IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 7, July 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Patient report Retrieval Using Semantic LDA with Cosine Similarity

Dr.Poonam Yadav

Assistant Professor, D.A.V College of Engineering & Technology, Kanina, Haryana 123027, India
poonam.y2002@gmail.com

Abstract

Now a day, the enlargement of information technology has altered the entire human life which includes medical and healthcare behaviors. A medical information storage and retrieval system is a precious tool used by the healthcare professionals for investigating the medical cases. Extracting the similar cases from the case database can aid the doctors to recognize the same kind of patients and their treatment details. This paper presents the case retrieval method for extracting the similar cases from the collection of patient's reports. Initially, the patient's details are gathered and divided into 100 reports. Then, the stop words and delimiters are removed from these reports, and each report is applied to the SLDA model. The SLDA model generates the topic keywords for the reports, and the patient-topic matrix is constructed. Then, the query report is sent to the proposed system, the stop words removal and the stemming process is performed. Then, the keywords in the query report are matched with the retrieved topic keywords, the most similar topic is extracted and the corresponding patient report is retrieved from the database. The matching process is performed by the cosine similarity. The performance of the proposed method is analyzed with the existing methods, such as LDA and LDA+CS for the performance measures DCG and fall-out. The experimental results show that the proposed method attains the higher DCG of 20.154 and the minimum fall-out 0.032 when compared to the existing methods.

Keywords: *Medical Case Retrieval, SLDA, Cosine Similarity, Patient-Topic Matrix, Query Matching.*

1. Introduction

In the medical field, the expert's knowledge includes the both experience and the textbook knowledge. It also includes cases, exceptional and typical, and the reasoning of the physicians [10]. There are two types of knowledge in the medical knowledge based systems, namely objective knowledge and the subjective knowledge [20]. The objective knowledge is derived from the textbooks, and the subjective knowledge is derived from the experience which is reduced in time and space changes. The subjective knowledge can be updated by integrating the new up-to-date cases [10]. The objective knowledge can be represented as functions or rules, and the subjective knowledge can be represented as cases [1].

One of the triumphant techniques in knowledge-based systems is the Case-based Reasoning (CBR) while in medical domains number of problems occurs to employ this method. CBR utilizes the past experience in the form of cases to recognize and resolve new problems [1]. The CBR consists of four phases, namely retrieval phase, reuse phase, revise phase and retain phase from which the retrieval is the important phase because the victory of CBR systems depends on the performance of this phase [11] [2]. The major goal of the retrieval phase is to retrieve the cases from the case database which are similar to the query [18]. If the retrieval phase does not retrieve the similar cases for the query, then the CBR system provide the incorrect solution to the new problem. CBR systems retrieve the similar cases based on similarity retrieval of knowledge, named as SBR [11].

The SBR determines the similarity among the old case and the new problem by the knowledge which is encoded in the form of the similarity measure. It also utilizes the searching and measurement similarity to find the most similar cases which are utilized to solve the new problems [6]. There are two problems occur in SBR. The first problem arises due to the dependency on the domain experts to represents the Similarity Knowledge (SK) [12]. The second problem is that the similarity measures definition of is static hence the definition is highly possible to be applied to every target problems. This leads to a problematic situation where a similarity criterion defined in a given domain is useful for some target problems but not useful for most of the problems [2]. Besides, information retrieval is facilitated by advanced compression methods [17] and modern data transmission technologies [15] [16] and intelligence methods used in many applications [14].

This paper presents the case retrieval method for extracting the similar cases from the collection of patient's reports. Initially, the patient's details are gathered and divided into 100 reports. Then, the stop words and delimiters are removed from these reports, and each report is applied to the SLDA model. The SLDA model generates the topic keywords for the reports, and the patient-topic matrix is constructed. Then, the query report is sent to the

proposed system, the stop words removal and the stemming process is performed. Then, the keywords in the query report are matched with the retrieved topic keywords, the most similar topic is extracted and the corresponding patient case is retrieved from the database.

The rest of the paper is organized as follows; Section 2 presents the motivation of this research work, Section 3 presents the system model of the proposed case retrieval model. Section 4 presents the proposed case retrieval method for retrieving the similar cases from the collection of patient's reports. Results and discussions are presented in Section 5 and Section 6 concludes the paper.

2. Motivation

This section presents the review of the existing research works for retrieving the similar cases from the patient report database and the challenges of the case retrieval process.

2.1 Literature Review

Here, five research works in case retrieval are discussed, the advantages and the disadvantages of each method is described. Yong-Bin Kang *et al.* [2] have presented a retrieval strategy for case-based reasoning using similarity and association knowledge. The authors utilize the different association rule mining methods for extracting the associative knowledge. The advantage of this method is that the AK can be built from the case base in a simple manner. This method was not suitable for complex structures, like semantic web-based cases, hierarchical cases, and object-oriented cases. André Mourão *et al.* [3] have proposed a medical information retrieval system for multimodal medical case-based retrieval with unsupervised rank fusion. This method helps the users to retrieve the relevant information and minimizes the frustration. The results generated by this method were poorer because of the performance difference among the image runs and text.

Israel Alonso and David [4] have proposed a straightforward information retrieval system for the biomedical field depending on the semantic similarity metrics and on the UMLS Metathesaurus. This method could apply to any similarity search process performed on the biomedical documents, such as X-Rays, CTscans, clinical reports, and patient histories. This method was not suitable for real-life environments. Yanshan Wang *et al.* [5] have proposed two NLP-empowered IR models, POS-BoW and POS-MRF, which integrates the automatic POS-based term weighting techniques into Markov Random Field (MRF) IR models and bag-of-word (BoW). This method was efficient than the existing frequency based and heuristic based weighting techniques. This method had two drawbacks; the first one is it depends on the tagging performance of POS, the second one is this method did not consider the external data resources. Corey W. Arnold *et al.* [8] have proposed a clinical Case-based retrieval using Latent Topic Analysis. This method could apply to all the type of clinical documents, and it did not require customization. This method provides poor results regarding of recall and precision because of the additional requirements of producing a gold standard.

2.2 Challenges

Medical case retrieval is crucial search application which provides several unique challenges.

- *Vocabulary Gap:* Medical domain utilizes the high-level languages including abbreviations, term-order variations, long multi-word expressions, and so on. Medical cases which are similar have various keyword variations. Hence, the keywords utilized in the query do not match with the variants in the documents, and the variants are morphologically different and conceptually similar. This problem is called as the vocabulary gap problem.
- *Non-Optimal Query Term Weighting:* Case retrieval queries consist of information about the background of patients, symptoms, medical test observations and results, and so on. Therefore, most of the keywords do not correctly determine a case. The significance of the query keyword is determined by the primary heuristic method which is based on the IDF (Inverse Document Frequency) which does not perform well [7].

3. System Model

This section presents the system model of the proposed case retrieval method for retrieving the similar cases from the collection of patient's information. The proposed method uses the Semantic Latent Dirichlet Allocation (SLDA) [9] for characterizing the reports. SLDA accommodates a diversity of response types. Initially, the hospital reports are collected from the different patients, and the collection of reports forms the patient's report database. Each report consists of a number of words. Then, the stop words removal and the stemming process are performed on these reports. After these processes, the SLDA model is applied to each report. The SLDA model estimates the Dirichlet parameters (alpha and beta), the phi and gamma value, and generates the topic for reports. Then, the input query which represents the hospital information of the new patient is applied to the proposed system, and the stop word removal, stemming processes are performed. Then, the keywords in the query document are matched with the retrieved topics, the most similar

topic is extracted and the corresponding patient case is retrieved from the database.

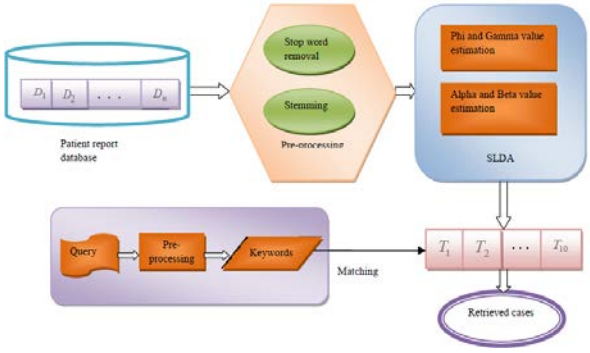


Fig.1. System model of the proposed case retrieval method

4. Proposed Case retrieval method for extracting the similar cases from the collection of patient reports

This section presents the proposed case retrieval model for retrieving the similar cases from the collection of patient reports. Initially, the medical reports are collected from different patients and divided into a number of groups. Then, the stop words removal and the stemming process are performed on every medical document. After this process, each document is applied to the SLDA model which generates the topics for the reports. Then, the query document is applied to the proposed case retrieval model; the stop words removal and the stemming process are performed on the query input. Then, the keywords in the query document are matched with the retrieved topics, and the most similar topic is extracted, and the corresponding patient case is retrieved from the database.

4.1 Collecting the hospital reports from patients

At first, the hospital reports are collected from 100 different patients, and the medical reports of every patient is divided into groups depending on the report types. The collection of reports is stored in the patient report database which can be represented as,

$$C = \{D_1, D_2, \ldots, D_n\} \quad (1)$$

where, C represents the Patient report database, $\{D_1, D_2, \ldots, D_n\}$ is the number of reports in the patient case database. Each document has a number of words which can be represented

as, $D = \{W_1, W_2, \ldots, W_m\}$ (2)

where, D represents the reports and $W_1, W_2, \ldots, W_m$ represents the number of words in the document.

4.2 Pre-processing

In this step, the document dependent attributes are retrieved from the patient report database, and the words from the reports are extracted by reading the reports via programming. The delimiters, like {(!",$&? are removed from the words while reading the report. Then, the stop words, such as it, a, he, she are removed from the reports. After removing the delimiters and stop words from the reports, a set of keywords are collected from each report in the patient's report database. Here, the major goal is to extract the document dependent attributes and to construct the corresponding matrix. This matrix is constructed by defining the vocabulary list and the vocabulary list is defined by determining the unique words from the set of keywords. Then, the next step is the semantic matrix construction in which the synonyms and it's semantic meaning are determined for each keyword. The semantic meaning of the word is extracted via MS word API. Each element in the semantic matrix represents the frequency of semantic words.

4.3 Topic Extraction using SLDA

Here, the topics are extracted by the SLDA model in which the input to the SLDA is the semantic matrix, and the output from the SLDA is the number of topics for the reports in the patient report database. Fig. 2 shows the block diagram of the SLDA model for retrieving the topic. Initially, the phi matrix is computed and this phi value is updated, and the synonyms distribution is determined by the semantic matrix. Then, the word distribution matrix is determined and the gamma value is estimated. After these processes, the Dirichlet parameters α and β are calculated and the topic words are extracted from the reports.

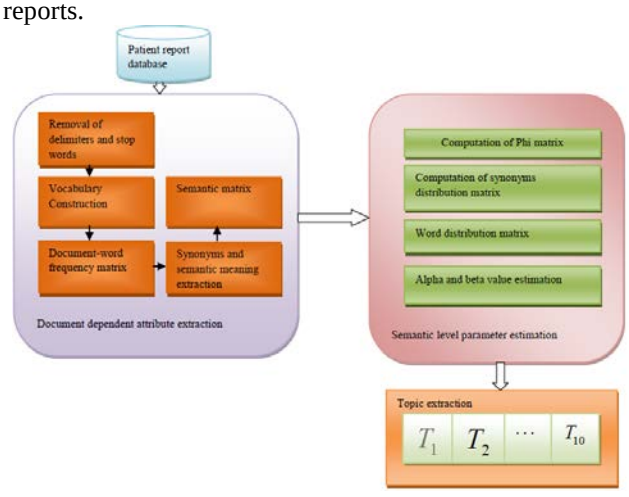


Fig. 2. Block diagram of the SLDA model

a) Phi and Gamma value estimation

At first, the random mixture of gamma and multinomial topic distribution is generated and the new matrix is generated. Then, the topic distribution matrix (Phi) for multinomial distribution is computed. This phi value is updated, and the synonyms distribution $\left(\mu^{sy}\right)$is attained by the semantic matrix. μ^{sy} specifies the probability distribution of all the semantic words among the topics. Then, the topic membership matrix is determined by the similarity function. Here, the similarity function is calculated by the four parameters: 1) Keywords are presented in both reports and have different frequencies, 2) Keyword is presented in any one of the report, 3) Keyword is not presented in both of the reports, and 4) Keyword is presented in both reports and have same frequencies. Depending on these parameters, the weight values are assigned. Then, the topic membership matrix is updated via the semantic word and then updated via data matrix. Then, the gamma value is determined by the word distribution $\left(\mu^{x}\right)$.

b) Estimation of alpha and beta values

The Dirichlet parameters α and β are significant for determining the topic membership function. A first, the beta value is calculated by the random mixture of normal distribution and the normalized form is calculated. After determining the beta value, the alpha value is calculated by determining the difference among the weight value of delta and the alpha value in the previous iteration. The alpha value is updated until the termination criteria reach. Here, the parameter utilized for evaluating the model is perplexity. At last, the membership matrix groups the reports depending on the report's probability and ten topic words are retrieved by determining the frequent words in the report groups.

4.4 Construction of Patient-Topic matrix

Here, the patient matrix is constructed with the topic words and the reports of patients which have the size of 100*10. Every patient's report is matched with the ten topics using cosine similarity. Every element in the patient-topic matrix is the matching similarity between the corresponding patient report and the topic keyword. The matching is performed by the cosine similarity which can be represented as,

$$\cos ine\ similarity = \frac{P \bullet T}{\|P\|_2 \|T\|_2} \qquad (3)$$

where, P represents the patient's report and T represents the extracted topic. Fig. 3 shows the patient-topic matrix in which $\left(T_1, T_2, \ldots, T_{10}\right)$ represents the topic words and $\left(P_1, P_2, \ldots, P_{100}\right)$represents the patient's reports.

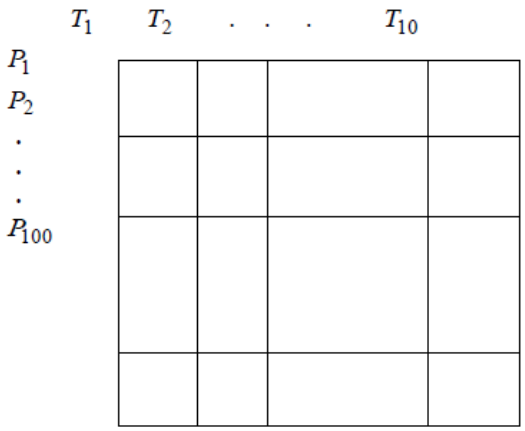


Fig. 3. Patient-Topic Matrix

4.5 Topic vector of query report

The keywords should match with every topic keywords using cosine similarity to generate topic vector which has the length of 1*10. Now, this topic vector is matched with every vector of a Patient-Topic matrix using cosine similarity. Then, the k-number of reports having the most similarity taken as retrieved cases.

4.6 Query Matching

In this step, the input query is applied to the proposed case retrieval model. At first, the stop words are removed from the input query, and the stemming process is performed. Then, the keywords in the query document are matched with the topic keywords in the patient-topic matrix and the most similar topic is extracted and the corresponding patient report is retrieved from the patient report database.

5. Results and Discussion

This section presents the experimental results of the proposed case retrieval method and the comparative analysis of the proposed method with the existing case retrieval methods, such as LDA [8] and LDA+CS for two different data sets, namely breast cancer and breast cancer wins.

5.1 Experimental Setup

Platform: The proposed case retrieval technique is experimented in a personal computer with 2GB RAM and 32-bit OS and implemented using MATLAB 8.2.0.701 (R2013b).

Datasets used: The datasets, such as breast cancer and breast cancer wins are taken from the UCI machine learning repository [13] for experimenting the proposed method.

Evaluation metrics: Evaluation metrics considered for analyzing the performance of the proposed method are Fall-out and Discounted Cumulative Gain (DCG).

Fall-out: It is defined as the proportion of retrieved non-relevant reports and the total number non-relevant reports.

$$fall-out = \frac{\left|\{non-relevent\ reports\} \cap \{retrieved\ reports\}\right|}{\left|\{non-relevent\ reports\}\right|} \quad (4)$$

DCG: It is defined as a measure of ranking quality. In information retrieval, it is utilized to quantify the usefulness of web search engine algorithms or related applications.

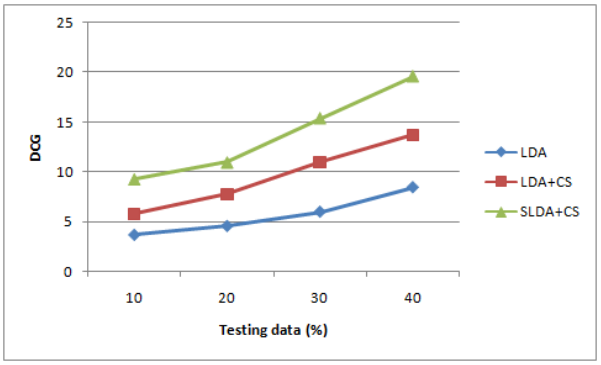
5.2 Performance Analysis

Here, the performance of the proposed method is analyzed with the existing methods, such as LDA [8] and LDA+CS for the performance measures DCG and Fall-out. The performance analysis is performed in two various data sets, namely breast cancer and breast cancer wins.

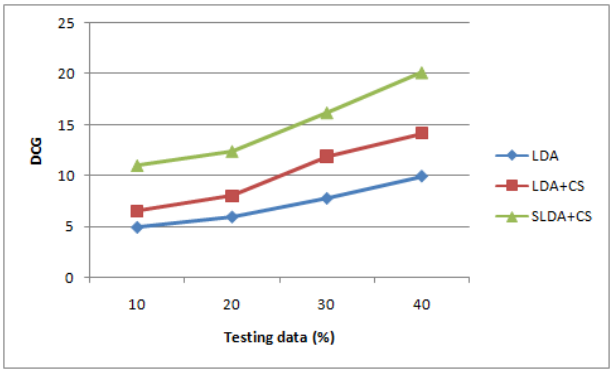
a) DCG

Fig. 4 shows the DCG curve of the proposed case retrieval method and the existing methods, such as LDA and LDA+CS for breast cancer and breast cancer wins data set for various sizes of the testing data. Fig. 4 (a) shows the DCG curve of the proposed method and the existing methods for breast cancer data set. For 10% of the testing data, the DCG of the proposed method is 9.255 while the DCG of the existing methods, such as LDA and LDA+CS is 3.654 and 5.789 respectively. The proposed method has the DCG of 10.948 on the other hand LDA method has the DCG of 4.546, and the LDA+CS has the DCG of 7.754 when the size of the testing data is 20%. Similarly, the proposed method has the higher DCG than the existing methods, such as LDA and LDA+CS for various sizes of the testing data.

Fig. 4 (b) shows the DCG curve of the proposed method and the existing methods for breast cancer wins data set. For 20% of the testing data, the DCG of the proposed method is 12.387 while the DCG of the existing methods, such as LDA and LDA+CS is 5.947 and 8.026 respectively. The proposed method has the DCG of 16.197, on the other hand, the LDA method has the DCG of 7.762, and the LDA+CS model has the DCG of 11.873 when the size of the testing data is 30%. Similarly, the proposed method has the higher DCG than the existing methods, such as LDA and LDA+CS for various sizes of the testing data. From Fig. 4, it can be shown that the proposed method has the maximum DCG than the existing methods for both data sets.



a) Breast cancer



b) Breast cancer wins

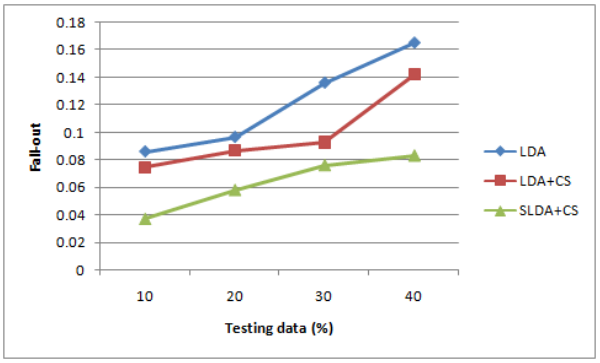
Fig. 4. Illustration of DCG curve of the proposed case retrieval method and the existing methods, such as LDA and LDA+CS

b) Fall-out

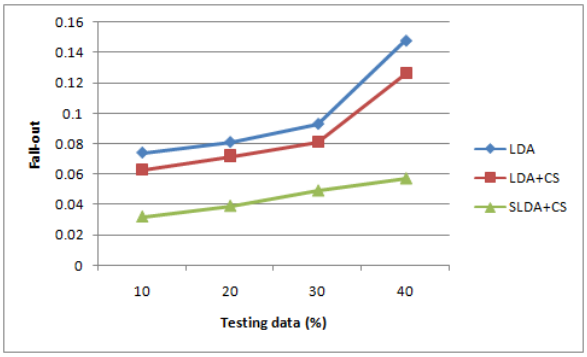
Fig. 5 shows the fall-out of the proposed method and the existing methods, such as LDA and LDA+CS for the breast cancer and breast cancer wins data set for the different size of the testing data. Fig. 5(a) shows the fall-out of the proposed method and the existing methods for breast cancer data set. For 10% of the testing data, the existing methods, such as LDA and LDA+CS has the fall-out of 0.086 and 0.075 respectively, on the other hand, the proposed method has the fall-out of 0.037. When the size of the testing data is 20%, the fall-out of the proposed method is 0.058 while the fall-out of the existing methods, such as LDA and LDA+CS is 0.0965 and 0.087 respectively. Similarly, the proposed method has the minimum fall-out than the existing methods for 30% and 40% of the testing data.

Fig. 5(b) shows the fall-out of the proposed method and the existing methods for breast cancer wins data set. For 10% of the testing data, the existing methods, such as LDA and LDA+CS has the fall-out of 0.074 and 0.063 respectively, on the other hand, the proposed method has the fall-out of 0.032. When the size of the testing data is 20%, the fall-out of the proposed method is 0.039 while

the fall-out of the existing methods, such as LDA and LDA+CS is 0.081 and 0.072 respectively. For 30% of the testing data, the fall-out of the proposed method is 0.049 while the fall-out of the existing methods, such as LDA and LDA+CS is 0.093 and 0.081 respectively. Similarly, the proposed method has the minimum fall-out than the existing methods for 40% of the testing data. From Fig. 5, it can be shown that the proposed case retrieval method has the minimum fall-out than the existing methods, such as LDA and LDA+CS for both breast cancer and breast cancer wins data set.



a) Breast cancer



b) Breast cancer wins

Fig. 5. Illustration of fall-out of the proposed case retrieval method and the existing methods, such as LDA and LDA+CS

6. Conclusion

This paper presents the case retrieval method for extracting the similar cases from the collection of patient's reports. Initially, the patient's details are gathered and divided into 100 reports. Then, the stop words and delimiters are removed from these reports, and each report is applied to the SLDA model. The SLDA model generates the topic keywords for the reports, and the patient-topic matrix is constructed. Then, the query report is sent to the proposed system, the stop words removal and the stemming process is performed. Then, the keywords in the query document are matched with the retrieved topic keywords, the most similar topic is extracted, and the corresponding patient case is retrieved from the database. The matching process is performed by the cosine similarity. The performance of the proposed method is analyzed with the existing methods, such as LDA and LDA+CS for the performance measures DCG and fall-out. The experimental results show that the proposed method attains the higher DCG of 20.154 and the minimum fall-out 0.032 when compared to the existing methods.

References

[1] Rainer Schmidt and Lothar Gierl, "Case-based Reasoning for Medical Knowledge-based Systems," Health Technology and Informatics, 2000.

[2] Yong-Bin Kang, Shonali Krishnaswamy, and Arkady Zaslavsky, "A Retrieval Strategy for Case-Based Reasoning Using Similarity and Association Knowledge," IEEE Transactions on Cybernetics, vol. 44, no. 4, pp. 473-487, April 2014.

[3] André Mourão, Flávio Martins, and João Magalhães, "Multimodal medical information retrieval with unsupervised rank fusion," Computerized Medical Imaging and Graphics, vol. 39, pp. 35-45, January 2015.

[4] Israel Alonso and David Contreras, "Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach," Expert Systems with Applications, vol. 44, pp. 386-399, February 2016.

[5] Yanshan Wang, Stephen Wu, Dingcheng Li, Saeed Mehrabi, Hongfang Liu, "A Part-Of-Speech Term Weighting Scheme for Biomedical Information Retrieval," Journal of Biomedical Informatics, vol. 63, pp. 379-389, October 2016.

[6] Wiwin Suwarningsih, Iping Supriana, and Ayu Purwanti, "Indonesian Medical Retrieval Case Based on Knowledge Association Rule Similarity," In Proceedings of the IEEE International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), pp.142-147, 2015.

[7] Parikshit Sondhi, Jimeng Sun, ChengXiang Zhai, Robert Sorrentino, Martin S. Kohn, Shahram Ebadollahi, and Yanen Li," Medical Case-based Retrieval by Leveraging Medical Ontology and Physician Feedback: UIUC-IBM at Image CLEF 2010," In Proceedings of the CEUR Workshop, vol. 1176, 2000.

[8] Corey W. Arnold, Suzie M. El-Saden, Alex A.T. Bui, and Ricky Taira, "Clinical Case-based Retrieval Using Latent Topic Analysis," In Proceedings of the Annual Symposium, AMIA, pp. 26-30, 2010.

[9] Sunil Bhutada, V. V. S. S. S. Balaram, and Vishnu Vardhan Bulusu, "Semantic latent dirichlet allocation for automatic topic extraction," Journal of Information and Optimization Sciences, vol. 37, no. 3, pp. 449-469, 2016.

[10] Spyropoulos, "On-line Educational Means supporting the tutoring of the Physical Principles applied in modern Biomedical Equipment Technology," Centennial Meeting of the American Physical Society, Atlanta, 1999.

[11] R. Lopez De Mantaras, D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw, B. Faltings, M. L. Maher, M. T. Cox, K. Forbus, M. Keane, A. Aamodt, and I. Watson, "Retrieval, reuse, revision and retention in case-based reasoning," Knowledge Engineering Review, vol. 20, no. 3, pp. 215–240, 2005.

[12] Y. Guo, J. Hu, and Y. Peng, "Research on CBR system based on data mining," Applied Soft Computing, vol. 11, no. 8, pp. 5006–5014, 2011.
[13] Datasets from UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.
[14] K. S. S. Rao Yarrapragada and B. Bala Krishna, "Impact of tamanu oil-diesel blend on combustion, performance and emissions of diesel engine and its prediction methodology", Journal of the Brazilian Society of Mechanical Sciences and Engineering, pp. 1-15, 2016.
[15] Kavita Bhatnagar and S. C. Gupta, "Investigating and Modeling the Effect of Laser Intensity and Nonlinear Regime of the Fiber on the Optical Link", Journal of Optical Communications, 2016.
[16] K Bhatnagar and SC Gupta, "Extending the Neural Model to Study the Impact of Effective Area of Optical Fiber on Laser Intensity", International Journal of Intelligent Engineering and Systems, vol.10, 2017.
[17] B.S. Sunil Kumar, A.S. Manjunath, S. Christopher, "Improved entropy encoding for high efficient video coding standard", Alexandria Engineering Journal, In press, corrected proof, November 2016.
[18] S Chander, P Vijaya, P Dhyani,"Fractional lion algorithm– an optimization algorithm for data clustering", Journal of computer science, 2016.
[19] P. Vijaya, Satish Chander,"Fuzzy Integrated Extended Nearest Neighbour Classification Algorithm for Web Page Retrieval", Proceeding of the International Conclave on Innovations in Engineering and Management, 2016.
[20] P Yadav and RP Singh, "An Ontology-Based Intelligent Information Retrieval Method For Document Retrieval", International Journal of Engineering Science, 2012.