GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

COMPARISON OF FILTERING RESULT FROM DECISION TREE BASED TECHNIQUE AND COUNT OF THE SEARCH WORDS

Neelu Tiwari [1], Sujeet Kumar Tiwari[*2]

[1]Computer Science Department ,Hitkarini College of Engineering and Technology, Jabalpur,MP,India

[2]Computer Science Department ,Lakshmi Narain College of Technology, Jabalpur,MP,India

neel_31@rediffmail.com, sujeet.tiwari08@gmail.com

ABSTRACT

Maximum search engines use only the search keywords for searching. Due to the ambiguity of semantics and usages of the search keywords, the results are noisy and many of them do not match the user's search goals. In general the search keywords are used by multiple search engines for searching. Semantics may be produce ambiguous and noisy results, which may result in non-matching search target. Here we shall discuss a two methods that shall filter the results in such manner that the results are well arranged and more intensive.We have two methods Decision Tree Based Technique and the new method shall make use of the page rank along with count of the search words that increase the weight age of the page URL.

Keywords: search intention, information retrieval, data mining, machine learning, search intention

I. INTRODUCTION

Growth of number of websites, the page size and the number of pages has increased more Frequently. There is a necessity to upgrade the search engines. There are two types of methods prevailing in the market on which the search engines work, one is Page Rank and other is HITS. Page Rank is used by Google to display the results of the query; while HITS is used in Clever System of IBM. A search query is entered by the user in the hope of desired results. Since the searched keyword or phrase may have several meanings and usages, many unwanted results are expected to be present in the search results returned by the search engine. Even if the results displayed correspond to the correct meaning of the phrase, still many unwanted and irrelevant results are expected. Moreover, a search result for a given search query might be relevant for one user and irrelevant for another. Thus there is a need of a user-centric search, which sorts results according to user specific search goals.

The PageRank algorithm, introduced by Page et al. [4, 5], precomputes a rank vector that provides a priori"importance" estimates for all of the pages on the Web.The vector is computed once, offline and is independent of Search Query. At query time, these importance scores are used in conjunction with query specific IR scores to rank the query results. PageRank has a clear efficiency advantage over the HITS algorithm, as the query-time cost of incorporating the precomputed PageRank importance score for a page is low.

Many web search engines use only keywords as queries.The users type in the search query hoping that they will get the desired results . Since the searched keyword or phrase may have several meanings and usages, many unwanted results are expected to be present in the search results returned by the search engine. Even if the results displayed correspond to the correct meaning of the phrase, still many unwanted and irrelevant results are expected. Moreover, a search result for a given search query might be relevant for one user and irrelevant for another. Thus there is a need of auser-centric search, which sorts results according to user specific search goals.

II. MATERIALS AND METHODS

We have two approaches –

1. Using decision tree approach.
2. Using word count approach .

1.Using decision tree approach

This approach follow below 8 step for filtering relevant result.

Step 1. User Training

Step 2. Inferring knowledge from the Training set

Step 2.1. Gathering Information from the Training Data

Step 2.2. Data Cleaning

Step 2.2.A. Filtering words of minimum length

Step 2.2.B. Removing parts of speech and inappropriate Words

Step 2.3. Decision Matrix

Step 2.4. Decision Tree

2. Using word count approach

The approach we are focusing over here is explained here.

1. The initial URL set is composed of the hot websites selected manually. These URLs are saved in the database and loaded into URL queue when the program starts.
2. The structured information is acquired by matching web content with keywords. New URLs, such as, the next page URL of the list, or the URL of each post, can also be extracted from the content.
3. The crawler displays pages based on the word count retrieved from the page URLs. The display is updated incase of recrawl. The structure for new Crawler is prepared in this manner:-
3. Get the word count for the page for which the keyword was supplied.
4. Save the record in the database.
5. Search for the phrase in the application by specifying in the search box.
6. It will now search for the whole phrase first.
7. Arrangement shall be done of the URLs in the order of the word count i.e. descending form.
8. Now first remove all those words that form part of the sentence.
9. Then retrieve all those URLs that contain those parts of the phrase that are not there in the list above mentioned.
10. The results of this word search shall again be displayed in the descending form of the word count which was attended at the training time.saved in the database and loaded into URL queue when the program starts.The structured information is acquired by matching web content with keywords. New URLs, such as, the next page URL of the list,or the URL of each post, can also be extracted from the content.The crawler displays pages based on the word count retrieved from the page URLs. The display is updated incase of recrawl.The structure for new Crawler is prepared in this manner.

III. IMPLEMENTATION DETAILS OF DECISION TREE APPROACH

The Search bot is a middleware between an existing search engine and the user. The search bot fetches the search results from a search engine, analyzes and filters the results, and displays the relevant results to the user. So an existing search engine like Google, Yahoo etc. can be used to fetch the regular search results. Application Programming Interface (API) can be used to fetch the search results into the search bot.The search bot can be developed in a .net framework. The tool used to test the results of the proposed technique was developed in Microsoft Visual Studio 2010. A database or a data store is used to handle the storage of data for the search bot. The user profiles, training data, knowledge representation and inference structures are stored here. SQL server 2008 was used in the test application for this purpose. Active Data Objects (ADO.net) was used for integrating the application with the database.

IV. IMPLEMENTATION DETAILS OF WORD COUNT APPROACH

word count approach The crawler shall be training the system for the results to be shown. The whole process can be described in following steps:

1. The user has to train the system for the search results.
2. Supply the URL's along with the keyword to search for in the page.
3. Get the word count for the page for which the keyword was supplied.
4. Save the record in the database.
5. Search for the phrase in the application by specifying in the search box.
6. It will now search for the whole phrase first.
7. Arrangement shall be done of the URLs in the order of the word count i.e. descending form.
8. Now first remove all those words that form part of the sentence.
9. Then retrieve all those URLs that contain those parts of the phrase that are not there in the list above mentioned.
10. The results of this word search shall again be displayed in the descending form of the word count which was attended at the training time.

V RESULTS & DISCUSSION

These screen shots show result of word count approach.

Figure1: System Training Screen

Figure 2: Shows one of the search result screen

Figure 3: Shows the search result of "a day"

VI. CONCLUSIONS

This paper looks into two techniques which can be used to sort the search results according to user's search goals. Explicit feedback, in form of User Training is given to the system. The system fetches the search results from an existing searching system, which gives search results on the bases of keywords in the search query. These techniques then filters these search results according to the user's requirements for the search query. Thus the search results that are finally displayed to the user are first filtered according to the keywords in the search query, and then according to the user's requirements from the search query. Both techniques delivers high accuracy in filtering and can improve its accuracy while usage. If the user is reluctant to explicitly train the system, implicit feedback will filter

the results to some extent.

VII. ACKNOWLEDGEMENT

I dedicate this work to my parents for their support and encouragement through all these

years. This work would not have been possible without their love, support and understanding.

I sincerely express indebtedness to esteemed and revered guide Prof. Sujeet kumar Tiwari,

Professor and Head, Department of Computer Science & Engineering, LNCT JABALPUR (M.P.) for his invaluable guidance, supervision and encouragement throughout the work.

VIII. REFERENCES

[1] Stuart Russell, Peter Norvig, 1995. Artificial Intelligence: A Modern

Approach New Jersey: Prantice Hall.

[2] Tom Mitchell, Machine Learning, McGraw Hill, 1997.

[3] J.R. Quinlan (1986): "Induction of Decision Tree" Machine Learning,

Vol. 1, pp.81-106.

[4] Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen, Xiaogang Wang

"IntentSearch: Capturing User Intention for One-Click Internet Image

Search" in IEEE Transactions On Pattern Analysis And Machine

Intelligence, Journal Of Latex Class Files, Vol.6, No.1, January 2010.

[5] Roman Y.Shttkh and Qun Jin "Enhancing IR with User-Centric

Integrated Approach of Interest Change Driven layered Profiling And

User Contributions" in 21st IEEE International Conference of

Advanced Information Networking And Applications Workshops

(AINAW'07).

*[6] Xujuan Zhou, Sheng-Tang Wu, Yuefeng Li, Yue Xu, *Raymond*

Y.K. Lau, Peter D. Bruza, "Utilizing Search Intent in Topic

Ontology-based User Profile for Web Mining", in proceedings of the

2006 IEEE/WIC/ACM International Conference on Web Intelligence

(WI 2006 Main Conference Proceedings)(WI'06)

[7] Roman Y. Shtykh and Qun Jin, "Enhancing IR with User-Centric

Integrated Approach of Interest Change Driven Layered Profiling and

User Contributions", in 21st International Conference on Advanced

Information Networking and Applications Workshops (AINAW'07)

[8] Takehiro Yamamoto1, Satoshi Nakamura1, and Katsumi Tanaka, "An

Editable Browser for Reranking Web Search Results", in IEEE

International Workshop on Databases for Next Generation
Researchers, 2007. SWOD 2007.
[9] Zhengyu ZHU, Jingqiu XU, Xiang REN, Yunyan TIAN, Lipei LI,
"Query Expansion Based on a Personalized Web Search Model",
Third International Conference on Semantics, Knowledge and Grid,
IEEE 2007.
[10] K.S. Kuppusamy, G. Aghila, ", FEAST - A Multistep, Feedback
Centric, Freshness Oriented Search Engine", 2009 IEEE International
Advance Computing Conference (IACC 2009) in Patiala, India, 6-7
March 2009
[11] Kinam Park, Taemin Lee, Soonyoung Jung, Heuiseok Lim, Sangyep
Nam, "Extracting Search Intentions from Web Search Logs", in 2nd
International Conference on Information Technology Convergence
and Services (ITCS), 2010.
[12] Bharat K. "SearchPad: Explicit capture of search context to support
web search" in Proceedings of the 9th International World Wide Web
Conference, pp. 493-501, 2000.
[13]]. E. Agichtein, E. Brill, S. Dumais, and R. Ragno,"Learning User
Interaction Models for Predicting Web Search Result Preferences," in
Proceedings of the ACM Conference on Research and Development
on Information Retrieval (SIGIR), 2006, pp. 3-10.
[14] Xiaomin Ning, , Hai Jin, Hao Wu, "SemreX: Towards Large-Scale
Literature Information Retrieval and Browsing with Semantic
Association", IEEE International Conference on e-Business
Engineering (ICEBE 2006).
[15] Mehdi Adda, Rokia Missaoui, Petko Valtchev, "Toward Feedback-
Based Web Search Engine", 2009 International Conference on
Advanced Information Networking and Applications Workshops.
[16] K.S. Kuppusamy, G. Aghila, "FEAST - A Multistep, Feedback
Centric, Freshness Oriented Search Engine", 2009 IEEE International
Advance Computing Conference (IACC 2009).
[17] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui
Zheng, "A New Algorithm for Inferring User Search Goals with
Feedback Sessions", IEEE Transactions On Knowledge And Data
Engineering, Journal Of Latex Class Files, Vol. 1, No. 8, August
2002
[18] W3C Resource Description Framework (RDF). http://www.w3c.org/RDF/
[19] The Semantic Web Community, http://www.semanticweb.org/.
[20] W3C Web Ontology Language (OWL), http://www.w3.org/2004/OWL/.
[21] Microsoft AdCenter Labs, http://adlab.microsoft.com/
[22] Google Scholar. http://scholar.google.com/
[23] http://searchsoa.techtarget.com/definition/bot
[24]http://en.wikipedia.org/wiki/ID3_algorithm
[25] http://www.123helpme.com/the-id-algorithm-preview.asp?id=154335

IX. AUTHOR BIBLOGRAPHY

Prof. Neelu Tiwari has received B.E. (Bachelor of Engineering) degree in Computer Science and Engineering from RGPV University "Bhopal" (M.P.), India in 2010. She is received her M.Tech. (Master of Technology) in Computer Technology and application from RGPV University, BHOPAL (M.P.), India in 2014 with honours. Her subjects of interest include Compute Networking ,Object orientation and technology,Data base management system and Analysis & Design of Algorithms.She has published 5 research papers in international journal and 2 paperspresented in national conferrence. Email: neel_31@rediffmail.com

Prof. Sujeet Kumar Tiwari has received B.E. (Bachelor of Engineering) degree in Information Technology from RGPV University "Bhopal" (M.P.), India in 2006.He has 5 year Industrial expireance as IT head in Share microfin Limited Hyderabad .He is received his M.Tech. (Master of Technology) in Computer Technology and application from RGPV University, BHOPAL (M.P.), India in 2013 .Presently he is Professor and Head, Department of Computer Science & Engineering, LNCT JABALPUR (M.P.) His subjects of interest include computer networking,Database management system, Distributed databases,Data Structure Email: sujeet.tiwari08@gmail.com