Topical Clustering of Tweets

CS 533: Information Retrieval

Project Description

Hakan Sözer, Muhammed Yağmur Şahin, Yağız Salor

1. Description of the Problem

TT (Top Tweet) concept of the twitter is working by counting the number of occurrence of the exact word. That situation causes a problem which is missing most of the tweets related to same topic but do not includes the word notes as TT. So by clustering the tweets this projects aims to more accurate and comprehensive TT concept. Take an election as an example. Lets consider word “election” as TT. For current TT mechanism “Obama rocks, 4 more years” tweet won’t be in the “election” TT. By using clustering, idea is not to miss any related tweets related to the phrase that people looking for.

2. Motivation/Importance

In modern era, Twitter has become one of the most important information sources. Not only single users make use of Twitter to learn what’s going on in their neighbourhood, but also the companies, politicians, celebrities and even TV. programs use Twitter. Let’s consider the TV. program case, in the two or three hours duration of a program, newly coming tweets is being read time to time. In the case of our clustering approach the tweets related to program can be easily analyzed, since the approach will not only list the TTs, under favour of this approach we will be able to cluster the tweets related tweets together which are related to the TV. program. This will provide Tv. program mentor to to see related tweets about the program easily even there is no mention about or direct relation to the program or its mentor, guests. The same motivation stands for the company or a product case.

3. Methodology

At first amount of the tweets to be cluster must be decided. For that purpose window approach will be used. Inıtial plan is to use time windows for collect some tweet collection. Further development can be done by considering some performance issues and for the times that the (no of the tweets / time) ratio increases rather than time windows tweet count windows can be used. Original plan is to use all phrase to doing clustering and define new TTs. Backup plan is to use hashtags for supervised clustering (1). Because of the volume of the data retrieving some information from all phrase can be challenging. After getting clusters, results can be checked by comparing the results with the TT founded from current TT concept which explained detaily in expected results section.

4. Expected Results

Results can be rated by comparing the clusters generated by the project and the real TT phrases. Expected output for the clusters, is very similar cluster and TTs but the anticipated difference between the number of the tweets in each cluster. As described above the goal is not to miss any related tweets so number of the tweets given by the project related to TT must be more than the original number resulted from the current TT concept in Twitter.

5. References

1. *Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, Robert Frederking*
Proceedings of the ACM SIGIR Special Interest Group on Information Retrieval's 3rd Workshop on Social Web Search and Mining (SIGIR: SWSM 2011). 2011