Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 566 – 573

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Concept Networks for Personalized Web Search Using Genetic Algorithm

K R Remesh Babu[a,]*, Philip Samuel[b]

[a]*Government Engineering College, Idukki - 685603, Kerala, India*
[b]*Division of IT, Cochin University of Science And Technology, Kochi - 682022, India*

Abstract

Web search engines gives users an initial point for their information hunt. The problem with the traditional search engine is that it retrieves the same set of web pages for all users even though each user has their own preference for a particular search. For retrieving web pages based on user's preference personalized search is needed. The main drawback of this method is that it cannot provide accurate result when user's preference changes or have to search something new. In the proposed method, a concept network is created to identify users search preferences. This concept network consists of list of related concepts based on the history of users' previous search. This policy helps to retrieve web pages related to the context in which the user needs information. Genetic Algorithm (GA) is used when user searches for something new. GA is used to compare the user's concept network with other user's concept network for similarity, thus helps to get a better search in their area of interest.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Concept Network; Genetic Algorithm; Information Retrieval; Web Search.

1. Introduction

The search engines are used to find information from the internet. As the quantity of web pages is rapidly increasing, it is a cumbersome task to retrieve relevant information related to the topic. The aim of Information

* Corresponding author. Tel.: +91-989-543-5562.
*E-mail address:*remeshbabu@yahoo.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.092

Retrieval (IR) systems is to helps user to properly arrange and stock up such information and retrieval of relevant information upon user requests. Retrieval of information relevant to each user's interests from a huge set of documents is the capability of IR systems. Users are in need of information depending on their area of interest[10]. There are several researchers studied the need to access information and its important benefits in several areas including marketing, socio-economic development, education, and healthcare[3].

Personalised search is the process of searching a set of documents or web pages by giving preference to the user's interest. User interest is identified by from the user profile, which is created by analysing the sites searched by the user. Accuracy of the system is improved as a personalized search which can predict the user's interest better compared to a normal search.

At present, the web search engines are built to serve all types of users, and independent of the particular interests of any individual user. Personalized search incorporates the personal interest of users in the search process, in order to retrieve relevant information required by a particular user. The overview of how personalized information can be gathered is shown in the Fig. 1.

The development in web search techniques grows rapidly; some search engines have already incorporated the personalized search service in their search engine. In the Google's personalized search the user's can specify their categories of interest. Some of the search engine technologies use relevance feedback to get user requirements. In few systems users have to register their demographic information earlier in order to get better service. All these methods require users to engage in extra activities to specify their preferences manually before search. So there is a need for new methods that are capable of implicitly recognizing user's information needs and to present required information. Several research attempts and their proposed solutions are available to provide the relevant information by considering the users situations. The various personalized web search approaches proposed by researchers are described in the section 2.

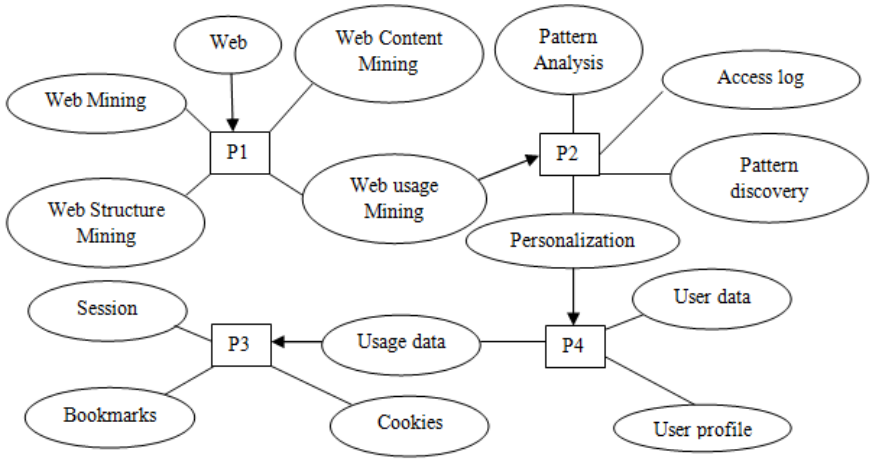

Fig. 1. Personalized Web Retrieval Process

1.1. Genetic Algorithm

Genetic Algorithms have been developed by John Holland[12]. Natural section and natural genetics are the underlying principles of these search algorithms. These probabilistic algorithms, which simulates the natural selection process of living organisms finally comes up with an approximate solution to a problem under consideration after several iterations[12,13]. They combine survival of the fittest offspring's among the string structures with a structured yet randomized information exchange to form a search algorithm. GA based algorithms efficiently and effectively exploits past information to conjecture on new search points with anticipated enhanced solution to the problem under consideration.

Genetic Algorithm starts with a population of possible solution sets. Each solution in the population is represented as chromosome, which is an abstract representation. In the beginning all possible solutions have to be coded into a chromosome. Then a set of reproduction operators has to be fixed. In the next step, it performs mutation and recombination operations on the population of chromosomes, the reproduction operators are applied directly on to the chromosomes. The fitness of every individual chromosome in the population is evaluated from current population based on their fitness value and modified to form a new population in every generation. The algorithm uses this newly created population in the next iteration. Based on the fitness value, each chromosome has an associated value. The fitness value shows the goodness of the solution. The algorithm terminates when either a maximum number of generations have been produced or an adequate fitness level has been arrived at for the population. The limitation of GA is that sometimes it is unable to produce a satisfactory solution with in the maximum number of generations permitted.

1.2. Concept Network

Concept networks are undirected graph in which the vertices represent some concept or a term. The edge weight varies based on the correlation between the concepts or most similar word appeared in the documents. If the edge weight is high between the query and another term then they are correlated and so query can be expanded using the term. The edge weight can be calculated by various methods. In the proposed method the term frequency–inverse document frequency (tf-idf) mechanism used fro calculating the edge weight. This is a numerical statistical method tf–idf, which shows how important a word is to a document in a collection or corpus retrieved by the browser. If the tf-idf value increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to control the fact that some words are generally more common than others[14].

2. Related Works

Personalized Search uses the pages browsed by the user to create a user profile. The user profile can be created based on two factors they are

- Links browsed by the user or click history
- Concepts searched by the user

In order to offer context-aware and personalized information to the users, intelligent searching techniques are necessary. Some Personalized Search Engine creates the user profile based on the links clicked by the user. The literature for comparative analysis and overview of the several web search concepts using web mining techniques are available for designing a more powerful search[1]. The fuzzy neural network approach for enhancing web search is also highlighted in this paper. The user profile contains users click history. When the user searches next time the search engine first searches the links stored in the user's profile and then it searches in other pages. The advantage of this method is that it is easier to implement and it returns the pages frequently accessed by the user, which could have been the page the user was searching for. The limitation of this method is that it cannot adapt when the user's interest changes. Cheqian Chen et al., (2010) was proposed a personalized search based on the concepts searched by the user[3]. In this method, user's positive as well as negative interest is considered while creating user profile. Negative interest denotes those interest that user are not interested in. In other methods only one user profile is created however it is not sufficient to detect user's interest as it can change. In this method Bipartite graph is created based on all user queries and concept extracted. Graph contains two types of nodes namely concept node and query node. In this, two types of clustering are applied to the graph. They are initial clustering and community clustering. Initial clustering clusters concepts and query with respect to one user. Initial clustering is a two step clustering process. In the initial step, similar query nodes are combined. In the next step similar concept nodes are combined. Community clustering is applied to concept node with respect to different users. Community clustering clusters queries and concepts that occur together regardless of the user. Community clustering is used to detect the changes in user's interest. When a user enters query first the community cluster is searched for the query and then the user cluster is used[3].

Personalized search engines based on concept network, creates the user profile based on the concepts browsed by the user. The concepts are extracted from the pages browsed by the user. Kenneth Wai-Ting et al., (2010) was developed a system which stores the concept searched by the user in the user's profile[4]. In this method, learning algorithms like SVM, naive Bayesian Classifier is used to learn the user's search pattern to predict whether a page is relevant to the user's interest. The first step in this method is to store user's click history. The pages are first reordered based on the document's relevance to the query. The relevance of document is calculated based on the term frequency of the word in the page with respect to total word count of the document. The pages are classified by the classifier as positive and negative, where positive denotes the page is related user's interest and negative denotes it is not related to user's interest. The classifier model is built using user's click history, and the words searched by the user. Query expansion is applied to all users and is useful for the new users. The problem with this classifier model is that it can get affected if the underlying search mechanism is not accurate.

Kyung-Joong Kim et al., (2001) proposed a personalized search which uses concept network to store user profile[5]. In this method fuzzy concept network was used to represent user's interest. Fuzzy concept network contains two types of nodes they are concept and document node. The edges between two nodes denote the relevance between the nodes. When a user enters queries the query is searched in the fuzzy network and retrieves those documents that have more relevance with respect to the query. If a concept is not directly connected to any document then we can take those documents that can be reached from the concept with least edge cost in the path as the edge cost between the concept and the document. The edge cost is determined based on factors like term frequency[2].

A fuzzy document retrieval system based on link structures, which search relevant and personalized web pages was proposed by Kyung-Joong Kim, et al., (2007). A personalized retrieval tools for searching educational resources on internet was available for academic use[7]. A personalized electronic news system (PENS) was implemented as the proof of concept network and to demonstrate how web pages are synthesized with different attributes from the same description and to show adaptation based on users' behaviour and client-side characteristics[8]. There are other methods such as agent-based architecture for context-aware and personalized event recommendations were proposed by researchers[9]. In this paper the proposed spreading algorithm learns user patterns by discovering user interests. The Filter Bubble FB effects drives growth in the popularity of alternative search engines that do not personalize results. Concerns about the Filter Bubble effects are hot research area, driving growth in the popularity of alternative search engines that do not personalize results. Unfortunately, there has been only little scientific quantification of the basis and extent of search personalization in practice[15, 16].

3. Personalized Search Using Concept Network and Genetic Algorithm

In this paper, we suggest a Personalized Search system which uses Concept Network and Genetic Algorithm. Here the Concept Network is used to represent the user's preference. Concept Network is stored in the user's profile. Genetic Algorithm is used to find concept network that might be relevant to the user's interest based on the user's present concept network.

3.1. Concept Network Based User Profile

The user profile that contains information about user's preference is modeled as a so called 'concept network' which is a networked structure of session interested concepts. The session interest concepts are determined with some keywords extracted from the documents retrieved by the users and it has some related features. A user's profile and related the concept network will be gradually complex as the user goes on to perform queries and consequent selection of the required documents from the search results.

In the development of an efficient personalized search system, it is highly necessary to design the user profile accurately so as to include user's information needs precisely. The Concept Network is constructed by extracting all the concepts from the pages searched by the user in a particular session.

TF-IDF method is used in the extraction of the concepts from the pages browsed by the user[18]. The tf-idf value of all the concepts is obtained from the pages browsed by the user. The concept with highest tf-idf value is considered as the most important concept or one that is most relevant to the user's preference. The tf-idf value is used as an edge value in the concept network created based on the concepts extracted.

Five concepts that have the highest value are used to create the concept network. The tf-idf value is used as the edge value of the concept in the concept network. Initially the query entered by the user is used to perform standard query based search and provide the result. The pages that are searched by the user in a particular session are used for extracting the concepts. The extracted concept is used in the construction of concept network. The first step in concept extraction is extraction of text from the web page. The extracted text is divided into individual words and each word is converted to their root using a process called stemming. Figure 2 shows the concept extraction procedure.

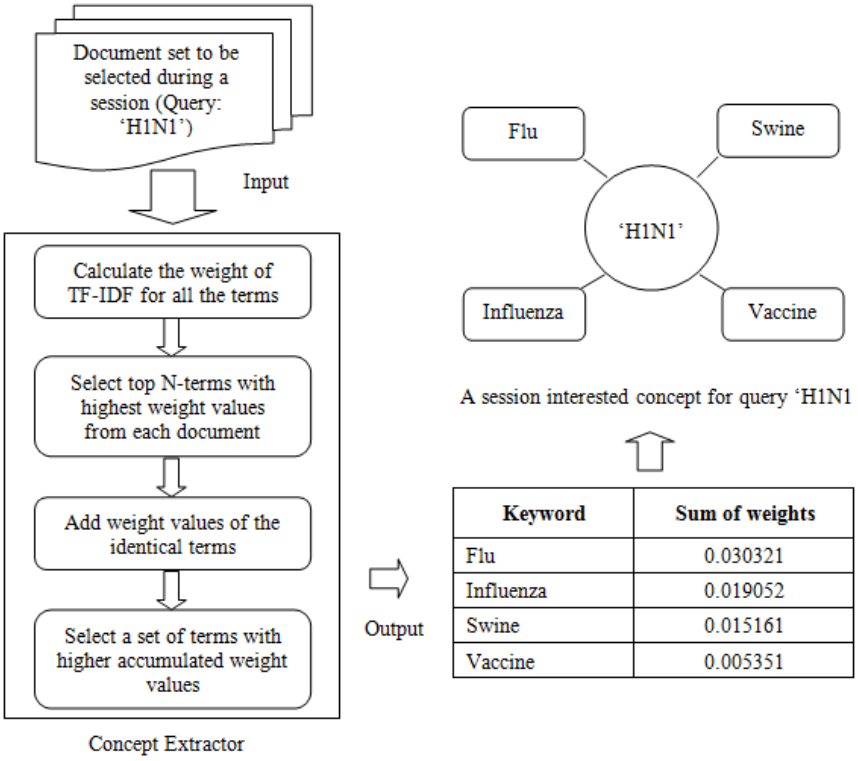

Fig. 2. Concept Extraction

3.2. Genetic Algorithm for clustering users

Genetic Algorithm is used to compare the user's concept network with other user's concept network. And it adds the concept networks that are similar to the user's preference into the users' profile. Genetic algorithm takes two users at a time and selects those concepts that are similar to each other. Genetic algorithm can also be applied for multiple users by taking two users at time. The concept networks that are similar to user concept network added to user profile if the concepts are above threshold value.

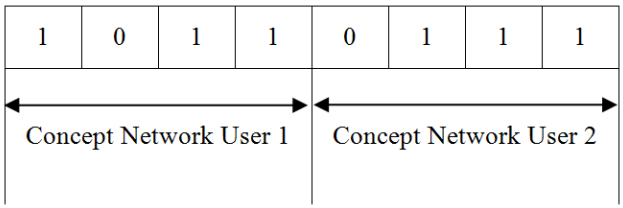

Fig. 3. Individual Representation

Genetic algorithm uses linear binary representations for representing solution. The most standard one is an array of bits. Binary representation is used to represent the population. In a binary representation each position of chromosome can contain 1 or 0. In this each bit represents a concept. The first n_1 bit represents the concept network of user 1 and next n_2 bit represents the concept network of user 2. Fig.3 represents that the individual representation of concept networks for two users, namely user1 and user 2.

In this proposed method GA uses the optimization function based on the number of terms in the concept network and their edge values. This function takes two concept networks as input and provides a value as output. The output value increases as the similarities between the two concepts network increases. This optimization function takes the two concept network and compares the concepts of the two concept networks. If there are any concepts that are equal, then their edge value will be added to the function value. The optimization function is given by

$$f_{12} = (n*(\sum ed(i) + ed(j)))/(N_1+N_2) \quad (1)$$

Where, $0 \leq i < n_1$, n_1- number of terms in concept network 1

$0 \leq j < n_2$, n_2- number of terms in concept network 2

$ed(i)$ and $ed(j))$ are the values of i^{th} and j^{th} edge respectively

Term i of concept network 1 = Term j of concept network 2 if the networks are similar

N_1- number of concept networks in user 1 profile

N_2- number of concept networks in user 2 profile

n - number of terms that are common between concept networks

In order to find most relevant documents, we have to omit some of the retrieved documents, which are unrelated to the concept under consideration. So we need a threshold value to check the documents. This threshold value is calculated by adding the least edge in the entire concept network created by the user and multiply by 2. This threshold value is calculated by adding the least edge in the entire concept network created by the user and multiply by 2 they will mean different things[19].

Crossover is a genetic operator used in GA to change the chromosomes from one generation to the next. It is analogous to reproduction and biological crossover, upon which GAs are based. In the cross over process it takes more than one parent solutions and produces a child solution from them. In the proposed method the Two Point cross over technique is applied. In two point crossover method children is obtained replacing the bits of one parent in between the crossover with bits in same position of other parent. The distance between the crossover points has to be constant. The distance between the crossover points is set based on no of bits in solution. The starting point of crossover is selected randomly. Crossover operation is applied in every iteration. The Fig. 4 shows the crossover operator.

The proposed method uses a mutation operator in order to maintain the genetic diversity from one generation to another. It is analogous to natural mutation process. When mutation happens the mutation operator alters one or more gene values in a chromosome from its original state. After mutation, the new solution may change entirely from the previous solutions. So a better mutation rate can result in better solution. During the evolution process the mutation happens according a mutation probability. For better performance the probability should be set low as

possible. This is because for higher mutation values, the search will turn into a primitive random search. Mutation helps to allow the algorithm to avoid or stuck in local minima by preventing the population of chromosomes from becoming too similar to each other, thus slowing or even stopping evolution process. One parent is selected randomly and mutation operator is applied. In the proposed method for mutation operation, Bit Flipping is used. The Bit Flipping operator replaces 1 with 0 and vice versa. After mutation is applied, the children are added to population. If there is any other children chromosome with lesser fitness value, then they are discarded. The mutation operation is shown in the Fig. 5.

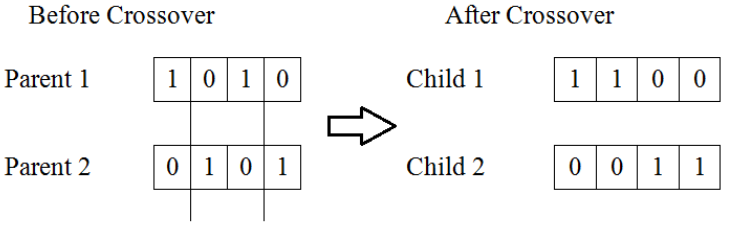

Fig 4.Crossover Operator

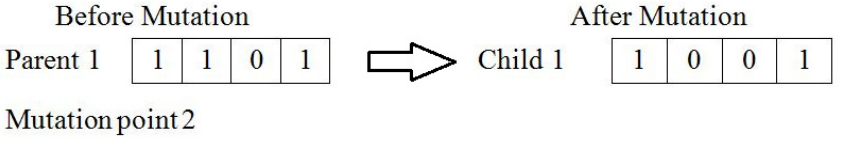

Fig. 5. Mutation Operator

The search relevance ratio of each retrieval process is calculated using the equation 2 as given below[17]. The result obtained for the query term 'H1N1' is shown in the Fig.6. From the graph we can see that as the search progresses, the relevance ration increases for the users.

$$\mathrm{Re}levance\_Ratio = \frac{Number\_of\_\mathrm{Re}levant\_URLs\_\mathrm{Re}trieved}{Total\_No\_of\_URLs\_\mathrm{Re}trieved} * 100 \qquad (2)$$

The document retrieval ratio indicates how good the documents are retrieved for the concept under search. It tells how the relevant is the text inside those documents? The Fig.6 depicts the ratio of relevant document to retrieved documents (1,000 results per topic). From the graph it is clear that the proposed method retrieves only useful documents.

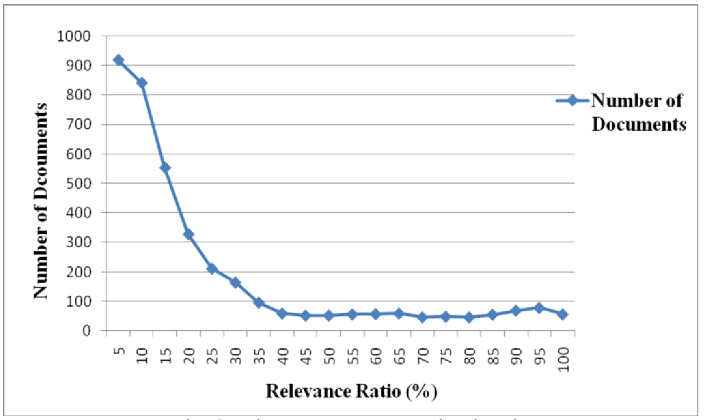

Fig. 6. Relevant Document Retrieval Ratio

4. Conclusions and Future Work

Normal search engine does not take into account the user's preferences and so the result may not contain the pages that user is searching for. A personalized search engine improves the efficiency of the search engine by creating user profile by analyzing his search pattern. This paper proposes a system that uses concept network and genetic algorithm to improve the efficiency of search process. Concept network is used to store the user's profile and Genetic Algorithm is used to compare user's interest in order to predict his interest. Concept network is constructed by extracting the concepts from the pages searched by the user. Genetic Algorithm calculates the similarities between the user's concept network and other user's concept network and merge these concept networks to the user's concept network and finally to the user profile. In this paper we are using TF-IDF value for extracting the concepts and creating the concept network for efficient personalized search. In future works we will investigate the possibilities of other methods to extract the concepts from the pages browsed by the user. Also we will come with metaheuristic algorithms like Ant Colony Optimization, Particle Swarm Optimization, etc., for getting better results.

References

1. Selvakumar K. Challenges and recent trends in personalized Web search: A survey. *IEEE Third International Conference on Advanced Computing (ICoAC)*: Chennai; 2011. p. 333-339.
2. Han-joon Kim. Sungjick Lee. Byungjeong Lee. Sooyong Kang. Building Concept Network-based User Profile for Personalized Web Search. *Proceedings of 9th IEEE/ACIS International Conference on Computer and Information Science*: Washington DC; 2010. p. 567-572.
3. Cheqian Chen. Kequan Lin. Heshan Li. Shoubin Dong. Personalized search based on learning user click history. *9th IEEE International Conference on Cognitive Informatics (ICCI)*: Beijing; 2010. p. 490-495.
4. Kenneth Wai Ting Leung and Dik Lun Lee. Deriving Concept Based User Profiles from Search Engine Logs. *IEEE Transactions on knowledge and data engineering*; 2010; 22:7. p. 969-982.
5. Kyung Joong Kim and Sung Bae Cho. A Personalized Web Search Engine Using Fuzzy Concept Network with Link Structure. *IFSA World Congress 20th NAFIPS International Conference*:Vancouver; 2001; 1. p. 81-86.
6. Kyung Joong Kim, Sung Bae Cho. Personalized mining of web documents using link structures and fuzzy concept networks. *Journal Applied Soft Computing (Elsevier)*; 2007; 7:1. p. 398-410.
7. Lihua Wu. JianPing Feng. Yunfen Luo. A Personalized Intelligent Web Retrieval System Based on the Knowledge-Base Concept and Latent Semantic Indexing Model. *IEEE 7th ACIS International Conference on Software Engineering Research, Management and Applications (SERA '2009)*: Haikou; 2009. p. 45-50.
8. Nadjarbashi Noghani M. Jie Zhang. Sadat KMH. Ghorbani AA. PENS: A personalized electronic news system. *Proceedings of 3rd Annual Conference on Communication Networks and Services*: Canada; 2005. p. 31-38.
9. Ana Regia de M Neves. Alvaro Marcos G Carvalho. Celia G Ralhab. Agent-based architecture for context-aware and personalized event recommendation. *Expert Systems with Applications (Elsevier)*; 2014; 41:2. p. 563–573.
10. R Baeza Yates, B Ribeiro Neto. *Modern Information Retrieval*. New York: Addison Wesley; 1999.
11. K Agbele. H Nyongesa. A Adesina. ICT and information security perspectives in E-health systems. *Journal of Mobile Communication*; 2010; 4. p. 17–22.
12. JH Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor. Mich. USA: The University of Michigan Press; 1975.
13. DE Goldberg. *Genetic Algorithms in Search, Optimization, Machine Learning*. Addison Wesley; 1989.
14. Gerard Salton. Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*; 1988; 24:5. p. 513–523.
15. D Sullivan. Why Google "Personalizes" Results Based on Obama Searches But Not Romney. *Search Engine Land*; 2012. http://selnd.com/PyfvvY.
16. A Hannak. P Sapiezynski. A Molavi Kakhki. B Krishnamurthy. D Lazer. A Mislove and C Wilson. Measuring personalization of web search. 22nd *International World Wide Web Conference*: Brazil; 2013. p. 527-537.
17. KR Remesh Babu. AP Arya. Design of a Metacrawler for Web Document Retrieval. *12th International Conference on Intelligent Systems Design and Applications (ISDA)*: Kochi; 2012. p. 478-484.
18. G Salton. A Wong. CS Yang. A vector space model for automatic indexing. *Communications of the ACM*; 1975; 18:11. p. 613–620.
19. Duncan A Buell. Donald H Kraft. Threshold values and Boolean retrieval systems. *Journal of Information Processing and Management: An International Journal Archive*; 1981; 17:3. p. 127-136.