CS512-Spring-2013-Research Project Recommendations

This page contains a set of project topics recommended by members of the Data Mining Research Group at CS, UIUC. Please contact the authors of those recommended projects for more information.

Thanks,

-JH

Recommended projects:

1. Role (People, Paper, Term) Discovery via Information Network Analysis by Chi Wang

   Introduction

   * Abstract: In real world social or information networks, the entities exhibit different patterns of communication and interaction with each other according to their roles and relations. It is a useful data mining problem of uncovering the hidden roles from a network that link entities with semantic interactions. For example, in the academic publication network, we can study the roles of researchers, papers and terms like: 1) what is the role of each researcher in a co-authored paper (main author, supervisor or helper); 2) what is the role of each cited paper among the references of a paper (a seminal paper, one of the combined approaches, victim for comparison); 3) what is the role of a terminology in a paper (a method, a task or a modifier) and so one and so forth. There are many interesting problems we can explore.

   * Required skills: taken CS412, skilled programming using one of the following: C++, Java, Matlab, Python or C#

   * Contact Info: chiwang1@uiuc.edu

2. Efficient Search on Large-Scale Text-Rich Heterogeneous Information Networks by Lidan Wang

   Introduction

   * Abstract: Many heterogeneous information networks are text-rich. Examples include the DBLP network, social networks, and Wikipedia. It is often desirable to allow users to issue keyword search queries in real-time to obtain relevant textual information. For instance, a computer science student may want to identify the most relevant papers on "learning fast models" on DBLP network in order to learn more about the subject; another computer science student, new to the field, may want to find a ranking of relevant papers on "big data" written by well-known authors; another student, specialized in education, may also want to get relevant papers on "learning fast models", but from the aspect of how to enable people to learn fast. We want to devise a unified ranking framework to answer these and other network-based queries on text-rich heterogeneous information networks.

   * Required skills: Interests in information network analysis and search. Programming skills: Java, Perl, Python, or C++.

   * Contact info: lidan101@gmail.com Thanks for your interest. We already found students to work on this project.

3. Hierarchical Summarization of Mobility Data: A Mobility Cube Based System by LuAn Tang and Tobias Lei

   Introduction

   * Abstract: The advance of object tracking technologies leads to huge volumes of trajectory data collected by the devices of GPS, cell phones, and so on. The users would like to analysis the data with spatial and temporal dimensions. A motivation example is shown as follows. Example 1. The movebank system collects the trajectory data of bird migration from 2001 to 2010. The animal scientist would like to learn from the data that: (1) what is the major migration paths and the rest places for the birds in North America? (2) what are the major differences of the bird movements between 2009 and 2010? (3) what are the outlier route/resting places in 2010? In the above example, the users asks for a summarization of the data, rather than an individual trajectory. However, such analytic query processing is not well supported by current system, mainly due to following difficulties. (1) The trajectories are sequence of spatial-temporal data. It is intrinsic difficult to be integrated and compared. An analytic query usually involves thousands of the trajectories, the result is not easy to be shown in an effective way. (2) The query result is highly sensitive to the query scale (i.e., the query constraints on spatial and temporal dimensions). (3) Query efficiency on big data. In this study, we propose a moving network based method to process the analytic queries on big trajectory data set. Our contributions include: (1) Retrieving a moving network from the trajectory data to represent the major places and paths; (2) Summarizing the trajectories as numeric vectors, the data are compressed but the major spatial-temporal features are stored; (3) Efficient processing the analytic query by calculating the numeric vectors. The results are represented effectively based on the moving network. In order to solve the above problems, we want to construct a mobility network in a hierarchical way from trajectories such that we can use the network to generate summarized information, compare two trajectory sets and find outliers.

   * Required skills: taken CS412, skilled at Java.

   * Contact info: <LuAn Tang> tang18@uiuc.edu, <Tobias Lei> klei2@illinois.edu Thanks for your interests. We already found enough students to work on the project.

4. Outlier Detection for Information Networks by Manish Gupta

   Thanks everyone. I already have students assigned to both of the projects below.

   * TopK Outlier Cuboid (Projected Graph) Detection
* Abstract: Given a weighted graph (like DBLP), one can create multiple projected views across various dimensions. For example, for DBLP graph, the dimensions could be research areas and years. For each value of each dimension, a projected graph can be built. For DBLP, a (DM, 2010)-projected graph is one such projected graph. Projected graphs could be at multiple levels of hierarchy like (DM=DB, 2010-2012)-projected graph. Thus, there are exponential number of projected graphs. Consider a subgraph query Q: In which research area and in which set of years, there were exceptionally high collaborations between Stanford, IITBombay and Berkeley authors? In other words, given a subgraph query Q, find topK projected graphs for which the percentage edge weight covered by matches is highest. Since, there are too many projected graphs, the problem is challenging. "Percentage edge weight covered by matches" does not seem like a monotonic function and so Apriori-like pruning algorithms may not help. Genetic algorithms have been known to perform well when the search space is huge. Can genetic algorithms be useful to solve this problem of TopK Outlier Cuboid (Projected Graph) Detection?

* Required skills: Java or Python. Background on graph query processing or genetic algorithms is a plus.

* Contact info: My illinois userid is gupta58

---

**Outlier Substructures in an Information Network**

* Abstract: Consider the DBLP network of authors, conferences and title terms. Consider a simple query: Find all outlier Data Mining researchers such that they are "out-of-place" in the network. Thus, the problem is given A heterogeneous information network and a heterogeneous subgraph query, find outlier matches of the query in the network. For example, here the query is simply a vertex such that research area=data mining. Usually, some patterns can be found for the neighborhood of data mining researchers like 1. They are connected to other data mining authors, conferences or terms. 2. They are usually connected to very few very popular authors. To solve the problem of finding outlier authors with an unusual neighborhood, two steps are essential. 1. Given the query, one can find all matches in the network. 2. For a match, given the usual connectivity patterns and the neighborhood for the match a. One can compute p = probability of generation of the match b. Outlier score can then be computed as 1-p. NetClus (Sun et al. KDD 2009) can be adapted for a node-centric graph generation model while block models can be adapted for an edge-centric graph generation model. Which of these is a better way of modeling for outlier detection? Will this methodology lead to interesting outliers from graphs? How do these outliers differ from any other method?

* Required skills: Java or Python. Some background on generative modeling or willingness/motivation to learn them independently.

* Contact info: My illinois userid is gupta58

---

5. News Article Analysis via Heterogeneous Information Networks by Hyunsul Kim [Introduction]

**Abstract:** There are many studies on the heterogeneous information network analysis in many different tasks including clustering, classification, outlier detection, trend analysis, and trustworthiness analysis. However, the applications are only limited to well-structured data like DBLP networks, movie database, Yelp database, and IMDB database. This project is to extend the application domains of the heterogeneous information network analysis to the unstructured documents, especially news articles. Constructing a heterogeneous information networks from a collection of news articles is very challenging because documents are noisy and ambiguous. This construction is different from information extraction because the result network should be useful in heterogeneous information network analysis for documents. In other words, the result information network is not the end product. We want to use such information networks to analyze documents in a different way from the traditional document analysis methods. Since this direction is very open, you are more than welcome to propose interesting research problems in this project.

* Required skills: Programming Skills(Java/C++/Ruby/Python), Some background on NLP or willingness to learn them independently.

* Contact info: hkim21@illinois.edu

---

6. Efficient Spectral Clustering for Heterogeneous Information Network by Jialu Liu [Introduction]