A Survey on Link Prediction and Analysis

Bhavya Agarwal
Department of Computer Science
Stony Brook University
Late Days Used - 1, Late Days Left - 3

1 Introduction

Links, or more generically relationships, among data instances are ubiquitous. These links often exhibit patterns that can indicate properties of the data instances such as the importance, rank, or category of the object. In this survey paper, I first go through the 2 very popular algorithms - PageRank[3] and HITS[4] and later on look at the 2 applications of link predictions[1,2]. The survey paper *Getoor et al*[5] explores the concept of link mining in detail and outlines 8 link mining tasks based on their focus on objects, links and graphs. These tasks are described in detail -

LINK-BASED OBJECT RANKING The objective of this task is to explore the link structure and and prioritize the set of objects within. The PageRank and HITS algorithm are the most famous in this field which will be explored further. Much of this research focuses on graphs with a single object type and a single link type.

LINK-BASED OBJECT CLASSIFICATION In the link-based object classification (LBC) problem, a data graph G = (O, L) is composed of a set objects O connected to each other via a set of links L. The task is to label the members of O from a finite set of categorical values.

GROUP DETECTION The goal of group detection is to cluster the nodes in the graph into groups that share common characteristics.

ENTITY RESOLUTION The goal of entity resolution is to determine which references in the data refer to the same real-world entity.

LINK PREDICTION Link prediction is the problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events, such as email, telephone calls.

SUBGRAPH DISCOVERY This work attempts to find interesting or commonly occurring subgraphs in a set of graphs.

GRAPH CLASSIFICATION Graph classification is a supervised learning problem in which the goal is to categorize an entire graph as a positive or negative instance of a concept. This is

one of the earliest tasks addressed within the context of applying machine learning and data mining techniques to graph data.

GENERATIVE MODELS FOR GRAPHS Generative model for graphs has been studied extensively and there are many major classes for same with Bernoulli graphs being the simplest

Out of these 9 tasks the main focus would be on link-based object ranking where we would analyse the 2 most famous algorithms of this field, PageRank which is used by Google and HITS which is used by Ask.com search engine

2 PageRank Algorithm

The PageRank changed the way we searched but its implementation in google. The algorithm was effective at keeping the spammers at bay which vastly improved the search results. While the spammers were still able to spam with the help of Link Spam Farms, there are many variations of PageRank like TrustRank that sprung up to counter these issues.

PageRank is a function that assigns a real number to that portion of the Web that has been crawled and its links discovered. The intent is that the higher the PageRank of a page, the more important it is.[3] Suppose we start a random surfer at any of the n pages of the Web with equal probability. Then the initial vector v_0 will have 1/n for each component. If M is the transition matrix of the Web, then after one step, the distribution of the surfer will be Mv_0 , after two steps it will be $M(Mv_0) = M^2v_0$, and so on. In general, multiplying the initial vector v_0 by M a total of i times will give us the distribution of the surfer after i steps.

However, this algorithm is makes 2 assumption of graph being connected and no dead ends being present in the graph. Hence, they introduced the concept of 'taxation'. We modify the calculation of PageRank by allowing each random surfer a small probability of teleporting to a random page, rather than following an out-link from their current page. The equation that gets modified to-

$$v = \beta v + (1 - \beta)e/n \tag{1}$$

where β is a chosen constant, usually in the range 0.8 to 0.9, e is a vector of all 1s with the appropriate number of components, and n is the number of nodes in the Web graph. The term βMv represents the case where, with probability β, the random surfer decides to follow an out-link from their present page.

Hence, in case of dead ends, we still have a probability that a few random surfers would jump out of the dead end and go to other random pages. In case of search engine, we consider many variables. First, we find out all the pages with all the terms contained on them. After that, many parameters are considered of which PageRank is one of the most important one of them.

In order to keep spammers at bay, a variation called TrustRank was introduced which gives higher weightage to outgoing links from more trusted pages, i.e. pages which are controlled (and not pages where content can be added like blogs and discussion forums). Google has also added a tag $< nofollow >$ to alert crawlers of potential spam prone locations. Also, links pages with higher PageRank are given preference in compare to lower PageRanks. This is one of the most used algorithms on web due to its high scalability and hence, has one of the biggest databases in form of google.

3 HITS Algorithm

This hubs-and-authorities algorithm, sometimes called HITS (hyperlink-induced topic search), was originally intended not as a preprocessing step before handling search queries, as PageRank is, but as a step to be done along with the processing of a search query, to rank only the responses to that query. We shall, however, describe it as a technique for analyzing the entire Web.
While PageRank assumes a one-dimensional notion of importance for pages, HITS views important pages as having two flavors of importance-
1. Certain pages are valuable because they provide information about a topic. These pages are called authorities.
2. Other pages are valuable not because they provide information about any topic, but because they tell you where to go to find out about that topic. These pages are called hubs.
We shall assign two scores to each Web page. One score represents the hubbiness of a page that is, the degree to which it is a good hub, and the second score represents the degree to which the page is a good authority.the normal way to describe the computation of hubbiness and authority is to add the authority of successors to estimate hubbiness and to add hubbiness of predecessors to estimate authority.[4] HITS uses a mutually recursive definition of two concepts: a page is a good hub if it links to good authorities, and a page is a good authority if it is linked to by good hubs.

Now after analysing the 2 algorithms on Link analysis, we can now move on to 2 applications which make use of the similar analysis tools and predict whether a link would be present between the nodes in future or not. The first paper we analyse deals with finding the missing links in wikipedia while the second one focuses more on social network link prediction analysis.
Ask.com has one of the biggest search databases on the web. The usage of this algorithm by Ask.com proves that it is scalable across the web.

4 Finding missing links in Wikipedia

Wikipedia, the free on-line encyclopedia, is a hypertext document with a rich link structure much similar to the web. However, due to manual editing of content it deals with the problems of missing links which result in non-uniform structure among similar links also.[2] Hence, this paper deals with the issue of identifying such missing links and replacing them in order to ensure uniform link structure across similar pages.
This paper follows a 2 step approach -
1. The first step concerns identification of topically related pages, i.e., clustering. Authors have proposed LTRank (Ranking based on Links and Titles) algorithm which applies widely used similarity measure from information retrieval and uses authors own version of the Lucene full-text search engine.
2. The second step involves identification of missing links. They identify candidate missing links and filter them through the anchor texts. Search for missing links is confined to set of similar pages found in first step. Their hypothesis is that similar pages should have similar link structure. They identify candidate missing links and filter them through the anchor texts i.e. the text with which the hyperlinks are placed in the wikipedia entry.

The first step reduces our search space by shortlisting the links that are similar to the given page. Then it replaces the missing links based on their presence in the similar pages. This approach seems to be more scalable for very large graphs, where it will be unfeasible to perform link prediction on global scale.

5 The Link-Prediction Problem for Social Networks

This paper uses a snapshot of the social network and uses the information provided to detect the presence of links in the future.[1] They use various approaches to for link prediction based on different metrics. The approaches can be classified as -

Methods Based on Node Neighborhoods A number of approaches are based on the idea that two nodes x and y are more likely to form a link in the future if their sets of neighbors N(x) and N(y) have large overlap. This approach follows the natural intuition that such nodes x and y represent authors who have many colleagues in common and hence who are more likely to come into contact themselves. This approach uses many different metrics like *Common neighbors*, *Jaccards coefficient* and *Preferential attachment*.

Methods Based on the Ensemble of All Paths These methods refine the notion of shortest-path distance by implicitly considering the ensemble of all paths between two nodes. This approach uses algorithms like *PageRank*, *HITS* and *SimRank* to measure probability of links.

Higher Level Approaches The paper lists 3 meta approaches, i.e. *Low-rank approximation*, *Unseen bigrams* and *Clustering.*. They can be used in conjunction with any of the methods discussed earlier.

The authors use a random predictor, which simply predicts randomly selected nodes which did not collaborate in the training interval. They found significant overlap in the predictions given by the methods. The data was taken from 5 different databases with information on different research papers and their citation. They found high prediction gains as compared to random predictor however, mostly there accuracy was limited to below 30% due to the unpredictable nature of scientific collaborations. However, many algorithms can be used in generalised social networks like Facebook where relationships are generally governed by the Small world rule which limits out search space a lot.

6 Discussion

All these papers can be linked in Link Analysis and Prediction.The first paper is a survey paper that elaborates all the steps that are frequently used in link analysis. The next 2 papers show the base on which most of the link prediction is based. They form very important metrics which are still in popular use across various domains. The other 2 papers mainly concerned themselves with applications in this field which can be applied to existing datasets to extract knowledge.

PageRank and HITS were meant to save the web searches from spammers who ruin the quality of results. Earlier "Term Spamming" was fairly easy and widespread across the web. These algorithms

came up with novel approaches to keep the spammers at bay and made it really hard for them to spam the crawlers. Moreover, there new variations have considerably reduced the amount of spam we see now a days in our searches. They are also hugely scalable as shown by their use in widespread products.

The last 2 papers concerned themselves mainly with prediction of links in a practical scenario. The first one strived to improve the content quality on wikipedia by clustering similar pages and anaylysing their link structure. The second paper was more concerned in mapping the future possible interactions between different entities connected in a Social Network. Both these papers in some way used the PageRank and HITS algorithm or their close variants. The paper on "The Link-Prediction Problem for Social Networks" were mainly about extracting the power of the many link analysis algorithms and comparing them on different datasets. While the paper on wikipedia missing links was able to find missing links with good efficiency, the other paper was more about critiquing the various approaches and analysing the results, hence guiding us to use them appropriately in future researches.

Hence, reading these papers gives you a holistic view of link related techniques in data mining and its application and ongoing research.

References

[1] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, in: CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 2003, pp. 556-559.

[2] S. F. Adafre and M. deRijke. Discovering missing links in wikipedia. In Proceedings of the 3rd International Workshop on Link Discovery atKDD05,Chicago, USA, August 2005.

[3] S. Brin and L. Page, Anatomy of a large-scale hypertextual web search engine, Proc. 7th Intl. World-Wide-Web Conference, pp. 107117, 1998.

[4] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J.ACM 46:5, pp. 604632, 1999.

[5] GETOOR, L. AND DIEHL, C. P. 2005. Link mining: a survey. ACM SIGKDD Explorations Newsletter 7, 2,312.