ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Profile Based Personalized Web Search Using LDA

Tinimol Andrews, Nimmy Manuel

M. Tech Student, Dept of CSE, Mangalam College of Engineering, Kottayam, Kerala, India

Assistant Professor, Dept. of CSE., Mangalam College of Engineering, Kottayam, Kerala, India

ABSTRACT: Personalized web search(PWS) gives better search quality while preserving privacy. However, user's private information can be exposed during PWS is an important issue. Bridging topic modelling and PWS enhance search quality and assure better privacy. Topic Modelling, such as Latent Dirichlet Allocation (LDA) was proposed to generate models to represent multiple topics in a collection of documents. Patterns are used to describe documents than a single term. User profile generation algorithm are used for better search. Client side privacy protection achieved by using UPS framework. Less number of frequent patterns are sent to server to achieve better result. User profile is maintained on client side itself. Experimental results shows improved performance by bridging PWS and topic modelling using LDA.

KEYWORDS: Personalized search, LDA, Topic modelling.

I. INTRODUCTION

Web search engine become a popular tool for ordinary people looking useful information on web. The information on the web is increasing day by day. When different users submit same query, the search engine returns same result without knowing user interest. Personalized web search is a searching technique provide better search result by considering user needs. The goal of PWS is to provide search result based on user interest. The effective personalized search collect and aggregate user information, which often raise serious privacy concerns.
. Information filtering (IF) removes unwanted information from entire document. Pattern based approach utilize pattern to represent user's interest, since patterns carry more semantic interest than terms. Topic modelling has most popular text modelling technique. It automatically classify documents in a collection by number of topics. A representative approach is LDA, represent topics by patterns rather than single words through combining topic modelling with pattern mining technique.

II. RELATED WORK

A probabilistic topic model [1] aims to extract topics from user search history. the topics relevant to query update the query model which distinguish relevant and irrelevant parts and filter out noise from user search history. The main contribution is to modelling user search history with topics. Topic extraction and relevance feedback are the two main steps. An unsupervised way to find user preferences and explore a relationship between different query units via topic.

The existing retrieval models retrieve information based on query and actual documents; information about user and search content ignored. In context sensitive information retrieval using implicit feedback [2] aims to extract implicit feedback information, include previous queries and click through data , to improve retrieval accuracy. use KL-divergence as the basis, four statistical models such as FixInt, BayesInt, OnlineUp and BatchUp. Experimental results shows implicit feedback improve retrieval performance without having any user effort.

Personalized search improves the search quality while preserving privacy. User's compromised privacy requirement if there is a faster service from server. Thus a balance between privacy and search. Privacy enhancing personalized web search [3] build user profile, which contain user interest into hierarchical structure. An additional privacy measure expoRatio is proposed to estimate the amount of privacy.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Automated identification of user interest for personalized search [4] propose a framework to solve the problem of personalizing web search based search history without user effort. A user model is proposed to formalize user's interest on web page and correlate them with user's click-through data. An intuitive algorithm learn user interest based on this correlation.

Mining long-term browsing history to improve searching accuracy [5] explore the way of how long-term search history improve the search performance. Statistical language modelling based methods used to mine information from long term search history. Long term search history contain all search activities in the past. In an information retrieval system, return a set of result document with titles and summaries. Then the user select full text of some results.

Personalized search based on user search history [6] build user profiles based on the activity at search site and study the profile to provide a personalized search result. In particular collected the queries for at least one search result and the snippets for each examined results. User profiles are created by classify queries or snippets into concepts in a reference concept hierarchy. These profiles used to re-rank the search results.

Large scale evaluation framework [7] for personalized search based on query logs and evaluate two click based and three profile based ones using 12 days MSN query log. personalized search has significant improvement over normal search on some queries. click based strategy perform considerably well. Queries with small click entropy will return same result as that of normal search.

III. PROPOSED METHOD

Nowadays, for every user query, user enters the query using search engine and displayed an output may or may not accurate. Personalized web search is a searching technique aiming to improve the search quality. Profile based searching technique improve the search results with complex user models generated from user profile. The target is to provide a better search experience with tightly bounded security constraints. The main goal is to provide protection to the user's personal data using UPS framework.
Topic modeling is a text mining technique which can apply in personalized web search for faster query results. Topic modeling such as, Latent Dirichlet Allocation(LDA), generate models to represent multiple topics in a collection of documents. Dataset is loaded from directory and perform document processing. LDA divide the documents into number of topics and TDT(Topic Detection and Tracking) is done based on generated LDA. Equivalent patterns can be generated from frequent patterns for better query search.
Personalized web search using topic modelling create a user profile which contain equivalent patterns. Equivalent pattern means when one of the word in pattern comes, remaining words will always appear as a whole.

USER CUSTOMIZABLE PRIVACY PRESERVING SEARCH

UPS consist a non trusty server and number of clients. Each client can access the server. Privacy protection can be achieved by using *online profiler* implemented on search proxy running on client machine. The proxy contain complete user profile, user specified privacy requirements and set of sensitive nodes which is set by user.

UPS provide runtime profiling, customization of privacy requirements and doesn't require iterative user interaction. It generalize profiles for query according to user specified privacy requirements.

LATENT DIRICHLET ALLOCATION (LDA)

LDA is the statistical topic modelling tool currently in use. The main idea of LDA is that the document contain multiple topics and each topic can be a distribution over vocabulary of words that appear in the document. Topic modelling algorithms discover hidden topics from collection of documents, where topic is a distribution over words.

The main contribution of LDA algorithm is that topic representation using word distribution and document can be represent using topics. Topic representation means which words are important to which topic and document representation indicate which document are important for that particular document.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

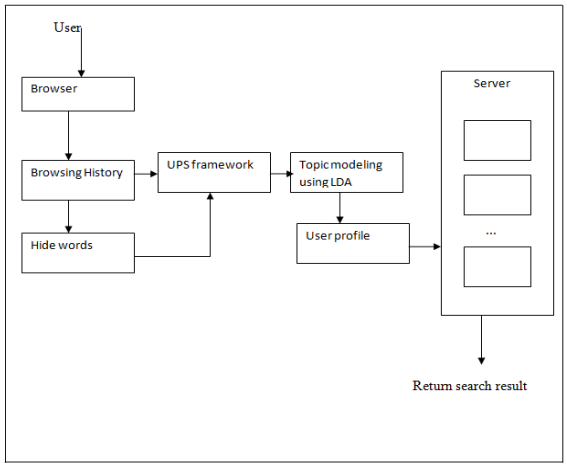
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

IV. ARCHITECTURE

The majority of query submitted by user are short and ambiguous. Query is given to the browser (Mozilla Firefox) and search using topic modelling. Topic modelling contain a sequence of steps such as Load dataset, query processing, LDA, TDT creator, Frequent pattern mining and Equivalent pattern mining. Based on the equivalent pattern search result obtained.

Server search result based on equivalent pattern, that have high support value. The resultant URL are less compared to natural search.



V. EXPERIMENTAL RESULTS

The UPS Framework is build on a PC with a Pentium Dual-Core 2.50-GHZ CPU and 8-GB main memory, running Microsoft Windows 8.

The topic repository uses the ODP web Directory. The click logs are downloaded from the online AOL query log, which is the most recently published data. The AOL query data contain over 20 million queries and 30 million clicks of 650k users over 3 month(march 1,2006 to may 31,2006). The data format of each record is,

< uid, query, time[, rank, url] > [8]

where first three fields indicate user uid issued at timestamp time, and last two fields are optional.

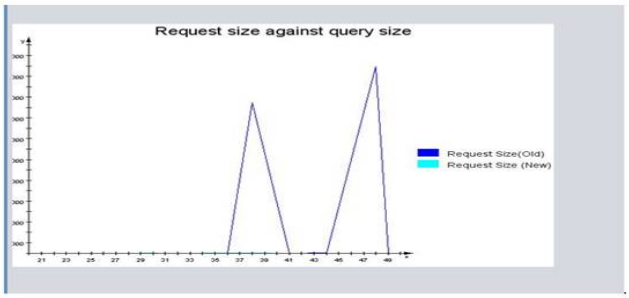


Fig: Request size against query size

Request size(old) means the request processed by the normal search while Request size(new) indicates the query processed by using personalized web search using topic modeling and LDA. The performance of new query processing is faster than Request size (old).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

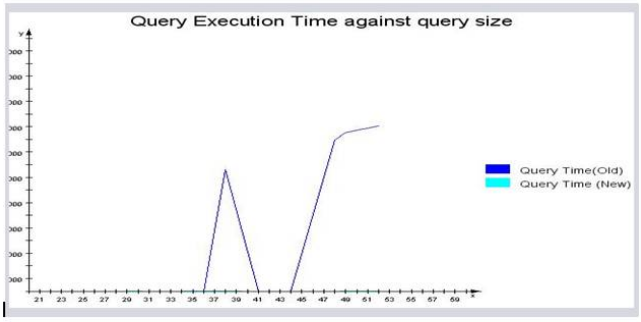


Fig: Performance execution

When query execution time is directly proportional to query size. For personalized search using Greedy algorithm, query size and execution time is too large but PWS using topic modelling and LDA gives less query execution time. PWS using LDA process only equivalent class patterns rather than entire search history.

VI. CONCLUSION AND FUTURE WORK

Profile based personalized web search with LDA improves the search quality and assure greater privacy constraints. Topic modelling such as LDA, load datasets from search history and then process the data and documents are classify into topics. Search result returned when the equivalent pattern are sent to server rather than entire URL visited. Customized privacy requirements can be specified by sensitive nodes.UPS procedure implemented for client side privacy protection. Profile based PWS with LDA improves performance against requested query and performance time against query size.
Further enhancement can be apply for image type personalised web search.

REFERENCES

1. WeiSong YuZhang TingLiu ShengLi "BridgingTopicModelingandPersonalizedSearch" School of Computer Science Harbin Institute of Technology
2. Xuehua Shen Bin Tan ChengXiang Zhai "Context-Sensitive Information Retrieval Using Implicit Feedback" Department of Computer Science University of Illinois at Urbana-Champaign
3. Yabo Xu* Benyu Zhang, Zheng Chen Ke Wang Simon " Privacy-Enhancing Personalized Web Search" Fraser University 8888 University Drive, Burnaby BC, Canada wangk@cs.sfu.ca .
4. Feng Qiu Junghoo Cho " Automatic Identification of User Interest For Personalized Search" University of California Los Angeles, CA 90095 cho@cs.ucla.edu
5. Bin Tan, Xuehua Shen, ChengXiang Zhai" Mining Long-Term Search History to Improve Search Accuracy" Department of Computer Science University of Illinois at Urbana-Champaign
6. Micro Speretta Susan Gauch "Personalized search based on user's search history" Electrical Engineering and Computer Science University of Cansas Lawrence
7. Zhicheng Dou Ruihua Song Ji-Rong Wen " A Large-scale Evaluation and Analysis of Personalized Search Strategies " Microsoft Research Asia Beijing 100080, China jrwen@microsoft.com
8. Lidan Shou, He Bai, Ke Chen, and Gang Chen" Supporting Privacy Protection in Personalized Web Search" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:26 NO:2 YEAR 2014

BIOGRAPHY

Tinimol Andrews doing MTech CSE in mangalam college of engineering. Receives bachelor of technology degree from mangalam college of engineering in 2013 . Area of interest is Datamining security.

Nimmy Manuel works as an assistant professor in mangalam college of engineering. Receives bachelor degree from Mahatma Gandhi University Kottayam in 2006 and master degree from M.S University Thirunelveli in 2012.She has 7 years teaching experience in Computer Science. Area of interest is datamining.