LINK ANALYSIS IN WWW

Prasad Sriram
Department of Computer Science and Engineering
University of Minnesota

SYNONYMS
Link analysis, Link Mining.

DEFINITION
The information age has made it easy to store large amounts of data. The proliferation of documents available on the Web is rapidly growing. Link Analysis is a new, exciting and rapidly growing area of research that tries to solve the information overload problem by using techniques from data mining, machine learning, Information Extraction, Text Categorization, Visualization and Knowledge Discovery. Link Analysis is the process of building up networks of interconnected objects through various relationships in order to discover patterns and trends [10].

Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks that are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fueled largely by interest in web mining, hypertext mining and social networks analysis.

HISTORICAL BACKGROUND
The main tasks of link analysis are to extract, discover, and link together sparse evidence from vast amounts of data sources, to represent and evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities. The relationships could be transactional, geographical, social or temporal.

Link Analysis involves the preprocessing of document collections (text categorization, term extraction, and information extraction), integration with structured information sources, the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules, etc.) and visualization of the results.

Consider for example, the World Wide Web. The World Wide Web is a huge repository of web pages connected to each other by hyperlinks. Each page points to another page or is pointed to by another page through hyperlinks. These web pages and the hyperlinks can be visualized as a graph model [5] in which all web pages are nodes or vertices of the graph and all hyperlinks are the edges connecting these different nodes (web pages).

When the entire WWW is viewed as a graph or a network, several properties in graph theory can be readily applied to many of the problems and applications that are prevalent in WWW. The process of analyzing these link structures is called Link Analysis.
Since the emergence of the Internet, a number of researchers have conducted various investigations and surveys of its distribution and size. A vast amount of statistical resources and numerous theoretical contributions on interpreting the growth of the Internet and especially web exist. Also the convenience of representing linked objects and data as a graph theoretic model gives more flexibility in understanding the properties of the data and also inferring new information from them. Right from the sixteenth century, network theory and applications started fascinating mathematicians and statisticians. There is a lot of published literature on the properties of graphs and networks, various properties of graphs, various laws and distributions that a graph can possess for example power law distribution, zipfian distribution etc. Viewing Internet and WWW as a graph led to a renewed interest in this community. To date, lots of research is being done and there are fifteen to twenty research conferences and workshops on link analysis and related topics like data mining and social network analysis.

**SCIENTIFIC FUNDAMENTALS**

As any linked data can readily be modeled as a graph structure, we will review some basic properties of graphs in the section.

![Directed graph with weighted edges](image1)

**Figure 1. An example directed and undirected graph**

A graph consists of a nonempty set of points or vertices, and a set of edges that link together the vertices. A graph can take on many forms: directed or undirected. A directed graph is one in which the direction of any given edge is defined. Conversely, in an undirected graph, one can move in both directions between vertices. The edges can also be weighted or unweighted.

An edge in a graph that joins two vertices is said to be incident to both vertices. Furthermore, the degree of a vertex is determined by the number of distinct edges that are incident to it. More specifically, the indegree and out-degree of a vertex represent the number of edges that terminate in and originate from a vertex, respectively.

Typically, a subgraph is a connected subset of nodes and links from a larger graph. A bipartite graph is a special graph where the set of vertices can be divided into two disjoint sets $U$ and $V$ such that no edge has both end-points in the same set.
Many studies have discovered patterns in static graphs, identifying properties in a single snapshot of a large network, or in a very small number of snapshots; these include heavy tails for in-degree and out-degree distributions, communities, small-world phenomena, and others. It has also been found that most of these graphs densify over time, with the number of edges growing superlinearly in the number of nodes and the average distance between nodes shrinking over time [7].

In 1999, [13] et al, came out with power law relationships for WWW based on Internet topology. Their power laws concisely described skewed distributions of graph properties such as the number of in-neighbors, number of out-neighbors, number of edges in a random graph. In addition, the power laws can be used to estimate important parameters such as the average neighborhood size.

**Small World Phenomenon** is the hypothesis that everyone in the world can be reached through a short chain of social acquaintances. This concept today widely used in social network analysis gave rise to the famous phrase ‘six degrees of separation’ after a 1967 small world experiment by social psychologist Stanley Milgram. This suggested that two random US citizens could be connected on average by a chain of six acquaintances. Similar to this hypothesis, many properties obtained by analyzing graphs serve as the foundation for link analysis.

**KEY APPLICATIONS**

The most popular applications of link analysis that would immediately come to one’s mind are Google’s PageRank [12] and Klienberg’s [6] Hypertext Induced Topic Selection or commonly known as Hubs and Authorities (HITS) algorithms. In this section, besides these two applications several other real world applications of link analysis are presented.

**PageRank**

PageRank is a numeric value that represents how important a page is on the web. This metric figures that when one page links to another page, it is effectively casting its vote for the other page. The more votes a page gets, the more important the page becomes. Also, the importance of the page that is casting the vote determines how important the vote itself is. PageRank is Google's algorithm to rank a webpage's importance. Though, it isn't the only factor that Google uses to rank pages, it is a very important one.

PageRank of a page is determined using the random surfer model described above. The PageRank of a page can be computed as follows:

\[
PR(A) = \frac{1-d}{N} + d \sum_{j \in S} \frac{PR(j)}{C(j)}
\]

Where:

- PR (j) is the PageRank of page j
- S is the set of nodes that have an in-link to page A
- C (j) is the out degree of page j
‘$d$’ is the dampening factor that is set to a value between 0 and 1. It is usually set to 0.85 for the web graph.

$N$ is the number of nodes in the graph.

The PageRank of all the nodes in the graph are computed using an iterative algorithm. Each node is assigned an initial value and the PageRank of all the nodes are then calculated in several iterations based on the equations determined by the PageRank Algorithm.

**HITS-Kleinberg’s algorithm**

The key idea behind HITS [6] is the filtering of broad search topics, through the discovery of “authoritative” information sources on such topics. HITS algorithm is a formulation of the notion of authority, based on the relationship between a set of relevant authoritative pages and the set of “hub pages” that join them together in the link structure.

HITS algorithm is a recursive algorithm in which each node is assigned an authority score and a hub score. The hub scores will be higher if it points to many nodes or nodes with high authority. Conversely, authority scores will be higher if it is pointed to by many nodes or pointed by good hubs.

**Web Document Categorization**

Consider for example a news website like New York Times. There are lots of pages uploaded daily along with lots of related advertisements appearing near those pages. It would be really useful if one can automate the process of deciding which advertisements should appear along which page. For this, the web pages need to be classified into categories and this is where link analysis plays a major role.

A hypertext collection [8,9] has a rich structure that can be exploited to improve classification accuracy. In addition to words, hypertext has both incoming and outgoing links. The category of a web page can be labeled based on features of the current page and features of linked neighbors. With the use of linkage information, such as anchor text and neighboring text around each incoming link, better categorization results can be achieved. Chakrabarti et al [9] proposes such a model to utilize both text and linkage information to classify a database of patents and a small web collection.

**Group Discovery**

One can find groups or communities [11] in WWW by extracting information from the link topology i.e., by understanding how the members in a community are linked with each other, one can identify and discover groups.

[11] relates this problem of extracting communities to that of extracting sub graphs from a large graph. [3] discusses an algorithm for finding large, dense subgraphs in massive graphs. The authors use their algorithm to characterize large, dense subgraphs of a graph showing connections between hosts on the World Wide Web. Through experiments, they present a very interesting result that many of these communities are websites which heavily ‘link spam’, a problem which we will discuss in the next section. Thus, dense
subgraph extraction in addition to discovering groups can also be a very useful way of detecting spam.

**Link Spam Detection**

Link spamming is a way of manipulating link structure of web sites in order to gain an undeservingly higher page rank. There has been a sudden surge of interest in this problem. Interestingly, link analysis provides a very handy way of detecting and thwarting this kind of spam. For instance, Fetterly et al. [14] analyze the in-degree and out-degree distributions of web pages. As described in section 3, most web pages have in- and out-degrees following a power-law distribution. Also for each page, the distribution of PageRank scores of pages pointing to it, obeys a power law. However, search engines occasionally encounter substantially more pages with the exact same in-degrees, out-degrees or pagerank scores than what is predicted by the distribution formula. Thus a vast majority of such outliers can be detected as spam pages.

Also, these link spam websites form a very tight community by colluding with each other spam websites in WWW. So, dense sub graph extraction can also be a very useful technique to identify link spam.

Thus link analysis serves as a really powerful tool in detecting large, automatically generated link spam structures with “unnatural” link patterns. However, they fail to recognize more sophisticated forms of spam, when spammers mimic reputable web content.

**Social Network Analysis**

Social Network Analysis is based on the importance of relationships among interacting units. The social network perspective encompasses theories, models, and applications that are expressed in terms of relational concepts.

The unit of analysis in social network analysis is not the individual, but an entity consisting of a collection of individuals and the linkages among them. Network methods focus on dyads (two actors and their ties), triads (three actors and their ties), or larger systems, subgroups of individuals, or entire networks. In the process of working in this field, network researchers have developed a set of distinctive theoretical perspectives as well.

One can also predict links between entities such as ‘Will there be an association between entity 1 and entity 2’ in future by inferring relationships between links of entity 1 and links of entity 2. Techniques like link prediction [4] can be used to predict this kind of future relationships between entities.

**Link Analysis for GIS Applications**

One can also find numerous applications of link analysis in Geographic Information Systems. Geographic Search Engines help the users focus their queries on a particular geographic region and allow users to constrain and order search results within their
geographic preference. We can incorporate geographic knowledge of the web pages and efficiently use link analysis.

Very often, when a user is performing a local search, he might be interested in local businesses, locally relevant news results, or tourism information about a particular region. In such cases, we might be interested in finding the geographical footprints of a page. The process of finding the geographical footprints of a web page is called geo-coding. During this process, we determine the IP address of a web page and other geographical details like the city the web page is talking about, its zip code and so on.

As this operation is expensive, it is done only on a limited set of seed pages. In another process called geo-propagation, with the help of link analysis, the footprints associated with other pages are determined. This is done based on the principle that if more than a certain threshold of pages having information about ‘New York’ have hyperlinks pointing towards a set of pages, then it is likely that the pages that are pointed to are also relevant to New York and hence should inherit such a footprint (though with lower certainty). It is based on the intuition that geographically related pages are closely connected in the web graph. Thus, we can improve the quality and coverage of the initial geo-coding by analysis of the link structure and site topology.

Also, while designing, distributed crawlers, different partitions of the web graph each belonging to one particular geographic region are created. Thus, in order to determine the geographical scopes of the web sites, some heuristics are employed to determine the geographic location. Based on the geodesic distance between two separated locations, a geodesic distances graph is built where host machines form the vertices and the respective distances form the edge weights of the graph. This graph is then partitioned and the various partitioned graphs are assigned to the distributed crawlers.

On one hand, the partitioning of the reduced IP Web link graph minimizes the inter-partitioning links between IPs, while on the other hand, the geodesic distances graph allow us to obtain geographically focused topics by minimizing the proximity between IPs. [15]

Using link analysis, we can also determine spatial information hubs. A spatial information hub is a web page which is related to a specified geographic area and has much local information and/or many hyperlinks to local web pages. By taking their geographic localities into consideration one can extract geographic information from web pages to create spatial nodes and spatial links, then conduct a link analysis based on the extended link structures. And some link analysis is performed. If enough links from a region point to a web site and these links are evenly distributed, the region is included in the site’s geographic scope. The approach was applied to the United States using states, counties, and cities for the geographic hierarchy [16].

Traditional travel guides provide the users with an opportunity to find information on various places of their interest, but they are often many clicks away via traditional hyperlink navigation. Moreover, travel portals are often out of date or incomplete due to the need for human editing. Thus, when a user wishes to see web resources that have
similar geospatial context, one can design an environment in which the user can browse by geographic proximity between web pages. Apart from WWW, we can also use link analysis to determine major places of military interest from a spatial network. A spatial network consisting of all places of military occupancy and showing its relations with other places. One can thus find places that are more critical by exploiting the properties of the spatial network. Similar to this, link analysis can also be used in transportation networks to determine places that serve as major hubs and also most congested places in the road network.

**CHALLENGES IN LINK ANALYSIS**

Some limitations of link analysis are [1]:

1. Search engines relying on meta tags in documents are often misled (intentionally) by web developers.
2. There is a bias introduced by search engines as organizations pay search engines and page rank.
3. Adding even a small number of nodes/edges to the graph has a significant impact to the link-based metrics arising questions about their stability.

**FUTURE DIRECTIONS**

Link analysis is a newly emerging research area at the intersection of research in artificial intelligence, social network analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming. The applications of link analysis are endless. Besides the above mentioned key applications, many interesting applications of using link analysis in citation network analysis, recommender systems, VLSI circuit design, entity resolution, record linkage, name deduplication [2] and spatial, transportation networks are still going on.

Computer-assisted or fully automatic computer-based link analysis is also increasingly employed by banks and insurance agencies for detecting frauds, by telecommunication operators in telecommunication network analysis, by medical sector in epidemiology and pharmacology, in law enforcement investigations, by search engines for relevance rating (and conversely by the spammers for link spamming and by business owners for search engine optimization), and everywhere else where relationships between objects have to be analyzed.

**RECOMMENDED READING**


