Predicting Employee Interaction in Large Enterprise Heterogeneous Information Network

Hill R., Nguyen V., McEnroe M.

University of Illinois at Urbana-Champaign

rwhill2@illinois.edu, vnnguye2@illinois.edu, mcenroe2@illinois.edu

May 10, 2016
Overview

1. Project Scope and Goals
2. Related Work
3. Problem Definition
4. Data Description
5. Classification Models
6. Results
7. Future Work
**Background**

**Link Prediction** - Given a snapshot of a social network at time $t$, we seek to accurately predict the edges that will be added to the network during the interval from time $t$ to a given future time $t_0$.

Link prediction is an important task for analyzing social networks

- Applications in domains like, information retrieval, bioinformatics and e-commerce.
- Used as the basis of recommendation system in social networks

Much previous work has focused on linked prediction studies on homogeneous networks

- Probabilistic Models (Markov Random Field)
- Feature-based Link Prediction

Realistically, most networks are of the heterogeneous form
Background

**Homogeneous Networks** - treat all entities as one node type: actor, director, producer all same type of node - compute ’six degrees of separation’ across all node types.

**Multi-relation and multi-mode Networks** - Link semantics can vary. Person A can like a tweet from Person B. Or Person A can follow Person B. Or Person A can retweet a tweet from Person B.

**Composite Networks** - a collection of independent networks interact because in the real world, the nodes (people) exist in both networks and behavior reinforces in the different networks. ”I like you in Facebook, I follow you in Twitter”

**Heterogeneous Networks**

- Multi-relational network - with single-typed node but links can be multi-typed
- Bipartite Network - has exactly two node types
- Star schema - connects multiple node types - target node serves as hub or center of schema
Goal: Predict future interactions among employees of a large corporation using data from an enterprise social network.

- Extend the previous work of PathPredict in heterogeneous networks to enterprise networks
- We show that the novel measure on hierarchical enterprise paths provides great precision of a co-worker prediction task

Implications: There are several reasons link prediction in enterprise networks is of interest.

- Large enterprises find success through collaboration among the members of the enterprise
- Becomes specific business goal to increase collaboration
Meta-Path captures a sequence of relations between different object types in a heterogeneous network. Some applications:

- Learn a **ranking model** to answer user queries based on similar queries as joined by a meta-path
- **Mining outliers** based on dissimilarity measure
- **Classification** of objects
- User guided **clustering**

**PathPredict** A meta path-based relationship prediction model that works on the problem of link prediction in the heterogeneous networks.

- Operates by extracting topological features from a network, a supervised model is used to learn the best weights associated with different topological features in deciding relationships
**Extract topological features** Traverse the network schema using standard graph traversal mechanism such as Depth-First or Breadth-First

**Apply a measure function:** weighted or unweighted - random walk, symmetric random walk, path count, normalized path count

Given pairs of nodes, **Learn weights associated with extracted features**

PathPredict uses logistic regression

\[ p_i = \frac{e^{x_i \beta}}{e^{x_i \beta} + 1} \]

\( \beta \) is the \( d + 1 \) coefficient weights associated with the constant and each topological feature

Maximum Likelihood Estimation (MLE) maximizes the likelihood of all the training pairs is then found to derive \( \beta^i \):
PathPredict and other meta-path based models fail to capture link strength in the task of link prediction. [Massung]

In the context of co-author relationship prediction on the DBLP bibliographic heterogeneous information network:

- **Topical features**: random walk, path count, normalized path count
- Consideration of weight based on similarity.
- Standard similarity measures: Jaccard, Cosine
- Similarity measure used in metapath creation
- **A P SP P A**: A and B are Authors and Papers, SP is a similarity node representing how similar two papers abstracts are.
- **New topical features**: Weighted random walk (WRW), Weighted symmetric random walk (WSRW), Weighted path count (WPC), and Weighted normalized path count (WNPC).
Heterogeneous Enterprise Network

Actual data from a large enterprise company from first quarter 2016

- Employee meeting invitations
  - A single organizer invites multiple people to a meeting, creating star networks
  - People are invited to meetings by different people, new links are added to the network
  - Sets of people invited to one meeting are invited to other meetings, giving their relationship weight

- Internal company social network
  - Each employee can have a web page where they can post information
  - Other employees can comment on a post, mention a person or like contributions
  - Interactions create a network between participants

- Employee-manager tree hierarchy
  - Branches represent the chain of command from worker to CEO
  - Traversing up the tree and down the tree creates a distance measure
The Co-Worker Prediction Task

**Task:** Predict future work relationships in a corporate network, using meeting, manager hierarchy, and tSpace data as a training and test set.

- Introduce a modified hierarchy based version of PathPredict.
- Acquire two snapshots of social network, meeting, and hierarchy data (Training and test)
- Define a set of features hierarchy features in conjunction with meeting and t-space social network data.
- Train a set of classification methods on the training set and verify on the test snapshot.
Large enterprises are extraordinarily concerned with employee, partner, and customer privacy.

- Our data is actual data about employee collaboration never before released outside the company.
- Project participants signed a non-disclosure agreement
- All user ids were anonymized
- Many data fields were not released to the project team
- Specific participants (vendors, partners, other companies, customers) were removed from the data
- Various data elements had to be aggregated to prevent data analysis outside the scope of the project
- Constant concerns about re-identification
- The company at this time wishes to keep its data donation anonymous
Data Fields Available - 62 total fields
- Organization descriptors (e.g., business unit number, business unit name)
- Individual unique descriptors (e.g., first, last, middle name, id, email)
- Geographical and location descriptors (e.g., country, street address)
- Job descriptors (e.g., level, title, job title)
- Hierarchy (e.g., supervisor id, chain of command string of ids)

Data selected for the project
- worker id
- supervisor id
- chain of command - string of ids from worker to CEO
Meeting Invitations Sent via Email

- **Data Description** -
  - 28 data fields available out of 50 in MS Outlook spec
  - 3 months of data - 27 GB containing 31 million entries
  - Removing duplicates leaves 16.8 million individual meetings
  - Finding all actual combinations of two people and weighting the number of meetings yields 78 million pairs

- **Limitations of the Data**
  - Some departments and merged entities have unusually formatted emails. They were excluded from this project
  - Emails from non-company personnel were removed to prevent inadvertent leak of proprietary or personal information. This could cause some meetings to look smaller, sometimes dramatically smaller.
  - No information about who actually attended meetings

- **Data Selected for Project**
  - Account Name
  - Start Time
  - Required Attendees
  - Optional Attendees
Data Fields Available

- Based on IBM Open Connections
- Has API, but simple XML decoding was sufficient for our purpose
- Ten different 'verbs' describing interactions, of which we used 'like', 'mention', and 'commented'

Limitations of the Data

- Project team was authorized to query the internally public activity feeds, which in general were limited to 30 days of activity.
- Rate limiting on the social network server prevented harvesting the entire company in the timeframe of this project and limited the length of data acquired.
Time Periods and Data Statistics

Data for feature preparation (training) from the month of February

- Contains 78 million lines of meeting records between a subset of all possible pairs of employees at the company
- Only a small fraction of all possible pairs (around 250 trillion possible pairs)

Dataset scaled down given the limited time and computing resource.

- Sampled around 160,000 pairs of employees
- Subset comprises 40,000 pairs of employees who have had interactions on tSpace (positive examples) / 120,000 pairs who have not had interacted before (negative examples)

Data for testing - actual prediction period - from the month of March

- Comprises of 4,500 positive examples and 15,500 negative examples
Features

- About 170 unique paths (length of up to 24) connecting any two employees (for e.g., see below table)
- Numeric feature counting the number of meetings a pair of employees share

<table>
<thead>
<tr>
<th>Path</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>A Reports to B</td>
</tr>
<tr>
<td>M</td>
<td>A Manages B</td>
</tr>
<tr>
<td>R-M</td>
<td>A and B are co-workers (both managed by C)</td>
</tr>
<tr>
<td>R-R-M</td>
<td>A is junior to B in rank</td>
</tr>
</tbody>
</table>

**Table**: Examples of hierarchical paths between the employees
Support Vector Machine In the (soft) SVM model, one tries to do the following optimization problem:

$$\arg \min \left\{ \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \right\}$$

subject to (for $i = 1, \ldots, n$)

$$y_i (w \cdot x_i - b) \geq 1 - \xi_i, \xi_i \geq 0$$

For nonlinear classification, the model uses the Gaussian radial basis function as its kernel

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

for $\gamma > 0$. 
A very popular technique among the ensemble machine learning methods

- Random Forests are a way of averaging multiple decision trees, trained on different parts of the same training set, with the goal of reducing the variance.

- The most important detail that distinguish Random Forests from bagging algorithm is that at each candidate split in the learning process, we consider only a random subset of the features.
Classification Models - Hyperparameter Tuning

SVM optimal parameters:
- $C = 3593813.66$
- $\gamma = 1.668 \times 10^{-8}$

RF optimal parameters:
- Number of trees = 200
- Max tree depth = 50

Figure: Support Vector Machine/Random Forest cross-validation grid search.
Classification Accuracy

- The SVM model does a good job identifying the negative examples (99.3% accuracy), however, it does poorly on the positive ones (17.7% accuracy). Overall accuracy is relatively low at 80.3%
- The Random Forest model, on the other hand, does extremely well identifying both positive and negative examples (100% and 97%, respectively). Overall accuracy is 97.8%

<table>
<thead>
<tr>
<th>Label</th>
<th>SVM 0</th>
<th>SVM 1</th>
<th>RF 0</th>
<th>RF 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15116</td>
<td>100</td>
<td>14754</td>
<td>462</td>
</tr>
<tr>
<td>1</td>
<td>3802</td>
<td>820</td>
<td>0</td>
<td>4622</td>
</tr>
</tbody>
</table>
The most important feature, possessing a little over 50% predicting power, is the meta-path 'manages', or 'A manages B' for a pair of employees.

In fact, most of the chained 'manages' meta-paths make top-10.

The feature 'numOfMeetings' does make top-10, however, is quite insignificant.
Future Work

- The techniques applied here can be reapplied to the same data sets, but with significantly longer time frames under study.
- Additionally, there are hundreds of additional unique data sets available in large enterprises and different meta-paths can be explored with those data sets. Future work can additionally use these features to further expand the semantics of metapaths.
- Another area of interest in our survey and in our initial proposal was to develop techniques to deal with network change.
References

Please see Project Report Reference page.
The End