Ontology in Text Mining To Cluster Research Project Proposals

Prof. Pankaj Chandre¹, Bharat Vishe², Hemant Vishe³, Pralhad Lengule⁴, Ankush Shah⁵

¹,²,³,⁴,⁵Department Of Computer Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune

Abstract—For government & private research funding agencies, Research project selection is important task. For large number of proposals, it is common to group according to their disciplines. This proposals are then sent to appropriate expert for peer review. Due to their subjective views & misinterpretations, the research discipline cannot be appropriately for more effectiveness in the proposals. The rich information uses full text. For classifying text document automatically, Text mining method is proposed. It generally accept English language. This paper represents ontology-based text-mining approach for clustering proposals based on similarities in research area. This method concerned with optimization model by geographical region. In government & private research funding agencies, the outcome can be used to improve efficiency of effectiveness of research project selection.

Keywords—Ontology, Text Mining, SOM, Sorting Algorithm.

I. INTRODUCTION

In computer & information science, Ontology as set of concepts i.e knowledge within domain & relation between the pairs of concepts. An ontology renders shared vocabulary & taxonomy which models a domain with the definition of objects or concepts & their properties. Ontology are used in artificial intelligence, the semantic web engineering, software, library science, enterprise bookmarking, Biomedical informatics as well as information architecture as a form of knowledge representation about the world. The ontology term has its origin in philosophy. It has been applied in many different ways. The core meaning of ontology within computer and information science is a model for describing the world that consists of a set of properties, types & relationships types. The creation domain ontology is fundamental to the definition & use for an enterprise architecture framework. There is also generally the state of expecting & there is a close resemblance between features of mode in ontology & the real world. Ontology have common in philosophy as well as computer science which is representation of events, ideas their properties & relations.

Research project selection is important task & which is used in government funding agencies. It is a challenge for multi process task that begins with a call for proposals by a funding agency. The submitted proposals are assigned to experts for review. The call for proposals(CFP) distributed in relevant communities & research are submitted to fund agencies & then that are assigned to experts. The assigned research papers results are collected & the submitted proposals are ranked based on the expert review result. This process is suitable for small information but large information or the area where information overload is more, this manual process get failed. Thus we proposed the Ontology in text mining to overcome this drawback.

II. LITERATURE SURVEY

The selection of research proposals in existing is done manually means the proposals are submitted to funding urgency and according to the name of research proposals or paper and the keywords the research proposals are classified into groups or domain this done manually means by the human. Following block diagram shows the process of manual clustering of research proposals.

![Figure No. 1 Existing System](image)

But this is not suitable for large data. It make misplacement of research proposals due to manual process and classification according only the name of research proposals. So this misplacement makes the reviewers or experts more confuse of the research proposals which are not from their area of research. There exists the software which also can not handle the large data and misplacement of research proposals.

III. PROPOSED SYSTEM

The proposed system is based on the ontology in text mining. we form four phases to process the selection work. The proposed system contains the representation of an agreed – upon conceptualization of the real word setting of domain ontology makes the knowledge explicit for computer which is implicit for human.
Thus ontology make it possible to automate. The information processing and specific domain facilitated by text mining, such as research project paper selection process. Ontology based text mining cluster the research proposals according to their domain. The unstructured text is processed and extracted interesting information and knowledge by applying text mining. The text mining patterns are extracted from natural language text it makes the text mining different from regular data mining. Which focuses on structured databases of facts.

1. **Construction of Research Ontology**

   The project which is funded in last five years are used to construct the research ontology according to keywords, A and it get updated annually. Research ontology is a set of research perfect management domain which is also public concept as a domain ontology. Research ontology expressed the topics of research of different disciplines more clearly to more understand.

2. **Classification of New Research Proposals using Sorting Algorithm**

   This uses the algorithm and according the keyword of the paper which is match with the started keywords of specific research domain and using this the research proposals are classified.

   
   \[
   \begin{align*}
   \text{For } n = 1 \text{ to } N \\
   \text{For } j = 1 \text{ to } J \\
   & \text{If } P_j \text{ belongs to } S_n \text{ then} \\
   & \quad \text{Then } P_j \text{ is added to } S_n \\
   \text{End} \\
   \text{End}
   \end{align*}
   \]

   **Algorithm No. 1 Sorting Algorithm**

3. **Clustering of Research Proposals According to Similarities Using Text Mining**

   Text mining technique is used to cluster the proposals in each discipline once the classification is done according to the domain. The five steps are performed to cluster the research proposals. Which are collection of text document, Encoding of text document vector dimension reduction and vector clustering. Self-organized mapping (SOM) algorithm is used cluster the new proposals.

   
   \[
   \text{Step 1) Document collection: After the classification of research proposal according to the discipline regions, the proposal documents in each discipline } R_n(n = 1, 2, \ldots, N) \text{ are collected for document preprocessing.}
   \]

   \[
   \text{Step 2) Document preprocessing: The proposal’s content are mostly non-structured. The research ontology is used to analyze, extract, and identify the keywords in the full text of the proposals. For example, “Research on Creational modeling and detection methods in financial fraud using ensemble learning” can be divided into set of word as {“behavior modeling,” “detection method,” “financial fraud,” “ensemble learning”}. Finally, a further reduction in the vocabulary can be achieved through removal of words only for few times.}
   \]

   \[
   \text{Step 3) Text document encoding: After text documents are segmented, they are converted into a vector representation: } W = (w_1, w_2, \ldots, w_F), \text{ where } F \text{ is the number of features selected and } w_j(j = 1, 2, \ldots, F) \text{ is the TF-IDF encoding of the keyword } y_j. \text{ TF-IDF encoding stands for a weighted method based on Inverse Document Frequency (IDF) with the Term Frequency (TF) to create the feature } w_j, \text{ such that}
   \]

   \[
   w_j = c_k j \ast \log(E/dkj), \text{ where } E \text{ is the total number of proposals in the discipline, } c_k j \text{ is the term frequency of the feature word } y_j, \text{ and } c_k j \text{ is the number of proposals containing the word } y_j.
   \]

   \[
   \text{Step 4) Vector dimension reduction. The dimension of feature vectors is too large. Hence, it is must to reduce the vector’s size by selecting a subset containing the most important term words in terms of frequency. To solve this problem, Latent semantic indexing (LSI) is used causes it not only creates the semantic relations among the keywords but also reduces the dimensions of the feature vectors effectively. LSI is a method for putting the original data vectors with short vectors in which the same information is preserved. For reducing the dimensions in a proposal of the document vectors without losing useful information, a term-by-document matrix is formed, where there is one column that corresponds to the term frequency of a document. The term-by-document matrix is decomposed into a set of eigenvectors using single-value decomposition with the help of eigenvalue. The eigenvectors that have the least weight on the matrix are then eliminated. Thus, the document vector formed from the term of the remaining eigenvectors & has a very small dimension and retains almost all of the relevant original features by using eigenvalue.}
   \]

   \[
   \text{Step 5) Text vector clustering. For clustering the feature vectors based on similarities of research Areas, this step uses an SOM algorithm. The SOM algorithm is a typical learning neural network model that clusters the given input data according to their similarities.}
   \]
Step 1: Initialization of weight vector $y_j$, initialize learning parameter $u$ & parameter $N_q$ where $q$ is winning neuron, define neighbor function, Set $n=0$

Step 2: Check the condition for stopping. If it is true then stop else continue.

Step 3: For each new vector $x$, Continue step 4 To 7

Step 4: For given input Compute best match of the vector $q(n) = \max \text{simi}(n, Y_j)$

Step 5: For all the units belong to their specified neighbor $j$ belongs to $N_q(n)$, Update weight vector as

$$Y_j(n+1) = \begin{cases} Y_j(n)+u(n)[x(n)-Y_j(n)] & j \text{ belongs to } N_q(n) \\ Y_j(n) & j \text{ not belongs to } N_q(n) \end{cases}$$

Step 6: Adjust learning parameter

Step 7: Approximately decrease topological neighbor $N_q(n)$

Step 8: Set $n=n+1$, then go to Step 2

Algorithm No. 2 SOM Algorithm

IV. CONCLUSION

This paper has presented an Ontology In Text Mining Method for grouping of research proposals. A research ontology is constructed to categorize the concept terms in different discipline areas and to form relationship among them. It facilitates text-mining and optimization technique to cluster research proposals based on their similarities and then to balance them according to the applicants characteristics. The proposed method can be use to expedite and improve the proposal grouping process. 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.

REFERENCE


