A Qualitative Analysis of Edge Closure in Information Networks

Hareendra Munimadugu
School for Electronics & Computer Systems
Machine Learning & Computational Intelligence Lab
School of Computing Sciences and Informatics
University of Cincinnati
Cincinnati, OH 45221-0030
munimah@mail.uc.edu

Anca Ralescu
Machine Learning & Computational Intelligence Lab
School of Computing Sciences & Informatics
University of Cincinnati
Cincinnati, Oh 45221-0030
anca.ralescu@uc.edu

Abstract

Social Networks Analysis is one of the cutting edge research areas which finds applications in Information Retrieval and Data Mining and which deals with very large amounts of data. A considerable amount of research in this field has focused on mining useful information patterns in networks. Other applications have focused primarily on structure of networks. Several models have been proposed to address both issues. Existing models have been developed and replaced with better ones. One direction of this research has focused on how to implement one method of analysis on several data associations in order to understand how different data models behave. Responses to such questions account for the underlying differences in properties of the data considered. In its broadest sense, a community is a large set of nodes that have been collected over a period of time. In the present scenario efforts are being made to develop a complete model that can correctly explain and predict how links are formed in networks and how a network of nodes dynamically "progresses" into a bigger and more diverse one. If an idea is implemented on networks derived from different domains the results can account for the underlying differences in the properties of the data. Such an approach is useful because it helps to find not only differences based on domain but also identify similarity and therefore to correlate one domain with another.

Introduction

The ideas of human relations in a community have received particular attention since time immemorial. Therefore community structure and expansion are two areas that have been studied by philosophers, sociologists, and, more recently, by computer scientists. It is an exciting venture to try to develop a method that can account for the way in which people form and maintain relations with one another. This brings us to the community structure analysis and link prediction and it has become one of the important research problems in analysis of social networks. In this paper we take up one recent method of evaluating link formation in social communities and we compare its application to two different networks. In other words we do this by applying the same method on two different collections of data and comparing results. We then identify some possible criteria due to which significant differences exist with the application of the same technique to the different networks. Pre-evaluating links is exciting; since it is interesting to be able to predict who someone's friends are before they actually form friendship in the network. In a similar context we can say that a particular user in a network will be able to efficiently develop and retain friendships because of his position in the network and the possibility of forming links in this manner. Many methods to evaluate community growth have concentrated on networks involving people and tried to explain about their relationships. However it is also quite interesting to see how some of these methods apply to similar networks not involving people, rather involving different aspects about people.

Keywords

Social Network, Information Network, Edge closure, Link Formation, Nodal Analysis, Closure ratio.

Problem Formulation

Computational analysis of Social Networks continues to be a subject of intense interest and. The problem of social ties is believed to be related to structure and function of social networks (Granovetter 1985). Positive and negative links play a major role in the same (Bogdanov, Larusso, and Singh 2010); this interplay has given foundation to many methods developed for efficient determination of influence and opinion (Leskovec, Huttenlocher, and Kleinberg 2010). The process of link copying which is implicit in all information networks and which explains the formation of new links is known as the directed closure process (Romero and Kleinberg 2010). The methodologies as mentioned above have been extensively used for understanding information networks consisting of people. Examples of such networks are Slashdot, twitter, Facebook and several others. Here we try to consider information networks not consisting exclusively of people though they contain relevant and important information about people. Here we are interested to see how this methodology applies to information networks having several different characteristics. We would also like to see how

differences in basic characteristics of networks account for observed differences in network formation. The idea of edge closure is one of the recent important developments in predicting link formation and community structure. Our objective here is to implement this method on different data sets or information networks with inherently distinct properties and concentrate upon what characteristics of the data sets might be influential in the difference of community structure by application of the same technique. As stated we focus our attention upon networks which involve not people but some practical aspects regarding people. One such practical aspect is to determine the interest and opinion of a group of people with respect to a certain trend or their attitude towards a topic. This type of treatment of networks can be useful to also determine closeness between various domains of knowledge.

Much research has been carried out on social networks consisting of undirected edges. Evaluation of an information network such as YouTube videos is comparatively challenging because closing of undirected triangles of nodes in a social network is relatively easier. It is instinctively known that the application of the same method will lead to different observation and result in different information networks. The short term goal of this research is to study the relevance and effect of directed closure on various information networks.

Towards Problem Resolution

It is a significant fact that in order to carry out such a research of an already developed technique, we have various kinds of online social communities available. Social networks (Facebook, Orkut, and connotation networks), Information networks (YouTube, Wikipedia, Web blogs, news blogs), Hybrid networks (twitter) and signed networks (Slashdot, Epinions) are among some examples of what communities we can possibly consider. However here we consider the information networks and try to provide a clear and concise analysis of such networks. We are interested in how the concept of edge closure is applicable to an information network. Some of the examples of such a network are E-book repositories, phishing corpus, Wikipedia pages, text corpus and YouTube. Because YouTube and Wikipedia are some of the largest networks available on the World Wide Web, and also because the information in these networks is freely available, we take these two information networks for the purpose of research in this paper. Therefore the data sets that we consider for this research are derived from YouTube videos and Wikipedia pages. These are essentially collection of tuples. In the data consisting of YouTube videos, each video is taken as a node in the information network. One node is directionally connected to another if the first video has the second in the list of top ten 'favorites' or 'suggestions'. Considering this order is important because the initial few videos from YouTube hits are going to be those that are closely associated with the main or desired result. Thus an edge is generated between nodes in a directed graph.

In the data set on Wikipedia, a directed edge exists between two nodes in the order of connection of one page and a page in the references list. We shall start with a test data collection of a hundred nodes where each node is a YouTube video. Two data sets are used simultaneously to test our hypothesis. The other data set consists of an equal number of Wikipedia links or pages.

Remark 1 *It is actually not essentially important that both data sets contain an equal number of data points because what we are primarily interested is the extent of closure or in other words the percentage of nodes exhibiting this phenomenon.*

For the sample data set of YouTube, we have a collection of tuples where each tuple has two nodes and the order in which they are connected is specified; it is the direction of the edge.

Data points for the Wikipedia information are also comprised of a collection of tuples. It is very clear that the number total of tuples in the flat file is the number of relations or the number of edges in the graph.

Experimental Design

As a starting example we begin with a specific context or theme in a query in YouTube.

An edge exists between two nodes in the graph if the source video has the destination video in its top ten favorites or suggested videos' list. This edge is going to be directional because it points towards one node starting from another as in the favorites list.

An edge exhibits closure if it completes a triangle between three nodes in order, or if it is going to be the edge that completes the triangle (Romero and Kleinberg 2010). As a working example, we shall consider "University of Cincinnati Engineering" as our theme or our concept for the data analysis. This means that is the query the user is interested in. Four of the related nodes in YouTube are "Tips to succeed in Engineering", "Is Engineering right for me?", "How an engineer folds a T-shirt" and "Advice for Engineering Students".

A similar query on Wikipedia yields four of the data points from Wikipedia data which are "University of Cincinnati College of Engineering", "University of Cincinnati", "graduate students" and "University of Cincinnati College of Design, Architecture, Art and Planning".

We can notice that, because the nodes are based on the same theme, they might be already linked, but what is more important is whether they exhibit closure.

When the results are taken as nodes in a directed graph, the nodes present in the graph formed from the YouTube results are as shown in Table while those present in the graph formed from the Wikipedia results are shown in Table .

Table 1: YouTube results

Node 1	:	Tips to succeed in engineering
Node 2	:	Is engineering right for me?
Node 3	:	How an engineering student folds a T-shirt
Node 4	:	Advice for engineering students

Table 2: Wikipedia results

Node 1	:	University of Cincinnati College of Engineering
Node 2	:	University of Cincinnati
Node 3	:	Graduate students
Node 4	:	University of Cincinnati College of Design, Architecture, Art and Planning

Directional connectivity exists in these networks in the following manner. In the first example, node 1 has nodes 2 and 3 in its top ten hits. Therefore the directed arrow in the graph exists from node 1 to node 2 and also from node 1 to node 3. Similarly node 2 and node 3 have node 4 in their hit list. The directed closure is satisfied if node 1 has node 4 in the list. In the example taken the closure does exist because node 4 is indeed present in the hit list of node 1. Similarly, in the second example directed connectivity exists between the nodes considered. A similar explanation holds for the second example also. In our testing phase over large data sets we make use of the idea that for large data sets nodal closure can be evaluated based upon analyzing ordered lists prepared for a node [4]. This is the additional information needed because we cannot always have time stamps for such data. We shall carry out this test of edge closure on both data sets taking an equal number of nodes at one time. For the testing example the result is that in both cases directional closure is satisfied. In the data sets we consider, the number of connections between nodes is fairly large. Many of the arrows are also bidirectional. The expected result of the experiment is that in the case of YouTube there will be a significantly more closure overall linkage when compared to Wikipedia pages. This, according to the hypothesis is based upon the search on a theme or idea. We can give the following reasoning to the expected result. In a search on YouTube the resultant videos or hits are based upon the overall query or collection of phrases or words. The videos will be displayed based on the theme of the query, which means combined meaning of words is important. In such a case there will be emphasis on certain words that are important. Unimportant words, though a part of query do not affect search results significantly. We might say that a change of these unimportant words does not cause the results to vary significantly. But results will change drastically should the main words change. However in case of the results displayed in Wikipedia pages are retrieved based on words, meaning that a change of a word might result in a possible difference in context in the search. In Wikipedia even though combined meaning of words is important sometimes a change of any one word might result in a related but different result. This analysis applies to both the information networks, but significance in words or phrases varies for each network. These are information networks and therefore dynamic; when a new node is added to the network, new links form between the node and its neighbors. However in some cases this leads to the change in links between the existing nodes. One example of this is as follows. When the new node is very closely related to some of the neighbors there can be a change because the new node might be included in the top hits, thereby removing an already linked node from the top ten favorites list. However this is obviously not always possible. Hence within the same context the inclusion of a new node in YouTube signifies two things: that a new node has really arrived, or that a video moves up in the list of hits of another video, causing a replacement.

Conclusion

The experimentation on online data or text might identify the closeness between features of the data sets or networks in general. It is useful to determine the degree of difference in results for a same theme for different networks. This can be useful to explain the general behavior of a certain network with respect to an idea. In other words, this might account for the way a certain network behaves with respect to a theme or context. We might also be able to predict how a network is going to behave with respect to a certain theme, based on its behavior to similar themes. In order to extend this we can consider other information networks like phishing corpus, E-book repositories, personal, and news and web blogs. There are many networks available online which can be used for such research. After such experimentation it might be possible to say that a particular network has similar results with another in a given context. This can lead to saying that the related communities in such networks might behave similarly under similar given situations. One interpretation is that for a new incoming node the two networks might result in a similar number of links from a particular community of nodes. It might be possible to group different networks based on their similarity in behavior towards one particular context. Further work might include other information networks. One type of networks which offers scope for such research is networks with weighted edges (Kunegis, Lommatzsch, and Bauckhage 2009). It might be possible to extend the idea to networks with weights assigned to edges; it might be very interesting to see if closure is applicable to a network containing mixture of signed links It can also be extended to compare several networks at once when considering a common theme, or it is possible to compare other social networks. It will be a more challenging work for signed networks. It will aid to clarify the understanding of web communities and information networks. This can find useful

applications in related fields such as Sociology, Linguistics, Mathematics and others.

References

Bogdanov, P.; Larusso, N.; and Singh, A. 2010. Towards Community Discovery in Signed Collaborative Interaction Networks. In *2010 IEEE International Conference on Data Mining Workshops*, 288–295. IEEE.

Granovetter, M. 1985. Economic Action and Social Structure: The Problem of Embeddedness. *American journal of sociology* 91(3):481–510.

Kunegis, J.; Lommatzsch, A.; and Bauckhage, C. 2009. The Slashdot Zoo: Mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, 741–750. ACM.

Leskovec, J.; Huttenlocher, D.; and Kleinberg, J. 2010. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, 1361–1370. ACM.

Romero, D., and Kleinberg, J. 2010. The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. *Arxiv preprint arXiv:1003.2469*.