International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 1, January-February 2015 **ISSN 2278-6856**

Clusterwise optimization of clicked web pages using Genetic algorithm for effective Personalized Web Search

Suruchi Chawla

Department of Computer Science, Shaheed Rajguru College of Applied Science for Women, University of Delhi, INDIA

Abstract

The Information Retrieval on the web retrieves huge collection of documents set and are ranked based on their relevance for a given user search query. It is found in the research that pages which are representative of some aspect of retrieve set P but appear with low position in ordering has a negligible chance of being looked and hence is responsible for low precision of search results. In this paper the genetic algorithm is used to perform the clusterwise optimization of clicked pages in a given domain in order to identify web pages that are not only relevant but also have high internal dissimilarity in order to cover the wider representation of cluster domain. This clusterwise optimization of clicked pages identify those relevant documents up in ranking which otherwise has low ranking and could not be clicked. During online processing, the subset of webpages associated with the cluster is used for the recommendation for effective personalization of web search. This recommendation of web pages continues till search is personalized to the information need of the user. Experimental study was conducted on the data set of web query sessions captured in three domains Academics, Entertainment and Sports. The experimental results which were verified statistically shows the improvement in the average precision of search results and hence it confirms the effectiveness of clusterwise optimization of webpages for better personalizing the Web Search of the user.

Keywords: Information Retrieval, Information Scent, Search engines, Genetic Algorithm, Personalized Web Search(PWS).

1. INTRODUCTION

Search engines on the web retrieve a large collection of web pages P for a given query where the web pages are ranked in order of their relevance. It is observed that web pages that are representative of some aspect of the retrieve set P but appear in a very low position in the ordering has a negligible chance of being looked at by the user and is therefore responsible for low precision.

Research has been done in [8] to determine the small set of pages, selected from the initial set P, that, besides containing pages with a high score, but also are different from each other and are chosen from different regions of some space represented by the set P for improving the precision of search results. This approach of using genetic algorithm for generating the subset of webpages, is implemented by clustering initial retrieved set of web pages to generate the clusters of web pages. The genetic algorithm is applied on the clusters of web pages to generate the subset of the original set of pages P. Thus this method identifies the subset of web pages which have the property of combining a good overall score along with a high internal dissimilarity. This provides the user with few non-duplicated pages that represent more correctly the structure of initial set of pages P.

The research has been done for effective personalized web search using Information Scent based on clustered query sessions in [10]. During web search, the user query is used to select the most similar cluster and the selected cluster is used to recommend the High Scent clicked URLs associated with the clustered query sessions. It is realized that use of Information Scent for identifying relevant clicked pages for recommendations during the personalization of web search could not retrieve those clicked web pages which are of low information scent but are representative of some subdomain of cluster C. Hence an approach is required that could find the subset of pages that not only have high relevance measured using Information Scent value but should also have high internal dissimilarity in order to have wide coverage of the subdomains represented in a selected cluster.

In this paper research is motivated to apply genetic algorithm for clusterwise optimization of clicked web pages in [10] in order to retrieve those documents which are both relevant and has sufficient dissimilarity in order to cover the wide representation of subdomains of a given cluster. The processing of the proposed approach of clusterwise optimization of clicked pages for the personalization of web search has been divided into two phase: Offline and Online. During offline processing genetic algorithm is applied on query sessions clustered using Leader-SubLeader Algorithm in order to identify the subset of pages associated with a given cluster. The subset of webpages contains the pages which not only have the high Information Scent but also have high internal dissimilarity in order to have wide coverage of subdomains of the cluster. Thus at the end of offline processing, each cluster is associated with optimal subset of webpages.

During online processing, the initial input query is used to select the most similar cluster and subclusters of a given selected cluster. The similarity measure of input query with selected cluster and subcluster are compared with each other. If the similarity measure computed for a

selected cluster is more than its subcluster then the selected cluster is used to recommend the associated optimal subset of clicked pages otherwise the high scent clicked pages associated with the more focused domain of subcluster are recommended. The recommended clicked pages are ranked in order of their information scent value. The user's clicks to the recommended clicked URLs are tracked and stored in his profile. As user request for next result page, the user profile so far captured has been transformed into keyword vector and is used for the selection of cluster/subcluster for the recommendation of associated clicked pages on the next result page. This process of recommendations continues till the search is personalized to the information need of the user.

Experimental study was conducted on the data set of user query sessions captured on the web in three domain Academics, Entertainment and Sports to test the effectiveness of clusterwise optimal subset of web pages generated for the Personalization of web search. The improvement in the average precision of search results confirms the effectiveness of Personalized Web Search using proposed method.

The rest of the sections are organized as follows. Section 2 describes the related work, section 3 explains the concepts required for understanding the proposed approach, section 4 describes the proposed work, section 5 presents the experimental study and section 6 concludes the paper.

2. RELATED WORK

In [27] ranking of Web search results is proposed from personalized perspective. In this common access patterns from user browsing activities are mined to automatically obtain user interests. According to the user interests mined and feedbacks of users, a new approach is proposed with the plan of dynamically altering the ranking scores of Web pages. In [2] a multi-agent based personalized meta-search engine using automatic fuzzy concept networks is proposed. An automatic fuzzy concept network is used to personalize outputs of a meta-search engine presented with a multi-agent architecture for searching and fast retrieving. In [1] personalized search engine using ontology-based fuzzy concept networks is proposed. The concepts of ontology are used to improve the common fuzzy concept networks built according to user's profile. The fuzzy concept networks are then used to personalize the search engine outputs. In [24] Fuzzy Logic was used for offline processing to recommend URLs to users. Fuzzy Logic testing shows slightly lower precision and is harder to program for the fuzzy part. In [25] Bee Colony Optimization was used for IR however this optimization technique is not a widely covered area of research.

In [33] Genetic Algorithm is found to be a powerful search mechanism and is suitable for information retrieval since the document search space represents a high dimensional space and GA is a powerful searching mechanism known for its robustness and quick search capabilities. In [36] GA is used with user feedback to choose weights for search terms in a query. In [6][19][35] GA is examined for Information Retrieval and a new crossover and mutation operator were suggested. In [13] an algorithm is proposed for index function learning based on genetic algorithm. The index function is used for key term weighting of a documentary collection in order to improve the Information Retrieval Process.

In [14] GA is also used to automatically learn a matching function with relevance feedback where classical generational scheme and the usual GA crossover are considered. In this the trees are used to represent a similarity function. In [31] Genetic algorithm is proposed for learning queries in Boolean IRS. The set of relevant documents are provided by the user and offline learning process is applied to automatically generate a query describing the user's needs. In [3] the modified GA was used for the optimal design of a website using the multiple optimization criteria such as download time, visuaization and product association level.

In [12] G- Search was implemented using GA-based search which is used to find other relevant home pages given some user-supplied homepages. In [22] the genetic based learning of importance factors of HTML tags has been described for web document retrieval where the importance of the tags is learnt from a training text set.In [5] query reformulation techniques are developed using GA in which several queries that explore the different areas of document space are generated to determine the optimal query. In [16] Genetic algorithm is used to derive a better description of documents. The document description is described as a set of indexing terms. The genetic operators and the relevance judgments are applied to these descriptions in order to determine the one which has best classification performance in response to a specific query. In [23] GA is used for automatic web page categorization and updation. In [10] an approach is proposed for Personalized Web Search based on clustered query sessions. During online web search, the user input query is used to select the cluster which is closest to the information need of the user and the selected cluster is used to recommend the clicked URLs associated the selected cluster. This process of recommendations continues till the search is personalized to the Information need of the user. The experimental results show an improvement in the precision.

It is found in [10] that performance of Personalized Web Search based on clustered web usage data depend on the quality of the recommendations. It is found that use of high Information Scent for web page recommendations could not retrieve some webpages that are representative of some subdomain of the selected cluster but are of low information scent. Hence it is responsible for low precision of search results. There is the need to identify the subset of pages associated with cluster in such a way so that not only they have high value of Information Scent but also has the wide coverage of subdomains of the cluster in order to bring more and more relevant documents up in ranking for improving the precision of search results. An approach has been proposed in [8] to apply the genetic algorithm on clusters of web pages in a

specific domain in order to identify the subset of webpages having both high score and wider coverage of region of the initial set of web pages.

The research in this paper is motivated to apply the given approach of genetic algorithm for web page optimization in clustered query sessions in order to identify the optimal subset of web pages in a domain of each cluster for effective Personalized Web Search.

The processing for Personalized web search based on clusterwise optimization of webpages using genetic algorithm is divided into two parts : Offline and online processing. During offline processing genetic algorithm is applied on clustered query sessions for optimization of clicked web pages in a given domain.

During online web search, user search query is used to select the most similar cluster/subcluster. The selected cluster/subcluster is used for the recommendations of the clicked URLs. This process of recommendation of clicked URLs continues till the search is personalized to the information need of the user. Moreover, there is no computation overhead during online web search since the processing involved for generation of optimal set of webpages for each cluster is done offline. The effectiveness of the proposed method is further confirmed with good experimental results.

3. BACKGROUND

3.1Genetic Algorithm

Genetic Algorithm is a search method based on the natural theory of evolution [7]. In GA, the decision variables of search problems are encoded into a finite length string of alphabets of certain cardinality. These strings representing the candidate solution to the problem are referred to as chromosomes. The alphabets of the strings are referred to as genes, the values of the genes are called alleles and the collection of chromosomes is called the population P1. The population size used in GA is a user specified parameter which affects the performance of the genetic algorithm. A small population size may lead to premature convergence and yield a suboptimal solution whereas a large population size would involve a lot of computational effort. So the actual population size selected should neither be too low nor too high so as to avoid both premature convergence and high computational overhead. The algorithm to evolve solutions to the search problem using genetic algorithm is given below. [26]

Algorithm 1:
Choose an initial population of chromosomes;
while termination condition not satisfied do
repeat
if crossover condition satisfied then
select parent chromosomes
choose crossover parameters
perform crossover
if mutation condition satisfied then
choose mutation points
perform mutation
evaluate fitness of offspring
until sufficient offspring created
select new population
endwhile

During the implementation of Genetic Algorithm, the sequence of steps is defined as follows. [15]

1. Initialization: In the initialization step, population of chromosomes is initialized using the problem specific domain knowledge. The chromosomes represent the different possible solution to the given problem.
2. Evaluation: After the initialization of the population, the fitness value is defined relative to the problem. The fitness value measures the degree of goodness of the chromosomes in representing the solution to the problem. The selection of population of chromosomes for reproduction in next generations is done on the basis of the fitness value evaluated in this step.
3. Selection: In the selection phase, chromosomes with high fitness values are selected and are allocated more copies in the mating pool for reproduction using recombination operators. This results in the survival of the fittest mechanism on the candidate solutions. There are number of selection methods such as roulette-wheel selection, stochastic universal selection, ranking selection, tournament selection and truncate selection.
4. Recombination: In the Recombination phase, the selected chromosomes are recombined using crossover operator which is a genetic operator for the reproduction of offspring from parent chromosomes. The selected chromosomes are used as parents to generate the offspring by swapping the part of the genes present in two parent chromosomes to generate the offspring. There are various types of crossovers like k-point Crossover, Uniform Crossover, Uniform Order-Based Crossover, Order-Based Crossover and Partially Matched Crossover (PMX).
5. Mutation: In this phase mutation is applied to the selected chromosomes. The mutation is the genetic operator which changes the gene at the specific position in the chromosome. The purpose of the mutation is to add diversity to the population of chromosomes in order to avoid local minimum while searching optimum solution to a problem. A common mutation type is bit wise mutation.
6. Replacement: In the Replacement phase, the offspring population generated using selection, recombination and mutation operators will replace the parent population. There are a number of replacement techniques such as elitist replacement, generation-wise replacement, steady-state-no-duplicates and steady-state replacement methods.
7. Steps 2-6 are repeated until a terminating condition is met.

3.2.Information Scent

Information scent is the sense of value and cost of accessing a page based on perceptual cues with respect to the information need of user. The users on the web tend to click those pages in the retrieved search results on the web which seem to satisfy the user's information need. More the page is satisfying the information need of user, more will be the information scent perceived by the user associated to it and more is the probability that the page is clicked by the user. The interactions between user need, user action and content of web can be used to infer information need from a pattern of surfing. [28][29]

3.2.1 Information Scent Metric

The Inferring User Need by Information Scent (IUNIS) algorithm is used to quantify the Information Scent of the pages clicked by the user in i[th] query session. [11][18]

The page access weight and are used to quantify the information scent associated with the clicked page in a query session. The information scent is calculated for each clicked page in a given query session i for all m query sessions identified in query session mining as follows

$$s_{id} = PF.IPF(P_{id}) \times Time(P_{id}) \forall i \in 1..m \; \forall d \in 1..n \quad (1)$$

$$PF.IPF(P_{id}) = \frac{f_{P_{id}}}{\max_{d \in 1..n} f_{P_{id}}} \times \log\left(M/m_{P_d}\right) \quad (2)$$

$PF.IPF(P_{id})$: PF correspond to the page P_{id} normalized frequency $f_{P_{id}}$ in a given query session i where n is the number of distinct clicked page in session i and IPF correspond to the ratio of total number of query sessions M in the whole data set to the number of query sessions m_{P_d} that contain the given page P_d.

$Time(P_{id})$: It is the ratio of time spent on the page P_{id} in a given session i to the total duration of query session i. [10]

3.2.2Generation of Query sessions keyword vector

Each query session keyword vector is generated from query session which is represented as follows

query session=(input query,(clicked URLs/Page)$^+$)

where clicked URLs are those URLs which user clicked in the search results of the input query before submitting another query ; '+' indicates only those sessions are considered which have at least one clicked Page associated with the input query.

The query session vector Q_i of the i[th] session is defined as linear combination of content vector of each clicked page P_{id} scaled by the weight s_{id} which is the information scent associated with the clicked page P_{id} in session i. That is

$$Q_i = \sum_{d=1}^{n} s_{id} * P_{id} \qquad \forall i \in 1..m \quad (3)$$

In eq (3) n is the number of distinct clicked pages in the session i and s_{id} (information scent) is calculated for each clicked page present in a given session i as defined in eq 1. The content vector of clicked page P_{id} is weighted using TF.IDF. Each i[th] query session is obtained as weighted vector Q_i using formula (3). This vector is modeling the information need associated with the i[th] query session.

3.2.3Clustering of Query session keyword vector using Leader-SubLeaders Algorithm

Clustering is the process of grouping the objects in the clusters such that the objects within the cluster are similar and objects in the different cluster are dissimilar. The classification of objects according to perceived similarities forms the basis for much of science. [21]

The hierarchical clustering groups the data objects into a hierarchy of clusters. The hierarchical clustering is broadly classified into two types first is agglomerative and second is divisive method. In agglomerative, the clustering is performed in the bottom-up manner starting with the singleton clusters and continues grouping till all the data objects are in single cluster. Divisive method is top-down clustering where clustering starts with all data objects in one cluster and continue splitting the cluster till each data object is in its own cluster. In both divisive and agglomerative it is difficult to perform modifications after a decision for splitting or merging is taken. The advantages of both the methods are 1) They are known for their quick termination. 2) the number of clusters does not required to be known in advance, 3) complete hierarchy of clusters is computed, 4) good result visualizations are integrated into the methods, 5) a "flat" partition can be derived afterwards (e.g. via a cut through the dendrogram). [20][4][30][32]

Some of the hierarchical clustering algorithms are: Clustering Using Representatives – CURE [17] Leader Algorithm [32] ART3 [9] , Balanced Iterative Reducing and Clustering using Hierarchies – BIRCH [37] and Leader-SubLeader Algorithm [34].

In leader algorithm entire data set is partitioned incrementally into clusters where each cluster is represented by the leader. Leader-SubLeader is an extension of leader algorithm which creates a hierarchical structure. In this data objects are clustered to create groups of data objects. Each cluster is further partitioned into subclusters where each subcluster is represented by subleader. It has low computation cost and clusters/subclusters representation improves the classification accuracy. It performs better than Leader algorithm

Leaders Computation Algorithm
Input: {Query sessions vector Dataset, Threshold}
Output: {Leader List associated with Clusters of Query sessions where each cluster is represented by the Leader}
Algorithm:

1. Select any Query session vector as the initial Leader and add it to the Leader List and set Leader counter LC=1
2. For all Query sessions vector not yet processed
   2.1. Select the Query session vector Q.
   2.2. Calculate the similarity of Q with all the Leaders.
   2.3. Select the Leader which has maximum similarity represented by max.
   2.4. If (max>=threshold)
      2.4.1. Assign it to the selected Leader.
      2.4.2. Mark the cluster number associated with the selected Leader for Q.
      2.4.3. Add it to the member List of this cluster.
      2.4.4. Increment the member count of this cluster.

   else
      2.4.1. Add Q to Leader list.
      2.4.2. Increment Leader counter LC=LC+1.

SubLeaders Computation Algorithm
Input: {Leader List associated with Clusters of Query sessions where each cluster is represented by the Leader, SubThreshold: SubThreshold > Threshold value for similarity of Query sessions vector in SubLeaders}
Output: {Leader-SubLeader List is associated with Clusters of Query sessions where each cluster is represented by the Leader and SubLeaders of the i^{th} cluster are associated with the corresponding subclusters.}
Algorithm:

1. For i= 1 to LC
   1.1. Initialize sub Leader List SL_i with any Query session vector Q in i^{th} cluster.
   1.2. set SubLeader Counter SL_iCount=1
   1.3. for all j= 2 to $|Cluster_i|$
      1.3.1. Calculate the similarity of Q_{ij} of the j^{th} session in i^{th} cluster with all subleader in SL_i
      1.3.2. Select the sub Leader with maximum similarity represented by max1
      1.3.3. If (max1>=subthreshold)
         1.3.3.1. Assign it to the selected SubLeader.
         1.3.3.2. Mark the subcluster number associated with the selected SubLeader for Q_{ij}.
         1.3.3.3. Add it to the member List of this subcluster.
         1.3.3.4. Increment the member count of this subcluster

      else
         1.3.3.1. Add Q_{ij} to the SubLeader List SL_i
         1.3.3.2. Set SL_i Count= SL_i Count +1.

4. Personalized web search based on clusterwise optimization of clicked web pages using genetic algorithm.

In this paper an approach is proposed where uses genetic algorithm on clustered query sessions to generate optimal subset of web pages for effective Personalized Web Search. The algorithm used for Personalized Web Search based on optimized clustered query sessions has been divided in two phases offline and online. During offline processing, the query sessions on web are processed to generate the query session keyword vectors. The query session keyword vectors are clustered in groups using leader-subleader algorithm where each cluster/subcluster contains clicked URLs satisfying similar information need. The genetic algorithm is then applied on each obtained cluster/subclusters to find the optimal subset of clicked pages which are not only relevant measured using high Information Scent but also have sufficient internal dissimilarity in order to have wide coverage of subdomains represented in the cluster.

For the implementation of genetic algorithm on cluster-subclusters, the population of chromosomes is formed. The chromosomes are generated in successions where first chromosome is built by taking the docid of clicked pages of the highest information scent from each subcluster of a given cluster and the second chromosome is built by taking the docid of next highest information scent web page from each subcluster and so on till clicked pages from all subclusters are covered. After generating the population of chromosomes for a given cluster, the fitness function is calculated for each chromosome of clicked pages by taking the sum of two values: first is the sum of information scent of clicked pages present in the chromosome and second is the euclidean distance between all pair of clicked pages represented in the chromosome. The first value gives significance to those clicked pages which have high relevance and second value gives significance to those combinations of web pages which have high internal dissimilarity thereby causing the wide coverage of the subdomains represented in the cluster. Thus high fitness value is obtained when both values are high thus identifying that combination of clicked pages that are both relevant and causes the wide coverage of subdomains represented in the cluster.

Thus chromosomes with high fitness value in the current generation are selected for reproduction using crossover and mutation to generate the next generation of population. This process of generating the generations of population continues till the specified number of iteration. Upon termination, the chromosome having the highest fitness value in the last generation identifies the docid of subset of pages for a given cluster. The stepwise execution of offline processing is given below.

Phase I:

Offline Preprocessing matic

1. Data Set Collected on the Web is preprocessed to get the Query Sessions. A query session is defined as the input query and the associated clicked URLs.
2. For each clicked URLs, the Information Scent Metric is calculated which is the measure of the relevancy of the clicked URLs with respect to the information need of the user using eq 1.
3. Query sessions keyword vector is generated from query sessions using Information Scent and content of Clicked URLs given by eq 3.
4. Leader-SubLeader clustering algorithm is used for clustering query sessions keyword vector where each cluster and subclusters of a given cluster are associated with the high scent clicked urls.
5. Use the **Algorithm 2** to apply genetic algorithm on each clusters/subclusters of query session keyword vector to generate the optimal subset of high scent pages associated with each cluster.
6. Thus each cluster is associated with the optimal subset of pages having both high Information Scent value and has the wide coverage of subdomains of the cluster.

Algorithm 2:

Optimization of Clicked Page set associated with clusters using Genetic Algorithm.

Input : Selected Clusters-Subclusters and the associated clicked pages with the high scent values.

Output: the optimal subset of pages associated with the clusters. having both high scent and belong to different sublcusters in the specific domain

For each cluster do the following

1. Initialization

During Initialization, for a given cluster/sublcusters initial population of chromosome is generated using the docid of clicked pages present in the subclusters of a given cluster. Each gene position in the chromosome represents the docid of clicked page belonging to a given subcluster. The length of the chromosome is equal to the number of subclusters present in it. The first chromosome is built by taking the high scent clicked page from each subclusters of cluster. Then the second chromosome is built by taking the clicked urls with next highest scent values present in each sub clusters of the cluster and so-on. Thus the presence of the clicked url in the chromosome is identified by its docid in the gene associated to it in the chromosome.

2. Evaluation

Once the population is initialized with the chromosomes, the fitness value of the candidate solutions represented by chromosomes is evaluated. Each chromosome C is evaluated using the Fitness Function. Fitness function is the weighted sum of two factors. First is the sum of information scent of the clicked pages present in the chromosome and second factor is the sum of the distances between the pair of pages in chromosome C and measures the total variability expressed by C. Thus the fitness function is defined as

$Ff(C)=\alpha.t_1(C)+\beta t_2(C)$

Where α and β are chosen such that contribution made by $t_1(C)$ and $t_2(C)$ are balanced. Where $t_1(C)$ is the sum of the information scent of the clicked pages present in the chromosome and $t_2(C)$ is the sum of the distances between the pair of clicked pages in chromosome C and measures the total variability expressed by C.

$$t_1(C)=\sum_{Pid\in C} s_{id}(P_{id})$$

$$t_2(C)=\sum_{P_{id}P_{jd}\in C} D\left(P_{id},P_{jd}\right)$$

Where $D\left(P_{id},P_{jd}\right)$ is Euclidean distance of vector representing clicked pages P_{id} and P_{jd}. and $s_{id}(P_{id})$ is the information scent of clicked page P_{id}.

3. Selection
   Tournament Selection and elitism is used to select the chromosomes with the highest fitness values for reproduction.
4. Recombination

Uniform Crossover is selected for recombination. The crossover probability lies in [0.2-0.8]. The single point mutation probability lies in [.005-.01] to generate the new population.

5. Replacement

The offspring population created by selection, recombination, and mutation replaces the original parental population. **Steady-state** technique with no duplicates deletes n old members and replaces them with n new members. The number to delete and replace, n, at any one time is a parameter to this deletion technique. Only the low fitness value chromosomes in the parent population are replaced with high fitness value chromosome from the offspring population in order to generate the next generation of population.

6. Repeat step 3-5 till the terminating condition is obtained. The terminating condition can be the fixed N number of iterations or the change in the best fitness value is less than .000001 in last 100 trials. The goal is to find the maximum fitness chromosome in the last population.
7. Upon termination store the clicked pages in the chromosome with the maximum fitness value in the optimal set associated with a given cluster – subclusters.

During online processing, the user input query is used to select the most similar cluster measured using cosine similarity measure. Once the cluster is selected, the sub-clusters of a selected cluster are searched to find the subcluster which is most similar to the information need of the input query. If the selected subcluster is more similar to the input query than the entire cluster, then high scent web pages associated with the subcluster are recommended otherwise the optimal subset of pages associated with the cluster covering the different subdomains of the cluster is recommended. The user response to the recommended clicked pages are tracked and stored in the user profile. As the user request for next web page, the user profile so far captured is transformed into keyword vector and is used to select most similar cluster/subcluster. The selected cluster/subcluster is used to recommend the associated web pages on the next requested result page. This process of recommendations continues till the search is personalized to the information need of the user. The stepwise execution of online processing is given below.

Phase II

Online Processing.
Input: Set of Clusters/Subclusters
Output : Recommended Set of Clicked URLs

1. The input query q is used to calculate the similarity with all leaders associated with their clusters.
2. Select the leader with the maximum similarity and the optimal subset of high scent pages associated with the cluster of the leader is recommended to the user.
3. The user 's interaction to the recommended results is tracked and stored the clicked urls in the current user profile which will be used to further infer the user information need in the selected domain .
4. If the user request for the next result page
   a. Model the partial information need of the current user profile using the information scent and content of the URLs clicked so far in his partial user profile and obtain the user session keyword vector current_usersessionvector$_t$.
   b. Find the cluster leader which is most similar to the information need associated with the current_usersessionvector$_t$.
   c. Select the leader with the maximum similarity represented as max. Calculate the similarity of the current_usersessionvector$_t$ with all sub-leaders of the selected leader.
   d. Select the sub-leader with maximum similarity represented by max1.
   e. If (max> max1)
      i. Select the cluster associated with the leader.
      ii. Store the associated optimal set of high scent pages associated with the cluster in set P which covers different subdomains in a given selected cluster and the pages in set P are ordered in decreasing order of their information scent value.

      Else
      i. Select the subcluster associated with the subleader.
      ii. Store the high scent pages associated with the selected subcluster in set P where high scent pages are the pages belonging to the specific subdomain of the cluster and the pages in set P are ordered in decreasing order of their information scent value.
   f. Recommend the set P to the users.
   g. Goto step 3.

5.EXPERIMENTAL STUDY

The experiment was conducted on a data set of user query sessions collected on the web. The data set of query sessions was generated using an architecture which has

been developed to capture the URLs clicked by users in the search results obtained using the Google search engine. In order to generate the dataset, the user is required to enter the input query through a GUI based interface of the architecture. This input query is passed on to the Google search engine API, and the search results are retrieved and displayed along with the check boxes on the user interface. A SnapShot of GUI interface of the architecture showing the Google search results for the input query "hindi song" is shown below in Fig 1.

The user clicks on the retrieved search results, are captured through the check boxes displayed on the GUI and stored in the form of query sessions in database. The captured user query sessions on the web are processed further to find the query session keyword vector using Information Scent and content of clicked URLs. The Leader-subleader algorithm is then applied to group the similar information need query session keyword vector in clusters/subclusters.

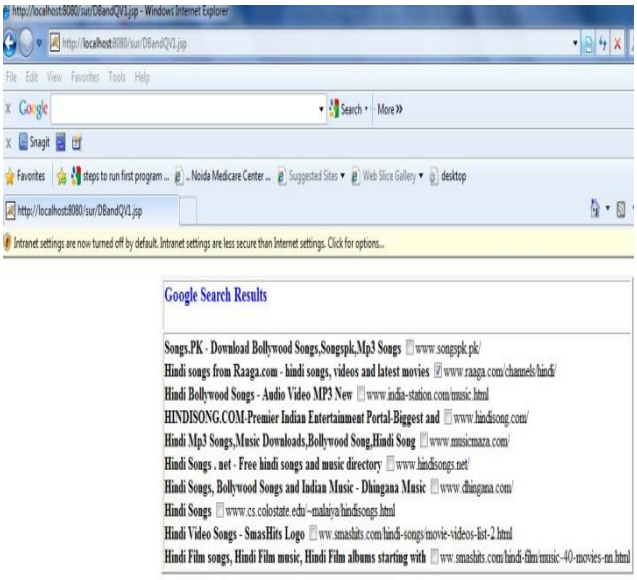

Fig 1 Screen SnapShot of architecture displaying Google Search results along with the checkboxes.

The experiment was performed on Pentium IV PC with 120 GB RAM under Windows XP using JSP, JADE , Oracle and genetic algorithm tool box of MATLAB. In the experimental set up for evaluating the performance of personalized web search based on clusterwise subset of webpages generated using genetic algorithm, the values of following parameters are used in the genetic algorithm : MAXGEN, length(P1), crossover rate, mutation rate, Tournament Size in the Tournament Selection method and the threshold value of Information Scent where MAXGEN is the maximum number of generations of population generated in the evolutionary process, length(P1) represents the number of chromosomes individuals in the population, crossover rate is the recombination rate of the selected chromosome individuals in the population and mutation rate is the rate of mutating the chromosomes in the population Since the genetic algorithm is a stochastic computational technique, it has to be iterated many times for a given problem so as to get a satisfactorily good result. In this study, the process of generating the population continues till the difference in the optimum fitness value of last 100 consecutive generations is less than the threshold value τ=0 .000001.

In this study, the experiment was conducted with the following values of selected parameters- the size of the population represented as length(P1) was m where m depends on the number and size of subclusters present in a given cluster, crossover probability was varied in the range of [0.6-0.8] in increment of 0.1 and the mutation rate was varied in the range in [0.1-0.3] in increment of .05.

The experiment was iterated for 100 generation for a given population P1 where length(P1)= 5040 and the size of the Tournament in the Tournament Selection was set to 4. The optimal results were obtained at the crossover rate of 0.8, mutation rate of 0.25, α=0.5, β=0.5 giving equal weightage to information scent and dissimilarity measure in calculating fitness function of chromosomes and threshold value of Information Scent (ρ) at 0.5 for the data set generated in this experimental study. The threshold value for Leaders computation was set to 0.5 and the subthreshold value for SubLeaders computation was set at 0.75.

During offline preprocessing, the tf.idf vector of the clicked URLs of the query sessions are fetched using the web sphinx crawler and loaded into database using Oraloader. The clustering agent developed in JADE is executed to generate the clusters/subclusters of query session keyword vectors. The Genetic algorithm is performed on each cluster and associated subclusters in order to generate the subset of webpages for a given cluster. The genetic algorithm tool box of MATLAB software package was used for applying the genetic algorithm on the clustered data set. The population generation function, single point mutation, uniform crossover, fitness function and output function are defined by the user in MATLAB. The output function is defined in MATLAB for storing the set of clicked URLs associated with a given cluster in the database for the later retrieval for personalized web search.

The approach proposed for PWS based on clusterwise optimal subset of webpages(using genetic algorithm) was compared with approach used for improving information retrieval precision using PWS based on clusterwise subset of webpages(without optimization) in [10] in order to compare and determine the effectiveness of clusterwise optimal subset of webpages for PWS.

During online processing, the input query is issued to GUI based interface designed for both PWS with/ /without optimization of clustered webpages. In PWS with clusterwise optimal subset of webpages, the input query is

used to select the cluster/subcluster most similar to the information need of the user. The set of clicked URLs associated with the selected cluster/subcluster are recommended and displayed in the GUI Interface for the current user input query along with checkboxes for capturing the user's clicks.

The user's clicks to the personalized search results are tracked to capture the user's profile and dynamically update the user's clicked profile during the search session of the user. When the user requests for next result page, this captured user's profile is transformed into keyword vector and is used to select the most similar cluster/subcluster for web page recommendations. This process of clicked URLs recommendations for personalized of web search continues till the user search is personalized to the need of the user.

In order to evaluate the performance, the 25 test queries were selected randomly in each of the domains Academics, Entertainment and Sports. The purpose of selecting the queries in these three domains is to cover wide range of queries on the web. The relevancy of the documents was decided by the experts in the domain to which the queries belong.

The test queries were issued in each of the selected domain to the GUI based interface to retrieve the personalized search results (with /without optimization of clustered webpages). The average precision is computed using the fraction of retrieved documents which are relevant in the personalized search results. The experimental results showing the average precision of test queries computed in the domains of academics, entertainment and sports using PWS (with / without optimization) are shown in Fig 2.

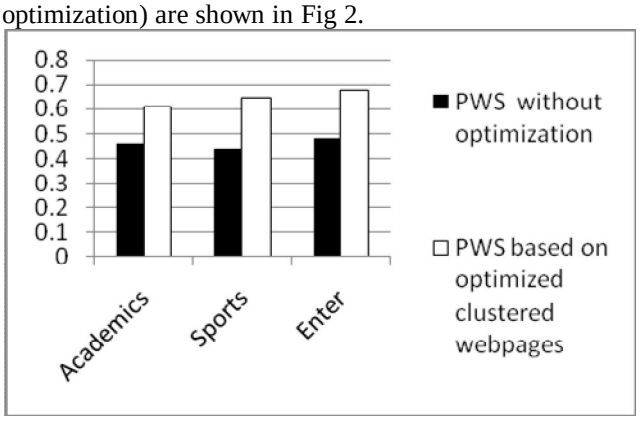

Fig.2. Shows the avgprecision of PWS without optimization and PWS with optimized clustered web pages in Academics, Sports and Entertainment.

The average precision is improved in each of the selected domains using personalized web search (with optimization of clustered webpages). The obtained results were analyzed using the statistical paired t-test for average pre-cision of PWS (with / without optimization of clustered web pages). In this the test data set of 25 queries in each of selected domain with 74 degrees of freedom (d.f.) for the combined sample as well as in all three categories (Academics, Entertainment and Sports) with 24 d.f each. The observed value of t for average precision was 33.98 for the combined sample. Value of t for paired difference of average precision was 15.2 for academics, 30.73 for entertainment and 28.75 for the sports categories. It was observed that the computed t value for paired difference of average precision lies outside the 95% confidence interval in each case. Hence Null hypothesis was rejected and alternate hypothesis was accepted in each case and it was concluded that average precision improved significantly using personalized web search with optimal subset of webpages in comparison to improvement in average precision of Personalized search results (without optimal subset of webpages).

This proves that use of genetic algorithm, for optimization of clustered webpages personalizes the web search more effectively since these set of webpages not only have high relevance measured using Information scent but has wide coverage of subdomains of a given cluster. Thus during online web search more and more relevant documents associated with selected cluster are retrieved up in ranking and is responsible for the improvement in the average precision of test queries in each of the selected domains.

6.Conclusion

In this paper an approach is proposed which uses the genetic algorithm for the optimization of clustered web pages for retrieving more and more relevant documents in a specific domain in order to improve the effectiveness of the Personalized Web Search. It is found that optimization of clustered web pages identify the documents which not only have the high relevance measured using Information Scent but also have high internal dissimilarity in order to have the wider coverage of the subdomains contained in the domain of the cluster. Hence this optimal subset of web pages brings more and more relevant documents up in ranking for PWS which otherwise could not be clicked due to low ranking. The effectiveness of the proposed method was confirmed with the experimental study on the data set captured in three domains mainly Academics, Entertainment and sports. The experimental results verified statistically confirm the effectiveness of the optimization of clustered webpages based on genetic algorithm for better catering to the information need of the user.

References

[1] Akhlaghian, F., Arzanian, B. , & Moradi, P. A. "Personalized Search Engine Using Ontology-Based Fuzzy Concept Networks", International Conference on Data Storage and Data Engineering, pp. 137 – 141, 2010.

[2] Arzanian,B., Akhlaghian, F., & Moradi, P. "A Multi- Agent Based Personalized Meta-Search Engine Using Automatic Fuzzy Concept Networks", Third International Conference on Knowledge Discovery and Data Mining, pp. 208 – 211, 2010.

[3] Asllani, A., & Lari, A. "Using genetic algorithm for dynamic and multiple criteria website optimizations", European Journal of Operational Research, Volume 176, Issue 3, pp. 1767- 1777, 2007.

[4] Berkhin P."Survey of clustering data mining techniques". Accrue software, Technical Report, 2002.

[5] Boughanem, M., Chrisment, C., Mothe, J., Dupuy, C. S., & Tamine, L. "Connectionist and genetic approaches for information retrieval", Soft Computing in Information Retrieval Studies in Fuzziness and Soft Computing, 50, pp. 173–198, 2000.

[6] Boughanem, M., Chrisment, C., & Tamine., L. "On using genetic algorithms for multimodal relevance optimization in information retrieval.", Journal of the American Society for Information Science and Technology, Volume 53, Issue 11 , pp. 934–942, 2002.

[7] Bremermann, H. J. "The evolution of intelligence. The nervous system as a model of its environment", Technical Report No. 1, Department of Mathematics, University of Washington, Seattle, WA, 1958.

[8] Caramia,M., Fellci, G., Pezzoli A. ,"Improving search results with data mining in a thematic search engine",Journal Computers and Operations Research, Volume 31,Issue 14,pp. 2387-2404,2004.

[9] Carpenter G. and Grossberg, S." ART3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures." Neural Networks, Volume. 3, pp. 129-152, 1990.

[10] Chawla, S. ,& Bedi, P."Personalized Web Search using Information Scent", International Joint Conferences on Computer, Information and Systems Sciences, and Engineering, Technically Co-Sponsored by: Institute of Electrical & Electronics Engineers (IEEE), University of Bridgeport, published in LNCS (Springer), pp. 483-488, 2007.

[11] Chi, E H., Pirolli, P., Chen, K., & Pitkow, J. "Using Information Scent to model User Information Needs and Actions on the Web", International Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 490-497, 2001.

[12] Crestani , F. , & Pasi, G. "Soft Computing in Information Retrieval: Techniques and Application",50, Heidelberg, Germany: Physica-Verlag, 2000.

[13] Fan, W., Gordon, M.D., & Pathak, P. "Personalization of search engine services for effective retrieval and knowledge management", International Conference on Information Systems, Brisbane, Australia, pp. 20-34, 2000.

[14] Fan, W., Gordon , M.D. , & Pathak, P. "Discovery of context-specific ranking functions for effective information retrieval using genetic programming", IEEE Transactions on Knowledge and Data Engineering, Volume 16, Issue 4, pp. 523-527, 2004.

[15] Goldberg, D. E. "Genetic Algorithms in Search,Optimization and Machine Learning", Addison-Wesley Longman Publishing Co, Boston, MA, USA, 1989.

[16] Gordon, M. "Probabilistic and genetic algorithms in document retrieval", Communications of the ACM, Volume 31, Issue 10, pp. 1208–1218, 1988.

[17] Guha S., Rastogi R. and Shim K.. "CURE: An efficient algorithm for clustering large databases". In Proceedings of ACM SIGMOD International Conference on Management of Data, Seattle, Washington, pp. 73–84, 1998.

[18] Heer, J., & Chi, E.H. "Separating the Swarm: Categorization method for user sessions on the web", International Conference on Human Factor in Computing Systems, pp. 243-250, 2002.

[19] Horng , J. T. & Yeh, C. C. "Applying genetic algorithms to query optimization in document retrieval", Information Processing & Management, Volume 36, Issue 5 , pp.737–759, 2000.

[20] Jain A.K, Murty M.N. and Flyn P.J.. "Data Clustering: A Review", ACM Computing Surveys, Volume.31, Number 3, pp 264-323, 1999.

[21] Jain K.A. and R.C. Dubes. "Algorithms for Clustering Data". Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[22] Kim , S. ,& Zhang, B. T. "Web document retrieval by genetic learning of importance factors for html tags", International Workshop on Text and Web Mining, Melbourne, Australia, pp. 13–23,2000.

[23] Loia , V. & Luongo, P. "An evolutionary approach to automatic web page categorization and updating", Conference on Web Intelligence: Research and Development, Springer-Verlag, pp. 292–302,2001.

[24] Nasraoui, O.& Petenes, C. "Combining Web Usage Mining and Fuzzy Inference for Website Personalization", International Conference on Knowledge Discovery and Data Mining, pp.37-46, 2003.

[25] Navrat, P., Kovacik, M., Ezzeddine, A. B., & Rozinajova, V. "Web search engine working as a bee hive", Journal Web Intelligence and Agent Systems, Volume 6, Issue 4, pp. 441–452,2008.

[26] Pal, S.K. , Talwar, V., & Mitra, P. "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE Transactions on Neural Networks, Volume 13, Issue 5, pp. 1163-1177,2002.

[27] Peng, Wen-Chih , & Lin,Yu-Chin. "Ranking Web Search Results from Personalized Perspective", The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, pp.12, 2006.

[28] Pirolli, P. "Computational models of information scent-following in a very large browsable text collection" , Conference on Human Factors in Computing Systems, pp. 3-10, 1997.

[29] Pirolli, P. "The use of proximal information scent to forage for distal content on the world wide web", Working with Technology, Mind: Brunswikian.

Resources for Cognitive Science and Engineering, Oxford University Press, 2004.
[30] Pujari A.K.. "Data Mining Techniques." University Press (India), Pvt. Ltd, 2002.
[31] Smith, M.P &, Smith, M. "The use of genetic programming to build Boolean queries for text retrieval through relevance feedback", Journal of Information Science, Volume 23, Issue 6, pp. 423–431, 1997.
[32] Spath H.. "Cluster Analysis Algorithms for Data Reduction and Classification". Ellis Horwood, Chichester,UK, 1980.
[33] Tamine, L., Chrisment, C., & Boughanem, M. "Multiple query evaluation based on an enhanced genetic algorithm", Information Processing and Management , Volume 39, Issue 2, pp. 215–231, 2003.
[34] Vijaya P. "Leaders–Subleaders: An efficient hierarchical clustering algorithm for large data sets": Pattern Recognition Letters, Volume 25, Number. 4., pp. 505-513, 2004.
[35] Vrajitoru, D. "Crossover improvement for the genetic algorithm in information retrieval", Information Processing& Management, Volume 34, Issue 4, pp. 405–415, 1998.
[36] Yang, J., Korfhage , R., Rasmussen, E.M. "Query improvement in information retrieval using genetic algorithms—a report on the experiments of the TREC project", 1st text retrieval conference (TREC-1), pp. 31–58, 1992.
[37] Zhang T., Ramakrishnan, R., Livny M.. "BIRCH: An efficient clustering method for very large databases". In: ACM SIGMOD Workshop on Data Mining and Knowledge, Montreal, Canada, pp. 103–114, 1996.