ISSN 1512-1232

WEBSITE SEARCH TECHNIQUE USING K-MEANS ALGORITHM

Mandal Shrabanti, Pal Anita

Departments of Mathematics,
National Institute of Technology Durgapur,
Durgapur-713209, West Bengal, India,
shrabmandal@gmail.com,anita.buie@gmail.com

Abstract

Because of rapid sharing of information internet is an important repository. There are millions of websites. It is impossible to remember all of them during the access time. So we adopt a network analysis method by which we can retrieve similar category of websites. To get the above result we cluster the websites. Website clustering can contribute to spam website, pornographic website and political sensitive website detection. So it can be applied to websites supervision.

Keywords: *Component, Website Clustering, Social Network Analysis Website supervision.*

1. INTRODUCTION

Clustering is one of the most relevant features of networks representing real systems [5]. There exist various networks, including WWW, Internet, social networks, economic network, power network, traffic network, and neural network and so on. It is vital, as more and more researches have shown that those appeared different networks have striking similarities between each other. The web can be seen as the largest intent to store all human knowledge, either explicitly or implicitly. Internet actually is stupendous graph, in which web pages are nodes and hyperlinks are edges. In our research, we summarize web pages to websites as the nodes and user relationship as the content of query logs instead of hyperlink as edges. Traditionally, researches based on complex network of Web or Internet extract relations from hyperlinks, which direct user to another URL by the hyperlink. In the graph of web network, pages represent nodes, as well as hyperlinks represent edges. This graph is as a basis for the following studies, such as social network analysis (SNA)[7].

A social network is a social structure made of individuals or organizations called "nodes," which are connected by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike relationships, or relationships of beliefs, knowledge or prestige. Social network has many different styles. Real community is a common kind of social network, and virtual community is at its opposition. It is a rear kind of social network. Social network analysis views social relationships in terms of network theory consisting of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors. The method of social network analysis is a method to study inner construction of social relationship from a perspective of quantization, it can be used to demonstrate and measure the relationship of actors or all kinds of visible and invisible matters, like information or resource through the relationship .The resulting graph-based structures are often very complex. There may be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in determining the problems, the way to solve problems, functioning of organizations and the degree to which individuals succeed in achieving their goals. In its simplest form a social network is a map of all of the relevant ties between all the nodes [6],[8],[9].

ISSN 1512-1232

These concepts are often displayed in a sociogram, where nodes are the points and ties are the lines. Social network analysis in the narrow sense, is refer to a set of routine methods to analyze the characters of social network .It is the mapping and measuring of relationships which flows between people, groups, organizations, computers,URLs and other connected information or knowledge entities.

The nodes in the network are the people and groups while the links show relationships or flows between then odes. Social network analysis provides both a visual and a mathematical analysis of human relationships. Management consultants use this methodology with their business clients and call it Organizational Network Analysis. There are many indexes which been used in social network analysis, for example Degree centrality, Betweenness Centrality, Closeness Centrality and Network Centralization.

a) Degree Centrality- Social network researchers' measure network activity for a node by using the concept of degrees, the number of direct connections a node has. Degree centrality is defined as the number of links incident upon a node, i.e., the number of ties that a node has. Degree is often interpreted in terms of the immediate risk of node for catching whatever is flowing through the network, such as a virus, or some information. If the network is directed, then we usually define two separate measures of degree centrality, namely in degree and out degree. In degree is a count of the number of ties directed to the node, and out degree is the number of ties that the node directs to others.

b) Between ness Centrality-Betweenness is a centrality measure of a vertex within a graph.Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

c) Closeness Centrality -In topology and related areas in mathematics, closeness is one of the basic concepts in a topological space. Intuitively we say two sets are close if they are arbitrarily near to each other. The concept can be defined naturally in a metric space where a notion of distance between elements of the space is defined, but it can be generalized to topological spaces where we have no concrete way to measure distances. In graph theory closeness is a centrality measure of a vertex within a graph. Vertices that are shallow to other vertices have higher closeness. Closeness is preferred in network analysis that is shortest-path length, as it gives higher values to more central vertices.

d) Network Centralization-Individual network centralities provide insight into the individual's location in the network. The relationship between the centralities of all nodes can reveal much about the overall network structure. A very centralized network is dominated by one or a few very central nodes. If these nodes are removed or damaged, the network quickly fragments into unconnected subnet works. A highly central node can become a single point of failure. A network centralized around a well-connected hub can fail abruptly if that hub is disabled or removed. Hubs are nodes with high degree and betweeness centrality. A less centralized network has no single points of failure. It is resilient in the face of many intentional attacks or random failures many nodes or links can fail while allowing the remaining nodes to still reach each other over other network paths. Networks of low centralization fail gracefully [4].

In this paper, we propose a new method for extracting website relations from query logs of search engines for website clustering in order to detect some latent sensitive websites, which are usually illegal for their pornographic or political sensitive content [7]. Nowadays, search engines do a great job of sifting through billions of pages and millions of websites and returning search results highly relevant to user queries. In the last decade search engines have improved their performance to the point of becoming a tool of everyday use for most Internet users [2], [8]. Each item of search log contains two parts, search terms and returning URL. At first we extract the domain URL of original URL [10]. We classify the terms into 21 categories, which is defined in table 1 using maximum entropy model[11]. Then, we use the amount of each category to construct multidimensional vector. Based on the vectors, we compute the matrix of similarities between every

ISSN 1512-1232

two different website. If the similarity is beyond the threshold, we link a line of the websites, which we look as nodes in the graph. On the graph, we use clique partition method to find the clusters or cliques. The websites in the same cluster always have high similarities, which mean they have something in common and belong to the same kind. With pre-defined patterns, we can detect some latent sensitive website by labeling the cluster.

2. BODY

Body is designed in four sections. First section describes the clustering technique and applying it. Second section contains the idea of correcting error of entered website. Third section describes the searching. Next section contains the analysis of result.

2.1 CLUSTRING TECHINQUE

Over the recent past organizations and other users have been capturing increasingly large amounts of data that they wish to analyze. The amount of data being collected in databases today far exceeds the ability to reduce and analyze data without the use of automated analysis techniques. Knowledge Discovery in Databases (KDD) is an interdisciplinary field that is evolving to provide automated analysis solutions. The core part of the KDD process is the application of specific data mining methods for pattern discovery and extraction.Among the various data mining techniques, clustering of data plays a major role in extracting knowledge from the existing database. In this paper we focus on the k-means clustering technique and some other concepts, called Meta-clustering Technique [5].

2.1.1 K-MEANS Algorithm:

Our technique is based on the k-means clustering method. The working principle of it is described below:

Step 1: Arbitrarily choose k objects as the initial cluster centers.
Step 2: Repeat step 3 to step 5
Step 3: Reassign each object to the cluster to which the object is the most similar based on
the mean value of the objects in the clusters.
Step 4: Update the cluster means .i.e., calculate the mean value of the object's for each cluster.
Step 5: Until means remain unchanged.

Figure 1: An algorithm for k-means clustering

For applying any clustering algorithm, input is coordinates of the data files. The coordinates of the file is decided by the word which appeared more frequently than others.So we may not identify the coordinates of the files .This is the one of the important task of our technique to calculate the correct coordinate. In our paper coordinates are obtained by vectorization technique between files in database and input categorical files.

ISSN 1512-1232

Here we are using the twenty one categorical files which are given bellow.

Srl.No.	Category File's Name	Srl.No.	Category File's Name	Srl.No	Category File's Name
1	ECONOMICS	2	EDUCATION	3	PORN
4	SEX	5	HOMOSEXSUAL	6	MEDICINE
7	FOOD	8	MOVIE	9	GAME
10	GIOGRAPHY	11	SUPERSTITION	12	MILITARY
13	GRAPH	14	IT	15	HISTORY
16	SPROTS	17	POLITICAL	18	ENTERTAINMENT
19	LAW	20	CELEBRITY	21	RACE

Table 1: Website Categories

The k-means clustering technique is used here for clustering the data base. Our database contains only the name of the website. Before apply the clustering technique website is read & represented them in twenty one coordinates system with categorical data files because we want to represent the data within cluster as hierarchical order .This gives the fast search time. We also observe the size of each cluster because large cluster does not give the efficient search time. Here we are using the dot(.) operation by which each of category of table 1 is searched to entered the website. If the match occurs an individual counter variable will be incremented by one.
again. Steps in this module are given below:

Step 1: Prepare database and categorical data files.
Step 2: Repeat step 3 to step 10 for each file.
Step 3: Repeat step 4 to step 5
Step 4: Perform dot (.) operation between file and all categorical data files.
Step 5: Output of dot (.) operation is a coordinate stored in a matrix called vector.
Step 6: Apply k-means algorithm to vector matrix.
Step 7: Newly generated cluster's size is checked.
Step 8: If cluster size is large then the one fourth of vector matrix goes to step 5.
Step 9: If a cluster contains a less amount of data then merge it with another cluster which has also less amount of data.
Step 10: Finally store the final clusters in a centroid_cluster matrix.

Figure 2: An algorithm for representing files in coordinate system

2.2. TEXT CORRECTION

This module helps us to correct spelling of keywords. For correcting the spelling it generates the possible combination of input keyword .In the next module we will search all the possible keywords from the file.To generate the all possible combination of the input keyword we use a matrix. This matrix has 26 rows and 3 columns .Every row contains that three alphabets which are similar sounding words and uses make frequently typing mistake because of simi appearance in nearest distance in keyboard. Here we input the similar matrix which is given bellow:

ISSN 1512-1232

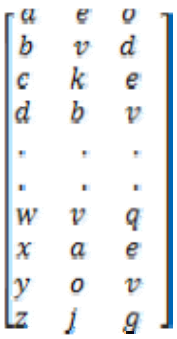
simi1= 

Figure 3: simi1 matrix

The steps which are followed in text correction module are given

Step 1: Accept the user website.
Step 2: Perform the dot (.) operation between the Input website name & simi1 matrix.
Step 3: the outputs of the dot (.) operation are store in website_ vector which represents the
coordinate of the entered website.
Step 4: Perform the dot (.) operation of the between each of the cat_web_matrix& simi1 matrixes.
Step 5: The output of the dot (.) operation are store at indivisual_cat matrix.
Step 6: Then calculate the distance between website_vector&indivisual_cat matrix.
Step 7: Identify the minimum distance.
Step 8: Fetch the data from the minimum distance of each of the cat_web_matrix files.
Step 9: Store all these website in web list vector.

Figure 4: An algorithm for text correction technique

The web_list vector gives the all combination of the input websites.

2.3. SEARCHING

In this module load the final_clusterfile which contains the final clusters. Then access all the possible websites which are stored in web_list matrix. Then represent all the website in the twenty one coordinate's system. There are some steps which are followed during this module:

Step 1: Load the final_cluster files.
Step 2: Represent all the websites in the twenty onecoordinates system & store the coordinates in
finput_vector.
Step 3: Calculate the distance betweencentroid_cluster matrix and finput_vector matrix.
Step 4: Find out the minimum distance and location of it.
Step 5: Fetch name of the websites from this location of final_cluster and display them.

Figure 5: An algorithm for searching technique

ISSN 1512-1232

2.4. RESULT ANALYSIS

The objective of our searching tool is to provide the correct result for all times even if the input keyword is wrongly spelled. When a user enters the website then our searching tool gives the list of websites which are similar to the entered website. We have used the database containing the name of the websites and cluster them after accessing the content of them. In this paper uses twenty one coordinates system. In our base paper the Girvan–Newman algorithm was user. This algorithm identifies edges in a network that lie between communities and then removes them, leaving behind just the communities themselves. The identification is performed by employing the graph-theoretic measure betweenness, which assigns a number to each edge which is large if the edge lies "between" many pairs of nodes. The Girvan–Newman algorithm returns results of reasonable quality and is popular because it has been implemented in a number of standard software packages. But it also runs slowly, taking time O (m^2n) on a network of n vertices and m edges, making it impractical for networks of more than a few thousand nodes. But here we use the k-means algorithm. The time complexity of it is O (n^2).

3. CONCLUSION

In this paper we have introduced a concept of website clustering using the k-means clustering technique, which avoids the misspelled input & generates the possible combinations of websites & searches all of them during the execution. This concept is built on the static database. In future we can apply this concept on the dynamic database.

REFERENCES:

1. Bin Wu, Qing Ke, Yuxiao Dong,"Degree and similarity based search in Networks", Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),2011.
2. Bo Yang,"Communities Detection with Applications to Real-World Networks Analysis",IEEE Seventh International Conference on Computational Intelligence and Security,2011
3. Huiqing Niu, "Social Network Analysis of University Online Forum",IEEE International Conference on Computational Aspects of Social Networks,2010.
4. Mandal S., and Silakari,"Meta-Clustering Tecniques for Document Searching",TRACE-2010.
5. Prantik BhaUacharyya, Jeff Rowe, Shyhtsun Felix Wu, Karen Haight, NiklasLavesson, and Henric Johnson, "Your Best might not be Good enough:Ranking in Collaborayive Social search Engines",7th International Conference on Collaborative omputing: Networking, Applications and Worksharing(CollaborateCom), Orlando, Florida, USA, October 15-18, 2011.
6. Wang, Bin Wu, Zhonghui Zhang."Website Clustering from Query Graph using Social Network Analysis",IEEE, 2010.
7. M.K.Mike Cassidy,"An update to google social search", http://googleblog.blogspot.com/2011/02/update-to-google-social-search.html,Febuary 17, 2011.
8. S.Wasserman,"Social Network Analysis: Methods & Application",Cambridge University press,1994.
9. P.P.Index,"Content term extraction using pos tagging",http://pypi.pyton.org/pypi/topia.term extract/,June 15 2011.
10. Yao-zong Liu ,Yong-li Wang, Wei Wei Hong Zhang , Feature Selection for Classifying Data Stream Based on Maximum Entropy, Pattern Recognition, 2009.

Article received: 2013-07-12