A STUDY ON SNA : TEXT MINING USING ACADEMIC SOCIAL NETWORKS

G.Ayyappan[1], Dr.C.Nalini[2], Dr.A.Kumaravel[3]
Research Scholar, Department of Computer Science and Engineering, Bharath University, Chennai[1]
Professor, Department of Computer Science and Engineering, Bharath University, Chennai [2]
Professor, Department of Information Technology, Bharath University, Chennai[3]
ayyappangmca@gmail.com[1], drnalinichidambaram@gmail.com[2],drkumaravel@gmail.com[3]

ABSTRACT - Social network are sharing the knowledge from one to others. Now a day's using social network is vast communications of people together. This is very useful in current era. Here in this paper mainly focuses on sharing the knowledge in research community. We have taken 2092356 research article and 80242869 citations among the researchers from various domains. This paper mainly focuses on the knowledge diffusion in research community. This knowledge diffusion not only homogeneous system but also heterogeneous system. Here measure the strength of research spectrum , authors contribution.

Keywords : SNA, meta, rules, misc, **t**rees, bayes

I. INTRODUCTION

Extraction and mining of academic social networks aims at providing comprehensive services in the scientific research field. In an academic social network, people are not only interested in searching for different types of information (such as authors, conferences,and papers), but are also interested in finding semantics-based information (such as structured researcher profiles).

Many issues in academic social networks have been investigated and several systems have been developed (e.g., DBLP, CiteSeer,and Google Scholar). However, the issues were usually studied separately and the methods proposed are not sufficient for mining the entire academic network. Two reasons are as follows: 1) Lack of semantics-based information. The social information obtained from user-entered profiles or by extraction using heuristics is some times incomplete or inconsistent; 2) Lack of a unified approach to efficiently model the academic network. Previously, different types of information in the academic network were modeled individually, thus dependencies between them cannot be captured accurately.

Heterogeneous networks are becoming prevalent in many real-world applications. For example in a heterogeneous academic network, there are different objects such as authors, conferences, and papers; in a product review system, there are objects like products, users, and reviews. The emerging complex networking data poses many fundamental challenges for search and mining of them. Here we have taken nominal attributes for classification of "Text Mining Using Academic Social Networks".

Traditional keyword-based search essentially matches the queried keywords with documents, and object-oriented search extracts attributes for specific objects (e.g., product) and performs search at the object level. In contrast with the traditional search, our topic level search tries to extract the "semantics" of documents/queries and match them at the topic level. The fundamental issue in the topic level search is how to model the object/document with semantic topics. The problem becomes more challenging with the prevalence of heterogeneous networks, which usually consist of different types of objects. In addition, we need also consider how to employ the learned latent topical information to help the search and mining tasks.

In Figure 1. displays the architecture of the flow of text mining classification process. We have collected this data set from ArnetMiner (http://www.arnetminer.org). In this massive real time dataset we have taken for text mining process only 15000 training set randomly. We have divided 15000 records into 3 different folders like as large, medium, small .each and every folder it contains 5000 individual text files records.

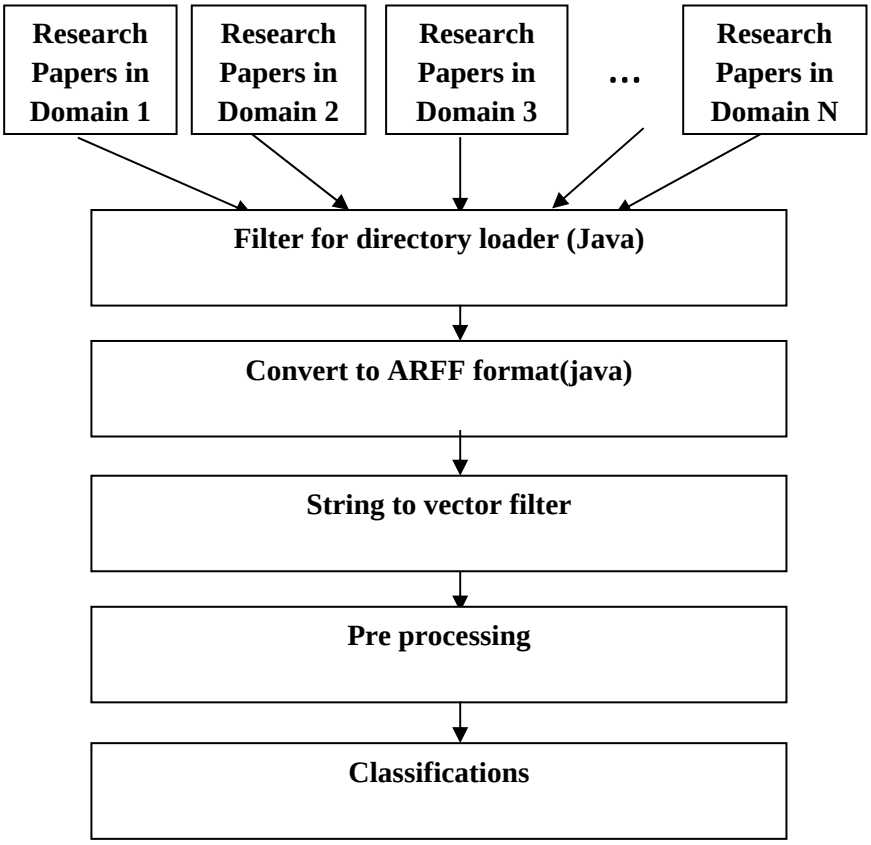

Fig:1 Layout of Text mining process

II. LITERATURE SURVEY

For academic search, several research issues have been intensively investigated, for example expert finding and association search. Expert finding is one of the most important issues for mining social networks. For example, both Nie et al. and Balog et al.[4] propose extended language models to address the expert finding problem. From 2005, Text REtrieval Conference (TREC) has provided a platform with the Enterprise Search Track for researchers to empirically assess their methods for expert finding . Association search aims at finding connections between people.

For example, the ReferralWeb system helps people search and explore social networks on the Web. Adamic and Adar have investigated the problem of association search in email networks. However, existing work mainly focuses on how to find connections between people and ignores how to rank the found associations.

In addition, a few systems have been developed for academic search such as, scholar.google.com, libra.msra.cn, citeseer.ist.psu,and Rexa.info. Though much work has been performed, to the best of our knowledge, the issues we focus on in this work (i.e., profile extraction, name disambiguation, and academic network modeling)have not been sufficiently investigated. Our system addresses all these problems holistically.

III. MATERIALS AND METHODS

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

In particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed

number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.[4] Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

In this paper we are going to experiment and find out the algorithm to find out the best accuracy in the several classification methods. Here main part is find out the best classification methods in our dataset. Now we will apply the different types of classifications like as bayes, meta, misc, rules and trees .

IV. EXPERIMENTS AND RESULTS

Here we are going to apply the algorithms which are suitable for the dataset to find out the best algorithm for text mining of academic social network dataset. Whatever we collected three different folders large medium and small . we were dividing in this folder we follow 2 steps.

1. Using random selection process we collected 15000 records from individual files 2092356 records contain dataset. Using this step randomly we select records using rand() function in excel. Then only we move to next step of forming three different folders.
2. We applied text pre processing algorithm for dividing 3 different folders. i.e., large, medium, small. Here in this step each and every folder contains 5000 individual text files. These text files are called individual records.

pseudocode :text preprocess mining

```
/* Conditions for inserting each records in each directory of folder
#index is the starting field of each record set
#% 0<size (number of % in each record) <5   then  small folder
#% 5<size(number of % in each record)<10 then medium folder
#%10< size(number of % in each record) then large folder  */

fileopen(ai)
for(#%)
size =size+1
size
        if (0<size<5) then store (ai) in small directory
        if (5<size<10) then store (ai) in medium directory
        else
                size store(ai) in large directory
Here, ai- each records
```

In this research work we classify several methods BayesNet, NaïveBayes, Attribute SelectedClassifiers, Dagging, DecisionStump, JRip, ZeroR, J48, HyperPipes, ComplimentNaiveBayes. Based on these classifications we find out the maximum accuracy of the result recommendation of the research work in our further research works in this dataset.

Table 1: Classification Methods

S.No	Methods	Accuracy
1	BayesNet	78.84%
2	NaiveBayes	74.46%
3	Attribute Selected Classifiers	77.63%
4	Dagging	87.20%
5	DecisionStump	62.98%
6	JRip	89.34%
7	ZeroR	33.33%
8	J48	98.32%
9	HyperPipes	66.75%
10	ComplimentNaiveBayes	78.69%

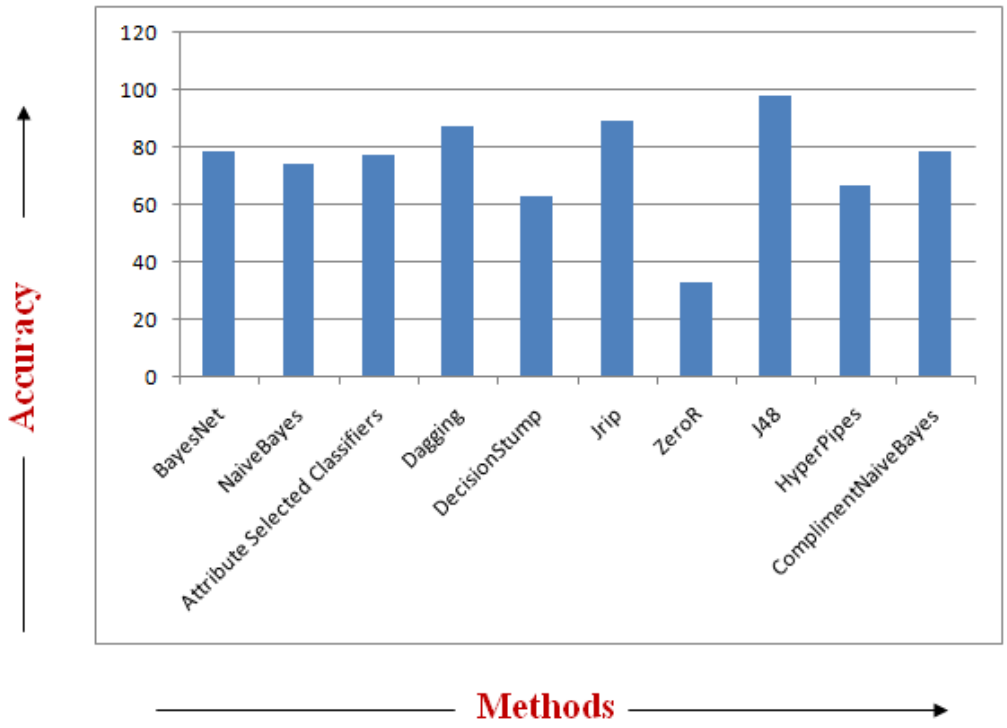

Fig1: Graphical representation between several classification methods and accuracy

After the applications of all the bayes, misc, meta, rules and meta functions now we are going to recommend or finalize the appropriate algorithm for the data set. That we can decide using the graphical representation of the tables.

V. CONCLUSION

In this paper mainly focuses on text mining process of Academic social networks. We classify J48 is the best classification method compare than other classifiers. In this research work J48 classification methods shows the maximum accuracy for the academic social network dataset. This can be extended to other datasets of different domains. Moreover one can be extend with other classifiers.

REFERENCES

[1] Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.
[2] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.
[3] L. A. Adamic and E. Adar. How to search a social network.Social Networks, 27:187–203, 2005.
[4] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A Combination Approach to Web User Profiling. ACM Transactions on Knowledge Discovery from Data (TKDD), (vol. 5 no. 1), Article 2 (December 2010), 44 pages.
[5] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning.Machine Learning,50:5–43, 2003.
[6] Jie Tang, A.C.M. Fong, Bo Wang, and Jing Zhang. A Unified Probabilistic Framework for Name Disambiguation in Digital Library. IEEE Transaction on Knowledge and Data Engineering (TKDE) , Volume 24, Issue 6, 2012, Pages 975-987.
[7] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In Proc. of SIGIR'06, pages 43–55,2006.
[8] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In Proc. of KDD'04, pages 59–68,2004.
[9] Jie Tang, Jing Zhang, Ruoming Jin, Zi Yang, Keke Cai, Li Zhang, and Zhong Su. Topic Level Expertise Search over Heterogeneous Networks. Machine Learning Journal, Volume 82, Issue 2 (2011), Pages 211-237.
[10] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In Proc. of WWW'05, pages 463–470,2005.
[11] D. M. Blei and J. D. McAuliffe. Supervised topic models. In Proc. of NIPS'07 , 2007.
[12] Jie Tang, Duo Zhang, and Limin Yao. Social Network Extraction of Academic Researchers. In Proceedings of 2007 IEEE International Conference on Data Mining(ICDM'2007). pp. 292-301.
[13] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In TREC'05, pages 199–205, 2005.
[14] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering.Communications of the ACM,40(3):63–65, 1997.
[15] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In Proc. of SIGIR'06, pages 43–55,2006.
[16] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma. Web object retrieval.In Proc. of WWW'07, pages 81–90, 2007.