ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Design and Development of Efficient Content Independent Genealogy

Rahul S. Dudhabaware[1], Mangala S. Madankar[2]

M. Tech Student, Department of Computer Science and Engineering, GHRCE, Nagpur, India[1]

Assistant Professor, Department of Computer Science and Engineering, GHRCE, Nagpur, India[2]

ABSTRACT: Whenever researchers go for a literature review for a new topic which is unfamiliar to them, they want to collect all relevant documents which influence the topic most, but while doing all these they need to collect huge number of papers, which makes them difficult to study each and every paper. To solve these problems, this paper proposed Design which focused on, how to find out relevant papers with respect to query, Discrimination of survey paper and Implementation paper, and creation of their genealogy. First it will find out all relevant matching papers with respect to keywords given in query which is preprocessed earlier, again there is a provision for a user to discriminate survey and implementation paper, and then it will create genealogy of those paper, by making association or interlinking among all matching documents on the basis of references of each paper have. This Genealogy will help user to get a quick look at, which papers are relevant for our topic of research, and association among them, so that user will focus on those documents only for literature review problem and not straying us on less important or unwanted documents.

KEYWORDS: Natural Language Processing (NLP); POS Tagging; Chunking; K-Nearest Neighbour (KNN); Genetic Algorithm (GA), Content Independent; Construction of Efficient Genealogy

I. INTRODUCTION

While doing research for a new topic, problem that researchers face is a literature survey for the topic which may know or unknown. Literature Review put a limelight on a desired topic for research, which shows, what are the hot issues are left to work on and also give us proper direction for future work. It requires spending some amount of time to extract high-quality and relevant papers from conferences, journals, or scholar search engines, for that purpose researcher can go for qualitative survey papers but because of their static nature, new invention cannot be disclosed. Researcher can browse papers from conferences or journals but, with conferences or journals, there are large number of papers need to extract and study them. Again the researcher may go for search engines like Google, CiteSeer, Google Scholar, etc. are useful to find papers with desired titles and keywords using keyword-based Query for review of research paper, but it displays large number of papers. Again sometime we don't want a survey paper as a part of our study; we don't want to include survey paper in research study. So, it is needed sometime to discriminate between Survey paper and Implementation paper before extracting papers on desired topic.

Therefore, it is necessary to find relevant papers on the research topic with relationships among them. And there should be a provision which isolates the survey paper from implementation paper. So, this paper is mainly describe to create research paper genealogy which lessen the difficulty of the literature survey at large extent and will help the researcher to easily grasp the inclination or movement of the research topic. In this paper, we specify three problems: 1) to find the relevant papers, 2) Discrimination of survey paper and Implementation paper, and 3) creation of their genealogy, from relevant papers belongs to same topic.

II. LITERATURE REVIEW

Keywords in a query need preprocessing before performing successive operations on that, so that is an important issue for survey while performing research of this paper. In [1], we can focus on so many Natural Language Processing (NLP) tasks which are useful for text preprocessing, such as Coreference Resolution [2], Discourse Analysis [3],

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Named Entity Recognition (NER) [4], Sentiment Analysis [5], Word sense disambiguation (WSD) [6], Stemming [7], Part of Speech (POS) [8], Chunking [9].

Classification techniques have been reviewed shown in [10], where it discussed about K-Nearest Neighbour (KNN), Naive Bayes, Term Graph Model which are used for text and document mining.

While doing research for Similarity measure which is an important issue in proposed design reviewed, Cosine Similarity [11] for text based measures, which is used for text clustering along with K-Clustering Method for similarity computation on the basis of terms in abstract, keyword and body of paper. Again for Link based measure, which takes help of citations in paper for similarity computation between papers, for that purpose reviewed [12]-[14].

Again there is a need to do survey of optimization techniques which enhance the performance of proposed work in the form of time context, for that purpose Neural Network techniques [15] have been reviewed different techniques such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization, Bees Algorithm.

Finding relationship between authors and papers is an important issue while doing research, for that purpose reviewed, Time-constraint Probabilistic Factor Graph (TPFG) model [16], used to expose hidden semantic knowledge in information networks which lead us to look at time-constrained advisor–advisee relationships. Another is [17] shows supportive measures for co-authorship network.

III. PROPOSED DESIGN

Proposed system is designed mainly to focus on problem of literature review for a new topic, this design shown in fig.1, proposed system consist of five modules which are described in this section. First module is Document collection, second is Pre-processing of Keywords in a query, third is Finding matching documents, fourth is Discriminating Survey paper from that Implementation paper, and fifth is Construction of Genealogy among research paper.

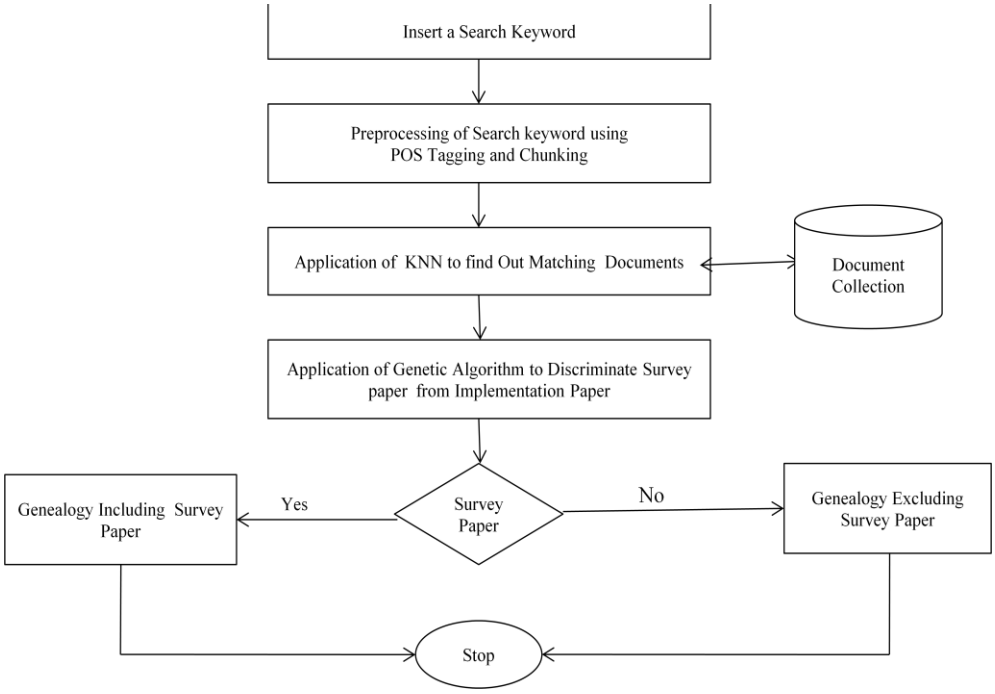

Fig. 1. Flowchart of Proposed Design

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

A. *Document Collection in the Database*

Here it is needed to create sample database, which consist number of documents, these documents are in form of pdf files. These documents are research papers which are not from same subject domain but are belongs to different subject domain. These research papers are includes both Survey papers as well as Implementation papers. There are some keywords in title that distinguishes survey paper from that Implementation paper and will be used for subsequent operation.

B. *Preprocessing of Search Keywords*

Text Preprocessing is an important part for preprocessing of input keywords for asking query, so that it will extract only important and relevant word from input keywords which will lead us to focus on extraction of only important documents and which are mostly closer to our query from the database on the basis of preprocessed input keywords, otherwise without preprocessing of input keyword, it will results huge number of documents, which is impossible for us to go through each and every documents. But preprocessing of search keyword in a query is helpful by not straying into wrong direction, by collecting lesser important or unwanted documents. Natural Language Processing (NLP) has good contribution in text preprocessing. After surveying so many NLP tasks [1], it has been discovered that POS tagging and Chunking will be the best option for text preprocessing. In that also application of POS tagging has to be made first on search keyword and subsequently Chunking has to be executed so that will give efficient preprocessing. POS Tagging will assign syntactic role to each keyword in a query by assigning them part of speech. Sometime group of words are taken as a single search keyword for matching documents, for that purpose Chunking has to use after application of POS tagging. Chunking will recognize the dependency of one keyword on another and according to that it will group the words.

E.g. Enter text for preprocessing: *The sheep is in the pen*.

After preprocessing of text using POS Tagging it will show tagging to each word: *Sheep-n, is-v, in-n, pen-n*

Here noun and verb are abbreviation for ‘n’ and ‘v’ respectively which can be set by tagset for tagging using tagging approaches.

Preprocessed Text after chunking will be: *sheep is in pen*

Here by using chunking, important and relevant word are extracted as efficient search keyword through which we can get exact and desired document in search result.

C. *Finding Matching Documents*

As K-Nearest Neighbor [10] is more preferable as compared to the Naive Bayes [10] and Term-Graph [10] but it has higher time complexity is high but gives a higher accuracy than others. The application of k-Nearest Neighbors (KNN) has been used to determine the categories of documents on the basis of query. To find out the category of a query is depends on both the documents which are nearest to it as well as categories of the K documents which are nearest to it. Vector space model [18] is used for document similarity calculation using a vector-based, distance-weighted matching function which is an instance of KNN method. The user first submits a query which is executed over the database. Then it will check out database of document collection and will results the matching document with respect to query. For each extracted document there will be Document Vector and for each class there will be Class Vector, which has to be created. After that, similarity will be calculated between Document vector and Class Vector, and then document belongs to that class for which it has maximum similarity.

Here we need to create two classes Survey paper and Implementation paper, according to that extracted document with respect to query which were discovered while matching will get classified. So, KNN has been used for two purposes, first is to match documents with respect to query and second is to classify the extracted documents with respected to query into two classes, survey paper and implementation paper.

D. *Discrimination of Survey paper from that of Implementation paper*

Genetic algorithm (GA) [19]-[21] is a powerful search mechanism which is discovered from evolution as well as natural selection and it is a randomized searching and global optimization technique, has been used in the proposed work to optimize the work of KNN. Basic operations of GA which are used in proposed system is Selection, Population, and Crossover for relevant document extraction. GA is appropriate for the information retrieval which

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

denotes relevant documents with respect to query from large document space. Document space denotes large dimensional space in proposed work.

KNN has some following drawback,

1. If there is a huge number of samples in the database, KNN will take higher time to calculate the similarities between the documents, which leads to higher calculation complexity.
2. KNN classifier is solely depends on training set and no other additional data, so if there is any changes in training set, it needs to recalculate all things again.
3. As there is no change in samples with large data and samples with small data it treated all samples equally and it will not clarify where exactly samples are distributed uncommonly.

To remove all these drawbacks, GA is planned to use with KNN, which in turn GA will optimize the KNN, by removing all above mentioned drawback. GA is comprised of following components:

A. Initial input given to algorithm are chromosomes, search keyword which is asking in query and all the documents in database are represented in the form of chromosome.
B. Similarity value is need to calculate the similarity between keywords in query to each of the documents in database, for that purpose Fitness Function is used.
C. Selection process is then select chromosome for next level or next generation in GA, which are having with highest fitness value, it may select chromosome with lower fitness value in a few or not at all.
D. Crossover is an important operation of GA, that transfer the information between two parents and exchange pair of genes with each other to generate two child chromosome.

By optimizing KNN, GA discriminate the paper into categories of Survey and Implementation in efficient way.

E. *Construction of Research paper Genealogy*

After getting all relevant matching documents, it moves for construction of genealogy among all matching documents. There is a provision for user to choose paper type. Here we have two paper types one is Survey and another is Implementation, It's on user which type of paper they want to include in their research paper survey and according to that Genealogy will be created. Now what does mean of Genealogy, in Genealogy it will show interlinking between all documents, which we get after extracting all matching documents in previous step. Whatever document we get, there may be documents which are referring other documents and all those documents also present in the database regarding to the same topic, then it will show interlinking among those documents and that is the Genealogy. This Genealogy is created with the help of Semantic matching [22]. This Genealogy will help user to get a quick look at which papers are relevant for our topic of research, and association among them, so that user will focus on those documents only and not straying us on less important or unwanted documents.

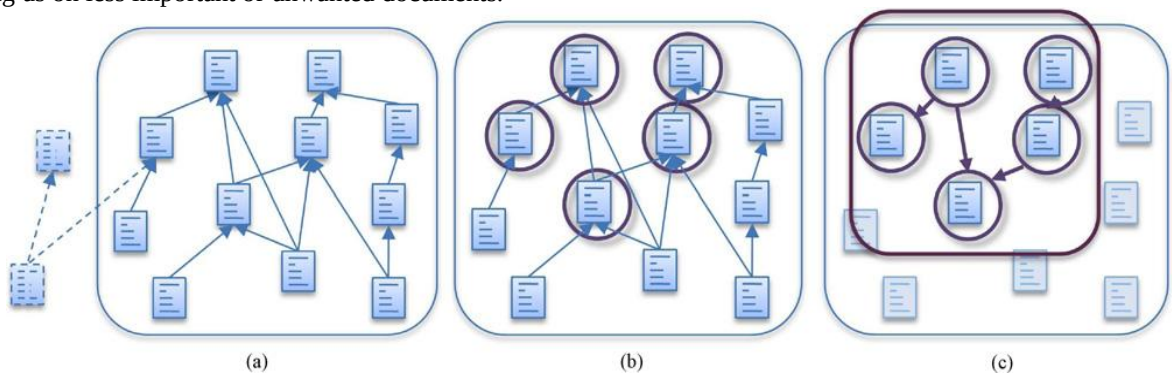

Fig. 2. Overview of relevant paper Genealogy Construction. (a) Extracting matching papers belonging to the same topic. (b) Finding relevant papers. (c) Constructing relevant paper genealogy.

IV. CONCLUSION

In this paper, proposed work is about collection of relevant matching research papers in accordance with preprocessed query related to any type of content, along with the provision of selection of paper type for user, whether to select Survey paper in literature review or not and then interlinking among all matching research paper to prepare an efficient content independent genealogy which will help the user to make quick look at which papers are relevant for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

our topic of research, so that user will focus on those documents only for literature review and not straying us on less important or unwanted documents.

REFERENCES

[1] Rahul S. Dudhabaware, Mangala S. Madankar, "Review on Natural Language Processing Tasks for Text Documents", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2014.
[2] Marcos Garcia and Pablo Gamallo, "Entity-Centric Coreference Resolution of Person Entities for Open Information Extraction", Procesamiento Del Lenguaje Natural, 53, pp. 25-32, 2014.
[3] Shun Shiramtsu, Tadachika Ozono, Toromatsu Shintani, Hirosh G. Okuno, "A Corpus-based Analysis of coreferential Recency effect in Japanese Discourse for Tracking Dynamic Topics", 9th IEEE/ACIS International Conference on Computer and Information Science, pp. 645-650, 2010.
[4] Fang Luo, Pei Fang, Qizhi Qiu, Han Xiao, "Features Induction for Product Named Entity Recognition with CRFs", Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, pp. 491-496, 2012.
[5] Soujanya Poria, Erik Cambria, Grégoire Winterstein, Guang-Bin Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment Analysis", Knowledge-Based Systems, vol. 69, pp. 45-63, 2014.
[6] Juan Wen, Ying Qin, Xiaojie Wang, "Chinese Verb Sense Disambiguation Using AdaBoosting", International Conference on Natural Language Processing and Knowledge Engineering, pp. 316- 321, 2007.
[7] Leah S. Larkey, Lisa Ballesteros, Margaret E. Connell, "Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis", Arabic Computational Morphology Text, Speech and Language Technology, vol. 38, pp. 275-282, 2007.
[8] Rana Forsati, Mehrnoush Shamsfard, Pouyan Mojtahedpour, "An Efficient Meta Heuristic Algorithm for POS-Tagging", Fifth International Multi-conference on Computing in the Global Information Technology, pp. 93-98, 2010.
[9] Guang-Lu Sun, Chang-Ning Huang,Xiao-Long Wang, and Zhi-Ming Xu, "Chinese Chunking Based on Maximum Entropy Markov Models", Computational Linguistics and Chinese Language Processing Vol. 11, no. 2, pp. 115-136, June 2006.
[10] Vishwanath Bijalwan, Pinki Kumari, Jordan Pascual and Vijay Bhaskar Semwal, "Machine learning approach for text and document mining", Cornell University Library, 6 June, 2014.
[11] S. Yuan and J. Sun, "Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management", IEEE Trans. Syst., Man Cybern., B Cybern., vol. 35, no. 5, pp. 1028–1040, Oct. 2005.
[12] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity", in Proc. Int. Conf. Special Interest Group Knowledge Discov. Data, pp. 538–543, Jul. 2002.
[13] S. Yoon, S. Kim, and S. Park, "A link-based similarity measure for scientific literature", in Proc. Int. Conf. World Wide Web, pp. 1213–1214, Apr. 2010.
[14] P. Zhao, J. Han, and Y. Sun, "P-Rank: A comprehensive structural similarity measure over information networks", in Proc. ACM Int. Conf. Inform. Knowledge Manage, pp. 553–562, Nov. 2009.
[15] R.Jensi and Dr.G.Wiselin Jiji, "A Survey on Optimization Approaches to Text Documents Clustering", International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.6, December 2013.
[16] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, "Mining advisor-advisee relationships from research publication networks", in Proc. ACM Int. Conf. Knowledge Discov. Data Mining, pp. 203–212, Jul. 2010.
[17] Y. Han, B. Zhou, J. Pei, and Y. Jia, "Understanding importance of collaborations in co-authorship networks: A supportiveness analysis approach", in *Proc. SIAM Conf. Data Mining*, pp. 1111– 1122, Apr. 2010.
[18] Manish Sharma, Mr. Rahul Patel, "Applying Genetic Algorithm in Text to Matrix Generator", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (1), pp.32-34, 2014.
[19] Manoj Chahal, Jaswinder Singh, "Effective Information Retrieval Using Similarity Function: Horng and Yeh Coefficient", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, pp. 401-406, August 2013.
[20] N. Suguna, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", International Journal of Computer Science Issues (IJCSI), Vol. 7, Issue 4, No 2, July 2010.
[21] Anubha Jain, Swati V. Chande, Preeti Tiwari, "Relevance of Genetic Algorithm Strategies in Query Optimization in Information Retrieval", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (4), pp. 5921-592, 2014.
[22] Hang Li, Jun Xu, "Semantic Matching in Search", Foundations and Trends in Information Retrieval, Vol. 7, No. 5, pp. 343–469, 2013.

BIOGRAPHY

Rahul S. Dudhabaware received the B.E degree in Information Technology from the Nagpur University, India, in 2012 and currently pursuing M.Tech degree in Computer Science & Engineering from G.H.Raisoni College of Engineering, Nagpur, India.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Mangala S. Madankar received the B.E degree in Computer Engineering from the Nagpur University, India, in 2004 and the M.E. degree (Distinction with CGPA 9.14) in Wireless Communication and Computing, in 2012. She worked with industry for 3 years 2005 to 2008 focused on php/mysql. She joined G.H.Raisoni College of Engineering, Nagpur, India, as Lecturer in 2008 and became an Assistant Professor in 2012 and currently working**.** Her area of specialization are Wireless Communication, Android, Mobile Computing, Theory of Computation, Security, NLP.