cs512 Assignment 2

Chia-Hao Hsieh (chsieh17)

Question 1: Phrase Mining (25 points)

- There are several phrase-based topic modeling methods, such as (1) Topical N-grams (TNG) (Wang, et al.'07), (2) TurboTopics (Blei & Lafferty'09), (3) KERT (Danilevsky, et al.'13), and (4) ToPMine (El-kishky, et al.'15). Compare and outline their major differences. What are the reasons that ToPMine may find quality phrases and phrase-based topics? (10 points)

  Topical N-grams (TNG) is a probabilistic model that creates n-grams by concatenating successive bigrams. TNG is a generalization of Bigram Topic Model. It infers topics and phrases at the same time.

  TurboTopics and KERT infer phrases after topics modeling.

  TurboTopics uses LDA to assign a topic labels to each word. It then merge adjacent unigrams with the same topic label recursively to get phrases.

  KERT uses LDA too. It then performs frequent pattern mining and phrase ranking on each topic.

  ToPMine mines phrases first and then models topics. It gets better result than previous methods. If topics are modeled before phrases like KERT does, then words of a phrase may be categoried as different topics. They never have a chance to be constructed as a phrase again in the later topics modeling stage.

- SegPhrase+ (Liu et al, SIGMOD'15) does not use any natural language processing (NLP) methods in phrase mining. Give two examples and show NLP methods can be integrated into the Segphrase framework to further improve the quality of phrase mining. (15 points)

  In the feature extraction stage of SegPhase, NLP methods such as POS tagging and chunking can be applied. These methods can be used to extract features from candidate set.

  For example, word type tags generated from POS tagging can be used as features of the candidate in the later classification stage. Similarly, chunking can also generate features.

  In the phrase segmentation stage, NLP techniques can be used to help compute the quality score for each phrases. With better quality score, we can partition a sequence of words into better phrases.

Question 2: Entity Recognition and Typing and Network Construction (25 points)

- Explain how ClusType (X. Ren et al., KDD 2015) can find quality types for news corpus by distant supervision and reason why such a method can be more effective than many existing typing methods. (10 points)

  ClusType has three stages. In the first stage, it extracts candidate entity mentions with minimal linguistic assumption, using NLP methods like POS tagging. A common limitation for traditional typing systems is that they have trouble adapting to new domains and new types. It requires labor-expensive human annotation. By relying on linguistic assumption as less as possible, ClusType is more generalized in terms of domain restrcition.

  In the second stage, ClusType constructs a heterogeneous graph of entity surface names, entity types and relation phases. Note that ClusType takes each mention with its context, instead of assuming all mentions of the same entity surface name mean the same thing. This addresses the name ambiguity problem from which other typing systems suffer.

  In the third stage, ClusType infers types and propogates types via synonymous relation phases. Using inferred types as features, it clustering relation phases. With better relation phases as result, we can propogates types again to get better result on types. By recrusively doing these two steps, we have a good final result on both of them. Using synonymous relation phases solves the context sparisity problem, in which types of entities with less context are hard to infer. We can have more context by connecting rare relation phases with their synonyms.

- Suppose a computer science research publication database contains millions of research papers generated in computer science (CS) research. You are required to construct a typed heterogeneous CS information networks from such a database. The network should contain not only bibliographic information (author, venue, title, year) but also detailed research theme information for each publication, such as "deep learning", "frequent pattern mining", etc. Outline your design of a set of methods that may construct such a network effectively. (15 points)

  We can easily extract bibliographic information from the paper, even if the datase doesn't provide them already. As for research theme information, ClusType can be used. Instead of entity types of words, we are looking for research themes of papers. So common research themes are used as the target types for ClusType. For each paper, we apply ClusType to get theme labels for all phrases in the paper. Then the most frequent theme is picked as the paper's theme.

  If we allow multiple themes for a paper, then we can simply pick the top few themes. Or we can set a

percentage threshold, say 10%. Themes with appearance above 10% in phrases are picked.

Question 3: Truth Finding (25 points)

- Explain what are the differences between the two truth finding mechanisms: TruthFinder (Xin et al., TKDE 2008) and LTM (Zhao, et al, VLDB 2012). (10 points)

  TruthFinder is a HITS-like random walk algorithm. It iteratively computes the trustworthiness of websites and the confidence of facts based on each other. Its limitation is that it treats websites by one single measure, quality, instead of measures like precision and accuracy. Some websites tend to provide false positive results while other may tend to ignore true attributes, which means false negative.

  LTM, on the other hand, models negative cleams and two-sided source quality with probabilistic model. Based on new assumptions that a source website makes more than one fact claims and most sources provide correct fact claims, LTM has better performance.

- Not every piece of news or tweets is truthworthy. Design a mechanism that may use both sources to identify what is likely to be the truth in news and tweets . (15 points)

  With news and tweets as sources, use LTM to find the truth. For tweets, each Twitter account is treated as a source. Tweets from one account are the fact claims made by that user. Similarly, For news, news sources such as news websites or newspaper publishers are treated as different sources in LTM. Every news from a publisher is a fact claim.

Question 4: Stream Data Clustering and Spatiotemporal Data Mining (25 points)

- We have discussed efficient methods for clustering data streams. Explain why Clustream can be used to study the evolution of clusters in dynamic data streams. (10 points)

  CluStream uses pyramidal tilted time frame to make sure it dones't lose dynamic changes. Depending on time, snapshots are stored at different levels of grannularity. CluStream has micro- and macro-clutering. And micro-clustering is better than k-mean in terms of quality since the k-mean approach is bad at detecting clusters over stream.

- Many tweets are geo-coded (i.e., their geo-locations are known). Suppose a tweet contains user-id, time, location, hashtag, and short text messages. Design an effective method that may detect an unusual local event from those that happen regularly in a local region. (15 points)

Use CluStream to do the mining process on stream data. During the online incremental updating, if a newly-created cluster is growing quickly and geo-locations of points within that cluster are close, then we know there is an unusual local event happening.