Recommended collaborative projects

2017 Autumn

1. Title: Automatic Extractive Text Summarization in Multi-dimensional Text Corpora

* Abstract: To systematically analyze large numbers of textual documents, it is often desirable to manage documents (and their metadata) in a multi-dimensional semantic space (hence text cube). Automatic text summarization has been studied extensively in the related fields. Given a single document or a set of them, a few sentences are used to capture their principle content. However, it has not been studied in the text cube scenario where semantic-rich dimensions are present, i.e., summarizing an ad-hoc subset of documents according to certain dimension values. An example is shown as:

Corpus: New York Times news articles

Query: <Topic: "Economy", Location: "China">

Summary: 1. [Bank lending] and [local government debt] have soared in recent years, and were a major driver of china’s [economic rebound] after [the global financial crisis]. 2. A notable slowdown from previous quarters shows that [china’s economy] continues to cool and indicating that Beijing may struggle to meet its [growth target]. 3. China’s [trade surplus] rose to its [highest level], while inflation remained under control.

* Required skills: some NLP background, good coding skills (python preferred), may read quite a bit research papers

2. Title: Automatic Rule-Mining for KeyPhrase and Entity Mention Extraction using Local and Global Features

* Abstract: When presented with large quantities of text data, automatic extraction of key phrases and entity mentions can provide valuable knowledge in about the overall content and meaning. In this research, we will investigate automatic methods for mining rule-based key-phrase extraction and entity mention detection leveraging global and local features. Such a method can provide interpret-able models that can be extended to a variety of text-segmenting tasks and more complicated models.

* Required Skills: Very strong with python with knowledge of basic algorithmic implementation. Ability to do data pre-processing / text cleaning. Some familiarity with NLP and Machine Learning toolkits in python (SciKitLearn, NLTK, TextBlob, etc)

3. Title: The Predictability of Relationship with Embedding Similarity in Heterogeneous Networks

* Abstract: Predicting relationships in a heterogeneous network is an important task, such as, co-authorship prediction in DBLP network and rating prediction in Netflix or Movielens data. The meta-path has been a popular method to generate powerful features for heterogeneous networks (PathPredict). However, the sparsity leads to a big challenge to bridge two nodes. Fortunately, the network embedding (PTE) represents different types of nodes with dimensional vectors, and thus, we are able to compute the similarity for every pair of nodes. Here comes the question: (1) Will the venue-venue similarity (SIGMOD-VLDB) improve the meta-path-based features

(Author-Paper-Venue-Paper-Author) for co-authorship prediction? (2) Will the paper-paper similarity improve the prediction, too? Or, will it bring too much noise because the case that authors do similar papers cannot be equal to a co-authorship. A general question is that given a specific type of relationship, can we generate the accurate prediction with statistical analysis and machine learning methods?

* Required Skills: (1) Programming with both C++ and Python. (2) Plotting nice figures with R/Matlab/Python. (3) Reading paper carefully.

4. Title: User Profiling with Massive Geo-Tagged Tweets

Abstract: With the proliferation of GPS-equipped mobile devices, more than 10 million geo-tagged tweets are published everyday. Such tweets provide an understanding of the user's activity from a 3W (what-where-when) view. Given a large collection of geo-tagged tweets, this project aims to perform effective user profiling. Specifically, we attempt to answer the following questions: What are the user's interests (sports, music, shopping, etc.)? Which regions does the user usually visit and when does he/she visit these regions? Can we find groups of users who share similar interests? The use profiling task is useful for various practical applications like targeted advertising and location recommendation.

Required Skills: (1) Ability to survey relevant research papers and learn about state-of-the-art techniques. (2) Basic knowledge in entity extraction, topic modeling, and classification. (3) Strong programming skills in Python/Java.

5. Title: User-generated Item Lists Recommendation

Abstract: User-generated item lists have become a popular functionality in many web services. For example, Foursquare users can create lists of venues (https://foursquare.com/lists), and YouTube users can create playlists of videos (https://www.youtube.com/view_all_playlist). A user can also subscribe lists created by others to explore new items. User-generated item lists help users organize and share items with others more efficiently. However, how to help users find the lists they might be interested in is still an open question. This project aims to study the properties of user-generated item lists and list subscription behavior of users, and more importantly, to design an effective recommender system for item lists. Specifically, given a target user, we rank candidate lists according to the user's preference.

Required Skills: Basic knowledge in recommender systems. Programming skills in Java (preferred) or Python.

6. Title: Data Mining for Precision Medicine: Drug Response Prediction for Cancer Patients

Abstract: The state-of-the-art drug development processes involve lengthy, expensive experimentation and testing. The average cost to develop a new drug that gains market approval is an astounding $2.5B. In silico, or computational, models aim to leverage existing knowledge and data to predict patient response without actual clinical experiments. Progress in this direction not only lowers the prohibitive R&D cost for developing new drugs but also helps guide clinicians in creating personalized treatment, which may lead to better chances of survival. We will investigate gene expression profiles and DNA sequences of real (anonymized) patients and create methods using various classification/embedding/pattern mining techniques for drug response prediction.

Required Skills: Proficiency in Python (and maybe Java) and experience using libraries/toolkits such as pandas and scikit-learn. Strong background in linear algebra. Familiarity with recommender systems and the learning to rank problem. Ability to find and understand relevant literature for biomedical background knowledge.

7. Location Based Social Network (LBSN)

* Abstract: The intrinsic rich features (location, time, textual description, category etc.) from location based social network (such as Foursquare) provide a more complicated setting to do research on. There can be new problems arising which do not exist in conventional social networks. Even the classic recommendation problem, when considered in this particular LBSN setting, becomes much more interesting and can be approached from many different perspectives due to a bunch of new dimensions. Location prediction/recommendation, user interest modeling, detect regularities in human behaviors, just to name a few, can also be cast into very diverse settings.
* Required skills: basic knowledge on DM and ML. Knowing statistical/probabilistic models is a plus. Programming skills: Python, C/C++, or Matlab.

8. Multi-typed Relational Data Clustering via Tensor Decomposition

Abstract: In this project, we aim to develop a clustering algorithm for multi-typed relational data based on tensor decomposition. We model the multi-type relational data as a high-order tensor (note that a matrix can be seen as a 2nd order tensor) and cast the data clustering problem as a tensor decomposition procedure. This project involves heavy theoretical analysis and light implementation.

Required skills: like theory, know Matlab

Reference: http://arxiv.org/abs/1010.2731
http://arxiv.org/abs/1010.0789
http://books.nips.cc/papers/files/nips24/NIPS2011_0596.pdf
http://arxiv.org/abs/1303.6370

9. Concept Definition Filling

* Abstract: Given a new concept/term like "Heterogeneous Information Network", the task is to find its definition through Web or a given corpus automatically. Traditional entity linking task prefers to link any existing concepts to a knowledge base like Wikipedia. For new emerged concepts or Wiki-uncovered concepts, it simply chooses to skip. A better solution is to augment their attributes including definition and other descriptions automatically. Based on our current research progress, a list of Wiki-uncovered concepts can be extracted from a scientific corpus. The proposed research problem is to enrich its content with mining and learning principles.
* Required skills: Good understanding of DM and ML. Good at programming especially Python and C++.

10. Outlier Detection

* Abstract: Outlier detection is an important task for data miners, with applications ranging from fraud/security breach detection to novelty discovery. The high-level

goal of our research is to extend classical outlier detection work in two directions: 1) leverage heterogeneous information from datasets to find better outliers, and 2) allow a way for the user to provide "hints" about what a meaningful outlier is (eg in the form of a query). Specifically, there are several ongoing research problems you could contribute to, depending on your interest: 1) given a network and a search query, how to find subnetworks that correspond to that query? 2) given a collection of small sub-networks, how to find which ones are outliers? 3) are there some parts of a network that are evolving "differently" (in an interesting way) from the rest of the network?

If any of these problems / directions sound interesting to you, send me an email and we can set up a time to discuss more.

* Required skills: Understanding of DM and ML. Strong programming skills in either Python or Java. Experience working with (or crawling) web datasets is a plus

2017-09-15