Ontology Based Search Engine

K.Suriya Prakash / P.Saravana kumar
Lecturer / HOD / Assistant Professor
Hindustan Institute of Engineering Technology Polytechnic College, Padappai, Chennai, TamilNadu, India
PS Muthu college of Arts and Science,Theni,TamilNadu,India

S. Ganesh Kumar M.E, (P.hd)
Assistant Professor (Sr.G)
Department of Computer Science and Engineering
SRM University, Chennai, TamilNadu, India

Abstract: This paper presents a novel ontology-based text-mining approach to cluster search proposals deep web search based on their similarities. The method is efficient and effective for clustering search proposals with English texts. Text-mining methods have been proposed to solve the problem by automatically classifying text documents. Current search methods for grouping proposals are based on manual matching of similar research discipline areas and/or keywords.

The advantages of this method are that it can extract three types of data records, namely, single-section data records, multiple-section data records, and loosely structured data records, and it also provides options for aligning iterative and disjunctive data items.

Keywords – Text classification, Web Usage Mining, Wordnet tool

I. INTRODUCTION

The paper deals with an overview of web page to cluster the text documents over the web page based on the user typed key term. To enhance deep web search (ontology) and overcome grouping of unrelated documents into the same cluster.

II.WHAT IS ONTOLOGY?

- **Ontology** is a specification of a conceptualization
- **Ontology** is the study of being alive and existing.

III. WHAT IS TEXT MINING?

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Text mining is required if organizations and individuals are to make sense of these vast information and data resources and leverage value. The resources need first to be processed – accessed, analyzed, annotated and related to existing information and understanding. The processed data can then be 'mined' to identify patterns and extract valuable information and new knowledge.

We originally focused on two key areas:

- Where text mining could potentially generate cost savings (and productivity gains)
- Where text mining use in UKFHE could potentially generate wider impact on the economy, for example by leading to wider innovation in products or services

IV. TOOL USED FOR ONTOLOGY

Wordnet : WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The main relation among words in WordNet is synonymy, as between the words shut and close or car and automobile. Synonyms--words that denote the same concept and are interchangeable in many contexts--are grouped into unordered sets (synsets).

V. PROPOSED SYSTEM & ONTOLOGY CLUSTERING

Our proposed text clustering has a frequent concept to cluster the text documents. Our proposed text clustering has a frequent concept to cluster the text documents. Our proposed text clustering has a frequent concept to cluster the text

documents. Our proposed text clustering has a frequent concept to cluster the text documents. Text-mining methods have been proposed to solve the problem by automatically classifying text documents.

*** Nym's Group :**

Words ending in **nym's** are often used to describe different classes of words, and the relationships between words.

- Hypernym: A word that has a more general meaning than another.
- Hyponym: A word that has a more specific meaning than another.
- Synonym : One of two (or more) words that have the same (or very similar)

*** Text Analysis:**

The Artificial-Intelligence literature contains many definitions of ontology (Wordnet).

- It includes machine-interpretable definitions of basic concepts in the domain and relations among them.

The featured results produced by the sentence-based, document-based, corpus-based, and the combined approach concept analysis have higher quality than those produced by a single-term analysis similarity.

VI. WHAT IS WEB MINING?

Web mining is the application of data mining techniques to extract knowledge from Web data.
Web data is

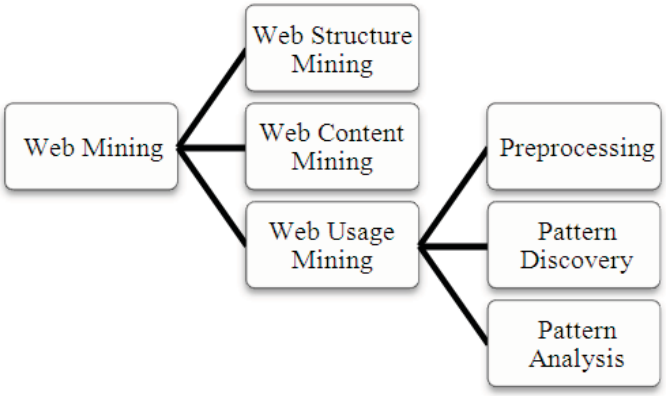
Web content – text, image, records, etc.
Web structure – hyperlinks, tags, etc.
Web usage – http logs, app server logs, etc.

Web mining - is the application of data mining techniques to discover patterns from the Web. web mining can be divided into three different types, which are **Web usage mining**, **Web content mining** and **Web structure mining**.

A. Web usage mining –

A Web is a collection of inter-related files on one or more Web servers. Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site some users might be looking at only textual data.

Figure 1. Web Mining Classification



B. Web Structure Mining –

Web structure mining is the process to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: A structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

C. Web Content mining –

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. Web content mining is differentiated from two different points of view. Information Retrieval

View and Database View. Web mining is an important component of content pipeline for web portals. It is used in data confirmation and validity verification, data integrity

VII. SCOPE

Text clustering is mainly used for a document clustering system which clusters the set of documents based on the user typed key term.Text-mining methods have been proposed to solve the problem by automatically classifying text documents.

VIII. METHODOLOGY

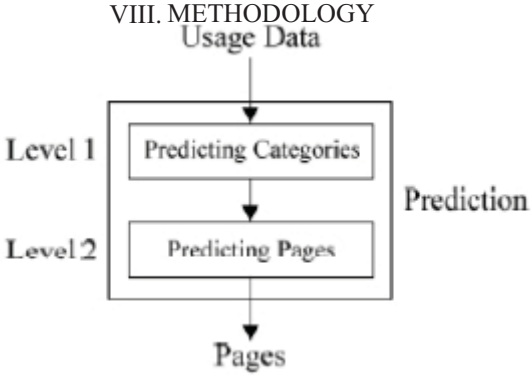


Fig. 1. Two Levels of Prediction Model

The Two Levels of Prediction Model are created by merging the Markov model and Bayesian theorem. In first level, Markov model is utilized for the purpose of filtering the highly probable of categories that will be surfed by user. In the second level, Bayesian theorem is utilized to assume precisely the maximum probability of web page.

In first level, it is to forecast the highly probable user'spresent state (web page) of group at time t, that depends on user's category at time t-1 and time t-2. Bayesian theorem is utilized to forecast the highly probable web pages at a time tbased on user's states at a time t-1. At last, the prediction result of two levels of prediction model is provided. The Two Levels of Prediction Model framework is provided in figure 2. In the first step, the similarity matrix S of category is created. The technique of creating similarity matrix is to collect statistics and to examine the users' behavior browsing that can be obtained from web log data. In the second step, it is to create the first-order transition matrix P and second-order transition matrix P2 of Markov model. The transition matrix of Markov is created by the similar technique, statistical method, from web log file.In the final step, the relevance matrix R is created from

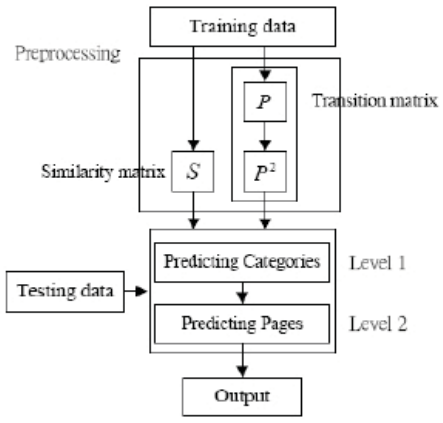
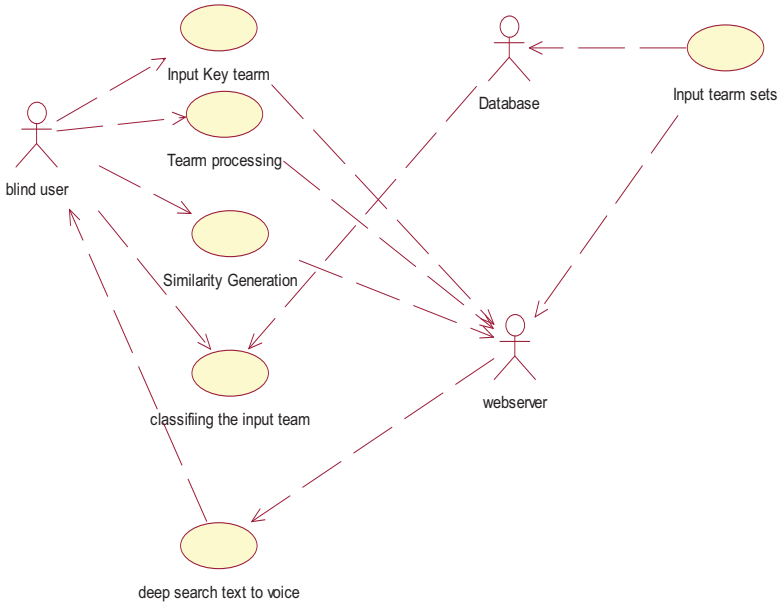


Fig. 2. Two Levels of Prediction Model Framework

first-order and second-order transition matrix of Markov model and similarity matrix. When the value of relevance is higher then the value of transition probability and similarity among categories is also higher. The relevance matrix is an significant feature of forecasting. Relevance matrix can be utilized to assume the users' browsing pattern among web categories.



Basically the web usage mining classifies the process into three categories.

1. DATA PREPROCESSING PHASE

The data pre-processing phase was performed using a reduced log file, which was "cleaned" by removing all useless, irregular, and missing data from the original common log file. After the initial pre-processing, a session filter was applied to the reduced log file for feature extractions. The purpose of the filter was to aggregate all user requests within a session into a single set of variables.

i) Data Collecting
ii) Data Cleaning
    a) Missing Data
    b) Noisy Data
iii) Data Integration
    a) Aggregation
    b) Normalization
iv) Data Reduction
    a) Data Cube
    b) Numerosity
    c) Data Discretization

2. PATTERN DISCOVERY PHASE

Statistical methods as well as data mining methods (path analysis, Association rule, Sequential patterns, and cluster and classification rules) are applied in order to detect interesting patterns. The data mining phase included two sub-phases: (a) descriptive analysis, and (b) artificial intelligence analysis. Descriptive analysis was used with summarizing, clustering, and association rules techniques to generate an overview on the dataset, to gain an insight into Users' characteristics, and to depict Users' browsing patterns. Artificial intelligence analysis was used for predictive purposes.

i) Statistical & Path Analysis

ii) Association Rules
iii) Sequential Patterns
iv) Clustering & Classification Rules
v) Boosting
vi) Bagging

3. PATTERN ANALYSIS PHASE

The pattern analysis phase included data interpretation and evaluation of the results. This phase was needed to identify meaningful results from outcomes of the data mining phase

i) Pattern Filtering
ii) Aggregation
iii) OLAP Analysis
iv) Rules, Patterns,& Statistics
v) Visualization Techniques
vi) Evaluation & Report Generation (Meta Data View, Data View, Plot View, 3D…Etc)

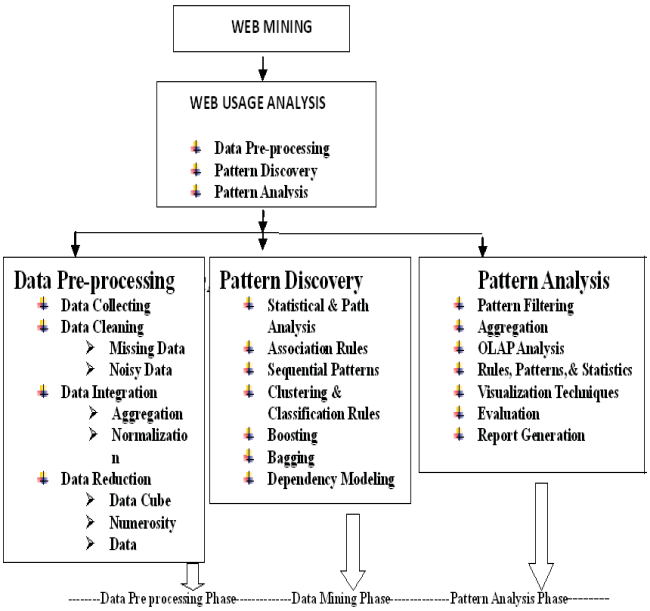


Figure 2. Web Usage Mining Phase

IX. ROLES AND RESPONSIBILITIES

Test	Requirement or Purpose	Action / Input	Expected Result	Actual Result	P/F
1	Validating the user information	Click the login button	Valid user	Same as expected	Pass
2	Searching for keyword	Submit Query	List of Meanings will be displayed	Same as expected	Pass

3	Clustering	Submit Query	Meanings are clustered	Same as expected	Pass
4	Searching for results	Click on the "search button"	Search results will be clustered	Same as expected	Pass

A.Key features of wordnet Tool

1. *Multiple Views of Multiple Wordnets*
2. *Freely Dened Text Views*
3. *Edit*
4. *Tree and RevTree*
5. *Query Result and External File Lists*
6. *Plain XML View*
7. *Synchronization*
8. *Editing Support*
9. *Tree Browsing*
10. *Consistency Checks*
11. *XML Conguration*

VII.System Architecture

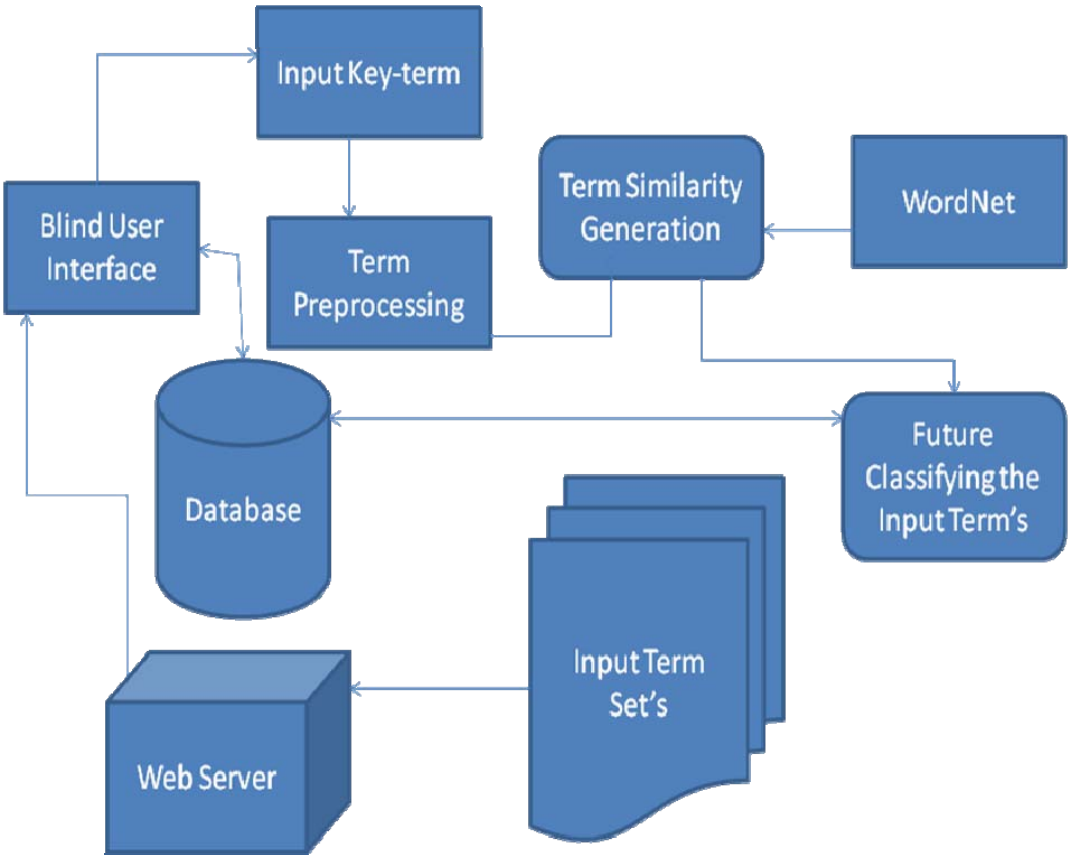


Figure 3. Overall system Architecture

Blind User Interface

- **Search space**
  - After user login process, web user can enter the search space page.
  - This is the environment for user to search the content from the web server.
  - This Search Space is the interface for user and web servers.
- **Input from User**

  Get the input text from the user for the search process.

VIII.CONCLUSION

This paper has presented an OTMM for grouping of research proposals. research ontology is constructed to categorize the concept terms in different discipline areas and to form relationships among them. It facilitates text-mining and optimization techniques to cluster research proposals based on their similarities and then to balance them according to the applicants' characteristics. The experimental results at the NSFC showed that the proposed method improved the similarity in proposal groups, as well as took into consideration the applicants' characteristics (e.g., distributing proposals equally according to the applicants' affiliations). Also, the proposed method promotes the efficiency in the proposal grouping process. The proposed method can be used to expedite and improve the proposal grouping process in the NSFC and elsewhere. It uses the data collected from a research social network (ScholarMate; http://scholarmate.com) and extends the functions of the Internetbased Science Information System (https://isis.nsfc.gov.cn). It also provides a formal procedure that enables similar proposals to be grouped together in a professional and ethical manner. The proposed method can also be used in other government research funding agencies that face information overload problems.

Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Finally, the method can be expanded to help in finding a better match between proposals and reviewers.

REFERENCES

[1] Q. Tian, J. Ma, and O. Liu, "A hybrid knowledge and model system for R&D project selection," Expert Syst. Appl., vol. 23, no. 3, pp. 265–271, Oct. 2002.
[2] K. Chen and N. Gorla, "Information system project selection using fuzzy logic," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 28, no. 6, pp. 849–855, Nov. 1998.
[3] D. Henriksen and A. J. Traynor, "A practical R&D project-selection scoring tool," IEEE Trans. Eng. Manag., vol. 46, no. 2, pp. 158–170, May 1999.
[4] F. Ghasemzadeh and N. P. Archer, "Project portfolio selection through decision support," Decis. Support Syst., vol. 29, no. 1, pp. 73–88, Jul. 2000.
[5] L. L. Machacha and P. Bhattacharya, "A fuzzy-logic-based approach to project selection," IEEE Trans. Eng. Manag., vol. 47, no. 1, pp. 65–73, Feb. 2000.
[6] WordNet Project Website, http://www.cogsci.princeton.edu/~wn/.
[7] Eurowordnet Project Website, http://www.illc.uva.nl/EuroWordNet/.
[8] Louw, M.: Polaris User's Guide. Technical report, Lernout & Hauspie Antwerp, Belgium (1998).
[9] Balkanet project website, http://www.ceid.upatras.gr/Balkanet/.
[10] Vossen, P., ed.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998).