Topic Mining With Generative Model

Arshad Mulla
Bachelor of Engineer(computer), PGMCOE Pune

Bhausaheb Kekan
Bachelor of Engineer(computer), PGMOCE Pune

Nikhil Gore
Bachelor of Engineer(computer), PGMCOE Pune

Dipak Kakade
Bachelor of Engineer(computer), PGMOCE Pune

ABSTRACT This paper focuses on discovery of valuable knowledge from the multiple text sequence. Previous extraction techniques give duplicate words of content. Extraction of text sequences content collects from different time stamps in same topic. Asynchronous text sequences are extracted with mining from multiple sequences. It is the trivial content of information extraction on multiple sequences. There is no correlation operation in between of multiple text sequences. It can give low accuracy of result in output content. Extract the common topics and joint topics of content without duplicates. These kinds of text sequences are come under synchronous text sequences. Synchronous text sequences are extracts as common topics in same amount of time. It comes under reliable content.

INTRODUCTION
Data mining is the process of knowledge discovery from the general data. The gathering and formulating knowledge from data using pattern extraction methods is referred as the data mining. Managing a huge amount of data has become one of the challenges now days. Data warehouses are used to store the huge amount of data. The process of data mining helps to extract the required amount of data from warehouse efficiently. Various techniques, algorithms have been introduced for data mining till date.

This paper targets the problem of mining common topics from multiple asynchronous text sequences and proposes an effective method to solve it. The paper introducing a principled probabilistic framework, based on which a unified objective function can be derived as a solution to the topic mining problem. Further it is focusing on optimization solution by exploiting the mutual impact between topic discovery and time synchronization.

The following figure shows the architecture of the proposed generative model.

**Figure 1: Generative Model for Topic Extraction**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>Document</td>
</tr>
<tr>
<td>t</td>
<td>timestamp</td>
</tr>
<tr>
<td>w</td>
<td>word</td>
</tr>
<tr>
<td>z</td>
<td>topic</td>
</tr>
<tr>
<td>M</td>
<td>Number of sequences</td>
</tr>
<tr>
<td>T</td>
<td>Length of sequences</td>
</tr>
<tr>
<td>V</td>
<td>Number of distinct words</td>
</tr>
<tr>
<td>K</td>
<td>Number of topics</td>
</tr>
</tbody>
</table>

ALGORITHM

STEP 1: Input the keyword or topic.

STEP 2: Find the related topics from all sequences.

STEP 3: Extract the common topics.

STEP 4: Synchronize the document according to the related topics with respective time stamps.

STEP 5: Repeat the steps 1-4 until the convergence.

TEXT SEQUENCES

Every document consists of some information in the textual format. Text sequences are nothing but the collection of words forming meaningful information.

COMMON TOPICS
Common topic data extraction is a process based on words distribution process. Using different asynchronous sequences, we extract the results with the help of interaction. Interaction process similar to mining. It can give the common topics data as a meaningful data at output end.

ASYNCHRONISM
Sometimes the documents in the database may contain same contents but different time stamps. This is the ambiguity in the database. While mining the database for that contents more time is consumed. This duplication of same topics with different time stamps referred as asynchronism in the database.

TOPIC EXTRACTION
It is the process of extracting common topics with respect to the given keyword or input. It gives efficient solution and displays the synchronous text sequences as a result in output environment. The same process is applied as a refinement process till reaches the good data results display as maximization features extraction. It can give the guaranteed solution as a meaningful result.

TIME SYNCHRONIZATION
Synchronization is the process of comparing the two or more documents with respect to their time stamps. According to the user input, common topics are extracted which are asynchronous in nature. Time synchronization is the process of rearranging the documents in the database according to the new refined results. Such that there is no ambiguity at output.

CONVERGENCE
Applying the given algorithm on any database we synchronize the database which makes the output more reliable and semantic. Convergence is the state from which there are no further refinements to the database. We can say it as the final state of algorithm's processing.

CONCLUSIONS
We tackle the problem of mining the common topics from the database using the proposed algorithm in this paper. The proposed algorithm is self-refinement process which iterates through database to mine the topics with relativity. It is novel method to mine the correct, reliable and meaningful data from the database. The synchronized results obtained after the refinements to the database help the end user to get semantic information at the end.

REFERENCE