Indian Journal of Science and Technology, Vol 9(25), DOI: 10.17485/ijst/2016/v9i25/86655, July 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Study on Some Tasks, Corpus and Resources of Medical Information Retrieval

P. Gayathri* and N. Jaisankar

School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India;
pgayathri@vit.ac.in, njaisankar@vit.ac.in

Abstract

Background/Objectives: This paper gives an overview of some tasks involved in the retrieval process, corpus and resources of medical information retrieval. **Methods/Statistical Analysis**: Inverted file representation method is used in the retrieval process for associating documents in the corpus with various search terms. Conventional statistical ranking functions such as Jaccard, Okapi and Euclidean have been widely used for ranking retrieved medical documents. An extractive informative generic mono-lingual single-document summarizer is used to produce medical domain-specific summary. Sentence ranking method is used to include most appropriate sentences in the final summary. **Findings:** Studies reveal that people are searching the web and read medical related information in order to be informed about their health. In the medical domain, richest and most used source of information is MEDLINE. Because of frequent use of acronyms in the medical literature, using the term that appears in documents as keywords for document indexing would not be effective. Also, using Bag of Words representation could not capture the semantic meaning of terms. Some domain-specific thesauri like UMLS, MeSH and Gene ontology are available for biomedical retrieval. These domain-specific thesauri can provide synonyms, hypernyms and hyponyms of a specific term but it does not look into the context. Therefore, the retrieval results of using domain-specific thesauri are somewhat conflicting. It is possible to identify which lexical variant of specific term should be used under specific context by using Wikipedia as resource for biomedical retrieval. Conventional ranking functions fail to capture the inherent features of natural language text. Evolutionary algorithm based ranking can enhance the retrieval performance. Any domain-specific summarizer must consider similarity between sentences as essential feature for summarization. **Applications/Improvements:** Improvements in retrieval results is achieved by using context-aware keywords as indexing keywords and highly robust hybrid evolutionary algorithm based ranking function for ordering the retrieved documents.

Keywords: Information Retrieval, Medical Information retrieval, Medical Document Corpus, Resources, Retrieval Process

1. Introduction

The foremost intention of retrieval system is to examine the documents that are appropriate to the information requirement of the user from outsized document group[1]. Every user has specific information requirement. User's requirement will be the phrase to which answer is essential to execute certain task. The phrase given by the user has to be converted into the form appropriate for retrieval system. The information requirement of the user transformed into the form suitable for retrieval system is called query. Based on the user query, the retrieval of documents will take place from large document collection. The retrieved documents are ordered and presented to the user as an answer list. The general representation of the retrieval system is shown in Figure 1.

In this retrieval process, the performance of the system depends on how far the retrieved documents are relevant to the information need of the user. Information Retrieval (IR) has been widely used in applications related to medical domain especially by the medical experts to practice Evidence-Based Medicine (EBM)[2]. It is also enormously used in agricultural, spatial, scientific, indus-

*Author for correspondence

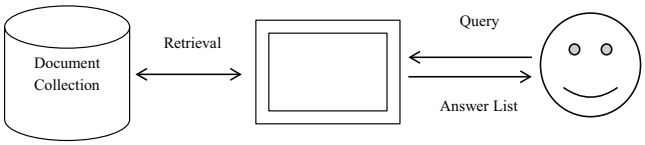

Figure 1. Information retrieval system representation.

trial, governmental, engineering, business and in many other applications.

2. Medical Information Retrieval (MIR)

MIR system is totally centered around and connected to medical domain[3]. Figure 2 depicts the general problems that exist in MIR.

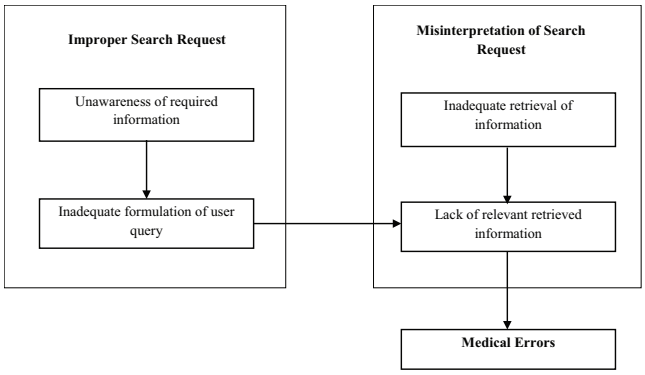

Figure 2. General problems.

The general problems that lead to poor performance of MIR are as follows:

I. Improper search request by the user:

The factors leading to improper search request are as follows:

I. Unawareness of required information

II. Inadequate formulation of user query

Due to lack of knowledge about the different keywords, they tend to input the symptoms, reactions instead of the disease itself and long lines of information need which lead to irrelevant information retrieval.

I. Misinterpretation of search request:

Due to the misinterpretation of search request and huge availability of data, the retrieval system faces the following problems:

I. Inadequate retrieval of information

II. Lack of relevant retrieved information

III. Medical errors:

Due to insufficient relevant data, medical errors may occur. For example, it may lead to wrong disease diagnosis, treatment procedure and so on.

3. Major Tasks Involved in Retrieval Process

3.1 Keyword Extraction

The keyword extraction is an essential task involved in the retrieval process which determines the performance of the retrieval system. The process of recognizing the most suitable terms in the document is called keyword extraction. A good IR system must have a collection of beneficial keywords of corresponding field[4].

3.2 Document Indexing

Indexing is the process of associating documents in the corpus with various search terms (keywords). It shrinks the time spent in encumbered information and also creates the internal representation for documents and keywords.

This internal representation can have set of useful weighted keywords for each document as given below:

$Doc_1 \rightarrow \{(term_1, wt_1), (term_2, wt_2), \ldots..\}$
$Doc_2 \rightarrow \{(term_1, wt_1), (term_2, wt_2), \ldots\ldots\}$
$\ldots..$
$\ldots..$
$\ldots..$
$Doc_n \rightarrow \{(term_1, wt_1), (term_2, wt_2), \ldots\ldots\}$

Where,

'Doc' represents documents, 'term' represents keywords, 'wt' represents its weight and 'n' represents the total number of documents.

Example:

$Doc_1 \rightarrow$ {(heart, 0.2), (infection, 0.3), …}
$Doc_2 \rightarrow$ {(kidney, 0.6), (blood, 0.4), …}
$Doc_3 \rightarrow$ {(heart, 0.4), (kidney, 0.3), …}
$Doc_4 \rightarrow$ {(heart, 0.5), (kidney, 0.2), (blood, 0.7), …}

The internal representation can also be in the form of inverted file. This representation is used during the retrieval process for achieving better efficiency. For each

keyword, it contains references to documents as given below:

$term_1 \rightarrow \{(Doc_1, wt_1), (Doc_2, wt_2), \ldots\}$
$term_2 \rightarrow \{(Doc_1, wt_1), (Doc_2, wt_2), \ldots\}$
.....
.....
.....
$term_n \rightarrow \{(Doc_1, wt_1), (Doc_2, wt_2), \ldots\}$

Here, 'n' represents the number of keywords.
Example:

heart → {(Doc_1,0.2), (Doc_3,0.4), (Doc_4,0.5), …}
kidney → {(Doc_2,0.6), (Doc_3,0.3), (Doc_4,0.2), …}
infection → {(Doc_1,0.3), …}
blood→ {(Doc_2,0.4), (Doc_4,0.7), …}

In the above examples, Doc_1, Doc_2, Doc_3 and Doc_4 represent the documents. The keywords are 'heart', 'infection', 'kidney', 'blood'. Each keyword has its own weight with respect to the document.

3.3 Document Ranking

Document ranking deals with ordering of all retrieved documents based on their relevance score to user query. In response to user query, ranked list of documents are returned by IR system. Ranking function is used to compute the relevance score between retrieved documents and user query. It is done by matching the keywords in the user query with those in the retrieved documents[5]. Each retrieved document is scored and ranked based on how well it matches with the user query.

3.4 Document Summarization

Document summarization is an IR task that deals with constructing lessened form of the document[6]. The summary made must safeguard the information content of the document and its meaning[7]. The summarization method can be categorized based on the nature of text obtained, kind of detail delivered, content offered, quantity of input documents and linguistic[8,9]. There are two types of summarizers under each category which is depicted in Table 1.

The summary can be genre-specific or domain-independent. In the former, documents belonging to one specific domain are considered and in the latter, documents belonging to any domain can be considered for summarization. These summarization approaches comprehends the bits of knowledge of information in the document, incase if a document does not contain the author written summary[10,11].

Table 1. Various categories and types of summarization methods

Category	Type	Description
Nature of text obtained	Extractive	Summary is formed by mining key sentences from original document.
	Abstractive	Summary is created by rephrasing the sentences in the original document. It is done by understanding the entire document content.
Kind of detail delivered	Indicative	Summary provides main idea of the original document.
	Informative	Summary is produced by minimizing the length of the original document without changing its meaning.
Kind of content offered	Generic	Summary is not based on the user interest. It gives the same level of importance to all sentences when producing a summary.
	Query-based	When producing a summary based on user interest or query, it gives importance to certain sentences.
Quantity of input documents	Single-document	Summary is produced for one document at a time.
	Multi-document	Summary is produced for collection of related documents.
Linguistic	Mono-lingual	Summary is created for documents written in one specific language.
	Multi-lingual	Summary is created for documents written in different language.

4. Medical Document Corpus for IR Experiments

The richest and most used source of information in medical domain is MEDLINE. It is the database of life science related articles. The process of identifying and disseminating relevant reliable information becomes a very difficult

task as all research discoveries come and enter this repository at very high rate. OHSUMED and PMC (PubMed Central) are the two subsets of MEDLINE that are commonly used for IR system performance evaluation[12].

OHSUMED is the standard corpus used for benchmarking IR systems evaluation which consists of collection of MEDLINE document abstracts from 270 medical journals from 1987 to 1991. It is the standard corpus for abstract indexing experiment.

PMC is the standard full document corpus used for benchmarking IR systems evaluation which consists of collection of MEDLINE documents. It is the standard corpus for full document IR indexing experiments.

5. Resources for Semantic based IR

There are various generic and medical domain-specific resources available for semantic based MIR. The free online fact file maintained by outsized quantity of volunteers collaboratively is Wikipedia[13]. It acts as a resource for IR in recent years. It contains inter-linked textual information, manually defined concepts and semantic relations. Therefore, the use of Wikipedia can provide not only facts, but also exact semantic information[14]. The open source software system that is used by researchers and developers to integrate Wikipedia's rich semantics into their own applications is Wikipedia Miner. Ontologies are resources that allow researchers to extract semantic based information[15]. The general ontology developed at Princeton University is WordNet[16]. It is an electronic lexical database which consists of set of synonyms called as Synset that represents one particular concept. Synset formation is based on the synonymy and polysemy. In medical domain, various domain-specific ontologies are available which allows medical professionals and researchers to process bio-medical data from countless sources. Among those, the widely used medical domain-specific thesaurus is UMLS which is established by National Library of Medicine (NLM). UMLS integrates many subdomains related to medical domain as depicted in Figure 3. MetaMap allows researchers and developers to use UMLS resource in their own applications[17].

The medical domain-specific ontology Generalised Architecture for Languages, Encyclopedias and Nomenclatures in medicine (GALEN) is the European Union project undertaken from 1992 – 1999 which

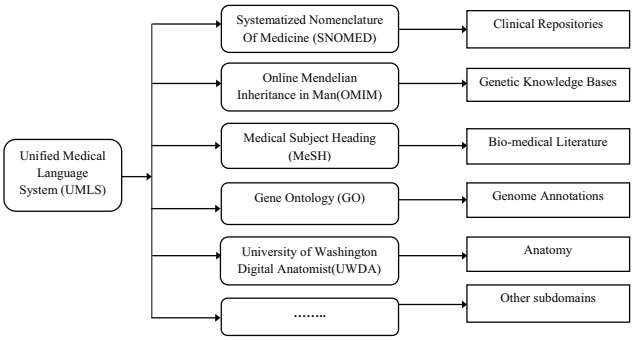

Figure 3. Various subdomains integrated in the UMLS.

provides reusable clinical terminology resources. It was designed before populating the ontology by defining the representation formalism and top level knowledge. OpenGALEN provides access to GALEN resource.

Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is another medical domain-specific dictionary established by the college of American Pathologists. It is the clinical repository which is now available as a part of UMLS. Therefore, it is widely used in medical information system.

The prior works mentioned here within Table 2, Table 3 and Table 4cover few of the existing works related to the indexing keyword extraction, document ranking and summarization respectively.

6. Observations

This section presents the observations made from the related work with respect to various tasks involved in the retrieval process.

6.1 Keyword extraction

MIR system achieves poor performance due to ambiguity in medical terms. Ambiguity means same keyword referring to different things in different context[18].

For example:

CHD may allude to 'Coronary Heart Disease' or 'Congenital Heart Defect'

HR may allude to 'Heart Rate' or 'Hormone Replacement'

LE may allude to 'Lupus Erythematosus' or 'Lower Extremities'

URI may allude to 'Upper Respiratory Infection' or 'Urinary Incontinence'

Also, frequent use of acronyms in medical domain reduces its retrieval performance.

Table 2. Summary of few works on indexing keyword extraction

Proposed Work	Method	Resource	Performance Metric(s)	Comparison with	Corpus
Multi-Terminology based Concept Extraction[18]	Voting	MeSH, SNOMED, GO, ICD-10	Mean Average Precision (MAP)	Baseline	TREC Genomics collections
Automatic MeSH Term Extraction (AMTEx)[12]	C/NC-value	MeSH	Processing Time	MMTx	OHSUMED, PMC
Single-SM, Multiple-SM19	Aspect detection and filtering	Wikipedia, UMLS	MAP	NLMinter, MuMshFd	TREC2006 and 2007 Genomics collections
Bio-DI[20]	Scoring and filtering	MeSH	Precision, Recall, F-measure	Various weight combinations	OHSUMED

Table 3. Summary of few works on ranking of documents

Proposed Work	Method	Features	Performance Metric(s)	Comparison with	Corpus
SimRank + CorRank + VoteRank[11]	Weighting method	Similarity, Correlation, Votes	Normalized Discount Cumulative Gain (NDCG)	SimRank + CorRank, SimRank and VoteRank	Reviews collected from http://www.androidpolice.com/, http://www.phonearena.com/, http://moneycontrol.com/.
Two level fuzzy logic based Ranking[21]	Term weighting method	Term Frequency (TF) and Inverse Document Frequency (IDF)	Precision, Recall and F-measure	Okapi-BM25 and fuzzy logic based approach	CACM and CISI
Fuzzy logic based Ranking[22]	Term weighting method	TF and IDF	MAP	state of the art search engine Apache Lucene	Financial Times, Federal Register 94, FBIS disk 5, LA Times
RankIP[23]	Immune programming	TF	Precision at n (P@n), MAP and NDCG	RankSVM, RankBoost and BM25	OHSUMED, TREC 2003 and 2004
GA1 and GA2[24]	Term weighting method	TF	Precision, Recall	Classic IR	CISI, NPL, CACM

For example:

BSL is an acronym for Blood Sugar Level

CAD is an acronym for Coronary Artery Disease

BP is an acronym for Blood Pressure

HB and HGB is an acronym for Hemoglobin

Likewise, there are many other acronyms and ambiguous terms in the medical field. As a result, using the term that appears in the document as indexing keyword may not be effective. Also, Bag of Words (BoW) representation of document cannot capture the semantic meaning of the document.

It is observed that the domain-specific thesaurus is widely used in the literature. These thesauruses can capture the synonyms, hypernyms and hyponyms of specific medical term. But it does not look into the context[31]. Domain-specific thesaurus based retrieval results are conflicting because it is very difficult to determine the lexical variant of specific medical term to be used under specific context. It is also observed that nearly about 50,000 words in UMLS, medical domain-specific thesaurus are ambiguous. For example, *'COLD'* is an ambiguous term in UMLS. It may allude to *'Common Cold'* disease or *'Chronic Obstructive Lung Disorder'* sickness. In this manner, the quality of the indexing keyword extraction relies upon the quality of the resource utilized.

Table 4. Summary of few works on document summarization

Proposed Work	Method	Features	Performance Metric(s)	Evaluation using	Corpus
Collective Message Summarizer (CMS)[25]	Sentence compression	Sentence relevance and redundancy	ROUGE-2 Recall	Manual summary	Enron E-mail collection
Individual Message Summarizer (IMS)[25]	Sentence compression	Sentence relevance and redundancy	ROUGE-2 Recall	Manual summary	Enron E-mail collection
Bernoulli based Sentence Similarity Model[26]	Sentence extraction	Probability of co-occurrence	ROUGE-1 and 2 Recall	Manual summary	DUC datasets
Wikipedia based Single Document Summarizer[27]	Sentence selection	Number of Wikipedia terms	ROUGE-1 Recall	Manual summary	DUC datasets
Fuzzy logic based Summarizer[28]	Sentence feature extraction	Cue-phrases, Legal vocabulary, Paragraph structure, Citation, Term weight, Named entity recognition, Similarity to neighboring sentences, Absolute location, Proper noun, Sentence position	Precision, Recall,F-measure	Neural Network based summary	Legal judgment documents collected from legal websites
Web Document Summarizer[29]	Sentence ranking	Sentence position, Cue-words, Document frequency,TF,IDF	ROUGE-1, 2 and3 Recall	Online summarizer based summary	Documents collected from websites
MA-SingleDocSum[9]	Genetic operators and guided local search	Sentence position, Relation of sentences with title, Sentence length, Cohesion, Coverage	ROUGE measures	UnifiedRank, DE, FEOM, NetSum, CRF, QCS, SVM, Manifold Ranking	DUC datasets
Machine Learning based Medical document Summarizer[30]	Bagging	Centroid, Similarity to first sentence, Sentence position, Cue-phrases, Position of cue-phrases, Acronyms, Sentence length	Precision, Recall, ROUGE-1,2 andSU Recall	MEAD, Baseline-lead, Manual summary	Medical news articles from online sources

6.2 Ranking of Retrieved Documents

There are many factors which affect the performance effectiveness of IR system, ranking function being the most. IR system uses ranking function to compute relevance score between retrieved documents and user query.

The traditional conventional ranking functions like Okapi-BM25, Euclidean and Cosine similarity are widely used in the literature[1]. These conventional measures do not capture the inherent features of the documents and user query due to uncertainty and vagueness present in the natural language text. Some researchers have designed evolutionary algorithm based ranking functions for enhancing the performance of the retrieval system.

These evolutionary algorithm based ranking functions except fuzzy logic based functions do not address upon the vagueness and uncertainty present in natural language. Fuzzy logic based ranking is found suitable in the literature to address upon the vagueness and uncertainty present in natural language text[21]. Fuzzy logic transforms the vagueness and uncertainty into fuzzy membership functions[32]. It is observed from the literature that these fuzzy based ranking functions are incapable to generalize i.e. they can answer to what is written in the rule base. Also, they are not robust enough to the topological changes of the system. Such changes need alterations in the rule base.

6.3 Summarization of Documents

Summarization systems lose their completeness due to similar sentences in the input document. The system that exploits domain-specific knowledge can produce high quality summary by considering domain's generic components for summarization. The system can produce highly informative summary if and only if it includes highly dissimilar sentences in the final summary. It is seen from the literature that all articles related to medical domain don't have author written summary. This builds the trouble for a user to confirm its significance. The user can characterize whether the retrieved document is relevant for the detailed study or not only after reading and understanding the whole document content. It is also observed that most of the summarization approaches create final summary by using sentence ranking method. Researchers have used few or more number of features for sentence ranking.

7. Conclusion

The accuracy of the retrieval system mainly depends on indexing keyword extraction. Context-aware keyword extraction can improve the retrieval results. Document ranking is another prominent factor that affects the retrieval performance. As conventional ranking functions do not capture the inherent features of the documents and user query, highly robust hybrid evolutionary algorithm based ranking functions can be used to improve the retrieval performance. Sentence ranking method is widely used in the literature for summarization. Any domain-specific summarizer must consider similarity between sentences as additional feature for summarization along with domain's other generic features. OHSUMED and PMC are the standard document corpuses used in the literature for MIR system evaluation. How to use Wikipedia to facilitate information retrieval became hot research topic over the last few years. As domain-specific resources are available for MIR, there is no much work done on investigating how to use Wikipedia to improve MIR performance. To shed light on the breadth of research on MIR, current state of art approaches and observations made from literature has been presented.

8. References

1. Singhal A. Modern information retrieval: A brief overview. IEEE Data Engineering Buletin. 2001; 24(4):35–43.
2. Frunza O, Inkpen D, Tran T. A machine learning approach for identifying disease-treatment relations in short texts. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(6):801–14.
3. Swartz K. Health care for the poor: for whom, what care, and whose responsibility? Focus. 2009; 26(2):69–74.
4. Coursey KH, Mihalcea R, Moen WE. Automatic keyword extraction for learning object repositories. Proceedings of the American Society for Information Science and Technology. 2008; 45(1):1–10.
5. Chou S, Chang W, Cheng CY, Jehng JC, Chang C. An information retrieval system for medical records and documents. 30th Annual International Conference of the Engineering in Medicine and Biology Society, Vancouver, British Columbia, Canada. 2008. p. 1474–77.
6. Gupta V, Lehal GS. A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence. 2010; 2(3):258–68.
7. Pimpalshende AN. Overview of text summarization extractive techniques. International Journal of Advanced Technology in Modern Engineering. 2015; 2(12):1–10.
8. Munot N, Govilkar SS. Comparative study of text summarization methods. International Journal of Computer Applications. 2014; 102(12):33–7.
9. Mendoza M, Bonilla S, Noguera C, Cobos C, Leon E. Extractive single-document summarization based on genetic operators and guided local search. Expert Systems with Applications. 2014; 41(9):4158–69.
10. Prakash S, Chakravarthy TC, Brindha GR. Preference based quantified summarization of on-line reviews. Indian Journal of Science and Technology. 2014; 7(11):1788–97.
11. Meghana Ramya Shri J, Subramaniyaswamy V. An effective approach to rank reviews based on relevance by weighting method. Indian Journal of Science and Technology. 2015; 8(11):1–7.
12. Hliaoutakis A, Zervanou K, Petrakis EG. The AMTEx approach in the medical document indexing and retrieval application. Data and Knowledge Engineering. 2009; 68(3):380–92.
13. Paci G, Pedrazzi G, Turra R. Wikipedia-based approach for linking ontology concepts to their realisations in text. International Conference on Language Resources and Evaluation. 2010. p. 33–8.
14. Milne D, Witten IH. An open-source toolkit for mining Wikipedia. Artificial Intelligence. 2013; 194:222–39.

15. Uschold M, Gruninger M, Ontologies: principles, methods and applications. The Knowledge Engineering Review. 1996; 11(2):93–36.
16. Shubhangi CT. An approach to single document text summarization and simplification. IOSR Journal of Computer Engineering. 2014; 16(3):42–9.
17. Milian K, Aleksovski Z, Vdovjak R, Teije AT, Harmelen FV. Identifying disease-centric subdomains in very large medical ontologies: a case-study on breast cancer concepts in SNOMED CT or finding 2500 out of 300.000, Knowledge Representation for Health-Care Data, Processes and Guidelines, Springer-Verlag: Berlin Heidelberg, 2010; 50–63.
18. Dinh D, Tamine L, Boubekeur F. Factors affecting the effectiveness of biomedical document indexing and retrieval based on terminologies. Artificial Intelligence in Medicine. 2013; 57(2):155–67.
19. Yin X, Huang JX, Li Z, Zhou X. A survival modeling approach to biomedical search result diversification using Wikipedia. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(6):1201–12.
20. Chebil W, Soualmia LF, Darmoni SJ. BioDI: a new approach to improve biomedical documents indexing. Database and Expert Systems Applications, Springer-Verlag: Berlin Heidelberg. 2013; 78–87.
21. Gupta Y, Saini A, Saxena AK. A new fuzzy logic based ranking function for efficient information retrieval system. Expert Systems with Applications. 2015; 42(3):1223–34.
22. Rubens N. The application of fuzzy logic to the construction of the ranking function of information retrieval systems. Computer Modelling and New Technologies. 2006; 10(1):20–7.
23. Wang S, Ma J, He Q. An immune programming-based ranking function discovery approach for effective information retrieval. Expert Systems with Applications. 2010; 37(8):5863–71.
24. Radwan AAA, Abdel Latef BA, Ali AA, Mgeid A, et al. Using genetic algorithm to improve information retrieval systems. World Academy of Science and Engineering Technology. 2006; 17(2):6–12.
25. Zajic DM, Dorr BJ, Lin J. Single-document and multi-document summarization techniques for email threads using sentence compression. Information Processing and Management. 2008; 44(4):1600–10.
26. Goyal P, Behera L, McGinnity TM. A context-based word indexing model for document summarization. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(8):1693–705.
27. Ramanathan K, Sankarasubramaniam Y, Mathur N, Gupta A, Document summarization using Wikipedia. Proceedings of the First International Conference on Intelligent Human Computer Interaction, Allahabad, India. 2009. p. 254–60.
28. Santhana Megala S, Kavitha A, Marimuthu A. Enriching text summarization using fuzzy logic. International Journal of Computer Science and Information Technologies. 2014; 5(1):863–67.
29. Reddy YS, Siva Kumar DA. An efficient approach for web document summarization by sentence ranking. International Journal of Advanced Research in Computer Science and Software Engineering. 2012; 2(7):221–25.
30. Sarkar K, Nasipuri M, Ghose S. Using machine learning for medical document summarization. International Journal of Database Theory and Application. 2011; 4(1):31–48.
31. Hliaoutakis A, Zervanou K, Petrakis EGM, Milios EE. Automatic document indexing in large medical collections. Proceedings of the International Workshop on Healthcare Information and Knowledge Management. 2006; 1–8.
32. Zadeh LA. Fuzzy sets. Information and Control. 1965; 8(3):338–53.