Exploring Sub Dominant Community on Web Graph: Using Link Structure and Usage Analysis

Nimisha Modi[1]

Submitted in June 2014; Accepted in March, 2015

Abstract - Information Retrieval (IR) process uses term based relevance measures to find relevant documents for a given query. Web IR utilities such as search engines tend to further process these relevant documents through link structure analysis and find rank score for each document within result set. The rank scores are used to sort the documents before presenting them to users for improving precision rate of top ranked results. Existing link analysis algorithms are using principal eigenvector of corresponding rank matrix for ranking. The limitation of using this approach is Web Local Aggregation (WLA). As an effect of WLA for the multi topic or polymorphic query, the dominant topic covers the major part within top ranked results and the sub-dominant topics are downgraded. The propose approach for link analysis serves for both ranking and grouping. Semantic analysis is incorporated along with link analysis to identify sub-dominant community through link analysis process and upgrade them according to the interest of the user searching on web. The paper suggests the model for search agent that categorizes the results based on their hyperlink connectivity and presents the groups of web pages that match semantic of user profile.

Index Terms – Eigenvector, Information Retrieval, Link Analysis, Usage profile, Web local aggregation, Web community

1.0 INTRODUCTION

Web contains huge amount of information. This information is available in form of millions of documents with varying degree of relevance. Specifically with respect to broad-topic queries, search engine returns a set of thousands or millions of documents as result. As the user's satisfaction requires few but highly authoritative results, the web search utilities encounter the challenge for retrieving only the most relevant resources in response to user query. This challenge raises the requirement to further process this result set before presenting it to user. The post-processing includes the expansion of result set, ranking of pages within result set, classification or clustering of result set etc... Basically, majority of the search engines assign the rank scores to web pages based on their hyperlink structure on web. As for example, the search engine Google pre-computes the PageRank [1] and presents the organized results to user in sorted order of PageRank scores.

Various research works are targeted towards statistical or machine learning algorithms for multi topic or polymorphic query. The broad topic query is a query for which various semantic are present on web documents. In response to broad topic query, normally the dominant topic or dominant semantic will appear among the top ranked results. Here, dominant topic or dominant semantic means the topic having more number of web documents that are highly connected with each other than that of documents with any other semantic.

For example, term 'java' is related with programming language java or island java. If we search web using query 'java', the index searching returns documents which contains term 'java' without concern of their semantic (either programming language or island or java coffee). These millions or billions of web documents returned through index searching of search engine are re-arranged by search engines in order of documents' ranking scores obtained through proprietary link analysis algorithm. Obviously, the major portion of result set present java in context of a programming language and these documents are even highly connected. As consequences, the documents with programming language 'java' cover the major portion in top ranked result set and user are not able to find those documents containing information about island 'java'.

My objective is to help users who search for diverse aspect of broad topic. I identify documents having various aspects of topic (i.e. various categories) with use of principal component analysis. Rather than using content of documents, I use their hyperlink structure for identifying groups of related documents.

2.0 RELATED RESEARCH

I analyze research works that is done in area of hyperlink structure analysis, web page categorization and semantic analysis based on user profile within information retrieval field. Hyperlink provides very rich information in addition to text that sometimes beats the text in form of quality and reliability. The role of hyperlink [2] is to confer the trust that one document puts on other document via the hyperlink to the later. Network of social interaction are found between web pages by hyperlink to other web pages. For exploitation and analysis of hyperlink structure on web graph [3], a number of algorithms have been introduced. All of them generally follow and possibly improve the concept of three basic algorithms: HITS [4], PageRank [1] and SALSA [5].

Basically these three algorithms work to assign a numerical weigh (rank) to each element within a hyperlinked set of documents on web graph.

These link analysis algorithms calculate rank based on varied type of neighborhood graphs that are represented by some adjacency matrix or transition probability matrix. They use the principal eigenvector [6] of matrix to assign the rank score for web pages. Google's PageRank is a variant of the eigenvector centrality measure.

[1]*Department of Computer Science, VNSGU, Surat, Gujarat.*
E-mail:nimishamehta@yahoo.com

Algorithms for web page categorization use different information containers such as - URL, unstructured text content, structure of the web page within markup tags, snippets i.e. short description of pages displayed with initial results of base search engine and linkage information in the form of incoming and outgoing links. Learning model for automatic classification [7, 8] is based on a combination of text and link analysis for distilling authoritative web resources.

Categorization of web pages is broad research area [9, 10, 11] where a huge variety of classification as well as clustering algorithms are introduced. As supervised learning or classification requires pre-specify topic taxonomy, the approach is limited for customized use only. Web directory are generally using such models to find the web page that match priory specified label. Majority of categorization algorithms are following unsupervised leaning i.e. clustering [10, 12, 13] to group web pages according to some intrinsic similarity within them either in term of content or structure.

Web usage mining [14] refers to the discovery of user access patterns from web usage logs. I analyze the research works which focus on use of usage profile before finding similarity of query phrase and documents. Hang and his collogues [15] proposed a method for query expansion by mining user logs, where the user profile is analyze prior to query is being submitted to the search utility and based on that the query is being expanded.

The limitation of such query expansion or refinement is that - the search becomes narrow (specific) search. Possibly user may require searching for some other perspective of topic. For example, a person in biology may require a search for type of virus that infected his laptop, in such case query expansion approach needlessly limits the search results to biological context and fails to satisfy user's requirement.

3.0 PROPOSED METHODOLOGY

I introduce some different approach that molds the application of link analyses algorithm to identify various community on web graph. For usage analysis, I prefer to utilize the user profile at post-link analysis stage rather than pre-processing stage as suggested in past studies [15].

3.1 Application of Link Analysis beyond Document Ranking

The limitation of ranking algorithms (PageRank [1] or HITS [4]) is that - they promote the highly interconnect dense set of nodes on web graph. As a consequence of web local aggregation (WLA) [16], the sub-dominant communities are downgraded by the dominant community. User could rarely find documents representing sub-dominant category in top 50 results and user are not even try to explore more than top 20-to-30 documents. My goal is to identify such sub-dominant communities and to upgrade them while presenting results based on interest of that user.

Existing link analysis algorithms are using only principal eigenvector of corresponding rank matrix. I am interested in higher order eigenvectors rather than only principal eigenvector. I take higher order eigen values and their corresponding eigenvectors. By looking at a larger set of eigenvectors, I find clusters of web pages that reflect through web local aggregation. The most important web community corresponds to the principal eigenvector and the component values within each eigenvector represent a ranking of web pages. The subsequent eigenvectors denotes corresponding minor communities on web graph

3.2 User Profile for Post-Query Processing

I suggest an approach of using users' profiles after the initial results are collected but before presenting these results to user. The approach is based on two objectives – first, we require broad search before any semantic analysis and then perform grouping of search results. The reason behind this approach is that we can present all the result as a broad query as well as specific to some semantic in different groups, as user's context of search query may be or may not be according to his previous search activity. With proposed approach, we use the user profile to make it convenient for user to find the topic of his routine at top results and to degrade the outliers that selected incorrectly within top results through link analysis process.

The second objective is to propose such a model that can be used by general purpose search engine. Evangelos traces [17] a popular search engine (Excite) to show the significant locality in the queries. More than 20%-30% of the queries have been previously submitted by the same or a different user. Search engines follow the practice to make a cache for query results for frequent queries. Research Experiments also shows the improvement in hit ratio via two level caching i.e. a cache of query result and a cache of inverted list of query terms [18]. So, in place of raw query result, results of link analysis can be stored as cache for frequent and broad topic query. When a search is performed on that query, it just needs to map the proper group of results. At time of user's search for such query, the semantic analysis is performed and results are collated from the cache copy.

3.3 System Model

Based on two approaches describe in section 3.1 and section 3.2, I propose the model for supporting system for web information seeker which now onwards will be referred as Search Assistant (SA). SA works as interface between user and search engine to monitor the search process and assist web users to intelligently retrieve information from the web. The model for the proposed system is described in figure 1.

This section outlines the flow of the system in three phases - collecting base result set, link analysis phase and semantic analysis phase.

- Collecting Base Result Set: User is provided the interface for interactive input for query that is submitted to search engines. It collects the initial results of user's query from search engines using the APIs from search engines like Google, Yahoo and MSN Search. Title and snippet returned by search engines is also stored along with URL.

This initial result set is referred as root set. The in-links and out-links of root result set are also collected. The collection of URLs in the root set, their in-links and out-links are collectively known as base set.

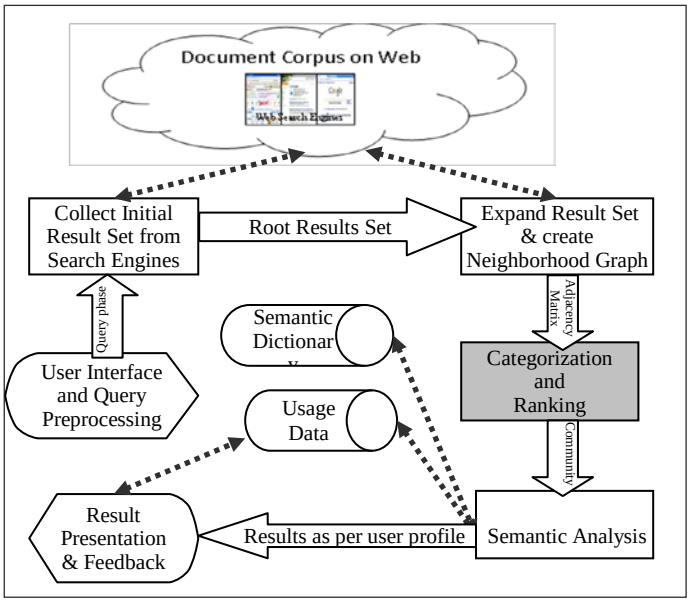

Figure 1: Proposed Model

- Link Analysis Phase: Link analysis algorithm visualizes this base set as the neighborhood graph N where each page (link) represents a node and a hyperlink from one page linking to another page is represented as directed edge. This neighborhood graph reflects linked structure of web pages in base set. We submit this neighborhood graph in form of adjacency matrix as an input for link analysis algorithm. Considering A as the adjacency matrix on neighborhood graph N, the authority matrix [6], AUT is derived from adjacency matrix A as $AUT = A^T * A$.
  SA applies principal component analysis (PCA) and make use of eigenvectors of adjacency matrix to identify the principal components i.e. web communities. So, next step is to calculate the set of eigen values and corresponding eigenvectors for authority matrix AUT.
  The eigenvector corresponding to highest eigen value is the principal eigenvector that denotes the dense set of web pages on web and thus interpreted as dominant community. The next higher order eigen value is corresponding to second dense set of web pages and thus second dominant web community and so on. We form the groups of web pages for each higher order eigenvector via collecting web pages having high rank score on corresponding eigenvector. Each group presents different web communities those pertaining to different semantics/topics.
- Semantic Analysis Phase: For personalization of results, SA filters the result set based on user profile. User profile basically describes the user's preference. The trivial approach to built user profile for IR is to create a set of words or word phrases that are frequently used by that user.

  SA collects the user profile using web access logs of users that are collected from web proxy server. In convention IR system, similarity measure is computed through matching documents to query phrase.

I suggest query post-processing technique ─ the set of documents that SA analyzes is already judged as query relevant by search engines. So to find similarity scores for documents, this IR model substitute query phrase with user profile. SA finds similarity of all groups with user profile and presents the set of web pages i.e. web communities according to the interest of user. Thus the search becomes general as well as specific search.

4.0 EXPERIMENTS AND FINDINGS

I select some broad topic or polymorphic terms and apply search using the user-id of users with different profiles. Form these experiments I pick 5 queries to discuss within this paper, which are polymorphic and leading to more than one semantic i.e. java, mouse, jaguar, virus, tree.
I identify dominant and sub-dominant set of results for selected topics. As per that, results of 'jaguar' on selected search engines categorized in to two categories: jaguar as well-known automobile manufacturing company as dominated category and jaguar as wild animal category. Similarly, categorization of results for query 'mouse' is animal mouse or an electronic input device. 'java' is categories as programming language and island. 'category of viruses' is classified as computer virus and biological virus. 'tree' is either woody plat or data structure .
I use the terminology RRS, LARS, SARS to evaluate the results of our experiments at different phases. RRS (Root Result Set) – precision rate for results returned by base search engines. LARS (Link Analysis Results Set) - precision rate for results set formed after link analysis and before semantic analysis. SARS (Semantic Analysis Results Set) - precision rate for results after semantic analysis.
The precision rates RRS, LARS and SARS of the results obtained for identified dominant and sub-dominant communities are given in table 1 and table 2.

Table 1: Search Precisions for dominant category

Query Term	RRS	LARS	SARS	LARS-RRS	SARS-RRS	SARS-LARS
java	90.00	86.00	98.00	-4.00	8.00	12.00
mouse	51.33	76.00	90.00	24.67	38.67	14.00
jaguar	52.00	78.00	98.00	26.00	46.00	20.00
category of virus	66.67	64.00	88.00	-2.67	21.33	24.00
tree	54.00	81.00	90.00	27.00	36.00	9.00
Average				14.20	30.00	15.80

The comparative analysis of table 1 and table 2 shows that for dominant community, link analysis phase shows average

14.20% increase in precision while semantic analysis boosts 30% increase in search precision than that of root results (RRS).

Similarly for sub-dominating community, link analysis shows average 42.27% increase in precision rate while semantic analysis boosts 58.87% increase in search precision than that of root results (RRS).

Table 2: Search Precisions for sub-dominant category

Query Term	RRS	LARS	SARS	LARS-RRS	SARS-RRS	SARS-LARS
Java	6.00	50.00	68.00	44.00	62.00	18.00
mouse	33.33	70.00	85.00	36.67	51.67	15.00
jaguar	22.67	66.00	90.00	43.33	67.33	24.00
category of virus	21.33	60.00	80.00	38.67	58.67	20.00
Tree	11.33	60.00	66.00	48.67	54.67	6.00
Average				42.27	58.87	16.60

Improvement from LARS to SARS is 16.60% in case of dominating community; while in case of subdominant community it is 15.80%. So the role of semantic analysis is stable in both the case. These findings illustrate the significance of link base categorization, as we conclude that the categorization of grouping using link analysis plays a major role to boost sub-dominating community on web.

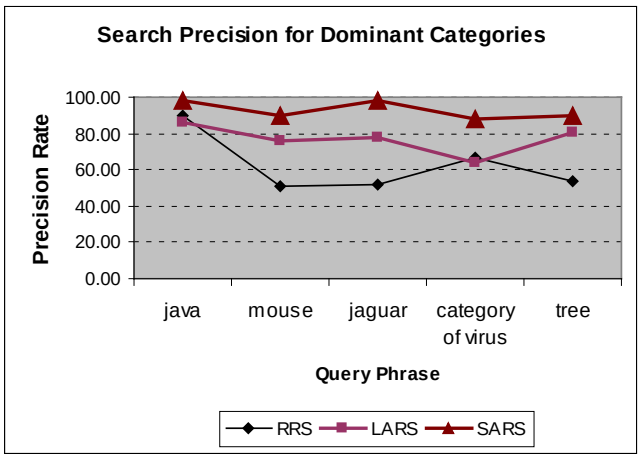

Chart 1: Search Precision for Dominant Category

Link analysis finds web communities based on hyperlink structure. It finds the sub-dominated communities. The role of semantic analysis is to identify the proper web community based on user's interest.

We can visualize the effect of both link analysis as well as semantic analysis on above results with chart 1 and chart 2.

Link analysis finds web communities based on hyperlink structure. It finds the sub-dominated communities. The role of semantic analysis is to identify the proper web community based on user's interest.

We can visualize the effect of both link analysis as well as semantic analysis on above results with chart 1 and chart 2.

5.0 CONCLUSIONS & FUTURE PATH

The proposed model is targeted towards broad topic query that applies PCA with eigenvectors for finding the group of web pages having common interest. Dominate category is having maximum resources than that of other. Dense linkage structure as well as rich text content of such dominating category suppressed the sub-dominating category via web local aggregation. SA uses the link structure of base result set to identify such set of highly linked group of web pages and using the principal component analysis we identify the dominated as well as some sub-dominated groups. Among these groups of documents, it performs the semantic analysis i.e. based on user profile.

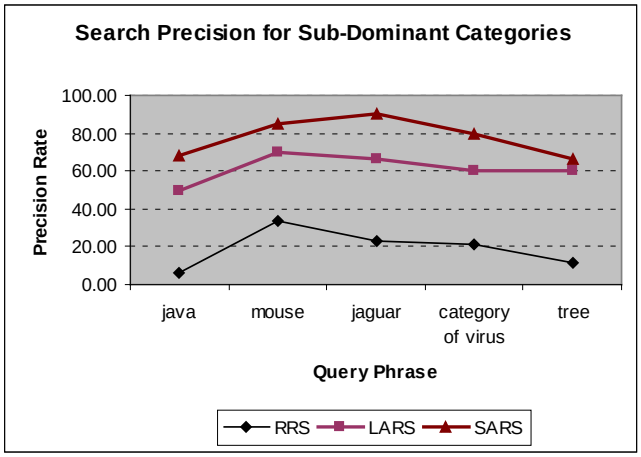

Chart 2: Search Precision for Sub-Dominant Category

The major limitation of algorithm is the central assumption that 'a hyperlink confers authority' which is largely applicable for social networks of the academic publications, but it is not guaranteed for commercial web pages. Sometimes, web sites are generally designed by commercial developers who link up their customers in densely connected cliques even though those customers have nothing in common

The algorithm presented in the paper generates the flat clustering on different themes. This can be enhanced towards hierarchical grouping in tree of topic [19] by analysis of inter-connection between documents on neighborhood graph.

Image, multimedia and other embedded objects are big sources of information, which are ignored except the text (e.g. anchor text, alternate text etc...) within their container markup tags. This indicates the other direction for future enhancement of the given model.

REFERENCES

[1]. Sergey Brin, Lawrence Page, "The anatomy of a large-scale hyper-textual Web search engine", In Computer Networks and ISDN Systems, 1998, 33:107-117.

[2]. Mandar R. Mutalikdesai, Srinath Srinivasa, "Co-citations as citation endorsements and co-links as link

endorsements", Journal of Information Science", v.36 n.3, p.383-400, June 2010

[3]. Lei Tang, Huan Liu, "Managing and Mining Graph Data, Advances in Database Systems", 2010, Volume 40, 487-513, DOI: 10.1007/978-1-4419-6045-0_16

[4]. Jon Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 1999, 46:604-632

[5]. R. Lempel, S. Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect", 9th International WWW Conference, 2000

[6]. Amy N. Langville, Carl D. Meyer, "The Use of Linear Algebra by Web Search Engines", Bulletin of the International Linear Algebra Society, 2005, 33:2-6.

[7]. Soumen Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan, "Automatic resource list compilation by analyzing hyperlink structure and associated text", Proceeding of 7th International World Wide Web Conference, 1998.

[8]. Ajay S. Patil, B.V. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", Proceedings of The International MultiConference of Engineers and Computer Scientists, IMECS 2012, March 14-16, 2012, Hong Kong, VOL I, Page No.

[9]. Xiaoguang Qi, Brian D. Davison, "Web Page Classification: Features and Algorithms", in Technical Report. 2007, Department of Computer Science and Engineering, Lehigh University: Bethlehem, PA. p. 1-31.

[10]. Gulli A., "On Two Web IR Boosting Tools: Clustering and Ranking", PhD Thesis, University of Pisa, May 2006

[11]. Ghanshyam Singh Thakur, Dr. R. C. Jain, "NFCKE: New Framework for Document Classification and Knowledge Extraction", BIJIT - BVICAM's International Journal of Information Technology, January-June, 2009, Vol.1 No.1; ISSN 0973-5658

[12]. Deepak P, Deepak Khemani, "Unsupervised Learning from URL Corpora", Proceedings of the 13th International Conference on Management of Data (COMAD-2006), Delhi, India

[13]. Anil Kumar Pandey, T. Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2009, Vol.1 No.2; ISSN 0973-5658

[14]. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations. Vol. 1–2. 2000

[15]. Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma, "Query Expansion by Mining User Logs", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 4, pp. 829-839, July/Aug. 2003.

[16]. Ziyang Wang, "Improved link-based algorithms for ranking web pages", proceeding of 5th International Conference of Web Age Information Management, 2004.

[17]. Evangelos P. Markatos, "On Caching Search Engine Query Results", In Proceedings of the 5th International Web Caching and Content Delivery Workshop, May 2000.

[18]. Amarjeet Singh, Mohd. Hussain, Rakesh Ranjan, "Two Level Caching Techniques for Improving Result Ranking", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2011 Vol.3 No.2; ISSN 0973-5658

[19]. Parul Gupta, A.K. Sharma, "A Framework for Hierarchical Clustering Based Indexing in Search Engines", BIJIT-BVICAM's International Journal of Information Technology, July-December, 2011 Vol.3 No.2; ISSN 0973-5658.