DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5505

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A SURVEY PAPER ON INFORMATION RETRIEVAL SYSTEM

Arpit Deo
Research Scholar
IES IPS Academy
Indore, India

Jayesh Gangrade
Associate Professor
IES IPS Academy
Indore, India

Shweta Gangrade
Assistant Professor
IES IPS Academy
Indore, India

Abstract: Information retrieval is the process of obtaining and presenting more related information from the largest collection of information resources according to the user's need. The tremendous growth in information resources on the Internet makes the information retrieval process a tedious and difficult task for users. Due to information overloading, there is a need for better techniques to retrieve most relevant information from web. This paper presents the information retrieval system by using the PSO algorithm. In presented system, to extract the text from web documents, all html tags are removed. After that stop words and special characters are removed from extracted text for recovering only meaningful contents. TF-IDF concept is used for feature selection. Now PSO optimization technique is used for identifying and refining the features set, these selected features are stored in a database which is used for information retrieval process. In other hand input query is converted into more than one similar semantic query strings. These query strings are compared with the obtained feature sets in the database by using the cosine similarity function. The most similar text is retrieved as an outcome of the information retrieval system.

Keywords: Information retrieval system; feature extraction; PSO optimization; similar query generator; similarity measure

I. INTRODUCTION

Today internet has turned to be the largest information sources. The World Wide Web is the collection of many interlinked hypertext documents. It provides the huge amount of information which is accessed via Internet by using hypertext transfer protocol. The web provides many types of informative data, such as text, images, videos and other multimedia data. The tremendous growth of information resources makes the information retrieval a difficult and tedious task for users. Because of that reason, user can't be able to access relevant information effectively [1].

Information retrieval plays a vital role in web search engines to access most relevant information according to the user's input query. It is a mainstream and the basics of web search engines.

Information retrieval is the process of obtaining and presenting more related information from the largest collection of information resources according to the user's input query. Whenever a user needs to access the information, it is necessary to enter a formal statement into a search engine. This formal statement, also known as a search engine's input query. A query does not obtain and present a single information resource in the largest collection of information resources. Instead, several information resources are presented those are matched by input query. Most relative to least relative information resources will be shown to the user [2].

Web search engines such as Bing, Yahoo, Google, Excite, AltaVista etc. are used by millions of users to access information across the world on any topic.

Information retrieval system is used in many application areas such as digital libraries, information filtering, recommendation system, media search, image retrieval etc. [3].

A. PSO Algorithm

PSO is an evolutionary computation method that inspired from the simulation of social behavior. It is based on birds flocking. It optimizes the population-based problems by iterative computation. It computes the initial population, which is random solutions of the problem and then provides the improved candidate results. It is also known as particles.

The algorithm initialized by potential solutions of a population-based optimization problem, each and every potential solution (particle) has randomized velocity.

Let S_w be the size of the swarm. Each particles i_k is initialized with random position P_k and velocity Vk. Fk is threshold objective function. It takes positional coordinates of particles as input. All particles are associated with best results called p_{best}, in the problem space. The global best value is represented by g_{best}. In every iteration p_{best} location and velocity of each particle is changed and also function is evaluated with changed positions and velocities.

Following steps of PSO algorithm:

1) Initialization:

a) initialize a population by potential solution with random position and velocities.

b) Evaluate the fitness of each population.

c) Stored the personal best position of each population in memory.

d) Choose the global best position.
e) update the iteration number.

2) Velocity updating: the velocities of all particles are updated in each iteration.
3) Position updating: position of all particles are updated in each iteration.
4) Memory updating: update the value of p_{best} and $g_{best.}$
5) Termination criteria evaluation: iterations are repeated until good result are not evaluated or maximum number of iterations reached [4].

B. Web page indexing

Indexing is a procedure used to give quick and precise data recovery from substantial gathering of data assets. Page indexing is the procedure utilized as a part of web crawlers to discover website pages on the Internet. It upgrades the term of seeking and increment the execution of data recovery framework. The web crawler will filter each record in the corpus, when ordering isn't utilized. It would require much time and all the more figuring power. For instance, a record of 2000 archives can be questioned rapidly while a consecutive output of each of the 2000 of reports could take much time. It is a very easy way of indexing, but this way of indexing can't provide accurate and good quality results by information retrieval system. To provide accurate and good quality results, it is necessary to parse the full document. This full document indexing provides better results by information retrieval system [5].

C. Web page ranking

The size of World Wide Web is growing exponentially and most of the peoples access the WWW to get information. Users and their queries are increasing to the web search engines. Therefore, it is necessary to process these queries properly and accurately by search engines to provide relevant information or documents. Thus, some web mining technique must be employed in order to extract only relevant documents from the database and provide the intended information to the users.

Page ranking algorithms are used to order the web pages in a proper and arranged manner. It uses web mining techniques to organize the web pages according to their importance, relativeness and content score etc. Web pages are presented to the user by their page rank score. Page rank of any web page is calculated by using backlinks, forward links, topic sensitivity etc. Some ranking algorithms use only links, some use only content of documents or some use both links and contents of documents to assign a rank value to the web page.

This algorithm is used by search engines to provide better quality of search results by ranking the web pages. [6].

D. Text feature extraction

Text data are non-structured or semi- structured, in order to convert these data into feature vector (one structured form), we need moderate-sized features collection. So that the text feature selection technique like TFIDF is used which select the least features that denotes the most information.

The former means that words are only considered as features, if they occur at least once in the training data. In feature selection process, non-content words are eliminated such as: the, and, or, for, etc. Additionally, the names of places and people are also eliminated. Then we could reduce some unnecessary repetitive operations.

To select a subset of n features, the n words with the highest TFIDF are chosen. By ordering these scores, we set up a threshold for getting expected information from a document and guarantee the percentage of information.

The term frequency-inverse document frequency is abbreviated by TF-IDF. It is a numeric measure that is utilized to assess the significance of a word in a report in view of how regularly did it show up in that record and a given gathering of archives. The instinct for this measure is: The word is essential on the off chance that it shows up as often as possible in a report and that word appointed by a high score. Be that as it may, the word may not be an interesting identifier in the event that the word shows up in an excessive number of different archives, in this manner, that word allocated by a lower score rather than a higher score. The math formula of TF-IDF:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad \text{...(1)}$$

Where t denotes the terms; d denotes each document; D denotes the collection of documents [7].

II. LITERATURE REVIEW

A. A. A. Radwan et al. [8] proposed a fitness function This function is used for information retrieval system. It worked very speedy and in a very flexible way. Proposed information retrieval system, where genetic algorithms used for document retrieval. A proposed IR system used for document collections which retrieve more relevant documents and presented it to the user. **P. Simon [9]** proposed a two-stage method, which is based on genetic algorithm. This proposed strategy is utilized for data recovery framework. Hereditary Algorithm utilized as a part of data recovery framework to create best blend terms from an arrangement of the record watchwords. This produced mix terms are utilized as watchwords. These catchphrases are utilized for recovering reports in an accumulation of archives. **M. Kc et al. [10]** proposed strategies for quality assessment and appraisal. It is utilized to survey the nature of site pages. An arrangement of value criteria utilized for quality assessment. Directed client's overview utilized for extraction of value criteria. Proposed the weighted algorithmic understanding of the most significant client cited quality criteria. Which used machine learning procedures to process an expectation of value for pages before downloaded. **H. C. Yang and C. H. Lee [11]** proposed a machine learning technique. This technique instinctively established a navigational map for the world wide web and it used in information searching. The web pages are mapped by self-organization map. Two feature maps are used to maintain the relationship between web pages and thematic keywords. Then they used these maps to develop a structure that may assist the users to finding the information they needed. **J. D. Rose et al. [12]** proposed a genetic algorithm based novel approach for information retrieval system to provide the web pages effectively and accurately. In the proposed approach, the inserted query is initially pre-processed and set of similar word is generated. It used the word net tool for semantic keyword set generation. The user has to select a word from these keywords set or select the pre-processed word. Then the weightage of that particular word is calculated in all the web pages. The webpage with highest weightage showed top most and then sequentially web pages showed in order to highest to lowest weightage.

III. PROPOSED WORK

In this proposed information retrieval system, it initially accepts the web pages as the input dataset for information processing and retrieval system. In the first phase the HTML tags are removed from the web documents and the text is extracted from documents. In next the text content is pre-processed for recovering only meaningful contents. In this step firstly stop words are removed and then special characters from the data are removed.

The pre-processed data is then used with next step where the feature selection is performed using TF-IDF concept. Now for identifying and refining the features set extracted from raw document is used to the PSO optimization technique. The selected features of the documents are treated as a feature set for information retrieval. Which is stored in database, i.e., illustrates in figure 1.

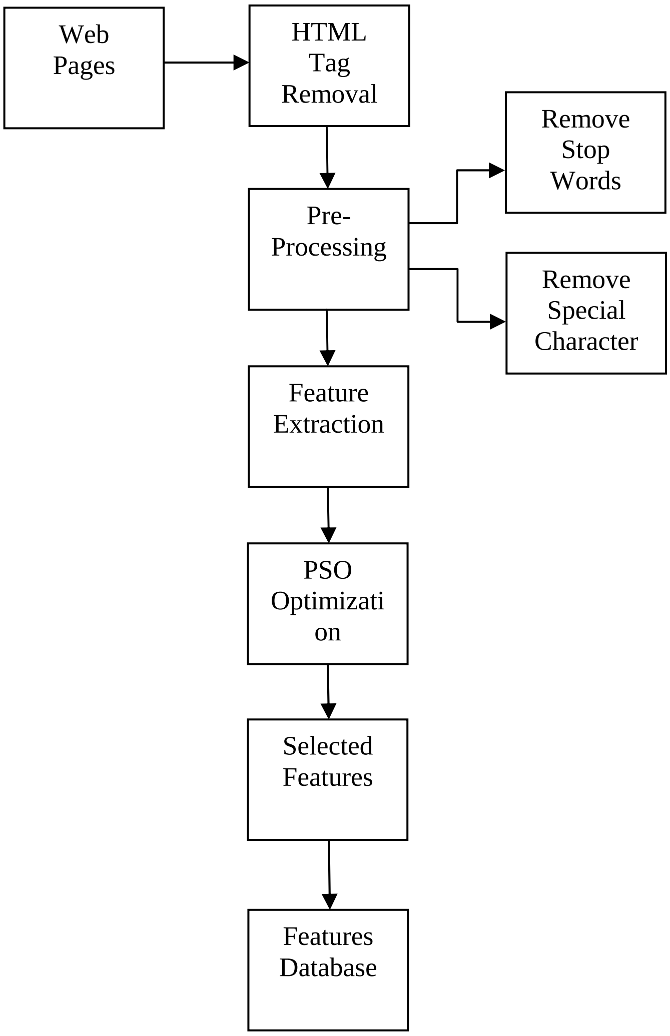

Figure 1.

On the other side in figure 2, user pass an input query for recovering information from a data source. That query is converted into more than one similar semantic query strings. These query strings are compared with the obtained feature sets in the database using the cosine similarity function. The most similar text is retrieved as an outcome of the information retrieval system.

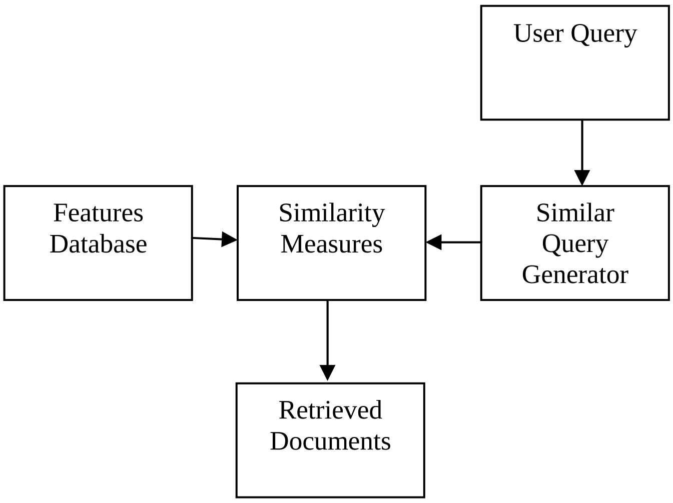

Figure 2.

IV. CONCLUSION

Information retrieval is an essential factor for all the people in the world, searching the content on the web if it comes to our relevant content means we can express our self for that information. Web pages are rich sources of information. Part of this information and features come through the link structure of the web, and the other part comes from the web content. However, since there is no specific discipline in web development, these features can also create profound challenges as web data are mainly semi-structured, unstructured and different structured data.

The information is need of current age human, therefore a number of information extraction and retrieval applications are developed recently that supports the current age information needs. During the implementation of different kinds of data search and retrieval techniques a number of methods for structured and similarly unstructured information processing is developed recently. Among most of the work is devoted to the structured data processing, but their limited efforts are made for retrieving information from the unstructured data sources. The traditional unstructured data processing techniques either not much efficient, or not accurate for adopting and using in real world application therefore a new technique is required to investigate and develop by which the user query relevancy and performance are both improved.

This research work includes the detailed description about the information retrieval system which is based on PSO. That information retrieval system provides the most relevant information on the basis of a user's input query.

V. REFERENCES

[1] V. L. Praba and T. Vasantha, "Evaluation of Web Searching Method Using a Novel WPRR Algorithm for Two Different

Case Studies", ICTACT Journal on Soft Computing, APRIL 2012, Volume: 02, Issue: 03, pp. 341-347.

[2] http://en.wikipedia.org/wiki/information_retrieval

[3] C.S. Naga Manjula Rani, "Importance of Information Retrieval", Oriental Journal of Computer Science& Technology, vol. 4, no. 2, pp. 459-462, 2011.

[4] K. Ammulu, T. Venugopal "Mining Web Data using PSO Algorithm", IJIRST –International Journal for Innovative Research in Science & Technology, Volume 4, Issue 2, pp.201-207, July 2017.

[5] https://en.wikipedia.org/wiki/Search_engine_indexing

[6] N. Duhan, A. K. Sharma, K. K. Bhatia, "Page Ranking Algorithms: A Survey", IEEE International Advance Computing Conference, IEEE, pp.1530-1537, 2009.

[7] L. P. Jing, H. K. Huan, H. B. Shi, "Improved Feature Selection Approach TFIDF in Text Mining", Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, IEEE, pp.944-946,2002.

[8] A. A. Ahmed, Radwan, A. Bahgat A. Latef, A. M. A. Ali, and O. A. Sadek "Using genetic algorithm to improve information retrieval systems", International Journal of Computer, Electrical, Automation, Control and Information Engineering, Volume 2, No. 5, pp. 1544-1550, 2008.

[9] P. Simon, "Two Stage Approach to Document Retrieval using Genetic Algorithm", International Journal of Recent Trends in Engineering, Vol. 1, No. 1, pp. 524-526, May 2009.

[10] M. Kc, M. Hagenbuchner, A. C. Tsoi, "Quality Information Retrieval for the World Wide Web", International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, pp. 655-661, 2008.

[11] H. C. Yang, C. H. Lee, "Mining Unstructured Web Pages to Enhance Web Information Retrieval", International Conference on Innovative Computing, Information and Control, IEEE, Volume 1, 2006.

[12] J. D. Rose, J. Komala, M. Krithiga, "Efficient Webpage Retrieval Using WEGA", Procedia Computer Science, Volume 87, pp.281 – 287, 2016.