Mining Evolving Topics

Web and Social Information Extraction

Project for a.y. 2018/2019

The scope of the project is to identify and trace topics over a temporal interval. A topic can be seen as a set of keywords. Hence, one can see changes in the keyword set of topics throughout the specified time period.

Problem statement: Given h-l+1 snapshots, G_1,...,G_{h-l+1}, of a temporal research network, identify topics in each snapshot G_i and trace them over the time interval $[l,h]$

Datasets:

DS-1: The keyword co-occurrence

This dataset contains one graph for every year [2000-2018]. Let V be the set of nodes that represents the different ***keywords*** used by articles in literature: e.g. *machine learning, crawling, statistic inference,* etc. Let E be the set of edges which represent the relationship of two keywords being used by two different articles. Every edge $e=\{k_i,k_j\}$ is decorated with an ordered dictionary of authors. The dictionary of authors is organized as follows: key-value ***(a,n)*** pairs where a represents an author, whereas n is the number of times author a uses k_i and k_j in his/her articles.

Each row of the dataset is formatted as follows:

y_q<tab>k_i<tab>k_j<tab>[a_0:n_0,....,a_m:n_m]<newline>

DS-2: The Co-authorship

This dataset contains one graph for every year [2000-2018]. Let V be the set of nodes representing those authors that have published in the year in question. Let E be the set of edges which depict the relationship of co-authorship: i.e. two nodes a_i and a_j have an edge $e=\{a_i,a_j\}$ if the corresponding authors have published an article together. The weight of an edge e corresponds to the number of collaborations between its two incident nodes.

Each row of the dataset is formatted as follows:

$\mathbf{y_q}$`<tab>`$\mathbf{a_i}$`<tab>`$\mathbf{a_j}$`<tab>`$\mathbf{n}$`<newline>`

T1: Topic Identification

1. Select *top-k* [k=5,10,20,100] (k is the number of generated topics) (according to a certain metric such as pagerank, hits, betweenness, brokerage) keywords in DS-1 **for each year.**
2. For every node in *top-k* apply a Spreading of Influence Algorithm to report the nodes influenced by them in each iteration of the algorithm (similar to Linear Threshold Model or Independent Cascade). *The influenced nodes represent a topic.*
3. Join the produced topics **in a given year** following a merging strategy which takes care of possible overlaps among them.

The students must decide[1]:

- which measure should be used to select the top-k starting nodes;
- which is the weights function for the edges;
- which is the threshold function for the nodes;

[1] *You can use DS-2 to define your weight and threshold functions.*

T2: Topic Tracing

1. Trace, over the timeline [2000-2018], any topic identified in task **T1**;
2. While analysing the topic temporal/structural behaviour the student must decide if two topics identified in two consecutive years can be merged together;
3. Create a final list of the merged topics;

The students must decide:

- how to determine that a certain topic $\boldsymbol{t_j}$ exposes a similar temporal/structural behaviour of another topic $\boldsymbol{t_i}$;