Tracking a specific local community within the web graph

Thesis proposal

Frank Koopmans (f.koopmans@student.ru.nl)
Information Retrieval & Information Systems (IRIS)
Radboud University Nijmegen, The Netherlands

April 2008

Abstract

A web community is a set of web pages created by individuals or associations with a common interest on a topic. There are different techniques that can identify this kind of (local) communities within the web graph. If we could keep track of a specific community on the World Wide Web, commercial opportunities arise. Advertisers could reach their target consumers better and product developers could follow the latest trend within a set of consumers. This research will focus on identifying the right community and then tracking it as it evolves through time.

Keywords: web graph; local communities; centrality; community tracking; target advertising

Introduction

The Internet is a rapidly changing and evolving environment. The size of the Internet is incredible, more than eleven billion web pages exist today and this number is ever increasing.

Several theories have been developed to study the Internet. One of these theories models the Internet as a graph, this graph is called the web graph (Broder et al., 2003). Web pages are represented as nodes in this graph and hyperlinks between pages as edges in this graph. The graph can both be directed, incorporating the direction of the hyperlink, or undirected. Since the Internet is now modeled as a normal graph, all know graph theory can be applied to it. Another important aspect of the web graph to note is that page content is not incorporated in the graph model.

This web graph can be used to derive all kinds of interesting properties. Several techniques exist to determine an importance value for a node (website) in the graph. This is known as the centrality of a node. In a more social context, prestige can also be used as a measure of importance (Rupnik, 2006). Identifying communities in this graph is also a well-studied topic (Hinne, 2007).

Problem description

A web community is a set of web pages created by individuals or associations with a common interest on a topic (Yang Nan et al., 2006). For example, such a community could be a set of related and connected websites that discuss a specific football club.

When viewing the web as a graph, one can identify these communities by analyzing the nodes and edges using regular graph theory. Researchers have found several techniques to do this; it is quite a popular field of research. Each technique has its pros and cons. Popular algorithms include HITS, bipartite cores algorithm, MCL algorithm and the minimum cut framework.

After finding a community on the web using such algorithms, it may be necessary to filter the result. Advertising links on the web done by content matching may cause "unrelated" links within the web graph. For instance, the football club Ajax may be mistaken for the JavaScript technique Ajax. So if a football fan site has a lot of topics on Ajax, advertisers may have created links to the Ajax/Javascript sites. If that kind of problems surface some kind of filtering will be necessary (Bekkerman et al., 2005).

Finding a specific community on the World Wide Web is one thing, keeping track of that community after it changes, shrinks or grows is another. A community will change over time. Users might talk other topics and relate to other websites, but although there are changes we would still like to talk about the same community. Can we define a certain set of core features that can be used to identify the same community in another dataset?

There is commercial interest for this topic because if one could find and track a certain group of people on the Internet, one could better target advertising on the right people or try to predict the next trend for a certain set of consumers/local community.

So the problem is twofold. In order to track a specific community within the web graph one needs to identify the proper community first. Then, when a new dataset is available, we need to find a way to identify the same community again. If we manage to do that, we could compare the first community with the latter and see the mutations.

Research question

The problem described earlier leads to the following main research question:

How can a specific (local) community be identified and tracked through its evolution in the World Wide Web?

This question can be broken down into a set of questions that will need to be resolved in order to answer the main question. These will help break the research into manageable smaller problems.

1. How can we identify a (local) community within the web graph?
2. Is there need for extra filtering after identifying the community, like disambiguation?
3. How do we keep track of a community after it evolves?
4. Can we show the mutations of a community?

Products

These products will be the final result of the research.

1. Master Thesis

   The master thesis will be a written report describing the results of the research project. The study is conducted by two members but they will both study a separate subject. This will result in a master thesis with a shared introduction and an individual section and conclusion about this research.

2. Presentation

   A presentation aimed to present the key achievements of the research project.

Global planning

The following table shows the global project planning:

Orientation:	March+April
Sub questions 1+2	May
Sub question 3	June
Sub questions 3+4	July
Finishing up + Presentation	August

Overview of holidays during the research project period:

31 April – 5 May
30 May – 2 June
27-29 June
12-18 July

The research project yields three deliverables.

- A project proposal after the orientation period.
- A master thesis after the research and writing period.
- A final presentation when the project is finished.

Project conditions

The research is conducted partly together with Willem Elbers and the research project is supervised by Theo van der Weide.

Meetings are scheduled roughly every two weeks. In the beginning there will be more meetings, once a week, to prevent problems. During the research there will be less meetings and during the end of the project the two week interval will be used again.

During these meetings we discuss the progress and the current state of the research and try to identify possible problems. Besides these meetings any problems or questions can also be discussed with the supervisor by email.

References

Bekkerman, R., McCallum, A. (2005). Disambiguating Web appearances of people in a social network, Proceedings of the 14th international conference on World Wide Web, May 10-14, 2005, Chiba, Japan

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2003). Graph structure in the web (Tech. Rep.).

Hinne, M. (2007). Local identification of web graph communities.

Rupnik, J. (2006). Finding community structure in social network analysis - overview (Tech. Rep.). Department of Knowledge Technologies, Jozef Stefan Institute.

Yang Nan, Meng Xiaofeng (2006). Identify implicit communities by graph clustering. Wuhan University Journals Press. Volume 11, Number 5 / September, 2006