CS: 8803 Advanced Internet Applications Development

Mining Social Networks

Abhishek Saxena, Ankit Kharadi, Chirag Rajan
{asaxena3,akharadi3,crajan3}@mail.gatech.edu
Objectives:

The growth of social networks has been nothing short of phenomenal. Every day social networks such as Facebook add literally hundreds of users. These users bring with them an enormous amount of data that if mined effectively could prove useful both from commercial and scientific purposes. However attempts to mine social networks have all encountered significant problems due mainly to the sheer scope and diversity of data available. Thus while mining social networks is lucrative from both a monetary and scientific point of view, it is easier said than done. For our term project in the course Advanced Internet Application Development we have built a tool which mines social networks and helps visualize relationships between different entities. Our system uses DBLP as the social network to mine and finds the relationship between two authors and tries to establish the closeness between them based on a coefficient based on regression analysis. This coefficient helps us infer how two authors are related and how close they are in terms of research publications.

Research problems addressed:

One of the most active research areas today (as a function of the number of papers published) is in the field of mining social networks. However we find that most of the papers propose several interesting new algorithms with mathematical proofs justifying their effectiveness. We feel that there is a need for a more 'systems' oriented approach to this problem. Thus we have built a system which applies data mining techniques to social networks. This tool uses a variety of different components and can easily be adapted to run new algorithms thus ensuring flexibility and usability in the future.

Goals of our Project:

Having defined our objective as to build a system to mine social networks, we set about defining a clear set of measurable goals. Our goals were as follows:

- To mine the social network DBLP. We chose DBLP because it is a collection of research papers by authors from universities around the world. Mining DBLP had several advantages over a conventional social network such as Facebook, namely:
  - DBLP did not have any privacy restrictions like conventional social networks, information was available freely for anyone to access.
  - DBLP offered much more structured data for us to use than other conventional social networks. This made mining and formatting much easier.
  - Researchers always had their own web pages which made information about them much easier to extract.
- To visualize the results of our mining using graphs so that users can easily understand the various complex relationships between authors.
Key Concepts:

Before proceeding further the reader should be familiar with the following key terms for a clear understanding of our system:

Crawler:
A web crawler (also known as a web spider or web robot or - especially in the FOAF community - web scutter) is a program or automated script which browses the World Wide Web in a methodical, automated manner.

Parser:
In computer science and linguistics, parsing is the process of analyzing a sequence of tokens to determine grammatical structure with respect to a given (more or less) formal grammar. A parser is thus one of the components in an interpreter or compiler, where it captures the implied hierarchy of the input text and transforms it into a form suitable for further processing (often some kind of parse tree, abstract syntax tree or other hierarchical structure) and normally checks for syntax errors at the same time. The parser often uses a separate lexical analyzer to create tokens from the sequence of input characters.

Named Entity Recognizer:
Named entity recognition (NER) (also known as entity identification (EI) and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

For example, a NER system producing MUC-style output might tag the sentence,

\[
Jim \text{ bought } 300 \text{ shares of Acme Corp. in 2006.}
\]

\[
<\text{ENAMEX TYPE="PERSON"}>Jim</\text{ENAMEX}> \text{ bought } <\text{NUMEX TYPE="QUANTITY"}>300</\text{NUMEX}> \text{ shares of }<\text{ENAMEX TYPE="ORGANIZATION"}>Acme Corp.</\text{ENAMEX}> \text{ in }<\text{TIMEX TYPE="DATE"}>2006</\text{TIMEX}>.
\]

Regression:
Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data.

Visualization:
Visualization is any technique for creating images, diagrams, or animations to communicate a message.
System Architecture:

Our system has been constructed as follows:

The major components of our system are:

Crawler:
As described previously, crawler is that component of a search engine that automates the process of visiting web pages. For our project, we used a combination of crawlers. Initially we had used the crawler Web Sphinx developed at Carnegie Mellon University because of in addition to crawling web pages it allowed us to save the contents of those web pages. In addition to this crawler we also wrote our focused crawler which applied a classifying algorithm and had a higher harvest rate (ratio of number of relevant web pages crawled to total number of web pages crawled) than a conventional crawler. This enabled us to save on disk space.
**Parser:**
We have used a parser to transform HTML output into plain text. This is achieved using a standard HTML parsing library which we use to cycle through a given folder in which we store all the researchers web pages. We then parse into text one file at a time through the entire folder. This process is repeated for every new file. The plain text thus obtained is now written into a new file.

**NER:**
NER stands for Named Entity Recognizer. We have used the NER library developed by Christopher Manning's group at Stanford. Through this library we are able to extract the most vital information related to a given researcher such as the papers he has published, in what conferences and in what years.

**MATLAB Regression Module:**
The MATLAB regression module is at the heart of our system, it performs regression analysis on a given set of author matrices to predict their relationship. We have collected information of around 4000 authors, and built a matrix relating each author to each other author thus having a total of around 16 million entries thus performing regression on a matrix of this size required significant computational power which MATLAB provided. The information matrices represented data concerning 4 popular conferences such as IEEE,SIGMOD,KDD.

**Visualization Module:**
Our visualization module is used to represent the results of our mining algorithms. We investigated several powerful JAVA libraries but ultimately settled on JUNG because it offers more powerful and well equipped architecture for visualization than the others. We also considered using JgraphT and Jgraph.
Working:

The working of our system is as follows:

➔ First we gather data from DBLP website about researchers and their interests. This is done by visiting individual researchers web pages and also looking at their associations on DBLP.

➔ We then save the above data in HTML format in a folder on our hard disk so that it can be asked later.

➔ Now we use an HTML parser to convert the HTML output into plain text which we can use subsequently for computation.

➔ The resulting text is now used as input for a program written in JAVA which applies the Stanford NER library to find names of authors, paper titles, conference names and years of publication

➔ Once this is done, we now apply regression and expectation maximization algorithms to the various researchers to get a matrix output representing the relationship between authors at their various conferences

➔ We now take from the user as input the authors between whom the relationships are to be determined

➔ Using a JUNG library we represent the above matrix results in a visual format so that users can see the exact relationships between authors
Screen shots:

The diagram below represents a sample output of our application. The red nodes represent authors and the black edges represent relationships.
Results:
Given an input of 3 researchers we can find the coefficients associated with each conference as:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIGMOD</td>
<td>0.145</td>
</tr>
<tr>
<td>KDD</td>
<td>0.602</td>
</tr>
<tr>
<td>ICDE</td>
<td>0.181</td>
</tr>
<tr>
<td>VLDB</td>
<td>0.072</td>
</tr>
</tbody>
</table>

Thus from the above output we can conclude that the researchers are connected more by KDD conference than any other.

Applications of our project:

Friend Suggestion:
One of the most popular applications of our project could be applying it as a friend suggester in a social network such as Facebook or MySpace. Both these networks have several thousand users and many of the relationships are very complex. Using our mining techniques it is possible for users to have access to relationships they did not know about.

Targeted marketing:
With the increasing awareness of todays consumers to marketing, it has become increasingly important for advertisers to target consumers intelligently and display advertisements that are catered specifically to a particular user. Using our tool marketers can easily refine to a greater extent the tastes of consumers and target the marketing more effectively.

Network Prediction:
Using mining techniques we can find the topology and structure of networks and predict future congestion points and design the topology accordingly.

Future Work:

Statistics Display:
As the first extension of our project we plan to display statistics about different authors and their publications. At this point our tool does not display the data about how the author data was gathered, we are currently working on this.

Reverse Engineering:
We can use reverse engineering and extend the functionality to our project. As an input we can take the different coefficients related to various conferences and return the authors who are associated with that value.
Conclusion:

In conclusion we are grateful for getting the opportunity to work on such an interesting research problem. Through out this project we have tried to approach the problem with a 'systems' perspective. Our approach has been to further the cause of research in the social mining area by building tools that will help researchers apply their ideas quickly and effectively. Our solution has a flexible and modular architecture that we hope researchers will build upon. Through this project we have been fortunate to work with a variety of different technologies from crawlers to parsers to matrix computation languages. We also learned much from discussions held in class, with several of our ideas on crawlers coming from reading suggestions on the class reading list. Having worked in the realm of social mining for the last few months we sincerely believe that it will be a challenging and worthy research problem for any graduate student to take up and we would only be too happy to assist future graduate students in any way we can should they choose to do so.
References:

1. Inferring Trust Relationships in Web-based Social Networks, JENNIFER GOLBECK, JAMES HENDLER University of Maryland, College Park, Computer Science Department {golbeck, hendler}@cs.umd.edu

2. Friends and foes: Multiple relationship types in online communities Tad Hogg, Gabor Szabo, Dennis M. Wilkinson, and Michael J. Brzozowski Information Dynamics Lab, HP Labs


