EXTRACTION OF LEADER-PAGES IN WWW

An Improved Approach based on Artificial Link Based Similarity and Higher-Order Link Based Cyclic Relationships

Ravi Shankar D, Pradeep Beerla
Tata Consultancy Services, Hyderabad, India

Abstract: WWW is the most popular and interactive medium to disseminate information. It creates many new challenges. Several initiatives have been taken to extract different kinds of knowledge from web. In our paper "Leader-page Resources in the World Wide Web" we defined a new method to rank the web pages entirely using the hyperlink information. The notion of "Leader-Page" is extended from the concept of leader from the leadership theory and the social networks. In a community, a leader is a person who interacts the most with other members of the community and whose characteristics are most similar to the characteristics of other members of the community. We have extended these properties of leader to identify leader-pages in WWW. In this paper we propose an improved approach to measure the "leadership score" of a web page based on artificial link based similarity and higher-order link based cyclic relationships it establishes with other web pages of the cyber community.

Key words: Search, Leadership, Artificial Links, Higher Order Cyclic links.

INTRODUCTION

The WWW is the single largest global repository of information and human knowledge. It continues to grow at a remarkable pace with contributions from allover the world. The knowledge discovered through navigation of this complex heterogeneous collection of text (content) and hyperlinks (that lend it a structure) is enormously benefiting the mankind. However owing to the hugeness and diversity of the web users are drowning in information and are facing information overload. It is very difficult to index all the information available on the web. Creating new knowledge out

Please use the following format when citing this chapter:

Shankar D, R., Beerla, P., 2006, in IFIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston: Springer), pp. 151–160.

of information available on the web is another problem. So, the ranking of searched results is very important. One can observe that the web is growing as a socializing medium to connect a group of like-minded people independent of their geographical location and time. Web has turned into one of the most important distribution channels for private, scientific and business information. With the time, web started to behave like a complex society. The social network theory, leadership theory and Mauss's gift exchange theory gave us insights for the concept of leader in the context of WWW. These concepts help in understanding the thought process of the sociology of web and other related issues. We earlier extended these concepts of leadership to World Wide Web [1] and proposed a new approach to rank the results of a search query. We defined a leader-page based on the cyclic and similarity relationships a web page establishes with other web pages of the cyber community. We considered only the hyperlink information to rank the web pages. Now we propose an improved approach to measure the "leadership score" by including the impact of content similarity and higher-order link based cyclic relationships in addition to the existing formulations. The central issue we would address within our framework is the application of this artificial links (content based similarity) and higher-order link based cyclic relationships to modify the "Leader-page algorithm" and rank the web pages accordingly.

The paper is organized as follows. In section 2, we discuss the related work. Here we explain the leadership theory, social network theory and Mauss's gift exchange theory. In section 3, we explain how we extended the concept of leader to define a "leader-page" in the context of WWW and the "leader-page" algorithm. In section 4, we define the "artificial links", "higher-order link based cyclic relationships" and present the modifications in the "leader-page" algorithm. In the last section, we present the summary and conclusions.

1. RELATED WORK

Search engines perform both link and text based analysis to improve the quality of search results. We have used the concepts from several other theories to propose and improve the leader-page approach. Here we present the background for the social network theory, Mauss's gift exchange theory and leadership theory. We also discuss how leaders evolve in a community.

Social network theory [9] views actors and its relationships in a society as nodes and edges. A social network is a map of all of the relationships between the nodes. This relationship indicates the existence of information exchange among nodes.

Hierarchy in social networks is stated strictly in terms of position of a given node relative to other nodes, without assuming any content to position. The content is given by the nature of exchange and connection. Information exchange happens through the acquaintances in the zones to other nodes. So, a node can be connected to many other nodes by virtue of its own actions or preferences. In social networks, one's immediate zone of neighbors is connected to the immediate zone of those neighbors and so on, which allow a node to reach other nodes in very few steps. Thus nodes draw information, which it would not otherwise know. Thus information is not directly an attribute of individuals, but rather their ability to draw up to their position in a network. Another form of effect of networks is the concept of "threshold point". This idea refers to the extent to which a given phenomenon is allowed to spread through the network. Once a certain level has been reached, all the nodes join in the phenomenon. The probability of any individual node acting is a function of the number of other nodes in the network that have acted in a given way.

Mauss's Theory of gift exchange [10] says that when people give a gift, they are expecting a return gift and when they receive a gift they have a duty to give something in return. The gift embodies some kind of relation of economic reciprocity. So in a network of interactions, when someone gives a reference they expect to be referred by. Elaborating on these observations, given a phenomenon, it will flow through the network because of the interactions being reciprocated among the nodes. The more the nodes interact, the more they will like each other. And the more the nodes like each other, the more they interact (link based cyclic relationship). The more the nodes interact, the more their characteristics become similar (similarity based relationships). Once each node attains a certain level of information (i.e. greater than the threshold) the nodes join in the behavior of the phenomenon. So after some optimum level they start behaving similarly. The nodes, which are similar, form a community.

In any community, the phenomenon of leadership has a great prominence. A leader is interpreted as a person who sets direction in an effort and influences other members of the community to follow that direction. The leaders have strong mutual relationships with other members of the community. A community can be analyzed by studying the leadership phenomenon in the community. A scan of various theories of leadership can help to comprehend the leadership phenomenon. The phenomenon of leadership has been studied since Aristotle. Trait theory is one of the earliest theories on leadership. This theory of leadership focuses on the traits of the leader that make him a leader. The focus has shifted towards the behavior of the leader. Studies have led to the notion of "Charismatic leadership". A charismatic leader continually assesses the environment. He/she

communicates with other people, and builds trust and commitment. Finally he/she is the role model of the whole community. Finally, we draw upon the work done by George C Homans [11] in the area of social exchange theory. According to Homan, the leader is a person who interact the most with other members of the group, both initiates and receives the communication, has more social contacts within the group and whose actions and sentiments are most similar to the group's own sentiments and actions. So, interactions (exchanges) among the members of a community help in increasing the similarity among its members finally leading to the evolution of leaders.

2. LEADER IN WWW & LEADER-PAGE ALGORITHM

We have seen the evolution of leader with the foundations of social networks and gift exchange theories. We can see the analogy with information as flow in the context of WWW. In this section we see how we extend the concept of leader to the World Wide Web.

Creators of web pages exchange hyperlinks to other pages to express some relationship. When the creator of a web page Pi places a hyperlink to a page Pj while there is no hyperlink from Pj to Pi, we say that the creator of P has established an association with the creator of Q. At a later stage, a hyperlink placed from Pj to Pi will create a cyclic relationship between them. Our hypothesis is that the quality and credibility of the content of each of the two web pages of different creators is of higher value if the creators place hyperlinks to each other's web pages than in the case where only one of them places a link to other's pages. This relationship is the basis for the flow of information (exchange of hyperlinks) among the pages. The flow of information spreads a phenomenon among the pages of the web. The web pages that cross the threshold limit for the phenomenon (pages which are similar and have noticeable characteristics) start behaving similarly. This behavior of the information flow of the web pages leads to the development of a cyber Community [8]. There exist some web pages which interact the most with other web pages of the cyber community, both initiates and receives links, has more hyperlinks within the cyber community and whose characteristics are the most similar to the cyber community's own characteristics. Such web pages are called the leader-pages of the cyber community. There can be any number of such leader-pages for a cyber community.

In our paper on leader-page resources in WWW, we have identified the relationship between pages that contribute to the evolution of leader-pages entirely based on the hyperlink analysis. For a web page Pi, the leadership

score L[Pi], is determined based on the direct link-based cyclic relationship, indirect link-based cyclic relationship, cocitation based similarity relationship and coupling based similarity relationship with other pages of the web. L[Pi] is the weighted sum of leadership scores of all the other pages which participate in preceding relationships with Pi. Leadership score of a page Pi is defined as $\mathbf{L[Pi]=k_{dl}(DL(Pi))+k_{indl}(INDL(Pi))+k_{coct}(COCT(Pi))+k_{coup}(COUP(Pi))}$, where **DL(Pi)**= sum of leadership scores of all pages in Direct Link based Cyclic relationship with Pi. **INDL(Pi)**= sum of leadership scores of all pages having Indirect Link based Cyclic relationship with Pi. **COCT(Pi)**=sum of leadership scores of all pages having Cocitation based similarity relationship with Pi. **COUP(Pi)**=sum of leadership scores of all pages having Coupling based similarity relationship with Pi. Here, k_{dl}, k_{indl}, k_{coct} and k_{coup} are the parameters that determine the weights of corresponding relationships in the measure of leadership score.

Given search query, the leader-page extraction algorithm first builds the focused sub-graph. The search query is given as input to a search engine. It takes a reasonable number of top pages in the output list and forms corresponding root-set. For each web page in the root-set, corresponding parents and children are extracted. A base-set is formed with the root set, its parents and children of pages in root set. Pre-processing techniques are applied on the base-set.

For each web page Pi in focused sub graph S, the leadership score of page L[Pi] is calculated in the following way

If Pi forms a Direct Link based cyclic relationship with Pj in S, then $\mathbf{L[Pi]=L[Pi]+k_{dl}(L[Pj])}$, If Pi forms an Indirect Link based cyclic relationship with Pj&Pk in S, then $\mathbf{L[Pj]=L[Pi]+k_{indl}(L[Pj]+L[Pk])}$, If Pi forms a Cocitation based similarity relationship with another Pj in S, then $\mathbf{L[Pi]=L[Pi]+k_{coct}(L[Pj])}$ and If Pi forms a Coupling based similarity relationship with Pj, then $\mathbf{L[Pi]=L[Pi]+k_{coup}(L[Pj])}$. After updating the leadership scores of all web pages, we normalize the leadership scores. The web pages are sorted based on corresponding leadership scores. The values for parameters k_{dl}, k_{indl}, k_{coct} and k_{coup} should be selected based on the corresponding influence on the leadership score. The web pages with high leadership score are identified as the leader-pages. The results have proved that leader-page approach is a potential approach to rank the web pages as compared to hubs-authorities and google's page rank [3,4].

3. IMPROVED LEADER-PAGE APPROACH

The Leader-page algorithm essentially concentrated on the importance of hyperlinks to calculate the leadership score of a web page. Using the analogy

between the web community and a social community it identified the essential properties of leader-page to be the cyclic and similarity relationships it can establish with other web pages. The leader-page algorithm considered only the direct link-based cyclic relationship and indirect link-based relationship. But we can define higher-order link based cyclic relationships of order N. Also we can define artificial links based on content similarity between web pages and calculate the leadership score.

3.1 Higher-order link based cyclic relationships

Pair of web pages P0 & P1 participates in a direct link based cyclic relation ship if P0 establishes a link to P1 and P1 establishes a link to P0. This can be referred as 1st order link based cyclic relationship. Similarly web pages P0, P1 & P2 participate in an indirect link based cyclic relationship if P0 establishes a link to P1, P1 establishes a link to P2 and P2 establishes a link to Po. This can be referred as 2nd order link based cyclic relationship. We can generalize this kind of cyclic relationships. A set of pages P0, P1, P2...Pn participate in a kind of relationship such that P0 establishes a link to P1, P1 establishes a link to P2, ... Pi establishes link to Pi+1...Pn-1 establishes a link to Pn and finally Pn establishes a link to P0. Existence of such a relationship can be called as nth order link based cyclic relationship. In general we can define any number of higher-order link based cyclic relationship of order N. As we use more levels the leadership score become more accurate. So while applying the leader page-algorithm, we can calculate the leadership scores using higher-order link based cyclic relationships till some order. Now for a web page Pi, If Pi participates in a higher-order cyclic relationship of order N P0, P1, P2...Pn then $L[Pi]=L[Pi]+ k_{cycN} (L[P0]+L[P1]+L[P2]+...L[Pn])$ where k_{cycN} is the parameter that determines the impact of higher-order link based cyclic relationship of order N. A web page participating in many such relationships increases its potential to become a better leader-page.

3.2 Artificial Link based similarity relationship

Artificial links are links introduced between web pages based on content similarity irrespective of the presence of actual hyperlink between them. The objectives are to embed artificial links [5,6,7] among web pages based on text analysis methodologies and extract the required values based on these artificial links. For two web pages Pi and Pj, in this process all the hyperlinks and stop words present in the web pages are removed. Stop words are trivial words with no significance. A dictionary is used to exhaust them. All the words are stemmed and sorted alphabetically. A vector representation

of the page with words and frequencies is made assuming all the dimensions are orthogonal. For each page in the data set, which is in vector format, a cluster can be formed with the page having a cosine similarity greater than the specified threshold value of other pages in the data set. A group is formed out of these pages. For each group, artificial links are incorporated among the constituents of the group. In each group, for each of the pages the cosine similarity with other pages is computed. The sum of cosine similarities for each page is computed. This total sum of cosine similarities for each page in a group is called the weight of the page. A sorted set of pages based on the weights is formed. Then artificial hyperlinks are incorporated between all pages in the group with the first page in the group (Lenient strategy). Originally there will not be any link between the web pages Pi and Pj, but with the introduction of an artificial link based on the above-mentioned process, we calculate the leadership score. As the artificial link is given after analysis of the content it is more powerful than a normal link and it has more weight. We give more weight to the artificial link.

So while applying the leader page-algorithm, we can calculate the leadership scores using artificial link based similarity relationships. Now for a web page Pi, If Pi participates in a artificial link based similarity relationship with Pj then L[Pi]=L[Pi]+ k_{artl} (L[Pj]) where k_{artl} is the parameter that determines the impact of artificial link based similarity relationship. A web page participating in many such relationships increases its potential to become a better leader-page.

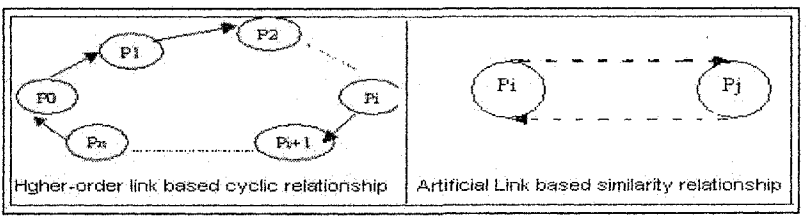

Figure 1. The improved Approach

3.3 Modification in the Leadership score formula

The leadership score formula is impacted by the higher-order link based cyclic relationship and artificial link based similarity relationship modified in the following way $\mathbf{L[Pi]=k_{cyc1}(CYC1(Pi))+k_{cyc2}(CYC2(Pi))+...+k_{cycm}(CYCm(Pi))+....k_{cycn}(CYCn(Pi))+k_{coct}(COCT(Pi))+k_{coup}(COUP(Pi))+k_{artl}(ARTL(Pi))}$,Where CYC1(Pi)=sum of leadership scores of all pages having Direct link based cyclic relationship with Pi (order 1). CYC2(Pi)=

sum of leadership scores of all pages having higher order link based Cyclic relationship of order 2 with Pi (order 2). CYCm(Pi)= sum of leadership scores of all pages having higher order link based Cyclic relationship of order M Pi (order M). CYCn(Pi)= sum of leadership scores of all pages having higher order link based Cyclic relationship of order N with Pi (order N). COCT(Pi)=sum of leadership scores of all pages having Cocitation based similarity relationship with Pi. COUP(Pi)=sum of leadership scores of all pages having Coupling based similarity relationship with Pi. ARTL(Pi)=sum of leadership scores of all pages having artificial link based similarity relationship with Pi.

Here k_{cyc1}, k_{cyc2}, k_{cyc3} ... k_{cycM} ... k_{cycN}, k_{coct}, k_{coup} *and* k_{artl} are the parameters that determine the weights of corresponding of higher order link based cyclic relationships of order 1, 2, 3, ..., N, Cocitation based similarity relationship, Coupling based similarity relationships and artificial link based similarity relationship.

3.4 Modification in the Leader-page algorithm

Given search query, the Leader-page algorithm extracts the corresponding "leader-pages" from WWW. The process of extraction of leader-pages is similar to the extraction of Hub and Authority web pages in HITS [2]. For the specific search query, we build a focused sub-graph. Next, we apply the leader-page extraction algorithm to calculate leadership scores to all the pages in the focused sub-graph. The pages with high leadership score are considered as leaders for the search query. The phases involved in the implementation are as follows.

Building the focused sub graph is build by giving the search query to a search engine. By taking a reasonable number of top pages in the output list corresponding root-set is formed. For each web page in the root-set, corresponding parents and Children are extracted. The parents and children of all the pages and pages of root set form a base-set. Pre-processing techniques are applied on the base-set. This is the focused sub-graph of WWW corresponding to search query.

The algorithm to calculate the leadership scores for the web pages in S is given below. L[Pi] denotes the leadership score of the page Pi and L denotes the leadership score vector for all the pages in S. The leadership scores of all the pages in S are initialized to one. For each web page Pi in S, we use the modified Leadership score formula. If Pi participates in a cyclic relationship of order M with P0, P1, P2...Pm then L[Pi]=L[Pi]+ k_{cycM} (L[P0]+L[P1]+L[P2]+...L[Pm]) where k_{cycM} is the parameter that determines the impact of link based cyclic relationship of order M (M can be any integer >= 1). If Pi forms a Artificial link based similarity relationship

with Pj, then L[Pi]=L[Pi]+k_{artl}(L[Pj]) where k_{artl} is the parameter that determines the impact of artificial link based similarity relationship. If Pi forms a Cocitation based similarity relationship with Pj, then L[Pi]=L[Pi]+k_{coct}(L[Pj]) where k_{coct} is the parameter that determines the impact of cocitation based similarity relationship. If Pi forms a Coupling based similarity relationship with Pj, then L[Pi]=L[Pi]+k_{coup}(L[Pj]) where k_{coup} is the parameter that determines the impact of coupling based similarity relationship. After updating L[Pi], the leadership score vector is normalized. The leader-page extraction algorithm repeatedly updates and normalizes the leadership scores. This process is continued till we observe not many variations among the values of leadership scores for the pages in the focused sub graph. Thus the modified leader-page algorithm calculates the leadership scores of all the pages. We filter out top c leader pages for the specific broad topic query and declare them as the results.

4. SUMMARY AND CONCLUSIONS

In this paper we brought in the concept of higher-order cyclic relationship and generalized the level of cyclic relationships that determine the leadership score of a web page. We also used the content-based similarity to define artificial links among web pages to contribute to the leadership score. We modified the leader-page algorithm. We presented a simple and efficient method to determine these leader pages using the modified leader-page algorithm. These leaders are web pages that we feel are more specific to be the results of a search query.

In our paper on "Leader-page resources in WWW", we have shown that the leader-page algorithm is an efficient measure of ranking the pages as compared to the hubs-authorities and Google's page rank. In this paper we have improved the algorithm by taking into consideration more information through higher-order cyclic relationships and content-based similarity (artificial links) along with the already defined formulations. So, this improved approaches promises better results.

As part of future work we plan to implement the improvements in the leader-page algorithm. The algorithm is still in its preliminary stages and it can be extended using various other algorithms. It can be used with already existing efficient measures to improve the results for a broad topic search query and the rankling methodology

In the improved approach, we plan to use the search results from the yahoo search engine for a search query by making a base set from it. Then we use our improved leader-page algorithm to identify the leaders. We have link based cyclic relationships and similarity based relationships. We will

take a relative approach based on the normal links to give the values to the parameters (k_{cycM}, k_{artl}, k_{coct}, k_{coup}), which decide the impact of a relationship on the measure of leadership score. The leadership score for a web page in the focused sub graph is the cumulative weight of all kinds of link based cyclic relationships and similarity based relationships. The web pages with the high leadership score can be identified as the leaders. The various formulations should be instrumental to efficiently identify the leaders in World Wide Web for a specific search query. This whole process involves lot of computations. These computations need to be done offline due to the extensive time input that is needed to calculate the leadership scores.

The optimal value allocation for the parameters (k_{cycM} , k_{artl}, k_{coct}, k_{coup}) is yet to be decided. These parameter values determine the net leadership score of a web page. So it is crucial to identify the optimal values for the parameters in terms of the normal link between the pages.

Finally, using the leader-page approach we can observe the evolution of web leaders in the web communities, the way in which the cyber communities react with the leader pages and also the trends and changes in the community with respect to the leader pages in the World Wide Web. The leadership score acts an efficient ordering metric to develop web directories and web groups.

REFERENCES

[1] D.Ravi Shankar, Pradeep Beerla, P.Krishna Reddy Leader-page Resources in World Wide Web, 12th COMAD 2005b [link]

[2] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. PageRank Citation Ranking: Bringing order to Web. *Stanford Digital Library Technologies Project*, 1998.

[4] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *7th WWW Conference*, 1998.

[5] K.Bharat and M. Henzinger, "Improved algorithms for Topic Distillation in Hyperlinked environments", Proc 21st SIGR Conference, (1998).

[6] Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, Prabhakar Raghavan, "Scalable feature selection, Classification and signature generation for organizing large text databases into hierarchical topic taxonomies", VLDB Journal, 1998.

[7] Sergey Brin, "Extracting patterns and relations from the world wide web", 6th EDBT, 1998.

[8] D. Gibson, J. Kleinberg, P. Raghavan, "Inferring Web Communities from Link Topology", 9th ACM Conf on Hypertext and Hypermedia, 1998.

[9] Charles Kadushin, "Intro to Social Network Theory", Brandeis University, 2004.

[10] Mauss,"The Gift Forms and functions of exchange in archaic societies", 1954.

[11] George C Homans. *The Human Group.* New York Harcourt, Brace & World, 1950.

[12] Northouse, P.G. *Leadership theory and practice.* Thousand Oaks, CA: Sage Publications, 2001.