A Programming Language for Mining Fuzzy Entity-Relation Graphs

Azadeh Shakery
shakery@uiuc.edu
Introduction

- **Fuzzy Entity-Relation Graph**
  - A graphic notation for representing knowledge in patterns of interconnected nodes and arcs.
  - The declarative graphic representation can be used either to represent knowledge or support automated systems for mining information from the knowledge.
  - Components:
    - Entities
      - Name
      - [Category]
    - Fuzzy relations (Represent the degree or strength of association)
      - Name
      - Confidence
Introduction (continued)

- Fuzzy ER Graph
  - In many problems, the knowledge can be represented as a fuzzy ER graph.
  - We need a way to query such a graph in order to extract information from the data.
Examples of Mining fuzzy ER Graphs

• Concept navigation

• How do we support this in a general way?
A Mining Language

• **Current Supported Operators**
  - Finding the neighbor set of a given set of nodes

  - **NBSet**(X, L, O, C, t)
    - **X**: Set of Nodes
    - **L**: Set of Labels
    - **O**: Operator
      - **Avg**: Find all the neighbors with average weight above threshold
      - **Min**: Find all the neighbors whose weight of the weakest link is above threshold
      - **Max**: Find the top t neighbors which have the strongest link to at least one of the items in the set
    - **C**: Set of Node Categories
    - **t**: Threshold when the operator is either avg or min
      - **K** when the operator is max
A Mining Language (continued)

- **Current Supported Operators**
  - Finding the shortest path between two sets of nodes

  - **shortestpath(**X, Y, L, C, t)**
    - X, Y: Sets of Nodes
    - L: Set of Labels
    - C: Set of Node Categories
    - t : threshold
  
  - Finds a path with the minimum number of intermediate nodes between two sets X and Y with these constraints:
    - The label of each edge in the path is in the set L
    - All intermediate nodes are among categories in C
    - The weight of each edge in the path is above the threshold t
A Mining Language (continued)

- Current Supported Operators
  - Finding the top k highest weight elements of a given set
    - \textbf{topk}(X, k)
      - \textit{X}: Set of Nodes
      - \textit{k}: An integer
  - Finding the union/intersection of two sets of nodes
    - \textbf{union}(X, Y)
    - \textbf{intersect}(X, Y)
      - \textit{X, Y}: Sets of Nodes
Our System

X = \{\text{bee}\}
Y = \{\text{flower}\}
l = \{
\}
c = \{
\}
r = \text{shortestpath}(X, Y, l, c, 0.0024)

Entity-Relation File
Pollen bee 100
pollen 0.409966
spermatophyta 0.0696191s
spermatophytes 0.0696191
vascular 0.0696191
plants 0.0687397
...

Entity Category File
behavior 171
solicitation
declaration
reproductive
signal
predator
feeding
attack
...

Data Processing

Index on relations

Index on categories

Miner

Result

flower 0.00773875
chelicerates 0.0143616
bee
Current Data

• Bee Data
  – 1200 records about apis mellifera (honey bee)
  – ~1.5 MB

• Fly Data
  – 3600 records about drosophila (fruit fly) genes
  – ~6.5 MB

• No Stemming

• No Removal of Stop Words
Current Data

- In our experiments, we use words as entities
- We used the mutual information between words as the relationship between the entities

\[ I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \]

- Compare the probability of observing \( x \) and \( y \) together with the probabilities of observing \( x \) and \( y \) independently
  - If there is a genuine association between \( x \) and \( y \), then the joint probability will be much larger than chance \( P(x)P(y) \) and consequently \( I(x, y) \gg 0 \)
  - If there is no interesting relationship between \( x \) and \( y \), then \( I(x, y) \approx 0 \)
Current Data (continued)

- Part of MI on the bee data

pollen 100
pollen 0.409966
spermatophyta 0.0696191
spermatophytes 0.0696191
vascular 0.0696191
plants 0.0687397
angiospermae 0.0677029
angiosperms 0.0677029
plantae 0.066736
dicotyledones 0.0651498
dicots 0.0651498
nectar 0.0572719
...

Demo