Enhanced Personalized Web Search using Pattern-based Topic Modelling

Ramitha A T, Jayasudha J S

***Abstract:** Personalized Web Search is a method of searching to improve the quality and accuracy of web search. It has gained much attention recently. The main goal of personalized web search is to customize search results that are more relevant and tailored to the user interests. Effective personalization needs collecting and aggregating user information that can be private or general. Personalized search results can be improved by information filtering. Information Filtering is a system to remove irrelevant or unwanted information from an information stream based on document representations which represent users' interest. Traditional information filtering models assume that one user is only interested in a single topic. In statistical topic modelling documents and collections can be represented by word distributions. But directly applying topic models for information filtering is insufficient to distinctively represent documents with different semantic content. In order to alleviate these problems, patterns are used to represent topics for information filtering. Pattern-based representations are considered more meaningful and more accurate to represent topics than word-based representations. Pattern-based Topic Model (PBTM) combines pattern mining with statistical topic modelling to generate more discriminative and semantic rich topic representations. In the proposed system, user information preferences are acquired as a collection of documents from user browsing history. Latent Dirichlet Allocation is used to perform topic modelling on the collected documents. Word-topic assignments from LDA are used for constructing transactional dataset. Frequent patterns are discovered from topic models. Maximum matched Pattern-based Topic Model is used to build user interest model representing the user preference information from the collection of documents and filter the incoming documents based on the user preferences by document relevance ranking.*

***Keywords:** Topic model, Information filtering, Pattern based mining, User interest model*

I. INTRODUCTION

Information Filtering (IF) is used to remove irrelevant or unwanted information from an information stream based on documents which represent user's interest. Traditional Information Filtering models assume that one user is only interested in a single topic and so concentrate on term-based approach. But term-based representation suffers from polysemy and synonymy. To overcome the limitations of term-based approach, pattern mining based techniques can be used. Pattern mining based techniques make use of patterns to represent user preferences and interests [1].Patterns are semantically meaningful than terms. Most of the text mining techniques assume that the user's interest is only related to a single topic. The user interests can change over time and new topics may be added to the document stream.

In the proposed work, user interests are modelled in multiple topics rather than a single topic. Topic modelling has been widely accepted in the areas of machine learning and text mining, etc. It was proposed to generate statistical models to classify multiple topics in a collection of documents, and each topic is represented by a distribution of words. Topic modelling has become a powerful tool for unsupervised analysis of large collection of documents. Topic modelling [2] can automatically classify documents in a collection by a number of topics. Probabilistic Latent Semantic Analysis (PLSA)[3] and Latent Dirichlet Allocation(LDA)[4] are two commonly used topic modelling methods. However, there are some problems in topic modelling when directly applied. The first problem is that, as number of topics is predefined topic distribution is not sufficient for document representation. The second problem is that word based topic representation may not distinctively represent documents having different semantic content. To overcome these problems, patterns are used to represent topics for information retrieval.LDA is considered as a traditional and effective topic modelling method that generates patterns from words. As patterns are group of words extracted from LDA based on occurrence of the words in the documents, patterns represent the topics well [5].

II. RELATED WORKS

User information needs are obtained by IF systems from user profiles. The main objective of IF systems is to map a set of incoming documents to a set of user relevant documents. Traditional IF systems make use of term based models for user profiling. The term based models like td*idf, Okapi BM25 and weighting schemes for the bag of words representation [6][7][8] are popular in information filtering. However, term based models have the limitations of polysemy and synonymy. IF systems that extract more semantic features like phrases and patterns were used for document representation. The n-gram was later used by data mining techniques for text mining and classification. n-gram is a continuous sequence of words collected from a collection of documents [9][10].But n-gram is not commonly used due to the low frequency of phrases. To overcome the drawbacks of term-based methods, pattern mining based methods have been used in IF systems to remove repeated or unwanted patterns. However, pattern mining methods return large patterns because sub-patterns of frequent patterns are always frequent.

Manuscript published on 30 August 2016.
* Correspondence Author (s)

Ramitha A T, Department of Computer Science & Engineering, Sree Chitra Thirunal College of Engineering, Thiruvananthapuram (Kerala)- 695018, India.

Dr. Jayasudha J. S, Department of Computer Science & Engineering, Sree Chitra Thirunal College of Engineering, Thiruvananthapuram (Kerala)- 695018, India.

Selection of reliable patterns is very important [11].

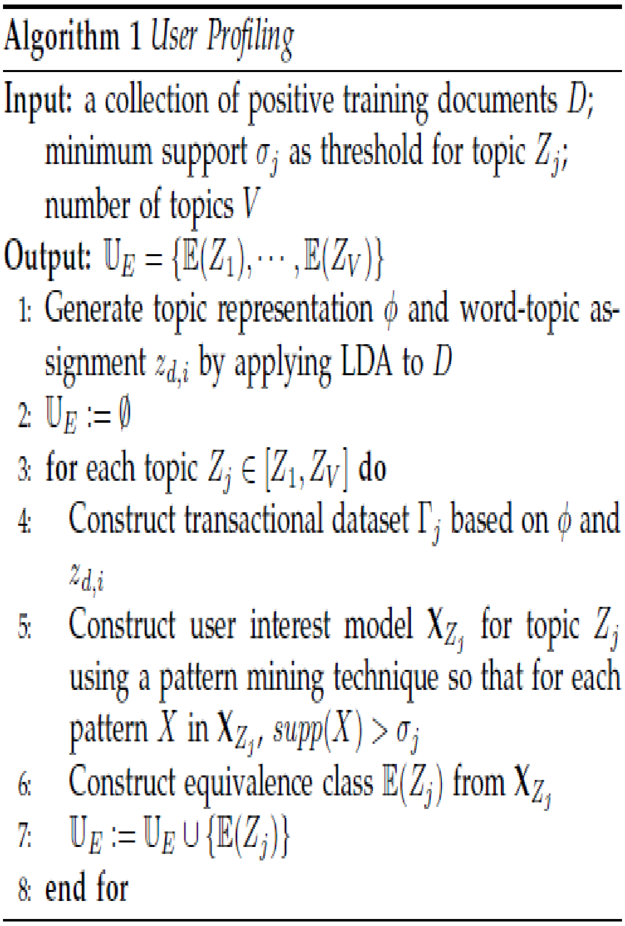
For mining frequent patterns efficiently, algorithms like Apriori, Prefix Span and FP tree have been proposed. Condensed representations of frequent item sets like closed itemsets [12], maximal itemsets [13] etc. have been proposed to enhance efficiency of frequent item sets without information loss. Frequent closed patterns efficiently represented user profile and documents. This efficiency of frequent closed pattern is because of the predefined support threshold, all closed patterns contain relevant information about all frequent patterns. The synonymy problem of term based document models was solved by the multiple topic models by representing each topic in the topic model by a group of semantically similar words. Language models achieved efficient search results when incorporated with topic models. This increased the relevance of retrieved documents. Long-term user preferences can be extracted by analysing content and representing in terms of latent topics found from user profiles using probabilistic topic modelling [14].

L.Shou et.al. Proposed a framework in which a proxy maintains a complete hierarchical user profile with user specified privacy requirements [15]. A publicly available taxonomy repository Word Net is used and the user profile is built as a rooted subtree of the repository. The user can customize privacy requirements in his profile. When the user submits a query for searching a generalized profile is send with the query to the server for search. The user profile is updated with the search results obtained. The user profile may become too large as user interest areas widen. The generalized user profile is send to the server with the search query, if user has some privacy requirements or the actual user profile is send to the server. So the time and space complexities increase exponentially with the profile size which is the limitation of this framework. In the proposed work, the size of the user profile send to the server is reduced by performing pattern based topic modelling.

III. METHODOLOGY

Latent Dirichlet Allocation (LDA) is implemented using Gibbs sampling for parameter estimation and inference.LDA was introduced by David Blei et al[4]. JGibbLDA is used as the java implementation of LDA using Gibbs sampling. It is found useful in application areas like Information Retrieval (IR) where LDA focus on inferring latent topic structures from a collection of documents. LDA considers each document to contain multiple topics and each topic as a distribution of words that appear in the documents. Topic representation using word distribution and document distribution using topic distribution are the contributions of LDA. From the collection of documents prepared from the user browsing history, LDA learn topics and break up the documents according to the topics. But LDA model produce single word based topic representations containing ambiguous semantics. So the results from LDA are used transactional datasets and then frequent patterns are generated for each transactional dataset representing topics. For a given minimum support, an itemset X is frequent if the support of X is greater than or equal to the minimum support. Then equivalence classes are constructed by collecting the frequent patterns with the same frequency into one group. In the proposed work, equivalence classes are used to represent topics rather than frequent patterns. When the user submits the search query, equivalence classes for each word in the query are constructed and the cross product of equivalence classes is computed and send to the server. The benefit of this topic based search is that, there is no need to send the hierarchical user profile as a whole to the server and so the request size and the execution time for the personalized search can be considerably reduced as shown in section 4.

The proposed model for Information Filtering can be described in two steps: User Profiling and Document Filtering. In User Profiling, user interest model is generated using algorithm1 as given below.



Algorithm 1 *User Profiling*

```
Input: a collection of positive training documents $D$;
    minimum support $\sigma_j$ as threshold for topic $Z_j$;
    number of topics $V$
Output: $\mathbb{U}_E = \{\mathbb{E}(Z_1), \cdots, \mathbb{E}(Z_V)\}$
Generate topic representation $\phi$ and word-topic assignment $z_{d,i}$ by applying LDA to $D$
$\mathbb{U}_E := \emptyset$
for each topic $Z_j \in [Z_1, Z_V]$ do
    Construct transactional dataset $\Gamma_j$ based on $\phi$ and $z_{d,i}$
    Construct user interest model $\mathbb{X}_{Z_j}$ for topic $Z_j$ using a pattern mining technique so that for each pattern $X$ in $\mathbb{X}_{Z_j}$, $supp(X) > \sigma_j$
    Construct equivalence class $\mathbb{E}(Z_j)$ from $\mathbb{X}_{Z_j}$
    $\mathbb{U}_E := \mathbb{U}_E \cup \{\mathbb{E}(Z_j)\}$
end for
```

The input is a collection of documents, a minimum support is used as threshold for a given topic and number of topics as given by the user and generates pattern based topic representations as output to represent the user preferences. In the second step, relevant documents are filtered from incoming documents based on the relevance of the documents to the user's needs. In the proposed work, for document relevance ranking Google Application Programming Interface is used.

The flow diagram of the proposed work is shown in Figure 3.1.The proposed system works by collecting documents from the user search history.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The collected documents representing the user preferences are undergone topic based modelling and result kept as user interest model. When the user issues search query, topic based search is done by constructing equivalence classes for each word in the query and the cross product of equivalence classes is computed and send to the server. Then relevant search results are retrieved and send to the user.

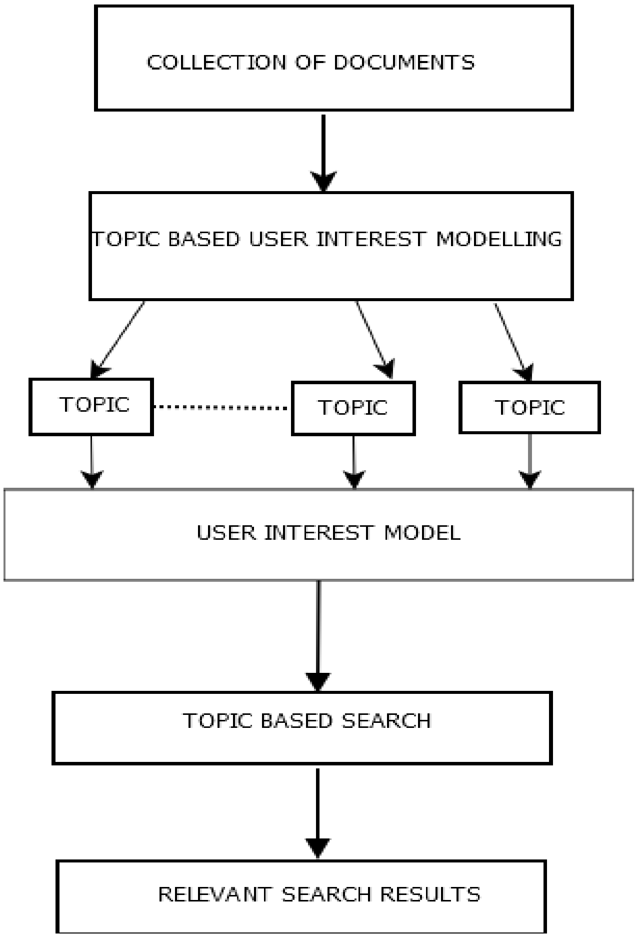


Fig.3.1. Flow diagram

IV. RESULTS AND DISCUSSION

The proposed system was implemented using Java Netbeans, SQL yog and Xampp server. The dataset used for experiment was collected from the browsing history of the user. The two personalized search methods compared were Word Net based search and topic model based search. As discussed in the methodology, first the hierarchical user profile was constructed and user search behaviour was analysed. Then the user was provided with Word Net based search and the analysis of search results were performed on the basis of relevance, request size and query time. The relevance was analysed by plotting frequency against profile size. The profile size is the number of distinct URLs in the user search history and frequency is the ratio of the sum of occurrences of each word in the visited page to number of words is the user profile. The request size is the number of words sent to the search server. The analysis of request size against profile size and query time against profile size were also performed. To improve the performance of searching, pattern based topic modelling was done and he user is provided with topic model based search and the search results were analysed. The search results of topic model based search showed more relevance, reduced request size and query time. The Figure 4.1 shows the plot of request size against profile size. The Figure 4.2 shows the plot of relevance against profile size.

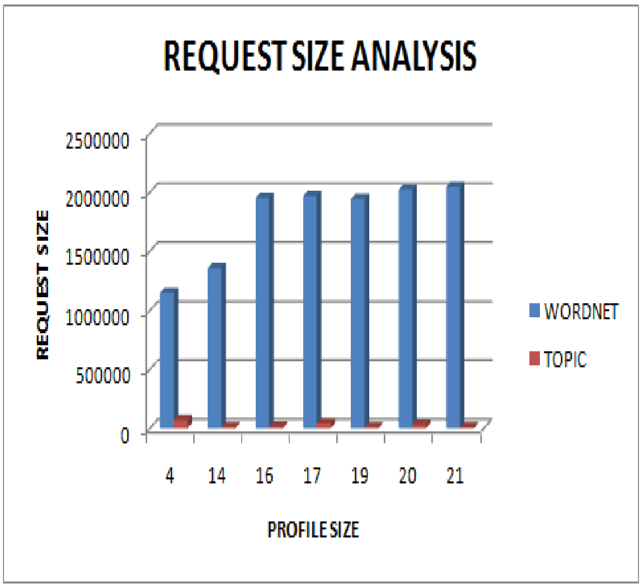


Fig. 4.1. Performance analysis1

The performance analysis1 (Fig.4.1) shows topic based modelling provides efficient search results compared with Word Net based search. The size of request send to the server in topic based search is reduced while compared with the request size of Word Net. This implies that the execution time of search will be reduced in topic based search.

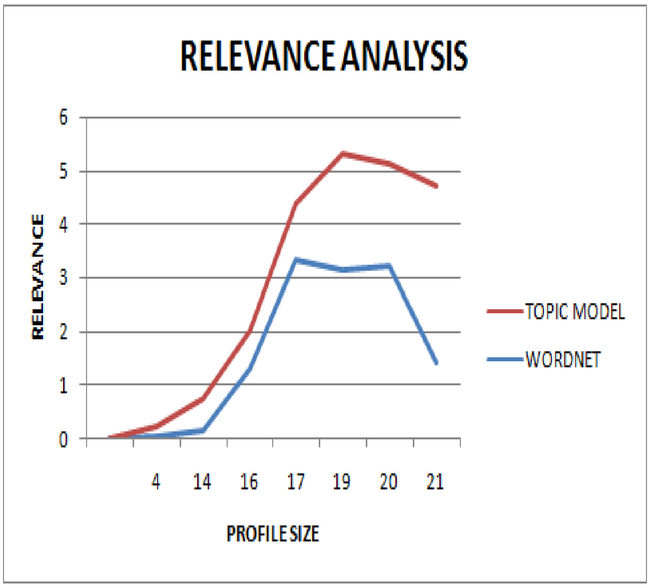


Fig.4.2 Performance analysis2

The performance analysis2 (Fig.4.2) shows that topic based search provides more relevant search results to the user compared to Word Net based search.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

V. CONCLUSION

The user profile was constructed as a hierarchical structure based on the user preferences collected from user search history. After getting keywords from each URL visited, word relations like hypernyms, hyponyms and synonyms were drawn from WordNet repository for each keyword. Related words are kept in a database and used for constructing user profile tree. The profile tree was built up to three levels in the proposed system. A pattern enhanced topic model is proposed for information filtering for improving personalized search efficiency. The proposed work models user preferences across multiple topics by generating pattern enhanced topic representations. Incoming documents are filtered based on relevance with user preferences. In this work, performance analysis based on request size send to the server and relevance of search results with user needs are done. It is concluded that the size of request send to server was considerably reduced in topic based search and the relevance of search result is also improved in topic based search.

REFERENCES

1. H. Cheng, X. Yan, J. Han, and C.-W. Hsu, “Discriminative frequent pattern analysis for effective classification,” in IEEE 23rd International Conference on Data Engineering, ICDE’2007. IEEE, 2007, pp.716–725
2. X. Wei and W. B. Croft, “LDA-based document models for ad-hoc retrieval,” in Proceedings of the 29th annual International ACM SIGIR conference on Research and Development in Information Retrieval. ACM, 2006, pp. 178–185.
3. T. Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999, pp.50–57
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.
5. Y. Gao, Y. Xu, Y. Li, and B. Liu, “A two-stage approach for generating topic models,” in Advances in Knowledge Discovery and Data Mining, PADKDD’13. Springer, 2013, pp. 221–232.
6. S. Robertson, H. Zaragoza, and M. Taylor, “Simple BM25 extension to multiple weighted fields,” in Proceedings of the thirteenth ACM International Conference on Information and Knowledge Management. ACM, 2004, pp. 42–49
7. Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, “Adapting ranking svm to document retrieval,” in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2006, pp. 186–193
8. X. Li and B. Liu, “Learning to classify texts using positive and unlabeled data,” in IJCAI, vol. 3, 2003, pp. 587–592.
9. J. Furnkranz, “A study using n-gram features for text categorization,”Austrian Research Institute for Artificial Intelligence, vol. 3, no.1998, pp. 1–10, 1998.
10. W. B. Cavnar, J. M. Trenkle et al., “N-gram-based text categorization,”Ann Arbor MI,vol.48113, no. 2, pp. 161–175, 1994.
11. Y. Xu, Y. Li, and G. Shaw, “Reliable representations for association rules,” Data & Knowledge Engineering, vol. 70, no. 6, pp. 555–575,2011.
12. T. Hofmann, “Probabilistic latent semantic indexing,” in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999, pp. 50–57.
13. Y. Gao, Y. Xu, Y. Li, and B. Liu, “A two-stage approach for generating topic models,” in Advances in Knowledge Discovery and Data Mining, PADKDD’13. Springer, 2013, pp. 221–232.
14. C. Wang and D. M. Blei, “Collaborative topic modeling for recommending scientific articles,” in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011, pp. 448–456.
15. L. Shou,H. Bai,K. Chen and G. Chen, "Supporting Privacy Protection in Personalized Web Search," IEEE Transaction on Knowledge and Data Engineering,Vol:26,No:2, 2014.