A segment-based approach to clustering multi-topic documents

Submitted to:

Rashmi Gulhane, 2015H103187P Dr. Poonam Goyal

Jishan Baig, 2015H103089P

Table of Contents

**Abstract3**

**Block Diagram**4

**Improvements**5

**Result and Analysis**6

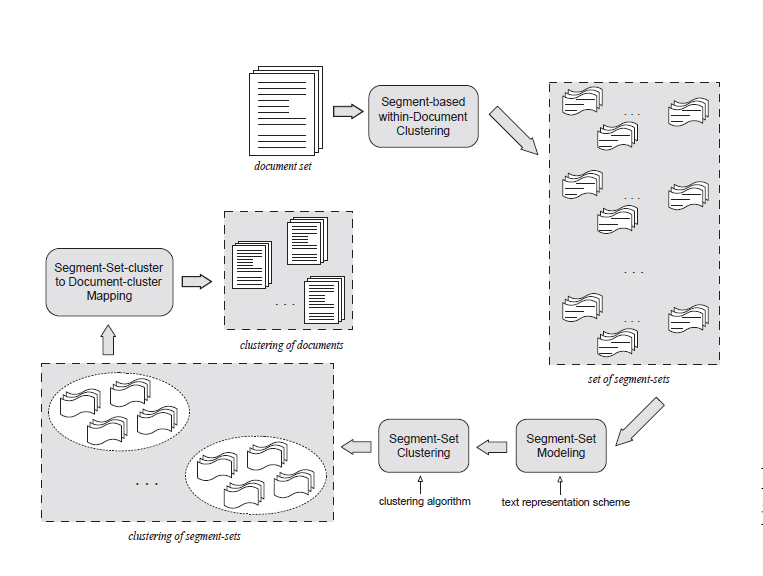
**Query Processing**9

**Abstract**

Document clustering and topic identification are well known problem in text mining. In case of research papers, many research papers contain sub topical information under the same time. Many sections in the research document contain background knowledge about the main topic. A document may contain information regarding multiple subtopics. All the existing methods used for document clustering assume document as an indivisible unit for text representation and similarity computation. These methods do not consider the underlying sub topics present in the corresponding document.

In the proposed problem, we need to cluster the acl dataset text documents using soft clustering algorithm. In soft clustering a document can belong to mutiple clusters according to its membership measure. To solve this problem, we propose a segment based approach by identifying the text segments that are coherent with respect to the underlying subtopics in the documents. First these segments are clustered within a document, which basically form segment sets. These segment sets belong to corresponding document, contain segments related to same topic. The segment set of all the documents are then clustered in a disjoint fashion. Then in the last phase these segment sets are mapped directly to the corresponding documents. Thus the segments corresponding to one document may belong to multiple clusters because of its underlying multi topic nature. For clustering within a document we have applied LDA where segments belonging to a single topic are appended together. Each Segment can at most be represented by 2 topics. Spherical K-means have been used for Segment-Set Clustering. Chi Square has been used to reduce dimension.LDA has been applied within a Cluster to identify topic and its Representative feature. Query Expansion has been done for better results. Some Internal and External Measure has been used for cluster Evaluation.

**Block Diagram:**

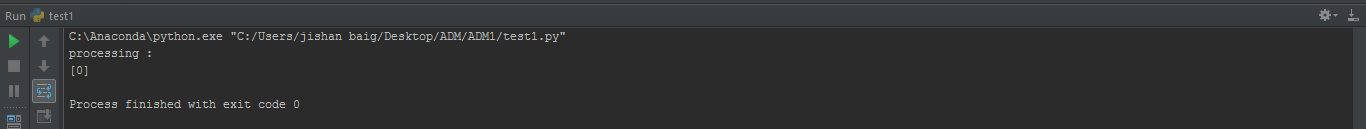
****

**Improvements:**

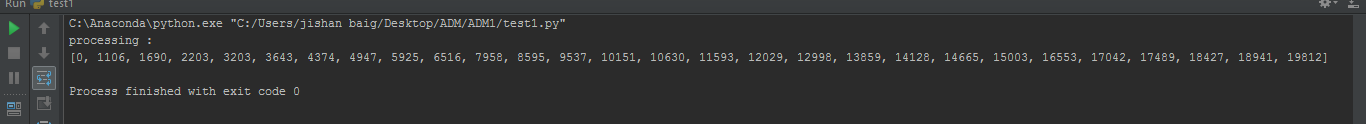
**1. Improvement in Text Segmentation Algorithm:**

The text segmentation algorithm is used in the 1st phase of the given clustering solution. By implementing text tiling algorithm the text is divided into multiple segments. Text tiling detects subtopic boundaries by analyzing patterns of lexical co-occurrence and distribution in the text. Terms that are discussing a particular subtopic tend to occur locally, and a switch to new subtopic is detected by ending of co-occurrence of a given set of terms and the beginning of the co-occurrence of another set of terms. All the pairs of adjacent blocks are compared by cosine similarity measure, and the resulting sequence of similarity values is examined in order to detect the boundaries between the coherent segments. At the end of the algorithm it normalizes the identified boundaries to the nearest paragraph breaks. The documents that are included in acl corpus have very less number of paragraph breaks. So the segment quality we are getting by basic algorithm was not so good and sometimes it’s unable to segment the input text. To solve this problem the boundary normalization has to be done according to the segment boundaries, not according to the paragraph breaks. So after modification the algorithm is giving the actual identified boundaries normalized to their nearest previous punctuation. By this we are getting actual segment boundaries.

Normalized boundaries before improvement:



Normalized boundaries after improvement:

****

**2. Improving Cluster Quality:**Goal: Using Title of the Research paper to improve the cluster Quality and its feature setProcedure: (Cluster 1: set of clusters formed by only title of paper)(Cluster 2 : set of Clusters formed by Documents)1) Titles of papers are clustered using spherical k means.2) If cluster1 has at least 80% document similar to cluster2 merge the cluster1 documents to cluster2 and add the cluster1 features to cluster2 list)Observation: has helped cluster2 to become softer and get better representative features

**Result and Analysis:**

Dataset Details Dataset used: ACL Anthology Reference Corpus

Total Documents: 717Total Complete Research papers: 97Total Segments formed: 3492Total No of Segment Set formed: 735

------------------------------------------------------------------------

3-way clustering: [I2=1.79e+02] [739 of 739]

------------------------------------------------------------------------

cid Size ISim ISdev ESim ESdev |

------------------------------------------------------------------------

0 176 +0.077 +0.038 +0.030 +0.010 |

1 219 +0.063 +0.017 +0.031 +0.010 |

2 344 +0.047 +0.012 +0.031 +0.010 |

------------------------------------------------------------------------The column labeled “Size” displays the number of objects that belongs to each cluster.

The column labeled “ISim” displays the average similarity between the objects of each cluster (i.e., internal similarities).

The column labeled “ISdev” displays the standard deviation of these average internal similarities (i.e., internal standard deviations).

The column labeled “ESim” displays the average similarity of the objects of each cluster and the rest of the objects (i.e., external similarities).

Finally, the column labeled “ESdev” display the standard deviation of the external similarities (i.e., external standard deviations).Criterion Function used is [I2=1.79e+02]

C:\Users\jishan baig\Downloads\formula.png

The mathematical definition of CLUTO’s clustering criterion functions. The notation in these equations is as follows: k is the total number of clusters, S is the total objects to be clustered, S i is the set of objects assigned to the ith cluster, ni is the number of objects in the ith cluster, v and u represent two objects, and sim(v, u) is the similarity between two objects.

**Omega index:**

Without improvement: 0.773442

with improvement : 0.792215

**Purity** without improvement 0.579**Purity** with improvement 0.858

**Accuracy** = (TP+TN)/ (TP+TN+FP+FN)

**Recall** = TP/TP+TN

**Precision** = TP/TP+FP

**Without Improvement:**

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Negative Cases | TN:2156 | FP:30 |
| Positive Cases | FN:519 | TP:1951 |

Accuracy = (TP+TN)/ (TP+TN+FP+FN) =4107/4656 = 0.88

Recall = TP/TP+TN=1951 / (1951+519) = 0.7898

Precision = TP/TP+FP=1951 / (1951+30) = 0.9848

**With Improvement:**

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Negative Cases | TN:1990 | FP:375 |
| Positive Cases | FN:108 | TP:2183 |

Accuracy = (TP+TN)/ (TP+TN+FP+FN)= 4173/4656 = 0.89

Recall = TP/TP+TN = 2183/ (2183+108) = 2183/2291 = 0.95

Precision = TP/TP+FP =2183 /(2183+375) = 2183/2558  = 0.85

**Query Processing:**

Preprocessing of Query: Query is preprocessed (stemming, stop word removal) to get better recallQuery Expansion: Synonym of term in query if requiredQuery Retrieval Algorithm:Part 1: Compare Query Term with Cluster Representative Terms (i.e. topics obtained by LDA for clusters and the cluster features)The cluster which has the most of terms of query its documents are selected for displayPart 2:For ranking of documents inside the cluster the following method is used1) For each term in the query get its occurrence in corpus with the help of inverted index. Rank higher the document in the cluster which has a occurrence in most of the inverted index of term.2) Documents which are in none of the inverted index would be rank lowerPart 3:3) If no Result is acquired from Part 1 then we find the synonym of words in the query and repeat the process part1 and part24) In case there is no match to query terms from previous 2 method we use the inverted index to get the match.Observation:This method of retrieval on clusters topic has helped in faster query retrieval. Ranking mechanism has helped to get better result first.