Cluster Computing (2019) 22:S12871–S12881
https://doi.org/10.1007/s10586-018-1789-8

Intelligent ontology based semantic information retrieval using feature selection and classification

B. Selvalakshmi[1] **· M. Subramaniam**[2]

Received: 29 November 2017 / Revised: 6 January 2018 / Accepted: 9 January 2018 / Published online: 3 February 2018

Abstract
Semantic information retrieval provides more relevant information to the user query by performing semantic analysis. In such a scenario, knowledge representation using ontology can provide effective semantic retrieval facility which is more efficient than representation using semantic networks and frames. The existing information retrieval systems have been developed to handle very large volume of data and information stored in text format. On the other hand, the information available in the current web based applications such as Facebook and twitter grow very fast and hence the existing information retrieval systems consume large amount of time for relevant information retrieval. Moreover, most of the existing search engines use syntactic approach for information retrieval and use page ranking algorithms to measure the relevancy score. However, such approach is not able to provide more accurate results in terms of relevancy. Therefore, a new semantic information retrieval system is proposed in this paper which uses feature selection and classification for enhancing the relevancy score which is performed in this work by proposing a new intelligent fuzzy rough set based feature selection algorithm and an intelligent ontology and Latent Dirichlet Allocation based semantic information retrieval algorithm. The main advantages of the proposed algorithms are the increase in relevancy, ability to handle big data and fast retrieval.

Keywords Ontology · Semantic-based retrieval · Map reduce · Multimedia big data · Big data retrieval and retrieval algorithm

1 Introduction

Ontology in semantic web is a knowledge representation technique which provides effective facility for the representation of concepts, roles and relationships more effectively. In the existing knowledge representation techniques such as semantic networks and frames are providing a facility for representing textual information with inheritance. Therefore, it is possible to search the semantic network or frames from the start node till the goal node is reached by applying class and sub-class hierarchies. However, the number of links and link types are restricted in semantic networks. This problem is resolved in the frame structure by using slots and fillers. The representation of knowledge using first order predicate logic to construct semantic networks and frames enable to represent semantic information more effectively. Moreover, retrieval using a knowledge representation technique is also possible by applying rules in the search procedure.

Heuristic search algorithms are useful to find the most promising route to reach the goal node. However, such algorithms are guided by the heuristic function to select the optimal path. On the other hand, for web information retrieval all these methods do not provide more relevant and optimal results. However, ontology with web languages is able to represent the information in both structured and semi-structured format in an efficient way. The ontology based approach for knowledge representation not only helps to represent the information with hierarchy but also maintains the inheritance hierarchy. Such information representation using ontology will retrieve information not only with single level inheritance but also with multilevel inheritance. Therefore, semantic information retrieval using an ontology based approach is more efficient for semantic analysis and relevant information retrieval.

Big data is defined as the data which cannot be captured, stored and processed by a normal computer in real time.

✉ B. Selvalakshmi
selvalakshmi856@rediffmail.com

[1] Department of CSE, Tagore Engineering College, Chennai, India

[2] S.A. Engineering College, Chennai 600077, India

Therefore, big data analytics models emphasize the need for pre-processing the data in order to reduce the volume of data by filtering unwanted and noisy data. Big data provides data generated from machines in the form of sensors or by human users who use the web based systems simultaneously from multiple locations leading to the generation of variety of information, with high volume at high velocity. In such a scenario, it is necessary to analyse the data and find the relevant information. All the irrelevant information must be identified and filtered. The relevant information must be analysed further to classify them based on their importance into different categories namely high relevant, medium relevant, less relevant information so that the search engine can provide more suitable results at a shorter period of time.

Pre-processing of information can be carried out in many ways. In natural language processing, it is performed by applying stop word removal, stemming and syntax analysis. However, pre-processing should not eliminate the relevant information from the dataset. Therefore, it is necessary to classify the data into two types of information namely relevant information and irrelevant information. In addition, the relevant information is also to be pre-processed by applying row reduction and column reduction if the data is stored in the form of a table. If the data is stored in unstructured format, first the information must be stored using an ontology by performing suitable semantic analysis. Based on this ontology, an information retrieval algorithm which uses ontology must be employed to select the more relevant information from the dataset. For this purpose, it is necessary to propose a new pre-processing algorithm which can perform data cleaning, data reduction and representation of data using ontology. In addition, a retrieval algorithm which performs classification of the available information in order to perform semantic analysis based on ontology and to retrieve the more relevant information which must be returned as the answer to the user query.

In this paper, a new feature selection algorithm called intelligent fuzzy rough set based feature selection algorithm and a new classification based information retrieval algorithm called intelligent ontology and Latent Dirichlet Allocation (LDA) based semantic information retrieval algorithm have been proposed for performing effective semantic information retrieval from big data datasets. Moreover, the proposed algorithms provide more relevant results to the user queries by performing ontology matching and hence they are very fast in retrieval and provide more accurate results. The proposed feature selection algorithm consists of the three major components namely feature analysis module for finding the relevancy by applying fuzzy rough sets in order to return the most significant features, feature and page ranking module for ranking the features and web pages and the semantic analysis module for providing significant features using Fuzzy Rough Sets, Ranking of the features and semantic analysis to form ontologies which will be used to represent the knowledge more efficiently. In addition, the newly proposed ontology and LDA based semantic information retrieval algorithm is useful to perform semantic analysis on the datasets and to retrieve syntactically and semantically relevant information from the dataset. The major advantages of the proposed model include increase in retrieval accuracy with respect to relevancy and semantics, reduction in retrieval time and semantics retrieval with multilevel inheritance.

Rest of the paper is organized as follows: Sect. 2 provides the related works in this direction. Section 3 explains the system architecture. Section 4 describe in detail about the proposed work. Section 6 shows the results and discussion. Section 7 gives conclusion and future works in this direction.

2 Literature survey

There are many works have been done by various researchers in this direction such as feature selection, classification, semantic analysis and ontology in the past. Among them, [1] designed a new inference model that is based on the variation approaches and they have introduced a new maximization algorithm that is used for the estimation of parameters which are to be used in the classification process using Bayes classification. [2] introduced a new LDA based document search model for retrieving the relevant information dynamically. [3] a set of topic detection approaches which is used for detecting relevant topics from documents that are available in online and offline. They have analysed the pros and cons of the traditional and non-probabilistic models.

Torii et al. [4] introduced a new architecture for search engine which is working based on rule for searching suitable information. [5] identified a relevancy based on the dependency by using their new model for text mining. Their model is used for retrieving the relevant information from web documents by applying Latent Dirichlet Allocation (LDA) model. [6] identified four factors which are to be used in their new recommendation system that has been developed to recommend the next location for the taxi drivers. In addition, they have performed an analysis on the location and times for developing a spatio-temporal based graph model. [7] identified the people's information seeking behaviour of information searching with time constraints. [8] introduced a new model for identifying new pattern that enhanced topic model which is filtered based on user interest and relevancy ranking.

Peng et al. [9] discusses about to provide the replies automatically and to the given queries in social networks. In their model, they have provided new techniques for noise

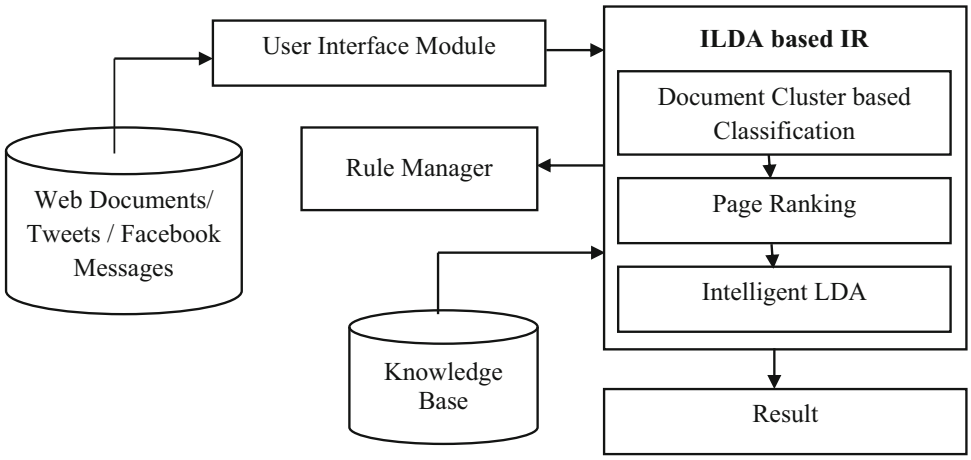

Fig. 1 System architecture

reduction, reduction of redundancy in messages and to perform text summarization. Hence, it is possible to improve the quality of the reply. [10] introduced a new solution which is systematic that referred as QDMiner and also to mine automatically by applying aggregation process frequently for items in the form of text or HTML tags in order to provide the top search results.A new model by [11] for semantic analysis and keyword analysis from the data available from data repositories. They developed a new model for analysing the meaning of input keywords by applying ontologies and by using different word sense disambiguation methods. In addition, they have performed an effective pragmatic analysis in order to connect the meaning from first sentence to the next sentence of the document. The main advantage of their work is that it refines the query and performs semantic analysis on the query first and then on the retrieved documents. Guo et al. [12] a new semantic ontology based optimization algorithm for effective information retrieval which uses the big data processing tools for storing and retrieving the ontologies from multimedia big data.

Xueke et al. [13] a new generative topic model which extracts aspects and aspect-dependent sentiment lexicons from online customer reviews. They used opinion words along with aspect-aware sentiment polarities for decision making. They applied the extracted aspect dependent sentiment lexicons to a series of aspect-level opinion mining tasks, including implicit aspect identification, aspect-based extractive opinion summarization and aspect-level sentiment classification. Claypo and Jaiyen [14] a new model to classify the opinions of customers for Thailand restaurants using opinion mining. In their method, a new algorithm for feature selection is adopted effectively to optimize the classification of Thailand restaurant reviews. In their model, neural networks are used to classify the opinions into positive and negative reviews. Lipizzi et al. [15] a new method to extract the contents of Twitter discussion to perform analysis on the reaction of social media before launching new products to prospective customers. Their method was supported by human analysts with respect to data collection and interpretation of social media data. The analysis considered both syntactic and semantic features. Their results showed significant differences in the structure of based on time.

Chung [16] in spite of the presence of all these works in the literature, most of the works do not consider and use the semantic analysis and fuzzy rules for making final decision. Moreover, the work considers ontology, time constraints, fuzzy rules and sematic for effective decision making. In this paper, a new fuzzy rough set based feature selection algorithm and a new ontology and LDA based classification algorithm.

3 System architecture

The proposed system architecture for the system developed for social networks analysis in this paper which is shown in Fig. 1. It consists of six modules namely web documents/tweets, user interface module, Intelligent LDA based information retrieval model, rule manager, knowledge base and result.

The collection of web documents, tweets and Facebook comments are considered as dataset. The user interface module collects the necessary web documents and tweets from the collection. Intelligent LDA and Ontology based IR module consists of three sub modules such as cluster based classification, page ranking and Intelligent LDA. The cluster based classification module make groups based on the relevancy and the ranking subsystem rank them. Finally, the intelligent LDA analyse and extract the suitable contents from the ranked documents that are based on the time of user queries. The knowledge base stored all kind of information for three times such as past, present and future. The pro-

posed model refers the knowledge base based on user query. The result module holds the resulted documents of the user query.

4 Proposed work

In this work, two new algorithms have been proposed for retrieving relevant information by applying feature selection and classification process with the help of rough sets, fuzzy rules and ontology.

4.1 Finding the relevance of each feature using fuzzy rough sets

In this work, fuzzy rough sets are used to perform feature selection. The feature selection algorithm proposed in this thesis computes the importance of each feature and selects a sub set of features which are more important for using as decision variables in the classification algorithms. Moreover, the fuzzy rough set provides a ranking of all features and their relevance score. Based on these parameters, important features are selected by this algorithm.

In the literature, the work on fuzzy sets was started by [17] who developed the theory of fuzzy sets which is use for the effective representation of incomplete information effectively. Even though, it was developed to handle semantic analysis by removing ambiguities, it is used in all types of applications now. However, the accuracy of decisions can further be improved by using fuzzy rough sets which was introduced by [18]. In their model, they discussed about fuzzy rough sets based on lower and upper approximation functions which are nothing but least upper bound (sup) and greatest lower bound (inf). They considered a non-empty universal set U, a relationship R and a fuzzy power set F(U).

$$\mu_{\emptyset} *_{(F)} F_i = \overset{sup}{x \in U} \min \{F_i(x), F(x)\} \tag{1}$$

$$\mu_{\emptyset} *_{(F)} (F) = \overset{inf}{x \in U} \max \{1 - F_i(x), F(x)\} \tag{2}$$

where $\emptyset = \{F_1, F_2, \ldots.., F_k\}$ is a fuzzy partition which is derived from R. Let r_i be the relevance of a feature A_i belonging to the feature set C. The proposed feature selection algorithm starts with the first feature and computes the value of r_i for each feature. After finding the relevance of all the features, each feature is assigned a weight based on its significance. Now, the weights W_i are multiplied with r_i to get the significance S_i . That is $S_i = W_i * r_i$. Apply the upper approximation and the lower approximation constraints on these S_i. Apply temporal constraints t_i to get $S_i(t_i) = W_i * r_i(t_i)$. Now, divide them into five groups namely low, low medium, medium, High medium and High depending upon the application. Select the features from high and high medium and rank them using their values. The ranked feature set is a reduced feature set which gives all the relevant information to the classifier.

The steps of the proposed Temporal Fuzzy Rough Set based Feature Selection Algorithm are as follows:

Intelligent Fuzzy Rough Set based Feature Selection Algorithm

Input: Set of all features.

Output:Set of relevant features.

1) Initialize the values of with 0.
2) Repeat the steps from 3 to 8 until the set D is empty.
3) Initialize maxto first feature.
4) Repeat the steps (4.a – 4.c) for each feature
   - 4.a Read one record from the dataset and form the parent class.
   - 4.b Identify the keywords and perform semantic analysis using ontology matching and find the importance of the feature.
   - 4.c Compute the rank of each feature by comparing the sub class relations for each super class present in the ontology.
5) Compute the weights W_i of each feature based on fuzzy rough sets scores.
6) Compute Semantic Score $S_i = F(W_i * r_i)$.
7) If S_i is greater than 10 and less than 40 then
   Set Relevancy score = less relevant
   Else If S_i is greater than 30 and less than 70 then
   Set relevancy score = medium Relevant
   Else If S_i is greater than 60 and less than or equal to 100 then
   Set relevancy score = High
8) Apply rough set constraints on S_i, rank S_i based on the current values.
9) Remove the irrelevant features from the dataset based on lower, medium and high thresholds.
10) Return the relevant features to the classifier.

4.2 Selection of threshold

The threshold selection and setting is a crucial step in feature selection process. Because of, threshold only will decide the selection process. In this work, the threshold is set using a user preference table where the user interests are ranked based on various features which are available in the document. The mean value of each feature from the user preference table has been considered in this work as the threshold values. Therefore, the thresholds are used in this newly proposed feature selection algorithm that consider the opinion of all the users and hence is an optimal threshold.

4.3 Information retrieval

Social networks are playing a major role in day today life due to the rapid growth of internet technology development and technological awareness. Many organizations and public are utilizing the social networks for advertising their business, sell and purchase. The young generations are living with social networks such as Twitter, WhatsApp and Facebook. They are exchanging their feels on social issues through posts, sharing their likes on various posts, shares their interest over the product, and shorter text messages through social network with their friends.

These popular social networks are having more than 1000 million users throughout the world today and they are also sharing their feel about the social issues, particular product, place, hotels, shopping malls and tourist places. People are receiving a lot of relevant information about above mentioned all through social networks, web through search engines. For example, if we want to go to any tourist place then

can get sufficient information regarding the beauty of the place, hotel and food facility in the place, culture of the living people, transport frequency and charges, minimum and maximum expected expenses and suitable time schedule for visiting the particular place. We can collect the necessary information from text documents which are uploaded by the experienced people, tweets and messages. We need an effective information model for retrieving the useful information from web and social networks. For that purpose, this paper introduced an effective information retrieval model for effective retrieval.

Information retrieval is the process of providing most relevant documents to the users from an existing collection which is available in web and social networks. In this fast world, time is most precious in every human being daily life. People likes and expects the fast information in short and correct time period. For example, people may need the past information, current information and future information depends on the weather and seasons. In such a scenario, temporal information retrieval is very useful for extracting the suitable information about the place, hotel, cost of living, transport and other expenses. This kind of temporal retrieval model uses the similarity search for finding the suitable information, document summarization for identifying the related documents and grouping the related documentation using clustering techniques for retrieving the effective information.

In this work, a new information retrieval model is proposed with intelligent rule, fuzzy rough set and ontology for effective information extraction from social networks. For this purpose, the proposed work has been collected tweets from five thousand users for a period of one month and hotel information related web documents. For example, consider the hotel related keywords such as charge, room, television, air condition, food, service, good, bad, excellent, transportation, nearest railway station, bus stand, airport and temples for identifying the user opinion. Based on synonym analysis, we can select features for positive sentiments and negative sentiments by proposing a new feature selection algorithm using keyword frequency and semantic analysis. In this thesis, a new Intelligent Ontology and LDA based Information Retrieval (IOLDAIR) algorithm is proposed for effective information retrieval in social networks. The proposed model uses a new cluster based classification technique which is used for grouping the relevant contents of web documents or short messages and rank them according to the relevancy of the given word and also use the Ontology and LDA for identifying the most suitable keyword for the particular time period and situation which are matched semantically with web contents in temporal nature. The main advantage of the proposed model is to retrieve the user expected information effectively. From the experiments conducted in this work, it is observed that the groups formed by relevance of documents provided better classification accuracy than the existing models.

In this work, a new Intelligent Ontology and Latent Dirichlet Allocation (TLDA) based information retrieval (IOLDAIR) model is proposed for effective information retrieval from social networks according to ([2] and [12]). It analyses the text semantically for extracting the exact words from web documents and short messages.

5 Intelligent ontology and latent Dirichlet allocation retrieval algorithm

The proposed Intelligent Ontology and Latent Dirichlet Allocation based Information Retrieval (IOLDAIR) model is designed according to LDA ([2]). The newly proposed model is a form of document model which used to search titles is categorized by a distribution over words, and documents are sampled from the collection of key terms. In this proposed IOLDAIR, the weights are assigned based on the users search words and multinomial parameters are random variables ϕ and $\theta^{(d)}$ with conjugate Dirichlet priors. The steps of the proposed algorithm as follows:

Input : User Query/ search word.

Output: Relevant Information

Step 1: Read the necessary web documents based on the user query/ search word and form the query ontology.

Step 2: Group the web documents from dataset based on the relevancy of user request and semantics using ontology alignment.

Step 3: Rank the web documents according to the closure of user request based on ontology hierarchy and by concept analysis using description logic based ontology matching.

Step 4: Form Documents based on a Dirichletcomputed using the parameter β for the multinomial $\vartheta_z$formed on the given words for each user query z;

Step 5: For each web document $doc(i) = \{D1, D2 \ldots . Dn\}$, i=1 to n, perform the following steps:

5.a Check for class and inheritance hierarchy based on properties present in the document.

5.bIf the documentis relevant (medium and high) for further analysis then

i) Read data and compute Dirichletscore with parameter α for the multinomial $\theta^{(d)}$ on the given query

ii) For n = 1 . . . n_{doc};

iii) Data from $\theta^{(doc)}$ for the title z_n;

iv) Data from φ_{z_n} for the word w_n;

Else

Display the suitable messagesby traversing through the ontology

Step 8: Stop the process and display the relevant information.

Each word win the web document doc is associated with a latent title z. A multinomial distribution $\theta^{(d)}$ on title/topic z is identified from a dirichlet distribution with parameter the given parameter α and the title or topic also selected based on the distribution of h. A word-level switch variable for the particular topic or title in the graphical model of LDA is introduced in this proposed model for successful information extraction. The weights are assigned for each documents of the group based on their relevancy. The relevancy is validated with the knowledge base of the system. This knowledge base has large volume of data or information with various combinations.

6 Results and discussion

In this paper, the proposed model is empirically evaluates in dynamic information retrieval and compare it with other existing models. This system is designed by using Java programming language.

Table 1 shows the relevancy score of the various number of web documents considered in this work based on the different groups such as Group 1 to Group 5 with 1000, 2000, 3000, 4000 and 5000 web documents and tweets.

From Table 1, it can be observed that the proposed model performance with different group of web documents and tweets in different five experiments. The relevancy score

Table 1 Relevancy score analysis

Exp. no.	Groups	No. of web documents considered	No. of Tweets considered	Relevancy score (%)	
				Web docs.	Tweets
1	Group1	2000	2000	98.32	99.32
2	Group2	4000	4000	97.54	99.31
3	Group3	6000	6000	99.31	99.60
4	Group4	8000	8000	98.17	99.14
5	Group5	10,000	10,000	98.22	98.97

Table 2 Comparative time analysis

No. of records considered	LDA		SOR		Proposed classifier	
	Full features	Selected features	Full features	Selected features	Full features	Selected features
1000	0.29	0.27	0.24	0.21	0.19	0.17
2000	0.38	0.32	0.30	0.25	0.23	0.20
3000	0.47	0.37	0.36	0.28	0.27	0.24
4000	0.55	0.42	0.41	0.32	0.31	0.27
5000	0.62	0.45	0.47	0.36	0.35	0.29

Table 3 Comparative performance analysis

No. of records considered	LDA		SOR		Proposed classifier	
	Full features	Selected features	Full features	Selected features	Full features	Selected features
1000	89.42	91.24	90.03	92.34	94.91	99.12
2000	89.51	91.26	90.06	92.37	94.95	99.15
3000	89.56	91.31	90.11	92.39	94.97	99.18
4000	89.62	91.33	90.15	92.42	94.99	99.21
5000	89.65	91.36	90.18	92.45	95.05	99.24

Table 4 Comparative performance analysis

No. of records considered	LDA		SOR		Proposed classifier	
	Selected features	Grouped features	Selected features	Grouped features	Selected features	Grouped features
1000	91.24	91.26	92.34	92.36	99.12	99.14
2000	91.26	91.28	92.37	92.39	99.15	99.16
3000	91.31	91.33	92.39	92.41	99.18	99.18
4000	91.33	91.35	92.42	92.44	99.21	99.22
5000	91.36	91.38	92.45	92.47	99.24	99.25

of these similar numbers of web documents and tweets are almost same in all the five experiments.

Table 2 shows the comparative time analysis for the various classification algorithms namely LDA ([2]), SOR ([12]) and the proposed classifier with full features and selected features for the various number of comments such as 1000, 2000, 3000, 4000 and 5000.

From Table 2, it can be observed that the proposed classifier takes less time than the existing classifiers with full features and the selected features. This is due to fact that the proposed algorithm uses ontology, temporal constraints and fuzzy rules for making effective decisions.

Table 3 shows the comparative performance (Accuracy of classification) analysis for the various classification algorithms namely LDA, SOR and the proposed classifier with full features and selected features for various numbers of comments such as 1000, 2000, 3000, 4000 and 5000.

From Table 3, it can be observed that the proposed classifier performs better than the existing classifiers when the experiments have been conducted with full features and selected features. This improvement in accuracy is obtained because the proposed classifier handles uncertainty effectively using fuzzy temporal rules. In addition, the proposed classifier also fires the spatial constraints for effective processing.

Table 4 shows the comparative performance analysis for the various classification algorithms namely LDA, SOR and the proposed classifier with selected features and grouped features for the various amounts of comments such as 1000, 2000, 3000, 4000 and 5000.

From Table 4, it can be observed that the proposed classifier performs better than the existing classifiers with selected features and the grouped features due to grouping effectively.

Figure 2 shows the comparative analysis in terms of relevancy accuracy of the proposed system model and the existing model on web documents.

From Fig. 2, it is observed that the proposed model perform well in various experiments with different number of

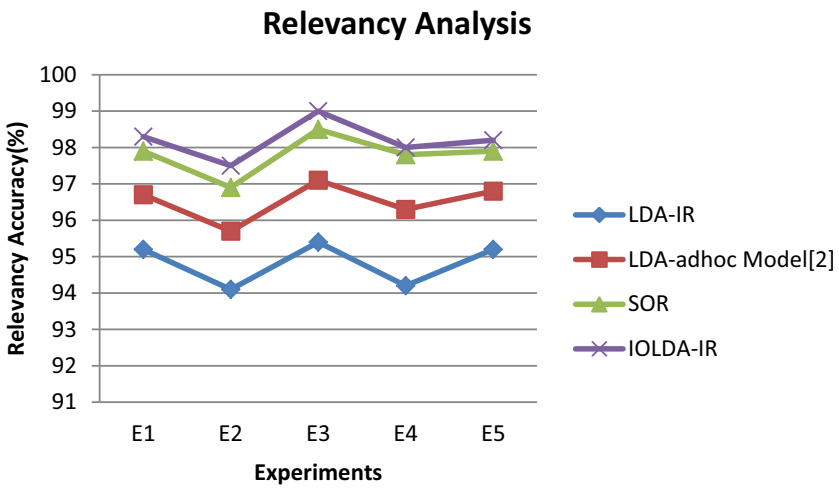

Fig. 2 Comparative analysis in relevancy score analysis (Web documents)

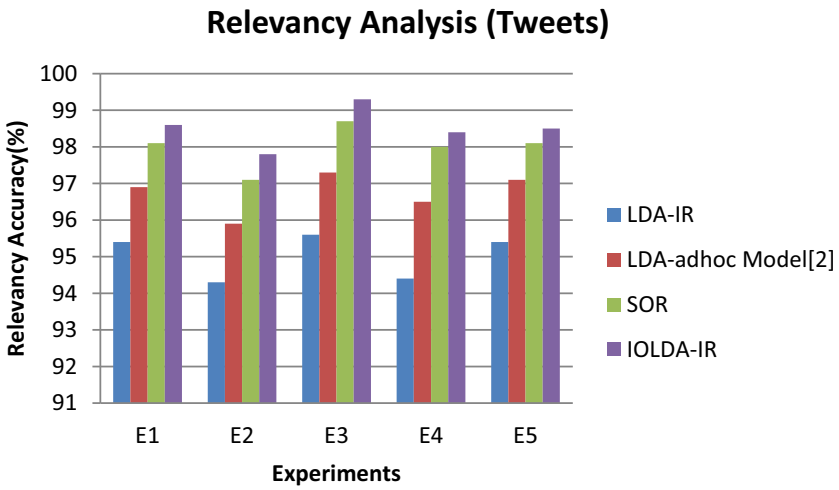

Fig. 3 Relevancy score analysis (Tweets)

web documents when it is compared with the existing system. This is because of the fact that the use of temporal features and clustering technique.

Figure 3 shows the comparative analysis in terms of relevancy accuracy of the proposed system model and the existing model on Tweets.

From Figure w, it is observed that this proposed model performs well in various experiments with different number of tweets when it is compared with other existing model. This is because of the fact that the use of temporal features and clustering technique.

Figure 4 shows the comparative analysis in terms of relevancy accuracy of the proposed system model and the existing model on Facebook comments.

From Fig. 4, it can be seen that this proposed model performs well in various experiments with different number of Facebook messages when it is compared with other existing model. This is because of the fact that the use of ontology, temporal features and fuzzy rules.

A new ontology and LDA based model called IOLDA-IR is proposed and implemented in this paper for effective relevant information retrieval in social networks. The proposed model use intelligent fuzzy rules for making different group of web documents and rank them according to the relevancy of the given word/term and also uses the Ontology, Latent Dirichlet Allocation (LDA) for identifying the suitable keyword which are matched semantically with web documents. The experimental results show that the importance and need for the proposed model for this fast world.

7 Conclusion and future work

In this paper, new algorithms for pre-processing and LDA based document classification to perform semantic information retrieval using ontology matching have been proposed. The pre-processing algorithm eliminates the irrelevant and noisy information from the documents and forms ontologies. The Intelligent Ontology and LDA based document retrieval algorithm constructs ontology on the retrieved documents and they are compared with the query ontologies. Both these algorithms together reduces the retrieval time, increase

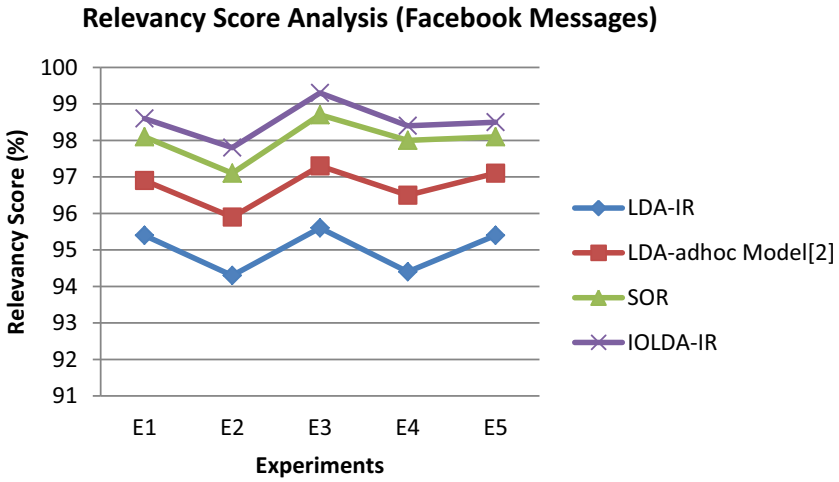

Fig. 4 Relevancy score analysis (Facebook messages)

the semantic relevancy by applying ontology and handle big data more efficiently using soft computing techniques namely rough sets and fuzzy rules for pre-processing and to perform effective grouping of the documents for returning more relevant documents to the user queries when they are compared with syntax based information retrieval systems and also the semantic based approaches which are used for semantic information retrieval. Future works in this direction can be the use of intelligent agents and agent based communication for fast and effective distributed processing of the user queries.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06), pp. 1–8 (2006)
3. He, Q., Chang, K., Lim, E.P., Banerjee, A.: Keep it simple with time: a reexamination of probabilistic topic detection models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1795–1808 (2010)
4. Torii, M., Arighi, C.N., Li, G., Wang, Q., Wu, C.H., Vijay-Shanker, K.: Rlims-P 2.0: a generalizable rule-based information extraction system for literature mining of protein phosphorylation information. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**(1), 17–29 (2015)
5. Wu, M.-S.: Modeling query-document dependencies with topic language models for information retrieval. Inf. Sci. **312**, 1–12 (2015)
6. Hwang, R.-H., Hsueh, Y.-L., Chen, Y.-T.: An effective taxi recommender system based on a spatio-temporal factor analysis model. Inf. Sci. **314**, 28–40 (2015)
7. Joho, H., Jatowt, A., Blanco, R.: Temporal information searching behaviour and strategies. Inf. Process. Manag. **51**, 834–850 (2015)
8. Gao, Y., Xu, Y., Li, Y.: Pattern-based topics for document modelling in information filtering. IEEE Trans. Knowl. Data Eng. **27**(6), 1629–1642 (2015)
9. Peng, M., Gao, B., Zhu, J., Huang, J., Yuan, M., Li, F.: High quality information extraction and query-oriented summarization for automatic query-reply in social network. Expert Syst. Appl. **44**, 92–101 (2016)
10. Dou, Z., Jiang, Z., Sha, H., Wen, J.-R., Song, R.: Automatically mining facets for queries from their search results. IEEE Trans. Knowl. Data Eng. **28**(2), 385–397 (2016)
11. Bobed, C., Mena, E.: QueryGen: semantic interpretation of keyword queries over heterogeneous information systems. Inf. Sci. **329**, 412–433 (2016)
12. Guo, K., Liang, Z., Tang, Y., Chi, T.: SOR: an optimized semantic ontology retrieval algorithm for heterogeneous multimeida big data. J. Comput. Sci. https://doi.org/10.1016/j.jocs.2017.02.005 (2017)
13. Xueke, X., Cheng, X., Tan, S., Liu, Y., Shen, H.: Aspect-level opinion mining of online customer reviews. China Commun. **10**(3), 25–41 (2014)
14. Claypo, N., Jaiyen, S.: Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection. In: 2014 International Computer Science and Engineering Conference (ICSEC), pp. 394–397 (2014)
15. Lipizzi, C., Landoli, L., Marquez, J.E.: Extracting and evaluating conversational patterns in social media: a socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. Int. J. Inf. Manag. **35**, 490–503 (2014)
16. Chung, W.: BizPro: extracting and categorizing business intelligence factors fromtextual news articles. Int. J. Inf. Manag. **34**, 272–284 (2014)
17. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
18. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. Int. J. Genetic Syst. **17**, 191–208 (1990)

B. Selvalakshmi is an Assistant Professor in Tagore Engineering College, Chennai. She received her B.E. Degree in Computer Science and Engineering from Madras University in 2008, M.B.A. Degree from Periyar University, Salem in 2001 and M.E. Degree in Computer Science and Engineering from Anna university, Chennai in 2013. She once worked as Sr. Lecturer in Vinayaga Mission Kirupananda Variyar Engineering College, Salem during 2001 to 2006 and System Analyst in L3

Info Solution during 2007 to 2010. She Joined Tagore Engineering College in 2013. Her research interest include big data, cloud computing and Networking. She has published around 5 academic papers.

M. Subramaniam (1974) is a Professor & Head for the Department of Information Technology at S.A. Engineering College affiliated to Anna University, Chennai, (INDIA). He obtained his Bachelor's degree (B.E.) in Computer Science and Engineering from University of Madras (1998), Master degree (M.E.) in Software Engineering and Ph.D. from College of Engineering Guindy (CEG), Anna University Main Campus, Chennai-25 in the year 2003 and 2013 respectively. His research focuses are Computer & Mobile Networks, Cloud, Big-data and Software Engineering. He is an active life member of the Computer Society of India (CSI) and the Indian Society for Technical Education (ISTE). He has six Research scholars perusing Ph.D. under his guidance. He published many research papers in reputed journals. He is also reviewer in IEEE- International Journal of Communication Systems.