ADDING UNIQUE TOPICS OF INTEREST WITH A NEWER INSIGHT

Problem Statement

Given a pool of 1 million or more unique documents (covering various topics), we want to ensure that a new document is added to this pool only if it is also unique (that is either it talks of a completely new topic(s), or brings in a fresh perspective on already discussed topics). What NLP approach should be taken to build such a system.

Introduction

With the ever increasing number of available resources on the internet, it becomes eminent that we optimise these resources. The objective is to create a system with houses documents each covering a unique topic or idea and any new document that has to be added to this pool, should either be covering an entirely new topic or proposes a new perspective to an existing topic. This kind of system can have applications when research communities want to generate topics of interests, or if one wants to build a database of highly optimised and useful information. One approach would be to use the classical Topic Modelling approaches (LDA, Topical N-gram, Phase-Discovering LDA etc.) in bibliographic data, they can be useful only for text categorization, but they will not be readable and organised as manually selected topics. In this write up I would like to propose an algorithm based on **Frequent pattern mining and Natural Language Processing** to create topics of interest which is very similar to manually created topics. Association mining for finding large number of possible topics and NLP to refine and select best topics. We could further use word embeddings to group semantically related topics.

Methodology

The following will be performed with the new set of documents and the existing documents:

1. Categorization of topics from the available documents
2. Generating candidate topics from the new set of documents
3. Prune topics that are redundant, ambiguous and uninteresting.
4. Refine candidate topics
5. Group semantically related topics (this is done to find topics which find a newer insight to an existing topic).

1. **Data Categorization**
   Entity Resolution algorithms can be used to clean the existing data and assign abbreviation to each unique topic.
2. **Candidate topic generation**

Association mining to find co-occurring words and phrases present in the new topics and the existing topics. The topics that are not frequent or rare are new topics.

- FP-growth association mining can be used.
- Can apply pruning to get well-formed topics.

This gives us a set of topics that could possibly be a new topic.

Now we will perform NLP to identify the best topics.

3. **Topic pruning**
   Compactness Pruning: Aims to remove those phrases whose words do not appear in a specific order.
   We also remove topics that are common and do not convey any useful information. (this can be done using the tf-idf score, uninteresting topics will have less idf scores).

   Redundancy pruning: Removes redundant words and phrases from selected list of topics.
4. **Topic Refining**
   Using some of the grammar rules such as "if a verb appears at the beginning of the phrase it will lead to a better topic" can help in topic refining.

5. **Topic grouping (topics which offer newer insight to already existing topics)**
   Finding similarities amongst topics by exploiting the co-occurrences of similar words.

<u>Evaluation Models</u>

1. **Precision Analysis**
   Using the opinion of experts can manually check if the topic is really a new topic. This can be done for a set of testing data and precision analysis can be done.
2. **Similarity Analysis**
   Create a base-line of gold-standard topics using existing online databases such as Wikipedia.
3. **Empirical Analysis**
   Compare the results of this algorithm with existing topic modelling algorithms such as LDA and phrase based algorithms.

<u>Why this approach is better</u>

1. **Association Mining Methodology**
   The full-text of a given document will contain a lots of trivial words or phrases. If we use full-text of the documents, we end up getting several unimportant words and phrases. The topic and a short description of the topic in the document (by-line or the abstract) should be useful. Association mining is used to find co-occurring words and phrases in the titles. **Thus the selected words and phrases from the topics and abstract should form baskets.** Now we can use **FP-Growth association mining**

algorithm for mining frequent patterns. The algorithm generates frequent topics by defining minimum support and confidence. Defining an appropriate minimum support and confidence can help us identify which documents from the set of newer documents are not unique, also the non-frequent topics are the unique topics and as described in my write up before, these topics would form the candidate topics.

2. **Why this approach is better for finding nuances**
   As far as the other topic modelling algorithms are concerned (mainly LDA, Topical N-gram, Phase-Discovering LDA), this algorithm performs better because it takes the semantics into consideration. As you have mentioned bag/phrase of words cannot differentiate between two documents that talk about different ideas but use same words.
   - **Why this approach is better than cosine similarity**
     Cosine similarity would fail in those cases wherein the ideas represented by two documents will have presented the same idea but since they have been written by two different people will use different vocabulary. This will lead to a low similarity but semantically they talk about the same idea.
     In the approach that has been proposed in the write-up uses **association mining** initially to weed out the topics that have similarity based on their title and the short description. After which we further **prune the topics** which are ambiguous using tf-idf scores. This is followed by **Topic Refining** by using grammar rules. And the last step **is finding the most dissimilar documents** by finding similarities (Ldatovec or word2vec model can be used here).
   - **Why this approach is better than just Word2Vec model or Ldatovec model**
     The approach described in the write-up can be considered as a hybrid model between a bag of words approach and a semantic approach. The documents available in the database are over 1 million and above. The set of new documents will be high in number as well. The first step is to immediately remove non-unique topics. **Association Mining** is used for the above. Defining a minimum support and confidence will get rid of those documents and generate a set of candidate topics. After which we further **prune the topics** which are ambiguous using tf-idf scores. This is followed by **Topic Refining** by using grammar rules. And the last step **is finding the most dissimilar documents** by finding similarities (Ldatovec or word2vec model can be used here). So this approach still uses the word2vec model but the proposed approach is more refined and using just word2vec will be more time consuming and will most probably give less accurate results. Also it should be noted that the proposed approach has more variations like using idf scores, using grammar rules and then similarity comparison. Thus this approach will be more refined in terms of finding nuances.

Conclusion

I believe that the topics generated by this algorithm will be more meaningful and human interpretable than the existing methods because it uses more semantics than just fining similarities or dissimilarities between words and phrases.