Intelligent & Adaptive Systems Research Group

Grouping of Blogposts

D. Dubois, Rp. Singh and Ali. A. Ghorbani

Problem Statement

Opinion mining is a particular area of artificial intelligence, it extracts knowledge from a set of existing document. However extracting knowledge from a huge number of text documents is practically unfeasible if we deal with a heterogeneous set. As amount of document on internet is getting bigger everyday, this issue is taking importance. To solve this issue, it is important to group documents according to their similarities. That is why Automatic document grouping methods become an increasingly important tools for helping our system to organize this vast amount of data in order to extract opinion out of them.

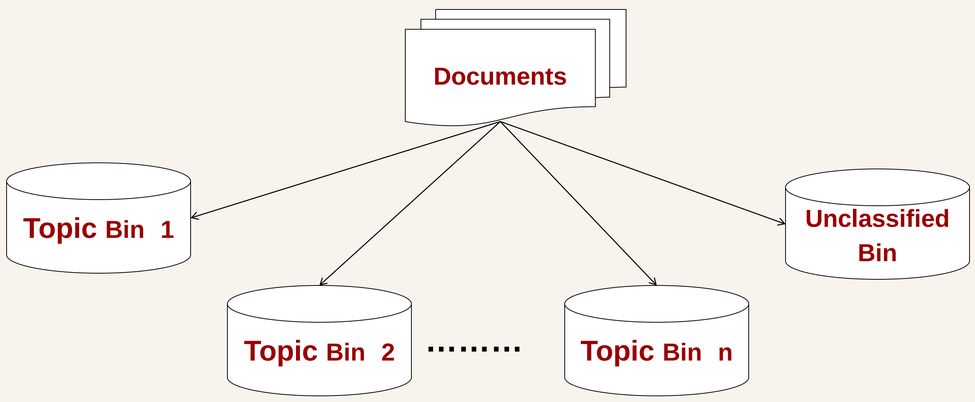
Solution

To group document according to their similarities , there are two existing area of research:

- if we know the categories before trying to group the document then we speak about *Classification*
- If we don't know these categories, then we have to discover them in the same time as grouping document. We talk about *clustering*.

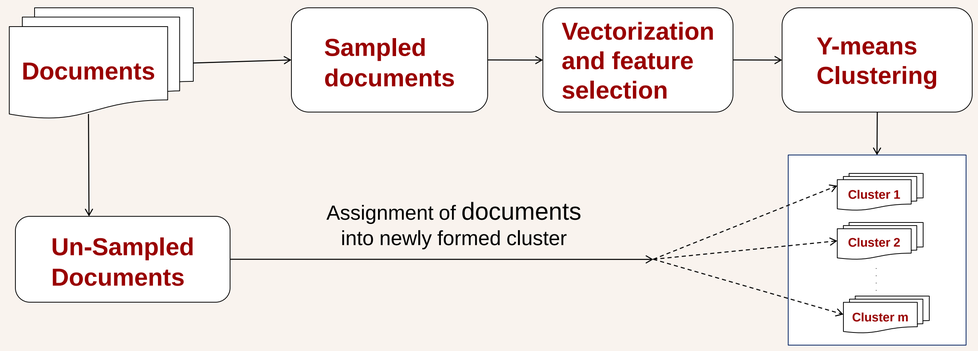
A lot of different categorization and clustering algorithms exist such as K-mean , Hierarchical clustering, Spectral clustering, Naïve Bayes classification, SVM and k-NN classifier. The two following methods are used in OMID(Opinion Mining and Issue Discovery) project.

Classification



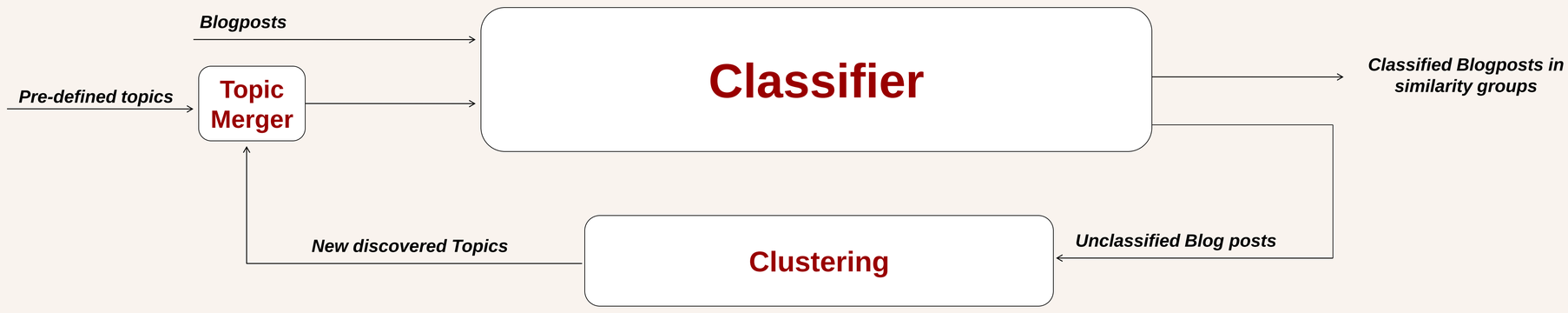
To classify a document into a predefined category, the document need to satisfy a constraint. That is what we call the selecting Criteria. In the case of OMID, a category is a set of word, and for a document to belong to this category, the document needs to contain at least a specific percentage (threshold) of this set of word. Other criteria can be applied but this one is not complex to compute, which is one of the most important characteristic in case of huge amount of documents.

Clustering



- As far as OMID project deals with a huge amount of documents, it is too long to apply clustering on the whole dataset. Sampling methods make us able to compute the new clusters on a reduced dataset and to extend these results on the unsampled documents.
- The input of a clustering algorithm is always a vector of value. In OMID project, Vectorization step convert a text document into a vector of TFIDF value which is a good representation of the importance of a word into a text among the text collection.
- Feature selection is one of the most important area of research, since it directly affects the quality of document groups. Feature selection in unsupervised domain is still an open area of research.
- OMID project use the Y-mean algorithm to cluster its documents.

Our Approach



Results and Issues

- The OMID project deals with hundreds of thousands of blogs, and in a time complexity point of view this algorithm perform well. However, for now, the Y-mean clustering doesn't find relevant enough new topics since after the classification based on the new topics we still have a lot of unclassified blogposts.
- Moreover some clusters, seems to be very close to pre-defined topic, it is not relevant to split documents of these categories in two different groups.

Future work

- Our experiment prove that the feature selection performed in OMID project is not good enough to represent the documents. Many new approach on this area could take the place of TFIDF methods in the future of this project.
- Another important area of research in this project could be to improve the sampling part occurring in the clustering step.
- To solve the problem of closed groups, a step of merging clusters according to similarities between then will be a good option.

2011 Research Expo, Fredericton, NB, April 15, 2011