Aiming at improving the representation of communities and their relations using local algorithms for community detection.

Name: Michael Kizito

Student number: s0535265

Supervisor: Prof. Dr. Th. P. van der Weide

Thesis Proposal

January 03, 2008

Abstract

A community has a number of varying definitions but we shall at define it in two ways. Firstly a community is a subset of nodes on the network such that nodes in the same community are more likely to be connected than nodes in different communities. Examples could be division of social networks in groups, division of biological networks, routing in communication networks and so on.

There are algorithms that are used for community detection and these are the ones we intend to look at in this thesis and later on come up with an improved way of how to represent communities and their relations. We shall look at various networks ranging from social networks to biological networks so as to come up with a more appropriate way of community detection.

Contents

1 Introduction

In this day and age many people have resorted to the World Wide Web to make themselves known or as a means of disseminating information about a given topic or subject. Individuals who choose to publish information about themselves do so by using web pages that are referred to as Weblogs or Blogs. Since many of these individual weblogs are for particular people it is such a big task to identify a set of weblogs that form a natural group simply because the content may differ considerable for social reasons. Furthermore, social influence can be used to explain the behaviour of individuals by their thoughts, feelings and actions either directly or indirectly.

Some traditional methods have been devised as a means of findings communities and they focus exclusively on topology analysis. The Web has gained interest in the scientific community and this has resulted in various studies concerning a wide variety of topics. Researchers in this area look at the structural properties of the World Wide Web. The contents of the websites are not of much importance and therefore the Web is seen as a graph.

1.1 Web Graph Theory

There are millions of HTML pages in the Net and most of them contain links. These links normally have a direction. Some pages receive much attention and have many pages referring to them: software repositories, international organizations, pages written by acknowledged experts in various fields etc. Conversely, there are pages with many links, such as resource lists. Some clever people realized that they could build better search engines by relating the two types that is the authority and the librarian as illustrated below:

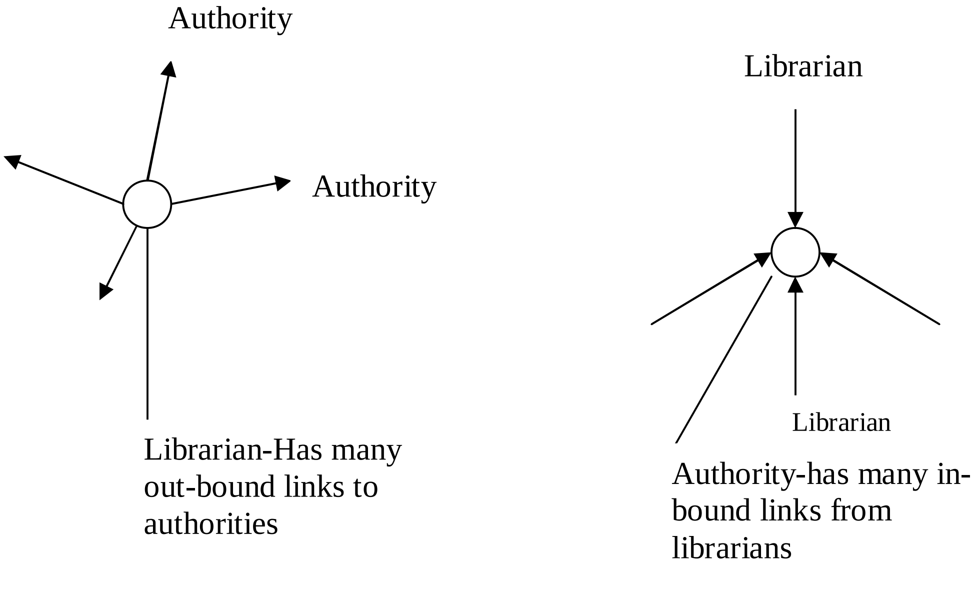
Authority
Authority
Librarian-Has many out-bound links to authorities
Librarian
Librarian
Authority-has many in-bound links from librarians

Figure 1 Illustrating the idea of an Authority and a Librarian

The above is an all encompassing definition but it works quite well. Iterative algorithms can take usual text-search results, thereby sorting out things and producing the "authorities". The commonly used search engine Google does something similar. Any extension of these ideas can lead to finding concepts and applications of Graph Theory in the Web map.

The Web is a huge directed graph G= (V, E) where the set V of vertices is the set of all pages and the set E of arcs corresponding to all pointers. Looking at the figure 2 the page X points to Y and Z and is pointed to by pages W and U. Vertex X has both outdegree and indegree equal to 2.

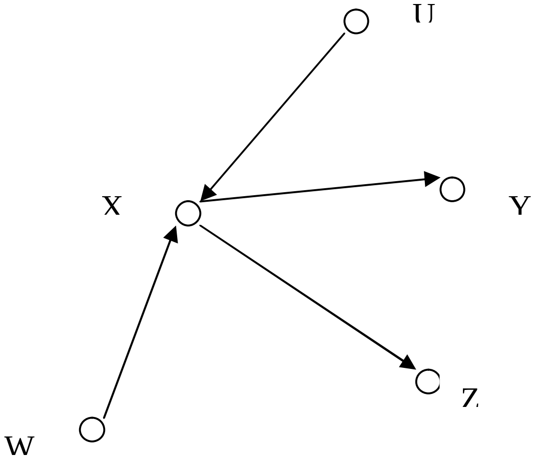

Figure 2 Illustrates the indegree and outdegree of vertex X

It has been noted that there is no link between two pages and so we can say that the Web is not a connected graph. It is rather a sparse graph. The pointers are much less in comparison to the maximum possible. One may wonder why we say that the Web is not a connected graph and yet from the definition of the Web almost every page is accessible to anybody. Our argument comes from the idea that everything is reachable from everywhere just by links. Furthermore there are isolated groups of pages with no link from or to the outside world. There can be isolated vertices not pointing to or being pointed to by anything. Examples of some these can be personal pages with just a CV, without pointers and only accessible only from their known URL.

After giving some little information about the Web we shall from now on stay within a subgraph G' of the Web. By this we mean some subset of the Web pages together with the arcs that have endpoints in this subset. The subset could be a “community” of pages on a particular topic, or all commercial pages in some domain. When we have four pages each pointing to or being pointed to by others, we say we have a complete subgraph in the Web. For instance, “if there is a rule that if I point to you and you point to P, then I will also put a link to P.” As a result we have complete subgraphs called tournaments with a pattern in the in-outdegrees of the vertices. We have an illustration in figure 3.

A minimal subset of vertices from which every other vertex in G' can be reached immediately is called a minimal dominating set of G'. This can be taken as an economic collection of bookmarks. We now arrive at the dominance number which is the number of vertices in a minimal dominating set of G'. In a specialized community with “outgoing” members, it is most probable that there are various sequences of pages each pointing to the next one, starting at P and ending at Q. If we take the arcs to have lengths of one, the shortest path should be the sequence with the minimum number of steps. Arcs may have varying lengths for example the distance between the hosting servers. The distance could be in terms of geographical distance or connection time and many more. The shortest path would approximate the fastest way to locate page Y starting from X. People have come up with a unique way of getting from every page to another. There is also the root which is not

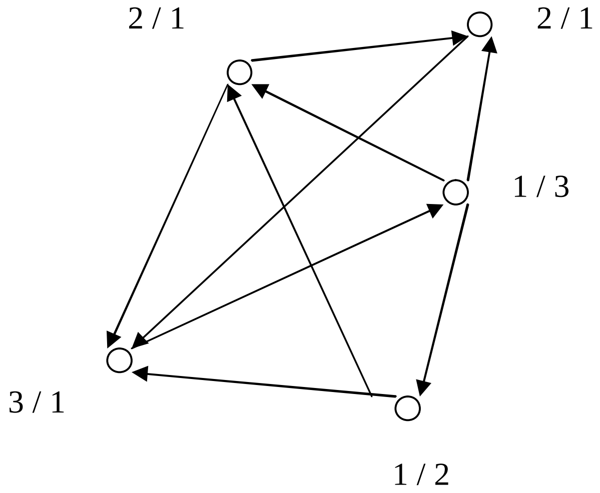

Figure 3 A pattern of in-outdegree of vertices

a page of particular importance per se but it is the best location to start building the optimal tree. Furthermore we can say that the root is not a resource list as it can point to only one other page for instance. In the collection of arcs, we cannot go back to the starting page and this comes from the definition of the tree.

As a way of building a bookmark list, we consider a time of fixed length say L. Each one of the non-bookmarked page has some connection-based measure from the list and this measure is taken from the nearest bookmarked page. In order to cover all the community in such a way as to minimize the sum of all these connection based measures (distances), we are in essence solving the minisum problem otherwise known as the L-Median problem. In practise this problem is seen as a way of installing facilities in an optimal way for customers, and this would mean locating of mirror sites.

On the other hand we have the L-Centres problem where we try to minimize the maximum connection based measure. It is slightly the same as the facilities location problem but here the idea is more or less avoiding extreme delays in serving individual customers rather than the community as a whole.

2 Thesis Statement

The problem of community detection is rather challenging and has been the subject of discussion in various disciplines. All existing methods intended to work out the community structure in complex networks, require a definition of community that imposes the limit up to which a group should be considered a community. A number of search engines use web crawlers and these normally collect all the information that is contained in a given blog or website in other words they create a copy of all visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

In this thesis we shall try to come up with the most appropriate way to come up with the different community networks. There is no step-by-step approach to managing complexities in the process of community detection but based on the already existing algorithms we shall aim at suggesting an appropriate way of coming up with the root or starting vertex or the entire node network.

The concept of community itself is qualitative—nodes must be more connected within its community than with the rest of the network. However there are some quantitative definitions that came from sociology and have been employed in recent studies [1], but in general the physics community has a widely acceptable measure for the strength of the community structure generated by an algorithm [2].

2.1 Objectives

After reviewing literature and getting a better understanding of the community detection process through studying some algorithms and how they work and the current short comings, and also understanding the concept of community through studying some of the available communities.

The main objective of this study is to come up with an improvement of some of the available algorithms or come up with some suggestions to help make the community detection process less tedious.

To achieve this objective, the study will be based on and guided by the research questions as follows;

2.2 Research Questions

- What is the quality of the community detected with a local algorithm?
- What group properties can be derived, for example is there a community leader?
- Can examples from biology be helpful, for example dominance detection in animal packs?

Using the first question, we intend to define, explain and describe what a community is, and evaluate the algorithms used in identifying the communities. This will be vital understanding the theme (community) of the study and we feel will uncover vital information like the algorithms' shortcomings and weaknesses that will be vital in the attempt to develop a better solution.

Since we intend to derive properties of the different blogs, the second question will be used to attain a deep understanding of the connectivity of the World Wide Web as well as come up with a tool which could be used to come up with the connectivity of a large graph (WWW).

The third question will be used to compare the available Weblogs. We shall look at examples to the biological environment to help us integrate knowledge attained using the other questions in order to come up with our conclusion.

3 Approaches/Methods

This section briefly describes the different methods that are to be employed for this research for data collection and analysis, subject familiarization and internalization.

3.1 Methods

3.1.1 Literature Study

Literature review will be the main knowledge reference method for this study and this will involve various scientific journals, articles and other publications about community detection, and the tools and methods that will be used for the research and solution development. Likely literature will be about Web Graph theory, Community detection algorithms.

3.1.2 Weekly meeting discussions

We shall have weekly communications by email with my supervisor Dr. Th. P. van der Weide that will involve guidance, correction and approval of the research progress and findings. Also in these meetings formal methods will be discussed to get more knowledgeable about the topic.

4 Work Plan

The plan that is shown below is a tentative one that may be review from time to time after consultation with my supervisor. But as it stands we shall try to follow it to the end of the thesis/project. .

Week	Task	Deliverable
1-2	Plan or proposal	Project plan
3-4	Community Detection	Write up answering research question 1
5-6	Come up with properties of Communities	Write up of answering research question 2
7	Comparison of the Weblogs	Write up answering research question 3
8	Combining the two	Draft thesis
9	Final write up	Final Thesis

References

1. F.Radicchi, C. Castellano, F.Cecconi, V. Loreto and D. Parisi. Proc. Natl.Acad. Sci. 101 (2004)
2. M.E.J Newman and M. Girvan. Phys. Rev. E, 69 (2004)
3. Max Hinne, Local Identifications of Web Graph Communities 2006
4. Jeroen Bulters, Maarten de Rijke, Discovering Weblog Communities 2007
5. Xiadon Song, Yun Chi, Koji Hino, Belle L.Tseng Summarization System by identifying Influential blogs.