic Keyword Extraction (RAKE) algorithm which uses the keyword – score matrix. The generated top scored keywords are added to the train dataset dynamically, which can be used further. This add-on feature of the system is considered to be one of the main advantages of the system since adding new born phrases is time-consuming and error prone. Cosine similarity is used for classifying the research paper into corresponding domain utilizing the extracted keywords. Tag selection concentrates on the selection of proper tags for the research paper. Tagging facilitates better search facility and determines the dynamics of research areas in terms of number of publications in a domain by each author. The system generates reports for statistical analysis of research papers in each domain and expertise of each faculty in the research community.**

Keywords*— Auto-tagging; cosine similarity; automatic keyword extraction; domain classification; graph-based method

I. INTRODUCTION

The evolution and dynamics of research community can be studied utilizing the research papers published by the community. Tagging gives an approach to share and associate relevant keywords or tags within a research paper. Nowadays, there exist many online research paper recommendation systems like CiteULike, Del.icio.us [1] which allows user to share and organizes resources. These websites allow user to specify tags for each uploaded document so that the document gets mapped to the given set of tags. The problem of tag recommendation arises when user gives no tags or user specified tags are irrelevant to the document. These kind of user-centered approaches are not very effective for tag recommendation.

In tagging mechanisms, when there are many research papers manual tagging becomes error-prone and time-consuming. In this paper, a document-centered approach for auto-tagging of research paper is proposed where accurate tags are returned that are linked to research paper. In order to perform tagging, each research document is classified into respective domain. As most of the research papers are composed of more than a single domain, it is required to identify the main domain as well as the sub-domains to which every uploaded research paper belongs to. Classification has no impact on tagging process as it makes tagging process more refined and efficient. Research papers of same domain are assumed to share more common tags when compared to papers over different domains.

Most of existing tag recommendation systems [2], [8] and [12] uses term frequency-inverse document frequency (TF-IDF) as the weight for information retrieval while our system uses keyword-score matrix generated by RAKE algorithm which is more efficient. Top-scored keywords are selected as the relevant keyword of the research paper and are added to the train dataset dynamically in order to reduce manual annotation of keywords. This add on feature of the system considers to be the one of the main advantages of the system since adding new born phrases is time consuming and error-prone. Cosine similarity is used for identifying the domains and sub-domains of each research article. Graph-based model [3] is used for auto-tagging process, where relationship among documents, word and tags are represented in bi-partite graph. Each tag is ranked using novel ranking method and the recommendations are done in the descending order of their ranks.

This paper addresses another issue of identifying the domain of an evolving new-born phrase that does not exist in past or in the man-made corpus used by the system. Topic modeling is an efficient technique for resolving this issue as it can be used for analyzing massive amount of data. Latent Dirichlet Allocation (LDA) [4] is a topic model used to find the mixture of topics within an article. The terms are categorized into different topics so that domain of the terms can be determined which can be used further.

II. RELATED WORKS

This segment of the paper describes researches and work related to automatic tagging methodologies. The point of these reviews is to improve understanding about various sorts of methods used in tagging.

Two major issues regarding tagging in recommendation system is examined in [1]. This paper also demonstrated tag-based research paper recommendation method which relies on

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

set of user-created tags which acts as user profile. A framework for research paper recommendation is presented in [2]. It captures the structural and architectural elements of a research paper. Relevant information is extracted from the research document and indices are created for each document. In this work, a corpus of research papers and TF-IDF weight is used for creating the indices as it shows the importance of every word in element or corpus.

Two document-centered approaches are explored in [3] for effective tag recommendation in recommender system. This paper evaluates the performance of user-centered approach and document-centered approach. The results suggest that document centered systems are efficient than user-centered approach. This paper explains graph-based and prototype method for tag recommendation as both the methods uses novel ranking method in order to rank the tags. Topic modeling [4] considered as one of the efficient techniques for analyzing corpus containing massive amount of documents. Latent Dirichlet Allocation (LDA) with Gibbs sampling is used for retrieving semantic data from the research paper. The results demonstrated that the extracted topics capture semantic structure in the data models which can be applied to tag the article abstracts into appropriate class.

An approach based on ontology for auto-tagging is proposed in [5]. Ontology is an information model in which terms are represented as a hierarchy. It includes classification and tag-selection. Term-weight matrix and cosine-similarity is used for classification process. Tag selection process depends on ontology weight. An input of large train dataset consisting of title, abstract is used and TF-IDF is used for building term-weight matrix. In this work, tags are ranked not only in terms of frequency but in terms of similarity also. A hybrid approach is used in [6] for extracting the header information which can be used for analyzing the dynamics in research area among research communities. Data integration and validation using extracted header information resources like GROBID, Parsit are used as it is argued that any one tool alone cannot gives efficient results on all sample research articles.

Clustering is used to group words based on similarity measures. An alternative approach is proposed in [7] to identify the relationship among words in a document taken from a specific topic. K-means algorithm classifies the input based on the sparse code generated by learning dictionary. SVD is used in [7] for mapping the latent relationship among words in the document. A combination of static ranking method to improve search results is explored in [8] The main focus of the paper is to improve the performance of search results of research paper recommendation system. The performance of Hybrid ranking and Static ranking is also evaluated. Subject classification of documents using naive bayes algorithm and inter-relationship analysis is demonstrated in [9] and [10] respectively. An approach based on Support Vector Machine (SVM) is used in [11] for identifying multiple feature elimination. [13] relies on genetic algorithm for scheduling panel for academic project reviews. PageRank algorithm is used in [14] to rank senses of words in order to determine the similarity of words using semantic processing namely Word Sense Disambiguation.

III. Proposed Methodology

The main contribution of this work is to propose auto-tagging methodology. It consists of automatic keyword extraction and tagging process. Fig.1 depicts the general architecture of proposed system. The system undergoes both training and testing phases.

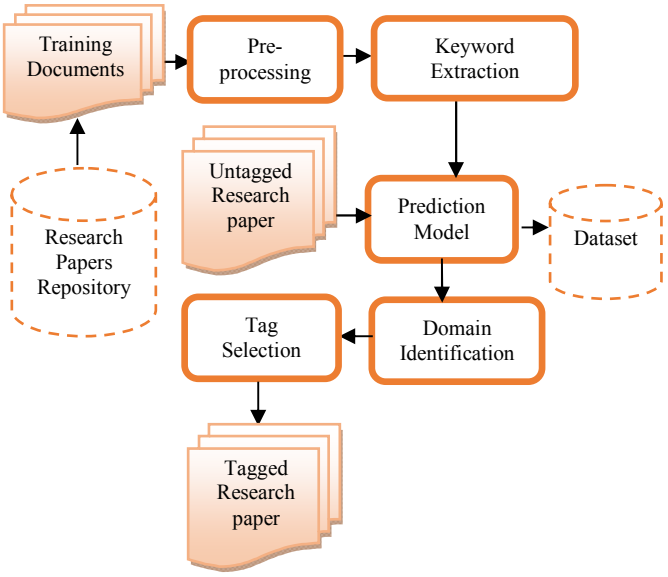

Fig. 1. System Architecture

For training, a document repository of untagged research documents with author specified tags or keywords are required. Each training document undergoes pre-processing and keyword extraction using RAKE algorithm. The author specified tags are extracted in order to analyse the capability of the system in generating relevant tags related to particular document. A prediction model is built in which each untagged research paper is classified into relevant domain and corresponding tags are generated. During the testing phase, this prediction model is applied on untagged test articles to determine the efficiency of the prediction model.

Automatic keyword extraction includes pre-processing and building of score-matrix. Pre-processing is the process of preparing the data prior to extraction. Extraction phase uses RAKE algorithm for extracting the relevant keywords from the research paper. RAKE selects keywords based on the score of each keyword. The score is calculated as the sum of the ratio of frequency and degree of keywords. Tagging method is composed of domain classification and tag selection process. Cosine similarity is used for classifying articles into relevant domains. Most of the research papers are comprised of more than single domain. Therefore finding the sub-domains is a secondary objective for the proposed system. The extracted keywords are tagged to the corresponding domains during the tag selection process. Graph-based method is used for tag selection process in which a rank is calculated for each keyword so that top-ranked keywords are selected as the relevant tag to a domain.

A. Automatic Keyword Extraction

As noted in Fig.1, prior to extraction, each document undergoes pre-processing. Obviating the requirement of having to create a man-made thesaurus, the RAKE algorithm is applied for the extraction of the keywords from document, based on score-matrix.

i) Pre-processing

The pre-processing method concentrates on preparing the unstructured textual data utilizing semantic components that serve to distinguish relevant candidate keywords. The document is split into tokens, and all stopwords, are removed in this phase. RAKE always maintains a stopword list. The words that are not in the stopword list are considered as candidate keywords, but might not be the actual keywords.

ii) Keyword Extraction

Rapid Automatic Keyword Extraction (RAKE) algorithm is used for extracting the keywords relevant to the research paper in absence of manual annotations. The benefit of using RAKE is that, it does not require man-made thesaurus for finding the relevant keywords. A score-weight is calculated for every candidate phrases. Here we try to model Score-weight matrix using degree and frequency of candidate phrases. The score-weight matrix is calculated on the basis of:

a. Word frequency: Number of times each keyword occurs in a document.
b. Word degree:- degree of co-occurrence of each word in the document.
c. Ratio of degree to frequency:- This ratio is considered as the score of candidate keywords.

The keywords with highest range of scores are considered as the candidate phrases. The score value of each keyword is continuous; therefore the scores are converted into discrete range of classes. Table I. depicts the discretization table of range of scores and classes to which the values falls.

TABLE I. DISCRETIZATION TABLE

	Discritization classes				
Class	1	2	3	4	5
Score	<3.5	3.6-7.5	7.6-11.5	11.6-15.5	>15.5

B. Topic Modelling

When the system finds it difficult to identify the domain to which the term belongs to, topic modelling is deployed to cluster the extracted terms. We adopted Latent Dirichlet Allocation (LDA) algorithm which generates topics based on word frequency from the document. LDA is particularly helpful in finding sensibly accurate mixtures of topics within a given research article.

The parameters used in LDA:

- Num of topics: user can determine number of topics to be generated.
- Id2word: Creates dictionary which map ids to strings.
- Passes: The number of iterations the model will take through corpus. The more prominent the number of passes, the more precise the model will be.

After topic modelling, we get N number of clusters consisting of set of related terms. In this stage we find the similarity of the terms in cluster with the terms in dataset of each domain. Domain to which the terms in cluster are more similar will be considered, so that all the related terms including new-born terms in the particular cluster will be dynamically mapped to that domain. This mechanism helps to identify the newly evolving terms in research documents so that the identified new-born terms can be used for further use. Therefore, the topic modelling is considered to be an update process since it updates the datasets with new terms without the requirement of manual annotations.

C. Tagging Process

The extracted keywords are fed into the tag selection process. All keywords of the document are tagged to be associated with their corresponding domain. This step helps accelerate the subsequent search process. Tagging process comprises of two stages: classification and tag selection process. Before tagging each untagged document undergoes classification process. Classification has no impact on tag recommendation yet it makes tagging process more refined on the grounds that tag terms are determined in more specific domain.

i) Domain Classification

Untagged research articles are classified into relevant domains by using cosine similarity. Computation of similarity of article is done with every train dataset and is classified into corresponding domain which is having higher similarity. Cosine similarity finds the normalized dot product of two vectors. It is computed by:

$$\text{Similarity (X, Y)} = \frac{\sum_{i=1}^{n} w_{X_i} w_{Y_i}}{\sqrt{\sum_{i=1}^{n} w_{X_i}^2 \cdot \sum_{i=1}^{n} w_{Y_i}^2}} \quad (1)$$

where X is the untagged research document and Y is the document dataset. w_{X_i} is the score of the extracted terms of document and w_{Y_i} is the score of the terms of document in dataset.

Some research papers were comprised of more than one topic or domain. Every article gets mapped to each relevant domain in the descending order of their similarity. Subsequent to classification of the papers to the pertinent domains and sub-domains, rank of every keyword of research paper in each

domain is computed. These steps help determine the influence of each term in its corresponding domain.

ii) Auto-tagging

In this work, graph-based method [3] is used for auto-tagging of research paper where each node is considered as tag and the edge between two nodes is represented as the association between the nodes in the documents. The graph-based method consists of two steps:

a. Represent the relationship among words, documents and tags in bi-partite graph.
b. Rank the tags within each class based on their scores generated by RAKE.

In the proposed framework, tagging of research papers was carried out in terms of authors and keywords. Application of interpreting bipartite graph node ranking method incorporated determination of the document-author relationship. This facilitated the definition of the social connection of authors in the same research area and selection of most illustrative research papers in the topic. The system we have proposed then extracted the author name from the research document so that the particular document was tagged to the name of author. This made possible the subsequent statistical analysis of the authors' contribution to a particular research domain.

Here graph G=(V,E,W) [3] is defined , where V indicates vertices, E the corresponding edges and W represents the weight of edges. W_{ij} denotes the score of the term *j* appears in document *i*. Fig 2 represents the relation between tags, documents and words.

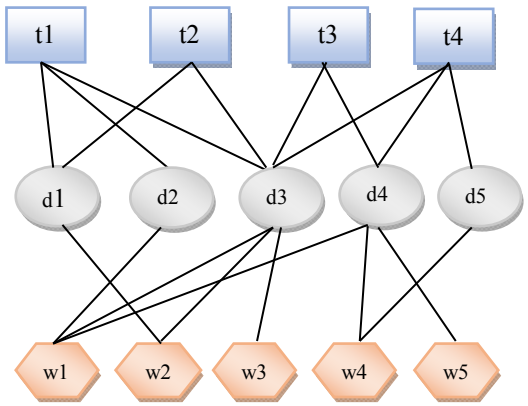

Fig.2. Bi-partite graph of tags, documents and words

During classification phase, each document was classified into predefined classes or domains. Class labels were assigned to document nodes and tag nodes. Tags were then selected in the descending order of their ranks. N-precision [2] was used to calculate the rank of each node. N-precision defines how important a node to each domain. It is the sum of the score of terms connected to associated documents in the same cluster or topic divided by total weight of terms in that cluster.

It was calculated as,

$$n_{pi} = \frac{\sum_{i=1}^{n} w_{ij} \prod[C(j)=C(i)]}{\sum_{j,k=1}^{n} w_{jk} \prod[C(j)=C(k)=C(i)]}, j,k \neq i \quad (2)$$

where, the value of $\prod[.]$ will be 1 if condition satisfies and otherwise 0. Weight w_{ij} indicates the score of word in the domain.

IV. RESULT AND ANALYSIS

A. Dataset

Our proposed system was evaluated with a heterogeneous dataset composed of 400 research documents published in four computer science research fields. The different topics covered in the dataset were Natural Language Processing (NLP), data mining, computer networks and network security, with each domain populated by 100 documents. Each of the research papers in dataset contained author specified tags. Author specified keywords were extracted and mapped to corresponding articles in dataset during the training phrase and score of each keyword was computed using RAKE algorithm.

B. Tagging Performance

Graph-based method was used for automatic tagging where each node was categorized into clusters or domains. Here the cluster names are the domain names where these names are assigned to the document nodes as their class labels and each tag is given rank in every domain. In order to determine whether the tags are relevant or not, ranked the tags. Tag selection was based on the descending order of their ranks. Top-ranked tags were used for recommendation.

Table II depicts the results after tagging of research documents, using graph based, and novel ranking method. From the table, domain1 and domain 2 depicts the main domain and sub-domain of the corresponding research papers, respectively. In the proposed work, the uploaded research document was classified into respective domains so that domain with highest cosine similarity would be the main domain; sub-domains were determined in the descending order of the similarity. Using the proposed ranking method, each keyword was ranked in associated domains. Rank 1 and rank 2 in the table depicts the ranking of the keyword in the domain 1 and domain 2 respectively. For example, the project document 1 was classified to data mining as the main domain and computer networks as the sub-domain. The keyword term *hidden markov model* had the rank 0.0834 in data mining and no rank in computer networks because the term was not associated to computer network domain. In project document 2, Network security and data mining were identified as the main domain and sub-domains respectively. Extracted term *information processing* had ranking in both the classes but it ranked higher in NLP class as it more associated to NLP domain than network security.

TABLE II. RESULT OF ARTICLE TAGGING UNDER VARIOUS RESEARCH DOMAINS

Research Document	Domain 1	Domain 2	Keyword	Rank 1	Rank 2
Document1	Data Mining	Computer Networks	hidden markov model	0.0834	0
			network monitoring	0	0.0458
			anomaly detection	0.073	0.0512
			data clustering	0.0831	0.0153
Document 2	Network Security	Data Mining	cloud networks	0.0383	0
			anomaly detection algorithms	0.0562	0.0812
			distributed system	0.0563	0
			information processing	0.0257	0.0573
Document 3	Network Security	Computer Networks	cloud network security	0.0816	0.0531
			optical networks	0.0138	0.078
			anomaly mitigation	0.0745	0.029
			intrusion detection	0.0891	0.0531

The main objective of ranking is to facilitate recommendation process. The recommendation is performed by classifying the research papers into domains and sub-domains and selecting top-ranked keywords as the tags. The query keyword given by user is compared with similar tags so that the article to which the tag has higher rank will be returned. During the testing phase, set of 200 papers are subjected for auto-tagging process where each research paper was tagged to its main-domain and sub-domains. Table III shows number of papers tagged to each research domain as main domains and sub-domains.

TABLE III. NUMBER OF TAGGED ARTICLES IN EACH DOMAIN

Number of articles / Research Domain	Correctly tagged as main domain	Correctly tagged as sub-domain	Total Related articles
Data mining	64	12	80
NLP	55	20	78
Computer Networks	39	16	59
Network Security	25	12	43

Fig.3 shows the pie chart demonstrating the influence of each domain in the research community in current year. From Fig.3, it is clear that more articles are published in data mining domain in current year.

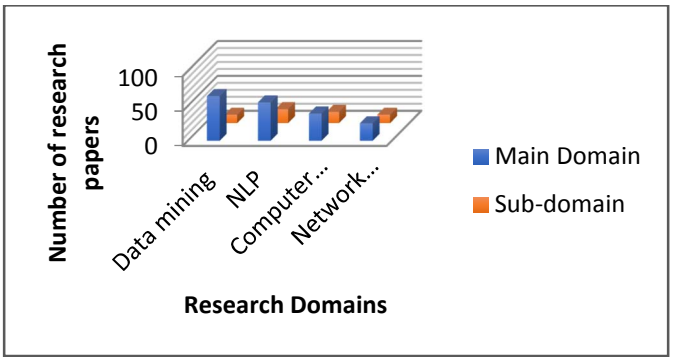

Fig.3 .Number of tagged papers in each domain

The main advantage of the graph-based model is to capture the author-document relationship. Fig.4. depicts contribution of each faculty of a research community in Natural Language Processing. This novel method of node ranking helps to determine the document-author relationship that is, influence of authors in each research domain.

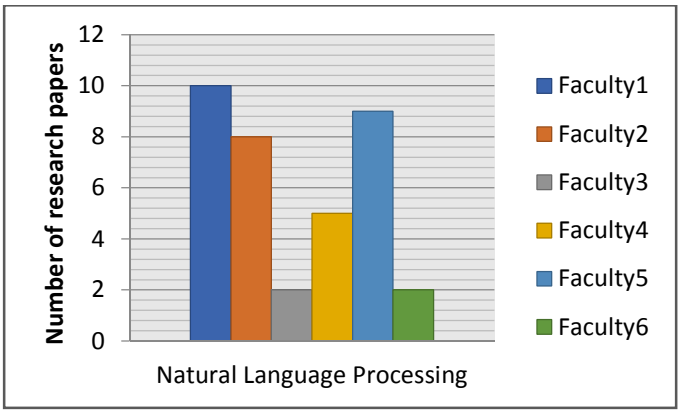

Fig.4. Faculty expertise in NLP

In order to examine the performance of the system, we need to measure the accuracy. For measuring the accuracy, we considered both the user tags and system generated tags among each. Table IV shows user tags for top 3 documents, as well as the tags generated by the system.

TABLE IV. TAGS GENERATED BY SYSTEM Vs USER GENERATED TAGS

Documents	Total number of relevant user tags	Number of system generated tags	Number of system generated relevant tags
Doc1	89	105	92
Doc2	99	116	109
Doc3	104	132	113

Accuracy is measured as,

$$Accuracy = \frac{TP + TN}{Total}$$

where TP indicates the number of correctly identified relevant tags by the system and TN is the number of correctly identified irrelevant tags. Accuracy rate thus obtained scores average of 91.23 % over changing number of generated tags.

V. CONCLUSION

Tagging provides an effective means to share and associate relevant keyword or tags within a research document. In this study, an accurate, flexible, domain-sensitive hybrid approach for auto-tagging of research articles has been proposed. In order to perform tagging, each article is classified into particular domain. Contemporaneous benefit of our approach includes classification of research papers into their relevant domains. Our scheme also generates reports showing the influence of each domain in research communities and the contribution of authors in each domain. Furthermore, the system can also help in the study of the dynamics and evolution of research domains in a specific research community. This system can be used as a tool for recommending research articles and for analyzing the evolution of research areas. Future plans of the proposed system include characterization of tagging accuracy and performance analysis with larger dimensional datasets.

ACKNOWLEDGMENT

We would like to thank the faculty of Department of Computer Science and Applications of Amrita Vishwa Vidyapeetham, Amritapuri for the strength and guidance. Our sincere gratitude to Dr. M. R. Kaimal, Chairman, Computer Science Department, Amrita Vishwa Vidyapeetham, Amritapuri for his support.

REFERENCES

[1] W. Choochaiwattana, "Usage of tagging for research paper recommendation," ICACTE 2010 - 2010 3rd Int. Conf. Adv. Comput. Theory Eng. Proc., vol. 2, pp. 439–442, 2010.

[2] P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A framework for tag-based research paper recommender system: An IR approach," 24th IEEE Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2010, pp. 103–108, 2010.

[3] Y. Song, L. Zhang, and C. L. Giles, "Automatic tag recommendation algorithms for social recommender systems," ACM Trans. Web, vol. 5, no. 1, pp. 1–31, 2011.

[4] Anupriya, P., & Karpagavalli, S. (2015). LDA based topic modeling of journal abstracts. Paper presented at the ICACCS 2015 - Proceedings of the 2nd International Conference on Advanced Computing and Communication Systems.

[5] G. Sriharee, "An ontology-based approach to auto-tagging articles," Vietnam J. Comput. Sci., vol. 2, no. 2, pp. 85–94, 2015.

[6] O. Saleem and S. Latif, "Information extraction from research papers by data integration and data validation from multiple header extraction sources," Proc. World Congr. Eng. Comput. Sci. WCECS 2012, vol. I, pp. 215–219, 2012.

[7] Menon, R. R. K., Gargi, S., & Samili, S. (2016). "Clustering of words using dictionary-learnt word representations." Paper presented at the 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, 1539-1545. doi:10.1109/ICACCI.2016.773226

[8] Jomsri, P., & Prangchumpol, D. (2015). "A hybrid model ranking search result for research paper searching on social bookmarking". Paper presented at the Proceedings of the 2015 1st International Conference on Industrial Networks and Intelligent Systems, INISCom 2015, 38-43. doi:10.4108/icst.iniscom.2015.258417

[9] M. Taheriyan, "Subject classification of research papers based on interrelationships analysis," Proc. 2011 Work. Knowl. Discov. Model. Simul. - KDMS '11, p. 39, 2011.

[10] S. Sathyadevan, P. R. Sarath, U. Athira, and V. Anjana, "Improved document classification through enhanced Naive Bayes algorithm," Proc. - 2014 Int. Conf. Data Sci. Eng. ICDSE 2014, pp. 100–104, 2014.

[11] Kavitha, K. R., Rajendran, G. S., & Varsha, J. (2016). "A correlation based SVM-recursive multiple feature elimination classifier for breast cancer disease using microarray". Paper presented at the 2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016, 2677-2683. doi:10.1109/ICACCI.2016.7732464.

[12] Thushara, M. G., & Dominic, N. (2016). "A template based checking and automated tagging algorithm for project documents". International Journal of Control Theory and Applications, 9(10), 4537-4544.

[13] Thushara, M. G., Jayaprakash, V., & Pranav Kumar, A. (2016). Panel generation as an application of genetic algorithm. International Journal of Control Theory and Applications, 9(10), 4509-4518.

[14] Veena, G., and U. B. Veni. "Improving the Accuracy of Document Similarity Approach using Word Sense Disambiguation." Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015.