Survey on An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection

Smita Bachal
Computer Department, DCOER,
Pune University
smitabachal@yahoo.com

Prof. S.M. Sangave
Computer Department DCOER,
Pune university
sunil.sangve@zealeducation.com

Abstract— Research project selection is an important task in many organizations such as government and private research funding agencies, universities, research institutes, and technology intensive companies. When a large number of research proposals are received, it is common to group them in accordance with their similarities in research disciplines areas. The grouped proposals are then submitted to the appropriate experts for peer review. Current methodology is used for grouping proposals are based on manual matching of similar research discipline areas and/or keywords. However, the exact research discipline areas of the proposals cannot often be accurately selected by the applicants due to their subjective views and possible misinterpretations. Therefore, rich information in the proposals' full text can be used effectively. Research ontology is constructed to classify the concept terms in different discipline areas and to form relationships among them. Text-mining methods have been proposed to solve the problem by automatically classifying text documents, mainly in English. In this paper ontology-based text-mining (OTMM) approach is used to cluster research proposals based on their similarities in research areas. The method also includes an optimization model that considers applicants' characteristics for balancing proposals by geographical regions. The OTMM method is tested and validated based on the selection process at the National Natural Science Foundation of China. The (OTMM) can also be used to improve the efficiency and effectiveness of research project selection processes in other government and private research funding agencies.

Keywords — Ontology, research project selection, Clustering analysis, decision support systems, text mining, SOM

I. INTRODUCTION

Research project selection is an important and persistent activity in many organizations such as government and private funding agencies, universities, research institutes, and technology intensive companies. This is a challenging multiprocess activity that starts with a call for proposals (CFP) by a funding agency. The CFP is circulated to relevant communities such as universities or research institutions. The research proposals are submitted to the funding agency and then are assigned to various experts for peer review. The review results from various experts are collected, and then ranked the proposals based on the aggregation of the experts' review results.

Fig 1. Shows the processes research proposals selection at National Natural Science Foundation of Chine (NSFC). A CFP process consists of proposal submission, proposal grouping, proposal assignment to experts, peer review, aggregation of review results, panel evaluation, and final awarding decision [1]. These processes are very similar in other funding agencies, excepting that there are a very large number of research proposals that need to be grouped for peer evaluation in the NSFC. The NSFC is the one of the largest government funding agency in China, with primary goal to fund and manage basic research.

When the large number of research proposals is received, four to five reviewers are assigned to review each proposal so that it assures accurate and reliable opinions on proposals. To deal with the large number of proposals, it is necessary to group proposals according to their similarities in research disciplined and then to assign the proposal groups to relevant experts for review. It might be possible that the reviewer/expert may not have adequate knowledge in all research disciplines, and the contents of many proposals were not fully understood when the proposals were grouped. Therefore, there was an immediate requirement for an effective and feasible approach to group the submitted research proposals with computer supports. An ontology-based text-mining approach is used to solve the problem.

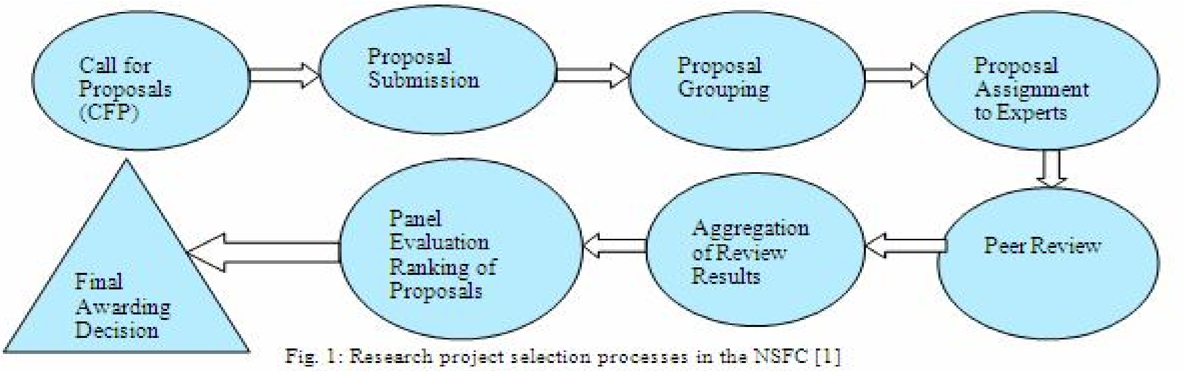
Call for Proposals (CFP)
Proposal Submission
Proposal Grouping
Proposal Assignment to Experts
Peer Review
Aggregation of Review Results
Panel Evaluation Ranking of Proposals
Final Awarding Decision

Fig. 1: Research project selection processes in the NSFC [1]

II. LITERATURE SURVEY

A text mining approach for automatic construction of hypertexts [3]

Most of the web pages were created manually by using some kind of the authoring tools. Although such authoring tools are easy to use and also provide sufficient functionality for individual users, if not impossible, a very hard work for those content providers that produce a large amount of web pages constantly or periodically. Manual construction is always unstructured and unmethodological. Solution to such inefficiency, we require a method to transform flat text to a hypertext. For this transformation we need to decide where to insert a hypertext in the text.

The important task of the hypertext construction process consists of creation of hyperlinks that connect source documents to their destinations. The hypertext construction process consist two parts: In the first part we concern about identifying the source of a hyperlink. In the second part we will try to find the destination of a hyperlink. To construct a hypertext, first perform a text mining process on a corpus of flat texts, and then analyse the Word Cluster Map (WCM) to obtain the sources of hyperlinks within every document. For each source, we have to decide its destination by analysing the Document Cluster Map (DCM) and create a hyperlink. After construction of every hyperlink within a document, then apply a text conversion program to transform a flat document to a hypertext document.

R&D Project Selection Using the Analytic Network Process [2]

The analytic network process (ANP) methodology is used as potentially valuable method to support the selection of R&D projects. A generic ANP model developed, which includes in its decision levels the actors are involved in the decision making activity, the stages of research, categories of metrics and individual metrics is presented. This model is suitable for small high-tech company, including data based on the actual use of decision making model. R&D Project selection using ANP is an effective and efficient decision making tool. The application of ANP, a multiattribute approach for decision that allows for the conversion of qualitative values into quantitative values and performing analysis on them. The ANP is a relatively simple, intuitive approach that can be accepted by other decision-makers and manages. A major limitation of ANP is the dependency on the decision maker.

An Ontology Based Text Mining Framework for R&D Project Selection [6]

In the R&D, after the research proposals are submitted, the next important decision making tasks is to grouping the proposals and assign them to experts for review. Grouping the proposals according to the similarities in research characteristics. If the number of proposals is small, then manual grouping based on keywords present in proposals can be used and then it assign to the reviewer manually. But, if the number of proposals is larges, it is very difficult to group proposals & assign them to reviewer manually. Solution for this problem, the proposals are classified using ontology and topic identification algorithm and then clustering the proposals using text-mining and lastly it is submitted to reviewer systematically.

In this paper, the project selection is based on following 5 modules:

Module 1: Research Ontology building

For building research ontology 3 steps are used:

1.1) Creating the research topics: The keywords in research projects each year are collected and their frequencies are counted.
1.2) Constructing the research ontology
1.3) Automatic topic identification

Module 2: Proposal classification

Proposals are classified by the discipline areas according to the keyword stored in ontology and the topic identification is done by using Topic identification Algorithm.

Module 3: Clustering the proposals

After the research proposals classification is done by their discipline areas, the proposals in each discipline are clustered using the concept based text-mining technique.

3.1) Sentence based concept analysis
3.2) Document based Text clustering using the concept based mining method

Module 4: Information retrieval

A knowledge-based agent and inference system is used for information retrieval. A knowledge based agent is used to retrieve grouped proposals & assign to external reviewer systematically.

Module 5: Assign to external reviewer

After information retrieval process, the proposals are given to reviewer according to their research area and experience. Assigning the proposals to reviewer is very challenging task. But there may be some ambiguity because reviewer may be specialized in more than one domain. So while clustering the reviewer, also consider their priority of research area.

A Fuzzy Ontological Knowledge Document Clustering Methodology [4]

This paper presents a novel hierarchical clustering approach for knowledge document self-organization, particularly for patent analysis. Current keyword-based methodologies for document content management tends to be inconsistent and inefficient when partial meanings of the technical content are used for cluster analysis. An ontology schema is used to automatically interpret and cluster knowledge documents are presented.

Traditionally, methodologies process knowledge documents using key phrases. A phrase can be used to represent many meanings, and many different phrases can represent the same meanings. In this communication, we analyse the grammar of the sentences & construct the ontology of documents. The fuzzy ontology-based methodology for clustering knowledge documents is presented and compared to the frequently used key-phrase K-means approach.

III. METHODOLOGY FOR CLUSTERING RESEARCH PROPOSALS [1]

A. Ontology-Based Text Mining To Cluster Research Proposals

Although there are various text-mining approaches that can be used for clustering and classifying the documents, they are developed with a focus on English text. TMMs which deal with English document are not effective in processing Chinese document. For example, Chinese text document consists of strings of Chinese characters, which English text document contain words. Also Chinese text has no delimiters to mark word boundaries, while English text uses a space as word delimiter. To process Chinese text document, TMM is not that much efficient or sufficient robust to process research proposals. To solve such problem, an ontology-based TMM (OTMM) is used. OTMM is able to handle English and Chinese text document efficiently.

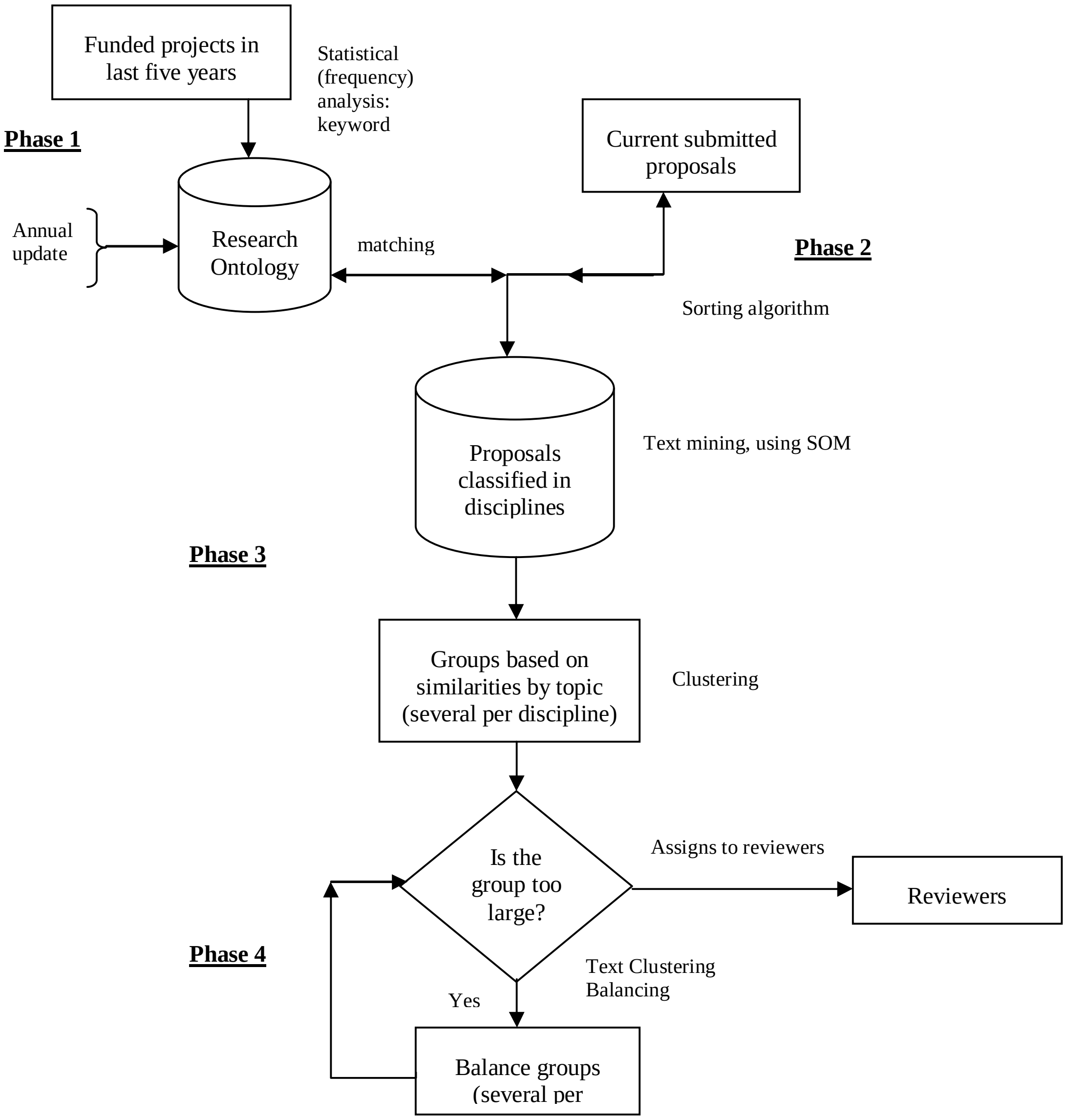
Funded projects in last five years
Statistical (frequency) analysis: keyword
Phase 1
Current submitted proposals
Annual update
Research Ontology
matching
Phase 2
Sorting algorithm
Proposals classified in disciplines
Text mining, using SOM
Phase 3
Groups based on similarities by topic (several per discipline)
Clustering
Is the group too large?
Assigns to reviewers
Reviewers
Phase 4
Text Clustering Balancing
Yes
Balance groups (several per

Fig 2. Proposed OTMM Process [1]

Ontology is a knowledge warehouse in which concepts and terms are defined as well as relationships between these concepts. [1] It consists of relationships, set of concepts, axioms that describe a domain of interests and represents an agreed-upon conceptualization of the domain's "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology. Ontology can automatically process the information and can facilitate text mining in a specific domain (such as research proposal selection). A statistical method and optimization models is used with proposed OTMM to improve the efficiency and effectiveness of clustering the research proposals.

OTMM consist four phases as follows:

1) **Construction Research Ontology:** Research ontology maintain the database of research projects which is funded in latest five years and construction of ontology is according to keywords, and it is updated annually. With the help of research ontology, research proposals of different disciplines can be clearly defined. Updating research ontology is done by manually. Construction of research ontology is complex task than just creating tree-like structure. First, there are some cross-discipline research areas (e.g. , "data mining" can be placed under "Information Management" in "Management Sciences" or "Artificial Intelligence" in "Information Sciences").[1] Second, Same concepts can be represented by some synonyms used by different project applicants, which have different names in different proposals. Therefore, the research ontology allows specifying more complex relationship between concepts apart from the basic tree-like structure.
2) **Classification of New Research proposals:** New research proposals are classified according to the keyword stored in ontology. For classification of research proposals Topic identification Algorithm (TIA) is used.
3) **Clustering Research Proposals:** After Classification of research proposals, clustering the proposals into discipline areas. Text-mining technique is used for clustering the proposals. Clustering process is a five steps process including text document collection, text document preprocessing, text document encoding, vector dimension reduction, and text vector clustering. The new research proposals are clustered in each discipline area by using a self-organized mapping (SOM) algorithm.
4) **Balancing Research proposals and regrouping them according to Applicants' Characteristics:** IT the number of proposals in each cluster is very large (e.g., more than 20), then it will be further divided into subgroups where the applicants' characteristics (e.g., Affiliated) universities are taken into consideration.

IV. Conclusions

Research project selection is done by OTMM for grouping of research proposals. Clustering quality and the performance for grouping research proposals by using Ontology-based Text-Mining (OTMM) is better than that of Text-mining method (TMM).The OTMM method improved the similarity in proposal groups, as well as took into consideration the applicants' characteristics for balancing proposals by geographical regions(e.g. equal distribution of proposals according to the applicants' affiliations). The OTMM can also be used to improve the efficiency and effectiveness of research proposals selection processes in other government and private research funding agencies.

References

1. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang"An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection",IEEE Trans an systems and humans vol.42,no.3 May2012.
2. L. M. Meade and A. Presley, "R&D project selection using the analytic network process," *IEEE Trans. Eng. Manag.*,vol. 49, no. 1, pp. 59–66, Feb. 2002.
3. H. C. Yang and C. H. Lee, "A text mining approach for automatic construction of hypertexts," *Expert Syst. Appl.*, vol. 29, no. 4, pp. 723–734, Nov. 2005.
4. A. J. C. Trappey, C. V. Trappey, F. C. Hsu, and D. W. Hsiao, "A fuzzy ontological knowledge document clustering methodology," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 3, pp. 806–814, Jun. 2009.
5. A. 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.
6. N. Arunachalam, E. Sathya, S. Hismath Begum and M. Uma Makeswari, "An Ontology Based Text Mining Framework for R&D Project Selection",International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 1, February 2013.
7. F. Ghasemzadeh and N. P. Archer, "Project portfolio selection through decision support," *Decis. Support Syst.*, vol. 29, no. 1, pp. 73–88, Jul. 2000.
8. 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.
9. J. Butler, D. J. Morrice, and P. W. Mullarkey, "A multiple attribute utility theory approach to ranking and selection," *Manage. Sci.*, vol. 47, no. 6, pp. 800–816, Jun. 2001.
10. Q. Tian, J. Ma, J. Liang, R. Kowk, O. Liu, and Q. Zhang, "An organizational decision support system for effective R&D project selection," *Decis. Support Syst.*, vol. 39, no. 3, pp. 403–413, May 2005.
11. W. D. Cook, B. Golany, M. Kress, M. Penn, and T. Raviv, "Optimal allocation of proposals to reviewers to facilitate effective ranking," *Manage. Sci.*, vol. 51, no. 4, pp. 655–661, Apr. 2005.
12. Y. H. Sun, J. Ma, Z. P. Fan, and J. Wang, "A group decision support approach to evaluate experts for R&D project selection," *IEEE Trans. Eng. Manag.*, vol. 55, no. 1, pp. 158–170, Feb. 2008.