NetMine Group Monthly Report

• First stage, process 4 datasets
  • Pubmed (Yucheng, Fangbo, George, Shan-feng, Haoyuan)
    • Construct the subset “Arrhythmias, Cardiac” to a data-cube => EventCube
    • Explore two topic hierarchies: MeSH/Gene Ontology. Treat them in different ways (associated to documents):
      • 1). Topic Hierarchy for Topic-Cube
      • 2). One dimension in EventCube / One type of node in Network
      • 3). Word-net for query expanding and relevant search (PUBMED)
    • Name disambiguation: 1). Missing full name, 2). Disambiguation.
  • DBLP dataset (Jialu, Fangbo, Marina):
    • Topic hierarchy construction with the help of Cathy/Wikipedia
    • Already get the latest dump of dataset from ArnetMiner group.
    • In the process of updating dataset and further cleaning
NetMine Group, cont.

• **Data.gov (George, Brad)**
  • Explore the dataset and design the schema.
    • Dates: single-date, date-range(?)
    • Geographical location: from lat/Ing coordinates to a hierarchy (city/county/state)
    • Publisher/Organization: A hierarchy for the dataset source.
    • Topic: First use “Community” and “tags” in data.gov
  • Parsed all the dataset already. Trying to generate a cube format first to put it into EventCube.

• **News Dataset (Xiang, Shi, some undergrads):**
  • Build the basic star-schema network based on Heng’s dataset.
    • Include: Year/Location/Agency/Person/Organization/Event Type
    • Running NetClus/PathSim/… on this initial network:
      • If Person A is relevant to Person B: on what kind of events? Example: Finding Nancy Pelosi and Scott Brown are related on topic “Health Care Bill”
NetMine Group, cont.

• Other directions:
  • OLAP over Heterogeneous Information Network (Yihan, Fangbo, Xiao, Lidan)
    • Defined the Concept: OLAP is based on the dimensions defined by Documents (paper, news articles). Those documents are in a network (DBLP, News, Data.gov, Pubmed). Each cell is a sub-network induced from the documents covered by this cell.
    • Seek for some network-based measures that traditional OLAP can’t solve
      • Local multi-typed ranking. Example: Cell<Topic=“HIN”>, find the local ranking for “HIN network”, only contain the papers and citations belong to “HIN”. Find most influential authors/conference/affiliations. “Yizhou Sun” vs. “Michael Jordan”.
  • Server issue (Jialu, Fangbo):
    • Have ordered a server with large memory, so that we can put all the four datasets into memory.