Semantic Retrieval of Web Documents using Topic Modeling Based Weighted Nearest Neighborhood Technique

R.Priyadarshini, latha Tamilselvan, N.Rajendran

Abstract— Information retrieval systems are used to retrieve documents based on the keyword search. Semantic-based information retrieval is beyond standard information retrieval and uses related information to get the documents from the corpus. But semantic retrieval based documents is not efficient enough in real time. Content from the user's profile is used for searching the web documents. The documents which exactly matches the user requirement is retrieved and it improvises the personalized retrieval. In this paper, a methodology based on topic modelling is proposed to determine the retrieval of information for user to increase the accuracy of documents using Latent Dirichlet Allocation (LDA) and Weighted Nearest Neighbor (WNN) models. LDA model is developed to retrieve documents based on topics. The topic based retrieval is improvised using personalization technique which uses WNN model. Experimental analysis on building personalization and semantic retrieval of documents shows the improved precision compared to existing topic modeling.

Index terms— Personalized Information Retrieval, Topic Modeling, User Profile, Personalization, Semantics, WNN.

I. INTRODUCTION

The retrieval of content from any repository is the activity of obtaining information resources related to the users requirement. The retrieval of information is based on content in web document, metadata and content in the repository. Indexing is one of the technique used for faster search in the repository. Query parser and query processing plays a vital role in searching the content from the repository. Queries are formal state statements of information needs, for example, search strings in web search engines. With the growth of technologies and services, large data is produced that can be structured, semi-structured and unstructured from different resources. There are many techniques in Information Retrieval System to retrieve information like Keyword based on Information Retrieval, Semantic-based Information Retrieval and Personalized based Information Retrieval. The keyword search is one of the conventional information retrieval systems. In the traditional Information Retrieval System, keywords play a major role in retrieving the desired documents. The relevant documents are retrieved by comparing the query with the content from the document in the repository. The documents are ranked and prioritized by popular methodologies used by search engines. The relevance values ae calculated by semantic analysis in the retrieval process.

Revised Manuscript Received on July 05, 2019

R.Priyadarshini, Assistant Professor (Sr.Gr.), Department of Information Technology, B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

latha Tamilselvan, Department of Information Technology, B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

N.Rajendran, Assistant Professor (Sr.Gr.), Department of Information Technology, B.S.Abdur Rahman Crescent Institute of Science and Technology, Chennai, India

The keyword search is a simple process for query analysis and retrieval display. But the query analysis and keyword search is not enough for accurate search and ranking. The conventional search method based on keywords are not accurate and quality of documents are not exact results. Hence there is a need for a semantical search and retrieval of documents based on semantic analysis. The sematic retrieval is used to improve the quality of the documents using context based semantic analysis. The semantic analysis is used to generate the relevant documents. The web documents consist of Resource Description Framework (RDF) and Extended Markup Language (XML). The semantic analysis and search use the above resources and enables the system to retrieve rich and meaningful knowledge essence. Semantic queries allow for queries and analytics of associative and contextual nature. The most relevant results are retrieved from the databases and other storage repositories by semantic search. It evaluates and understands reasonably using vocabulary and another relationship matching. Semantic search is a data searching technique in which a search query not only process the meaning of the words but also determine the context, purpose, and relationship of the words with the key topic. Semantic search systems consider various points to provide relevant search results including the context of search, location, intent, a variation of words, synonyms, generalize, specific queries, conceptual matching, and natural language. Semantic-based information retrieval is beyond standard information retrieval and uses related information to get the documents from the corpus. But semantic retrieval based documents is not efficient enough in real time. The individual user based retrieval of content varies when compared with the older techniques. The search is performed based on the content of the personal data in the profile document. The content and information in the profile is analyzed and matched with the web document topics. This is further used in query development and prioritizing of retrieved documents. In this last case, the priority of the documents and their matching with the query is done. The matching is based on user likes and their previous searches. In the traditional systems, the matching between user interest and the web documents. The collection of the information from the user profile is performed and it is done in the required integrated development environment. The precision of the model and the retrieval is identified by the relevancy of document that are retrieved and saved. The complex data associated with the bunch of people using the software or the single user is very well analyzed by the algorithm, which further helps in finding same kind of behavior or interests are identified.

The browsing pattern is also The overall goal of the process is a combination of semantic retrieval and personalized retrieval of documents which shows the improved precision in the retrieval. This paper is structured as shown here: in the Section 2 Literature Overview is given and in Section 3, Problem Definition. In Section 4, System Design is given and In Section 5, System Architecture and proposed system. Further, the results are given in Section 6 the Results of the Experiments and Conclude with perspectives in Section 7.

II. LITERATURE SURVEY

Wei, Barnaghi and Bargiela, [10] in thier paper described about topic model for ontologies and terminological probabilistic model, This paper explores relationship between topic in the words of the document and the query, Probability distribution is utilized ub this case. This is based on LSI model, extraction of topic is one of the popular topic for retrieval of information. Automatic text learning from the repository is performed by the popular methodologies as explained in the paper. Topic relationship and model based algorithms are introduced. The browsing and information search and retrieval are used by the libraries. The proposed model are reduces time and effort complexity. The topic hierarchy are not clear and learning is done parallely in this work. YueshenXua, YuyuYinb and JianweiYind, [9] The Natrual Language processesing steps are performed in this work and the models based on topics and latent topic models in documents are also described. Good topics are obtained from repository and TGWs are preprocessed to identify the stop words. In the existing system the stop words are removed, in the proposed work the same is performed automatically which reduces time and optimized and human intervention is also avoided. Automatic cleaning and extraction is not available for unstructured texts.

PankajdeepKaur, Pallavi Sharma and Nikhil Vohra,[2, In this paper Huge amount of real-time data from social media, is analyzed and Ontology is constructed which also results in text analytics. When the data is very huge there is a requirement of algorithms and techniques which turn over the valid information and truth from the data available. In this paper, the system is modeled to handle 1000 requests and text mining based analytics is successfully used for this purpose. The high-quality data is absorbed and data analytics is applied over it. The ontology framework introduced in this work is for a huge amount of text in social media. The system can interact with the users and also interact with the system. The time complexity and space utilization are elaborated in this method. M.Rani, Kumar Dhar and Vyas,[1] The ontology-based topic modeling provide characteristics like vocabulary, machine-readable format, replaceable and reusable content for agent interaction and ordering and structuring of knowledge using semantic web applications. The semi-automatic methodology is utilized in this work and this proves there is the unavailability of the automatic approach. The issue in the automatic ontology engineering from the different repository using machine learning algorithms. In this work, the algorithms such as Latent Semantic Integration, single value decomposition, and Mr. Latent Drichlet Allocation are introduced for understanding the algorithms the statistical relationships among the web documents. Manual intervention is reduced to a greater extent. The topic ontology is built for semantic retrieval and the impact of the approach with corresponding ontology is explained. M. Shams, A.Baraani, [4] In this work the co-word occurrence is dealt with. The authors have introduced Enriched LDA (ELDA): Combination of latent Dirichlet allocation. The context based extraction and aspect grasping is performed with the help of topic modeling. Therefore the models based on the topics generate irregular features. The co-occurrence of words decide the relationship of the components in model and LDA is used to obtain the previous knowledge which helps in generation of aspects. In the work explained here firstly the basic stages are generated based on Dirichlet model and secondly co-occurrence are used to generate aspects and lastly used to retrieve similar aspects of relevant topics. The model is iterated based on the previous results and it enhances the quality. The data sets form English and Persian languages are used for ELDA and the final results are regarding topic coherences and no specific topic based documents are used in the system. The unique algorithm used makes the proposed paper a different model from the traditional ones. There are many drawbacks like lack of sentimental analysis, entity detection and so on. Songjie Gong,[13] The personalized information retrieval model based on user interest, this literature describes the problem of word meaning disambiguation is often decomposed to the determination of a semantic similarity of words. ChahrazedBouhini, [14] The information retrieval model is personalized in this work. User profile data is analyzed and the annotations are taken from the document. Delicious data set is used in their work and user profile data is added advantage for the same data set. The user centric search is used in the PIR model. The result analysis shows that the proposed model gives the user data and annotations with different combinations and weightings. The difficult queries are brilliant to explore the components of the queries based on concepts of the terms. Dixon Prem Daniel R, [15], in the paper movie recommender system is used for latent factor model. The personalization results in referring preferences of smart phone browsing history

III. PROBLEM DEFINITION

There is a continuous growth of documents on the web on a daily basis. The local storage and the databases also show overflowing transactional data which is further transferred as historical data. The constraints and strategies followed to store these documents make the analysis, categorization and retrieval a challenging task. It is also very much challenging to interpret the semantic essence of a whole or part of the document. Keyword search Information Retrieval system is based on the occurrence of words in documents. A keyword search may retrieve too many results, including the one that aren't relevant to the topic searched. Moreover, the update to the entire dataset for every change in the resource is confusing and too much time-consuming. The challenges and restrictions present in the traditional retrieval system is astounded by semantic systems. The meaning of the query is analyzed and query synonym is provided and results are received based on the query expansion. The relevancy of the document is examined by the concepts in the text with the key terms in the document.

The semantic process is carried out by the creation of ontology and the relationship between the classes created via the ontology. Ontology is one of semantic analysis which generates approximate relationship scores. Topic modeling technique can also replace it. Topic hierarchy makes ambiguous in the ontology. And also information overload on the web, it is increasingly difficult for information retrieval to satisfy the individual user needs. Users with different backgrounds and domains will ask for different information for the similar query. The proposed system Semantic Key Topic and User Profile based Information Retrieval System (SKTUP-IRS) resolves the above-mentioned issues.

IV. SYSTEM DESIGN

There are several methodologies in the retrieval systems to extract information from documents but these techniques are responsible for identifying and analyzing annotation in a semantic and keyword based search. As the internet technology has grown to a greater level the web which is in the form of structured, unstructured or semi-structured should be addressed. Therefore, it is difficult to find out and identify the relevant information from conventional search engine. And also for the same query each user needs different information based on Personalization. Most of the existing retrieval systems use keywords search and directory search. There is no fully automated system which encourages the user centric retrieval.

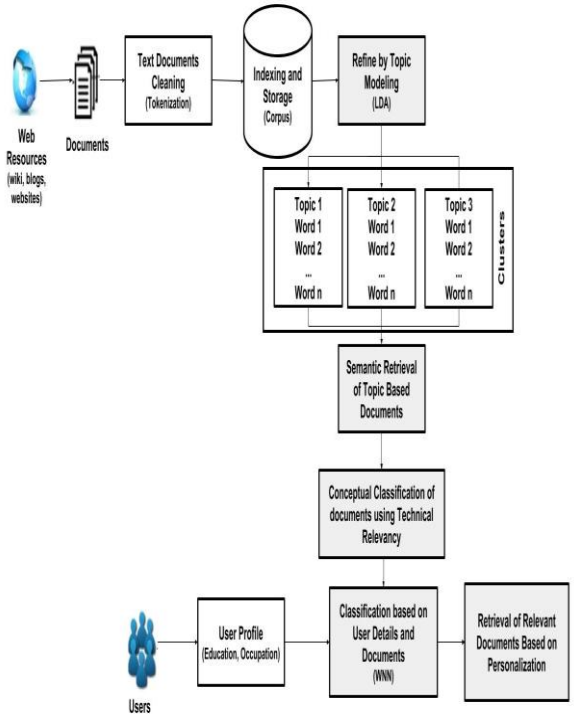

Figure 1: Semantic Key Topic and User Profile (SKTUP) based Retrieval System.

The proposed methodology aims to increase the accuracy of retrieved relevant documents based on user search topic. Semantic Key Topic and User Profile based Information Retrieval System (SKTUP-IRS) is proposed. This IRS comprised of a retrieval technique based on the topic modeling namely Latent Dirichlet Allocation (LDA) model and introduced Weighted Nearest Neighbor (WNN) model which utilizes Conceptual Classification of Documents using Taxonomy (CCDT) technique. LDA model is developed to retrieve documents based on topics.

Advantages of the proposed system are:
Retrieval of relevant documents based on personalized content will be available.
Processing time is reduced.
Automatic cleaning and extraction of unstructured data are possible.

V. SYSTEM ARCHITECTURE

The proposed architecture describes the overall process of the system. Documents are extracted from the web resources which cleaned and stored as a text document. LDA model is applied to cluster the documents based on topic. From each topic, documents are classified using taxonomy. WNN model used to retrieves documents based on personalization using user details. The modules includes the following Document Extraction

- Document Cleaning
- Indexing and Storage
- Topic clustering using Latent Dirichlet Allocation(LDA) model
- Personalization using Weighted Nearest Neighbor(WNN) model
  - Conceptual Classification of Document using taxonomy (CCUT).
  - Sentimental Analysis using Twitter Comments (SATC) and User Input based Feature Extraction.(UIFE)

VI. PROPOSED METHODOLOGY

Document Extraction

From the web resources like a wiki, blogs and websites documents are extracting by using web crawling. It detects and removes the surplus of the web page and retrieve only the main textual content of a web page. The textual content of webpage stores in the format of the text file. The images, tables and other data present there is removed by using web crawler. The web crawler works based on a step by step procedure

Step 1: Crawler is used to browse and retrieve the web pages and stored in the repository.

Step 2: Crawler extracts the keyword and matches with the web pages in the web and stores it.

Step 3: The count of the web pages is mentioned in the crawler with a conditional loop.

Document Cleaning

Document cleaning is performed along with stemming and stop words removal. The noise is removed by identifying outliers and inconsistencies in the data. Tokenization is used to change the paragraphs in the document into atomic segments. The meaningless words are removed and the words with the same meaning with different tense are also removed in this process.

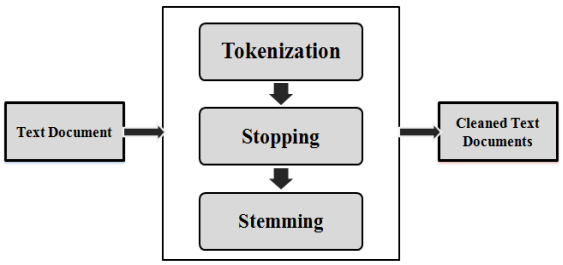

Figure 2 Block diagram of Document Cleaning

Indexing and Storage

Storing is an essential module which is used to store and index the several numbers of text documents in the corpus that are generated from the web resources. It is used to incorporate Latent Dirichlet Allocation (LDA) model to extract the exact information for user.

Topic clustering using Latent Dirichlet Allocation (LDA) model

Topic modeling can speed up the analysis significantly by reducing the number of documents for analysis. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Topic models provide a simple way to analyze large volumes of unlabeled text.

In LDA process, words in documents are the only observed variables while others are latent variables (θ) and hyper parameters (α). LDA assumes that each document can be represented as a probabilistic distribution over latent topics, and that topic distribution in all documents share a common Dirichlet prior. The documents are clustered according to the topics mentioned. The accuracy level of the clustering is evaluated by the metrics.

Personalization using Weighted Nearest Neighbor (WNN) model

The user profile is linked with the proposed system for prioritizing and ranking of the documents. Personalization based on Weighted Nearest Neighbor model use user profile and retrieval method of topic modeling to analyze user preferences and retrieves results. Weighted Nearest Neighbor model use to classify the user, based on the features like qualification, field of study and occupation with experiences. The feature extraction in this study is done by obtaining user inputs from the potential users. Sentimental analysis for the twitter comments posted by these users are accounted. The comments about the topics searched for is preserved and analyzed. The sentimental analysis proves that the comments of the topics searched is positive or negative. This also aids in finding out whether the user is familiar about the topic or not.

- Conceptual Classification using taxonomy (CCUT).
- Sentimental Analysis using Twitter Comments (SATC) and User Input based Feature Extraction.(UIFE)

Algorithm 1:

Input: Features extraction from user profile
Variables: Document Category (DC), User Category (UC), topic (T)

- **Step 1:** For each topic T, documents are retrieved based on topic modeling. The documents are classified as DC1, DC2, and DC3 using taxonomy based on technical relevancy which is implemented by AYLIEN API.
- **Step 2:** Each feature of user profile is represented with a set of numerical attributes (e.g. Software Engineer -9)
- **Step 3:** Each of the training data consists of set of features and a class label associated with each vector as UC1, UC2 and UC3.
- **Step 4:** Classification of user is done by comparing features of different k nearest points. (k=3)
- **Step 5:** Users are classified as Categories UC1, UC2 and UC3 based on the features extracted from the user profile.
- **Step 6:** Based on the query, user search retrieves relevant and personalized documents for user. For example, if UC1 category user searches for a topic cloud computing the system retrieves documents in the personalized order as DC1, DC2 and DC3.

For example:

Topic {T1,T2,T3} = { Cloud Computing, Python Programming, Big Data }
Doc {D1,D2,D3,D4,D5} = { cloud1.txt, cloud2.txt, cloud3.txt, cloud4.txt, cloud5.txt
DocumentCategory{DC1,DC2,DC3 } = { Basic, Intermediate, Advance }
DocumentCategory{DC1,DC2,DC3 } = { (cloud2.txt, cloud5.txt), (cloud1.txt, cloud4.txt), (cloud3.txt) }
User { U1, U2,….U10 } = { Alice, Bob,……Tom }
UserCategory{ UC1, UC2, UC3 } = { Non-Technical User, Intermediate User, Technical User} In this given example, according to the SKTUP-IRS model, the documents are categorized and retrieved based on user profile. For example, user (Bob) falls under the category of UC3 (Technical User). Bob retrieves the documents in priority order Documents {cloud3.txt, cloud1.txt, cloud4.txt, cloud2.txt, cloud5.txt}. The SKTUP-IRS model retrieves only the required documents and other fetched documents are prioritized and made available for future references. The kind of categorization implemented in this project enables the users to retrieve desired documents and store the ordered and less priotized documents for future use. It helps the user to have an instant conceptually classified and personalized search compared other models. The accuracy and quality of retrieval in this SKTUP-IRS is increased compared to the existing models like Keyword based model and LDA model.

VII. RESULT AND ANALYSIS

The metrics for evaluating the IRS are numerous. In this project, proposed system quality and accuracy is compared with the existing system like keyword search and topic modeling technique. All the extracted information from the web resources are analyzed and discussed in this chapter with comparison of keyword based retrieval and topic modeling. Recall is Number of relevant documents/Number of relevant documents exists in corpus

Table 1 : Recall and Precision of Documents

Set of Documents	Set 1	Set 2	Set 3	Set 4	Set 5
Precision	0.52	0.66	0.75	0.77	0.78
Recall	0.28	0.3	0.38	0.45	0.54
F-Score	0.182	0.281	0.412	0.504	0.644

The table 1 shows the performance levels of precision, recall and F-score for Keyword based model.The Precision and Recall are calculated here for keyword based model with different number of document sets for particular topic cloud computing.

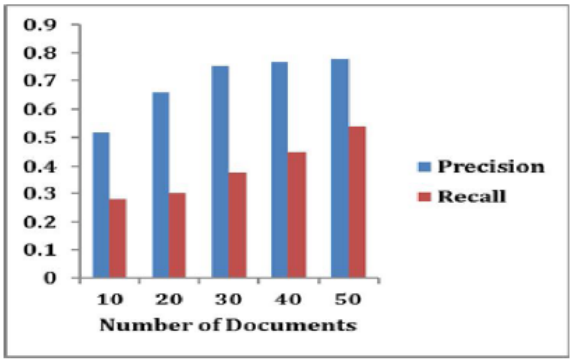

Figure 3 Precision and Recall for Keyword model

The figure 3 shows the performance levels of precision, recall and F-score for LDA model.

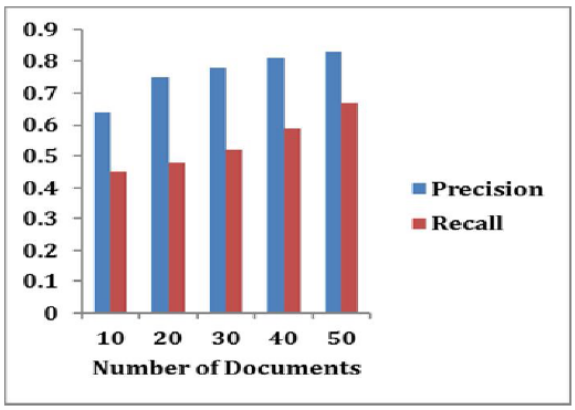

Figure 4 Precision and Recall for LDA model

The figure 5 shows the performance levels of precision, recall and F-score for SKTUP-IRS model.

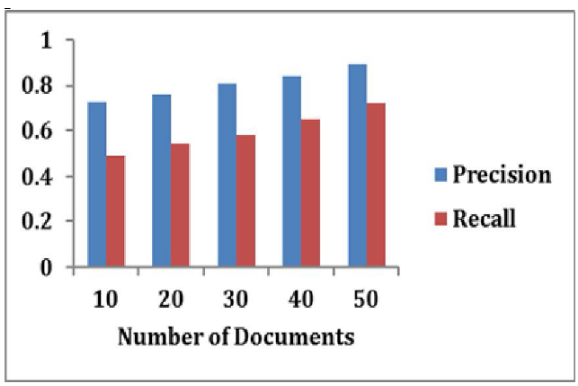

Figure 5 Precision and Recall for SKTUP-IRS model

The above graphs depicts that the precision and recall value of proposed SKTUP-IRS model is higher than the values obtained in Keyword based model. The relative recall values estimates the retrieval effectiveness between Keyword model and proposed system SKTUP-IRS model. The more relative recall values of proposed system shows that SKTUP-IRS model is more effective in retrieval than Keyword based model.

F- Score = 2*(Recall * Precision) / (Recall + Precision).

The accuracy of information retrieval is measured by F-score and performance is compared with keyword based model, LDA model and SKTUP-IRS. model.

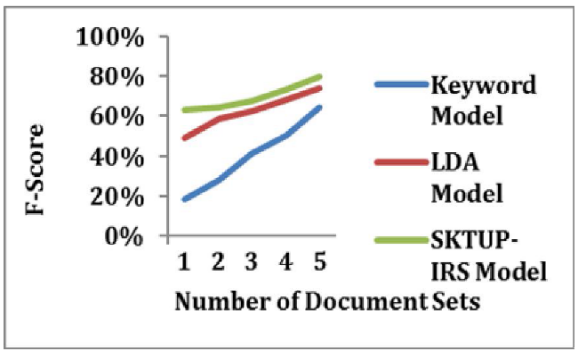

Figure 6 F-Score measures for models

The higher value denotes the best coverage of SKTUP-IRS model compared with Keyword based model and LDA model. This is because the combination of LDA model is realized with Weighted Nearest Neighbor (WNN) algorithm. The conceptual classification and WNN decreases and hampers the number of documents retrieved for the user in a particular domain. This limitation in retrieval of number of documents in turn increases the quality and accuracy of retrieval for an individual user as shown in the figure 6. The retrieval of documents using keyword method shows 40.2% accuracy and quality of retrieval, whereas in the LDA model the quality and accuracy is increased by 62.2% for the tested document sets. The proposed model shows the highest accuracy of retrieval with 69.2%. This proves that the proposed method stands best among the existing methods of retrieval.

VIII. CONCLUSION & FUTURE WORK

The information resources on the today's web are increasing day by day and there are number of problems in identifying the results are based on the users interest. In the conventional systems the data from the user's document are not taken in to account and the retrieval process has to be changed in the proposed work. There is a need to be changed. The whole of retrieval process is independent and exempted from query submission process and matching and retrieval process. To fix this above mentioned issue, in the proposed methodology, SKTUP-IRS is developed which comprises of Latent Dirichlet Allocation (LDA) topic model that is used to retrieve documents based on the key topics in the cluster. In addition to that personalization of content using Weighted Nearest Neighbor (WNN) model is incorporated.

Personalized retrieval of documents is performed by using features extracted from user profile details based on CCDT technique. The enhanced WNN and LDA are combined and a gain value is calculated to segregate the documents. The similarity index of the documents and recall, precision is enhanced compared to the typical LDA system and Keyword based models. The user interest based content retrieval system are applied on different kind of systems like knowledge representation systems, search engines and retrieval system can be used in knowledge management, search engine and other applications that require navigation and browsing of data with huge amounts of searchable information to store it for future use. Future work will be done based on a specific area and that takes up automatic semantic retrieval of web documents without manual work. The number of queries can be maximized and various experiments can be conducted for big data. Analytics can be applied over the obtained results to mine useful patterns from the search engines based on personalization.

REFERENCES

1. Amit Kumar Dhar, Monika Rani, O.P. Vyas, "Semi-Automatic Terminology Ontology Learning Based on Topic Modeling", Elsevier-Engineering Applications of Artificial Intelligence, Vol. 63, August 2017, pp 108-125
2. Nikhil Vohra, Pankajdeep Kaur, Pallavi Sharma, "An Ontology Based Text Analytics on Social Media", Elsevier-International Journal of Database Theory and Application, Vol. 8, February 2015, pp 233-240
3. KaisongSong, ShiFeng, WeiGao ,DalingWang ,GeYu ,Kam-FaiWong, "Personalized Sentiment Classification Based on Latent Individuality of Microblog Users", Engineering Applications of Artificial Intelligence, Vol. 7, June 2017, pp 124-133
4. Mohammadreza, Shams, Ahmad, Baraani-Dastjerdi, "Enriched LDA(ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction ", Elsevier- Expert Systems With Applications : An International Journal, Vol. 80 , February 2017 , pp. 83-93
5. Miha Pavlinek, Vili Podgorelec,"Text classification method based on self-training and LDA topic models", Elsevier- Expert Systems with Applications: An International Journal, Vol. 80, September 2017 pp. 83-93
6. Alhassan Bala, Mansur Zakariyya Shuaibu ,Zaharaddeen Karami Lawal, Rufa'i Yusuf Zakari, "Performance Analysis of Apriori and FP-Growth Algorithms", Elsevier- International Journal of Computer Technology & Applications, Vol. 7, July 2016, pp. 279-293
7. Andreia Dal Ponte Novelli, José Maria Parente de Oliveira," Simple Method for Ontology Automatic Extraction from Documents", Elsevier-International Journal of Advanced Computer Science and Applications, Vol. 3, July. 2012, pp. 121-131
8. A . Hogan , A . Harth , J. Umbrich , S. Kinsella , A. Polleres , S. Decker, "Searching and browsing linked data with SWSE: the semantic web search engine and Web Semantic" , Elsevier- Expert Systems With Applications : An International Journal, Vol. 9, September 2011, pp. 365–401
9. Wang Wei, Payam Barnaghi and Andrzej Bargiela, "Probabilistic Topic Model for Learning Terminological Ontologies", IEEE Transactions on Knowledge and Data Engineering, Vol.: 22, July 2010, pp. 246-253
10. Yueshen Xua, Yuyu Yinb, Jianwei Yind, "Tackling topic general words in topic modeling", Elsevier- Engineering Applications of Artificial Intelligence, Vol. 62, June 2017, pp. 124-133
11. Al-Mubaid, Nguyen," A cluster-based approach for semantic similarity in the biomedical domain", Elsevier-International Journal of Advanced Computer Science and Applications, Vol. 8, November 2006, pp. 273-277
12. Hyui Geon Yoon, Hyungjun Kim, Chang Ouk Kim, Min Song," Opinion polarity detection in Twitter data combining shrinkage regression and topic modeling", Elsevier-Journal of Informetrics, Vol.10, July. 2016, pp. 634–644
13. Songjie Gong," The personalized information retrieval model based on user interest", Elsevier- International Journal of Advanced Computer Science and Applications, Vol.24, June. 2012, pp. 817-821
14. Chahrazed Bouhini, Mathias Gery, Christine Largeron, "Personalized Information Retrieval Models Integrating the User's Profile", Elsevier- International Journal of Advanced Computer Science and Applications, Vol.22, Aug. 2015, pp. 628-634
15. Dixon Prem Daniel R, Rangaraja P Sundarraj, "A Latent Factor Model based Movie Recommender using Smartphone Browsing History", Elsevier- International Journal of Computer Technology & Applications, Vol. 8, July 2016, pp. 257-264
16. B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: A study of user queries on the web," SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998.
17. R.Priyadarshini, Latha Tamilselvan, Document based Semantic CMS, Information Technology Journal, Vol. 13, No.2, pp. 217-230.
18. R.Priyadarshini, Latha Tamilselvan, Semantic approach for documents in distributed framework with multi-node setup, International Journal of networking and virtual organizations, Vol. 19, No.2-4, pp. 321-340.