ISSN:0975-9646

Study of Various Algorithms on Vertical Search in Medical Domain

Roshini R Nair[1], Bhagyshri Shukla[2], Manjiri Lad[3], Prof. Ashraf Siddiqui[4]

[1,2,3]*BE Student,* [4]*Assistant Professor*
[1-3] *Department of Computer Engineering, Theem College of Engineering*
University of Mumbai, Mumbai, Maharashtra, INDIA

Abstract- The exponential growth of dynamic, unstructured data stored on web made it difficult to locate useful resources. Thus made imperative to identify effective means for clinicians, administrators and researchers to use data. Search engines like Google and Alta Vista, yahoo which gives huge amount of information many of which might not be relevant to the users query. In this paper vertical search engine accepts seed URL and dividethe URLs spidered on medical domain according to its page contents.It divides the web pages loaded by spider into particular domains.The spider checks accuracy of index choosen. The users search the keyword by firing query on database.The . It is based on page relevance for accurate domain and content to improve the quality of URLs thus removing irrelevant data.

Keywords **- Domain classifier, page ranking, vertical search, web crawler.**

I. INTRODUCTION

In the Vertical search engine it gives a list of documents were keywords are matched. Vertical search engine give accurate, relevant and faster search by indexing in specific domains. Examples of vertical search engines are Financial Search Engines, Law Search Engines, etc For example searches in google, yahoo, alta vista it doesn't gives relevancy to the topic whereas in this search engine it gives accurate information with upgrade link. thus saving time and space of users in this busy environment. It filters the pages and gives exact information according to ranking of webpages. User visiting vertical search engine must have basic knowledge on medical domain thus avoiding irrelevancy of search. One of the company started with 180 vertical search engine which provided them more ease to search on various domain.it provide online solution for all medical relevant data which is not provided in the horizontal search engine [1]. For example if we want to switch to a particular language program on television thus satisfying the needs of user such way vertical search engine works. It solve problem by indexing on particular domain by using BFS-traversal algorithm without refining using heuristic search.

2. PAGE RANK ALGORITHM

Page rank assigns a measure of "Prestige" or ranking to each web page, independent of any query.[1] The page rank has two distinct underlying motivations. The first motivation comes from the ranking academic citation literature which existed long before the advent of the web. It is defined using a digraph based on the hyperlink structure of the web called the web digraph. The web digraph W is the digraph whose vertex set V(W) consists of all web pages, and whose edges set E(W) corresponds to the hyperlinks that is an edge is included from page p to page q whenever there is a hyperlink reference (href) in page p to page q.

The network information produce many spam pages which is not even related to topic and is not desirable thus it adds page rank algorithm to fiter the websites whose authority is low each webpages measures the importance of authority of value The basic idea is referenced by a large number of high quality webpages

3. THE IMPLEMENTATION OF TOPIC RELEVANCE ALGORITHM

The software robot searches the URL link list and crawl to the relevant page. The index is analysed by searching pages which is needed to the topic.[5] It uses various method such as full text scan, Boolean model, probability model, vector space model. Here vertical search engine choose the space vector model. This model gives better effect and efficient application. Vector space model is a web page that take various keywords that gives the exact topic web pages is the main feature[6] The document feature the vector keyword in the following webpage Each feature have important by calculating the weight. the weighting is done by the features frequency in the web and the page number the pages have different position and weight. the calculation formula of the feature item in the webpage as follows are ;

Rt =mRt title+nRt meta +p Rt anchor+qRt normal +kRt(3)EM

The Rt is the frequency of feature keyword .Rt title, Rt meta Rt anchor, Rt normal Rt EM are the title of page,page meta information, web pages hyperlinked text, normal text and keyword frequency emphasis m, n, q and k.[7] greater than threshold it passes to spider index and when the value is less then it gets rejected. Basically the threshold value is 0.85.

4.HITS ALGORITHM

It gives textual information based on top ranked webpages by HITS algorithm such as frequency occurrence of the query string on the page, whether a query string occurs in bold or large font size, page rank and so forth.[13] The web pages choosen is based on predetermined number say r,r=200.HITS then augment the root set to obtain a base set.

HITS limits the number of pages it chooses to some reasonable number of pages choosen in arbitrary fashion from the neighbourhood pages. Also it can delete pages from popular sides which would tend to be the authorative pages with respect to any query string they contain. Web page would be brought into the base set B and focused subdigraph F_q. We define the hub value h[p] of page p to the sum of the authority value overall pages q in its out neighbourhood $N_{out}(p)$ of the focus of digraph F_q and the authority value to be the sum of the hub value overall pages q in its in neighbourhood and $N_{in}(p)$ of F_q.

$h[p] = \sum_{q \in Nout(p)} a[q]$,

$a[p] = \sum_{q \in Nin(p)} h[q]$.

A denote the adjacency matrix of F_Q, and a denote the column vectors whose entries corresponding to page (vertex)p are h[p]and a[p],respectively,

$a = A^T h$,

$h = Aa$.

Combining above equtions ,

$a = A^T h$ tical search engines,$= A^T(Aa) = A^T Aa$,

$h = Aa = A(A^T h) = AA^T h$.

$A^T A$ is the authority matrix and AA^T is hub matrix ,and page rank a and h are the principal eigen vector of $A^T A$ and AA^T respectively.

5. WEB CRAWLER

The different Web content and structure analysis technique to build software robot program for vertical search engines, with three versions of web crawlers are a) Breadth-First_Search(BFS) crawler. b) Page Rank web crawler c) Hop field Net Crawler.

The BFS follows simple algorithm.The PageRank web crawlers employs a Best –first search, using PageRank score of each unvisited URL as its heuristics. The Hopfield Net crawlers is a spreading activation algorithm. It weights all the nodes and is activated parallel and value from different sources are combined until it gets a stable state. It set a seed URLs as node and activate neighbouring URLs until it get a new nodes.[2][3]

$\mu(t+1) = f_s(\sum w_h)$ The weight is calculated and spider decide which node to visit first. By the weight it gives approximation of the algorithm.it gives in decreasing order of the other URL visited.The greater value of threshold decides the previous page quality and relevance of the downloaded page content.

6. WEB PAGE FILTERING

Web filtering is featured based approach in order to improve the efficiency of the webpage.

Using web page filtering algorithim instead of taking each document as bag it reduces,and represented by the limited number of features.

The relevance property and quality of webpage can be represented as follows:

(1)page content

(2)page content of neighboars

(3)link information.

Page releted content score can be represented based on a domain lexicon in which Two features will be used:

1. Title(*p*) = Number of content terms found in the title of page p in domain lexicon.
2. TFIDF(*p*) = Sum of TFIDF of the terms in page p found in domain lexicon.

Six features will be used in the web page fitering algorithim: The averages of the two scores for all incoming neighbours. The averages for all outgoing neighbors, and The averages for all siblings.[8]

Connectivity is used to determine the quality of a webpage. Link analysis (such as number of in-links, HITS and PageRank) have been useful in many other Web applications such as search result ranking but have not been used in text classification.

Six scores, namely

(1)hub score,

(2)authority score,

(3) PageRank score,

(4)number of in-links,

(5) number of out-links, and

(6) number of relevant terms in the anchor texts, will be used as features.

In the web page fitering algorithim there have 14 features which is used to input to a classifier. A Feedforward or backpropagation neural network will be accepted because of its robustness and wide usage in classification distribution.[10]. There will be one single output node determining the combination of relevance and quality of a page, in order to allow better comparison, a support vector machine also will be used because of its efficient performance in text classification . It will be used to perform classification based on the same feature quality scores.

7. HOPFIELD NET SPIDER

The Hopfield Net Spider algorithim have a spreading activation algorithm. In this Hopfield net spider algorithim the Web is derive as a Hopfield Net, a single-set weighted neural network Nodes are started in parallel and started values from different sources are collected for each single node until the starting scores of nodes on the network reach a stable state. The Hopfield Net Spider algorithim starts with a set of seed URLs link presenting as nodes, activates neighboring URLs, joined weighted links, and determines the weights of newly arrived nodes. At starting set of seed URLs is given to the system and each of them is presented as a node with a weight of value 1.[9] $\mu i(t)$ is value defined for the weight of node *i* at iteration value of *t*. The spider algorithim takes and analyzes these seed Web pages in iteration till 0. The new URLs link are found in these pages are automatically added to the network. Moving to the next iteration, the weight of each node is computed as follows:

$\mu i(t+1) = f_s$

$(\sum W_{n,i} U_h(t))$

where *wh,i* is the weight between two nodes and *fs* is the used as SIGMOID transformation function that normalised the a weight to a value satisfying the range from 0 and 1. After the weights of all the nodes in the current iteration value are computed, the spider algorithm should decide which URL node should be visited first. Using this it gives the weights decide the order in which URLs are to be

arrived it is very difficult to the effectiveness of the algorithm. The set of nodes in the current iteration are then visited and taken from the Web in decreasing order of weight. After all the pages with a weight higher than a threshold value have been visited and loaded, the weight of each node in the new iteration is upgraded to reflect the quality of relevance of the loaded page content. This process s repeated until found the needed number of web pages have been collected.

8.CONCLUSION

Finally, this paper concludes after surveying various algorithm as like hits algorithm,page rank,topic relevance,hopfield,web crawler, web filtering algorithms.we are coming with the best approach in order for cheaper and maximum throughput for the optimum and efficient search.

REFERENCES

[1] M.S.Aktas, M.A.Nacar, and F.Menczer,"Personalizing PageRank Based on Domain Profiles", WebKDD 2004.

[2] K. Bharat and G..A. Mihaila, "When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics",ACM Transactions on In-formation Systems, Vol. 20,No. 1, pp. 47-58, 2002

[3] Web Site: Google scholar paper search.http://Scholar.google.com

[4] Froogle. Google product search. http://froogle.google.com

[5] GoogleVideo. Google video search. http://video.google.com

[6] Google. Google image search. http://images.google.com

[7] Page L., Brin S., Motwani, R. & Winograd T. (1998) "The pagerank citation ranking: Bringing order to the web", In Technical report, Stanford Digital Libraries, pp. 1-17.GoogleNews. Google news search. http://news.google.com

[8] Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Qiang Wu, Ran Gilad-Bachrach, and Tapas Kanungo. On composition of a fed-erated web search result page: using online users to provide pairwise preference for heterogeneous verticals. In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pages 715–724, New York, NY, USA, 2011. ACM.

[9] Fangpeng Dong and Selim G. 2006Scheduling Algorithms for Grid Computing State of the Art and Open Problems.School of Computing, Queen's University Kingston, Ontario.

[10] Vijay Subramani, Rajkumar Kettimuthu, Srividya Srinivasan, P. SadayappanR 1994.Distributed Job Scheduling on Computational Grids using Multiple Simultaneous Requests,Department of Computer and Information Science. The Ohio State University.

[11] Amr Rekaby1 and Mohamed Abo Rizka 2013 A COMPARATIVE STUDY IN DYNAMIC JOB SCHEDULING APPROACHES IN GRID COMPUTING, Egyptian Research and Scientific Innovation Lab (ERSIL), Egypt Arab Academy for Science, Technology and Maritime Transport College of Computing & Information Technology, Cairo, Egypt.

[12] Jun Zhang and Chris Phillips, Job-Scheduling with Resource Availability Prediction for Volunteer-Based Computing, Queen Mary, University of London.

[13] A. K. Aggarwal and R. D. Kent, 2005 An Adaptive Generalized Scheduler for ranking Applications, in Proc. of the 19th Annual International Symposium on High Performance Computing Systems and Applications (HPCS'05), pp.15-18, Guelph, Ontario Canada.

[14] H. Kawamura, M. Yamamoto, K. Suzuki, and A. Ohuchi, 2000 "Multiple ant colonies algorithm based on colony level interactions", IEICE Trans. Fundamentals, Vol. E83-A, No.2, pp. 371-379.

[15] J. Jong, and M. Wiering, 2001 "Multiple ant colony system for the bus-stop allocation problem", Proc of the Thirteenth Belgium Netherlands Conference on Artificial Intelligence BNAIC'01, Amsterdam, Netherlands, pp. 141–148.

[16] Kruger.F, Middendorf.M, and Merkle.D," 1998 Studies on a Parallel Ant System for the BSP Model" Unpub. Manuscript.

[17] M. Ciubancan, M. Marinescu, O. Grigoriu, G. Neculoiu, V. Sandulescu, I. Halcu, "Computer aided learning with a ranked based approach technologies", 10th RoEduNet IEEE International Conference, STEF, pp 222-225, 20.

[18] M.S. Aktas, M.A. Nacar, and F. Menczer,"Personalizing PageRank Based on Domain Profiles", WebKDD 2004.

[19] Open Directory Project http://www.dmoz.org

[20] HITS ALGORITHM studied in Analysis of Algorithm.