Optimization of Clusters of Web Query Sessions using Genetic Algorithm for Effective Personalized Web Search

Suruchi Chawla, PhD
Assistant Professor
Shaheed Rajguru College of
Applied Science
University of Delhi, Vasundhra
Enclave,INDIA

ABSTRACT
Personalization of web search is used for effective Information Retrieval in order to better satisfy the information need of the user on the web. The web usage mining has been used widely in Personalization of Web Search(PWS). The effectiveness of the Personalization of Web Search based on clustered web usage data depends on the quality of clusters. It is found in research that there exist no clustering algorithms that produce clusters of 100% quality. In this paper the Genetic Algorithm(GA) is used for clusters optimization in order to improve the quality of clusters for effective Personalized web search. Experiment was conducted on the data set of query sessions captured on the web in Academics, Entertainment and Sports Domain. The search results confirm the improvement in the average precision of the PWS(with cluster optimization) in comparison to PWS(without cluster optimization).

Keywords
Web, Information Retrieval, Personalized Web Search, Genetic Algorithms, Clustering, Optimization, Information Scent

1. INTRODUCTION
The objective of Web Information Retrieval is to better cater to the information need of the user. Personalization of Web Search is used for Intelligent web search based on knowledge obtained by mining the web data.[21][22][40][23][28][2][25][7][6][8][9][10][11][12]. Clustering is one of the common data mining techniques applied on Web data for identifying the pattern of web usage by the users. The purpose of the clustering is to cluster the similar objects in a group and dissimilar objects in different group. These clusters are the bedrock on which most of the existing Personalization techniques for web search depends for effective Information Retrieval. It is found that performance of Personalized Web Search based on clustered query sessions depends on the quality of clusters. Research has been done in [31] in which it is found that there exist no clustering algorithms that produce clusters of 100% quality. There is always the chance of data being mis-clustered that is data get classified to wrong cluster. An approach has been proposed in [31] which uses genetic algorithm to improve the quality of clusters and produces better results. In this paper the genetic algorithm is used for improving the cluster quality for effective Personalization of web search based on clustered query sessions.

In [6] Personalization of Web Search is done using clustered user query sessions. The user query session containing the input query and the associated clicked URLs are transformed into keyword vector using Information Scent and content of clicked URLs.

These query sessions keyword vector are clustered using k-mean algorithm which cluster similar information need query session keyword vector in a group. During online processing, the user query is used to select the most similar cluster and the selected cluster is used to recommend the URLs clicked by the users with the similar information needs in the past. Thus this process of recommendations continues till the search is personalized to the information need of the user.

Thus an approach is proposed in this paper for personalized web search using genetic algorithm for cluster optimization in order to better cater to the Information need of the user. The entire processing of proposed approach is divided into phases: Phase I and Phase II. In Phase I offline processing is performed and Phase II online processing is performed. During offline processing, the query sessions keyword vector are clustered where keyword vector is generated from web query sessions using information scent and content of clicked URLs. The genetic algorithm is applied on the clustered data set for cluster optimization. Initially, the population of chromosomes is created where each chromosome is created taking one query session keyword vector id at a time from each cluster. Hence the size of each chromosome is equal to number of clusters in the data set and a fixed position is assigned to a given cluster in all chromosomes. After creating the population of chromosomes representing the entire set of clusters, the entropy is calculated for the population of chromosomes in a given generation and the global minimum is extracted. With this initial population, the genetic operators such as crossover and mutation are applied to produce a new population. While applying crossover operator, the cluster points will get shuffled and data point has been moved from one cluster to another.

For this new population, the local minimum fitness value is calculated and compared with global minimum. If the local minimum is less than the global minimum then this local minimum becomes global minimum and the next iteration is continued with the new population otherwise, the next iteration is continued with the same old population. This process is repeated for N number of iterations.

Once the required number of iterations is performed, the cluster points are reshuffled according to population of chromosomes having global minimum and is input to the online processing of PWS. The genetic algorithm used for

cluster optimization has no impact on the online performance of web search since the entire processing for cluster optimization is done offline. Experiment was conducted on the data set of query sessions captured on the web in the domains of Academics, Entertainment and Sports to compare the effectiveness of PWS(with cluster optimization) and PWS(without cluster optimization) in [6]. The results verified statistically confirmed the improvement in the precision of search results using PWS(with cluster optimization).

The rest of the sections are organized as follows. Section 2 explains related work, section 3 provides background knowledge required for understanding the proposed approach, section 4 describes the proposed approach, section 5 presents the experimental results and section 6 gives the conclusion.

2. RELATED WORK

In [28] 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 [25] 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 [26] Bee Colony Optimization was used for IR however this optimization technique is not a widely covered area of research.

In [7] a method is proposed for related queries recommendations based on clustered query sessions for improving the Information Retrieval precision. During online web search, the user search input query is used to select the cluster on basis of similarity measure and the selected cluster is used to recommend the related queries in specific domain. This process of recommendations of related queries continues till the search is personalized to the information need of the user. The effectiveness of this method is also confirmed with the experimental results.

In [33] Genetic Algorithm(GA) 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 [4][18][34] GA is examined for Information Retrieval and a new crossover and mutation operator were suggested. In [32] 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 [37] GA is used with user feedback to choose weights for search terms in a query.

In [14] G- Search was implemented using GA-based search which is used to find other relevant home pages given some user-supplied homepages. In [20] 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 [3] 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 [24] GA is used for automatic web page categorization and updation.

Thus the performance of Personalized Web Search based on clustered web query sessions is highly dependent on the quality of the clusters. It is found in the research that there exists no clustering algorithm that produces 100% cluster quality. There is always the probability of data being misclustered. Work has been done in [31] to improve the quality of clusters using genetic algorithm.

The research in this paper is motivated to perform the clusters optimization using Genetic Algorithm in order to improve the effectiveness of Personalized Web Search based on clustered query sessions. The quality of the cluster is measured as the fraction of total cluster points classified to a given cluster but actually belong to other cluster. Hence higher is the fraction of wrongly classified cluster points, lower will be the quality of cluster. Therefore the low quality clusters when used for Personalization of web search will be responsible for the poor performance of PWS. Hence with cluster optimization the quality of clusters is improved that ultimately improve the performance of PWS based on clustered query sessions.

3. BACKGROUND

3.1 Genetic Algorithms

Genetic Algorithm is a search method based on the natural theory of evolution [5]. 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 P. 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. [27]

Algorithm :

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. [29][30]

3.2.1 Information Scent metric

The Inferring User Need by Information Scent (IUNIS) algorithm is used to quantify the Information Scent s_{id} of the pages P_{id} clicked by the user in ith query session. [13][17]

The page access $PF.IPF$ weight and $Time$ are used to quantify the information scent associated with the clicked page in a query session. The information scent s_{id} is calculated for each clicked page P_{id} in a given query session i for all m query sessions identified in query session mining as follows

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

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

$PF.IPF(P_{id})$: 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. [6]

3.2.2 Generation 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 ith 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

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

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 ith query session is obtained as weighted vector Q_i using eq (3). This vector is modeling the information need associated with the ith query session.

3.2.3 Clustering of Query session keyword vector

The k-means algorithm [36][19] defines the centroid of a cluster as the mean value of points within a cluster. A centroid-based partitioning technique uses the centroid of a cluster, C_i, to represent that cluster. Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoids of the objects (or points) assigned to the cluster.

Algorithm:

k-means: the k-mean algorithm for partitioning, where each cluster's centre is represented by the mean value of the objects in the cluster.

Input: k: the number of clusters D: a data set containing n objects

Output: A set of k cluster Method:

1. Arbitrarily choose k objects from D as the initial cluster's centers;

2. Repeat

3. (Re) assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster;

4. Update the cluster means, that is, calculate the mean value of the objects for each cluster;

5. Until no change

k-means is easy to understand and implement and takes less time to execute as compared to other techniques. It can handle large data sets. The k-means algorithm is used for clustering query sessions keyword vectors since its performance is good for document clustering. [35][39]

The vector space implementation of k-means uses score or criterion function for measuring the quality of resulting clusters. The criterion function is computed on the basis of average similarity between vectors and centroid of the assigned clusters.

The criterion function I is defined as follows:

$$I = {}^{1}/_{M} \quad \sum_{p=1}^{k} \sum_{v_i \in C_p} sim(v_i, c_p) \qquad (4)$$

where C_p be a cluster found in a k-way clustering process (p∈1..k) , c_p is the centroid of p[th] cluster , v_i is the vector representing some query session belonging to the cluster C_p and M is the total number of query sessions in all clusters as defined below .[38]

$$M = \sum_{p=1}^{k} |C_p| \qquad (5)$$

The centroid c_p of the cluster C_p is defined as below:

$$c_p = \left(\sum_{v_i \in C_p} v_i\right) \Big/ |C_p| \qquad (6)$$

where $|C_p|$ denotes the number of query sessions in cluster C_p and $sim(v_i, c_p)$ is calculated using cosine measure.

4. OPTIMIZATION OF CLUSTERS OF WEB QUERY SESSIONS USING GENETIC ALGORITHM FOR EFFECTIVE PERSONALIZED WEB SEARCH

In this paper an approach is proposed using genetic algorithm for cluster optimization in order to improve the cluster quality for effective PWS. The processing involved in applying the genetic algorithm for clusters optimization is done offline and has no impact on the performance of online processing of Personalized Web Search using these optimal sets of clusters.

The processing of the PWS based on cluster optimization has been divided into two phases: Phase I and Phase II. In Phase I offline processing is performed. During offline processing, query sessions containing input query and clicked URLs are transformed into keyword vector using Information Scent and content of Clicked URLs. These keyword vectors are clustered using k-means algorithm. The genetic algorithm is then applied on clustered query sessions for optimization.

In order to apply genetic algorithm on the clusters, the population of chromosomes is created. Every chromosome in the population is constructed by taking one query session keyword vector id at a time from each cluster. Thus each query session id act as a gene and the number of genes in the chromosomes is equal to the number of clusters. Thus after generating the population of chromosomes, the fitness function is calculated using entropy function for the clustering solution(CS) where the CS is a set of clusters obtained by collecting all the query session id present at the specific position in all chromosomes in order to identify the points in a particular cluster where each cluster is allocated fixed position in all chromosomes. The Fitness function for the clustering solution CS is defined as given below.

$$E_j = -\sum_{i=1}^{m} t_{ij} \log t_{ij} \qquad (7)$$

$$\text{Fitness} = E_{cs} = \sum_{j=1}^{m} \frac{n_j \times E_j}{n} \qquad (8)$$

E_j is the entropy of the cluster j where t_{ij} is the number of data points belonging to cluster i and wrongly classified to cluster j where sum is taken overall m clusters. Fitness function is the weighted sum of entropies of the individual clusters where n_j is the number of query session vectors in cluster j and n is the total number of query session keyword vector present in all clusters.

This calculated value of entropy is the local minimum which is used to initialize the global minimum. The genetic operators like mutation and crossover have been applied on the selected chromosomes in order to generate the next generation of chromosomes. The fitness value is calculated using the clustering solution obtained from the given generation of population in order to find the local minimum. The local minimum is compared with the global minimum and if the current local minimum is less than global minimum that local minimum is assigned to global minimum and the next iteration is continued with the new population otherwise, the next iteration is continued with the same old population. This process will continue for fixed number of N iteration and upon termination reshuffled cluster points corresponding to clustering solution represented by population with global minima is used for online processing. The stepwise execution of offline processing phase is given below.

Phase I:

Offline Preprocessing

1. Data Set Collected on the Web is preprocessed to get the Query Sessions.
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.
3. Query sessions keyword vector is generated from query sessions using Information Scent and content of Clicked URLs using eq (3).
4. k-means algorithm is used for clustering query sessions keyword vector.$_j$.
5. Apply the algorithm **Genetic Algorithm based method for cluster optimization** on clusters of query sessions.
6. Each optimized cluster j is associated with the mean keyword vector $clust\_mean_j$.
7. For each optimal cluster j maintain the list of Clicked URLs in list L whose information scent >= threshold ρ.

Algorithm 1:

Genetic Algorithm based method for cluster optimization.

Input: Collection of Clusters C_j, and associated mean keyword vector $clust\_mean_j$

Output: Optimal Set of Clusters C_j

1. Generate the population of chromosomes where each chromosomes is built by taking randomly the query session vector id from each cluster$_j$ and each query session vector id represents the gene of the chromosome.
2. The length of the chromosome is equal to the number of genes which in turn is equal to the number of clusters. Each cluster is allocated fixed gene position in all chromosomes.
3. For a given clustering solution represented by set of chromosomes in a given population the entropy is calculated as fitness value using eq (8) and the global minimum is extracted.
4. In a given population, the genetic operators such as crossover and mutation operator are applied. The uniform crossover is performed where the crossover probability lies in [0.2-0.8] and single point mutation is applied with the mutation probability lies in [.005-0.3] to produce next generation of new population. While applying crossover operator, the cluster points will get shuffled means that a point can move from one cluster to another. The purpose of the mutation is to randomly select the gene in the chromosome and replace the query session id with the id not already present in the chromosome.
5. The clustering solution for a given generation of population is set of clusters obtained by collecting all the query session id present at the specific position in all chromosomes in order to identify the points in a given cluster where each cluster is allocated fixed position in all chromosomes.
6. The Fitness value is calculated for a given clustering solutions using eq (8).
7. If the fitness value representing the local minimum for the new generation of population obtained in step 6 is less than the global minimum, then this local minimum becomes global minimum and **Delete-all replacement** technique is applied which deletes all the members of the parent population and replaces them with the same number of chromosomes in the new population. Otherwise the next iteration continues with the same parent population.
8. Repeat step 4-7 till the terminating condition is obtained. The terminating condition can be the fixed N number of iterations or the change in the global fitness value is less than 0.0001 in last 100 trials.
9. The points in the clusters will be repositioned corresponding to the clustering solution having global minimum obtained in step 8.

In Phase II online processing is performed. During online processing, the optimal set of clusters obtained in Phase I is used for PWS. The user input query is used to select the most similar cluster. The selected cluster is used to recommend the clicked URLs associated to it. The user response to the clicked URLs on the current page are tracked and stored in his profile. As the user request for next page, the user profile is transformed into keyword vector and is used to select the most similar cluster. The selected cluster is used to recommend the clicked URLs and this process of recommendations continue till the search is personalized to the information need of the user. The stepwise execution of online processing is given below .

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 is 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.

Phase II

Online Processing.

Input: Optimal set of Clusters

Output : Recommended Set of Clicked URLs

1. The search query entered by the user is used to select the optimal cluster which is most similar to the information need of the keyword based user input query and is measured using cosine similarity measure.
2. The High Scent clicked URLs associated with the selected cluster i are recommended to the user.
3. The user's response to the recommended clicked URLs are tracked and stored in user's profile.
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. Select the cluster which is most similar to the information need associated with the $current\_usersessionvector_t$

c. Goto step 2.else

d. Current search session is terminated.

The user clicks on the retrieved search results, are captured through the check boxes displayed on the GUI and stored in the 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 k-means algorithm is then applied to group the similar information need query session keyword vector in clusters.

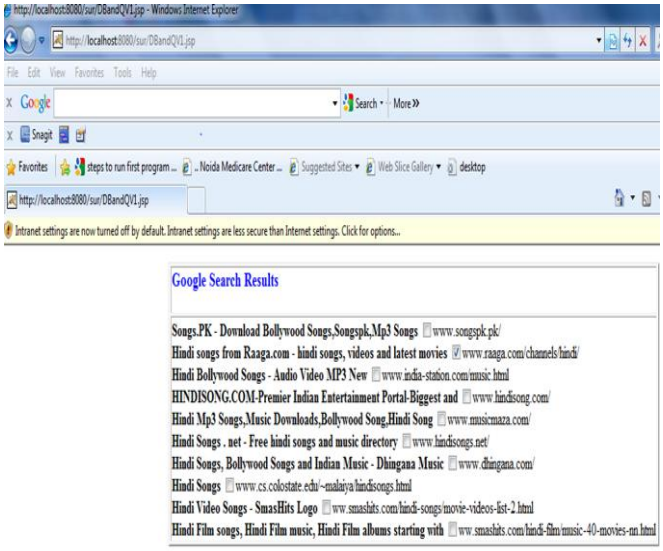

Fig.1: Screen SnapShot of architecture displaying Google Search results along with the checkboxes

The experiment was performed on Pentium IV PC with 120 GB RAM under Windows 8 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 cluster optimization, the values of following parameters are used in the genetic algorithm : MAXGEN, length(P), 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(P) 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 $\tau = 0.0001$.

In this study, the experiment was conducted with the following values of selected parameters- the size of the population represented as length(P) was m where m depends on the number and size of clusters, 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 P 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 and mutation rate of 0.25 and threshold value of Information Scent (ρ) at 0.5 for the data set generated in this experimental study.

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 of query session keyword vectors. The genetic algorithm tool box of MATLAB software package was used for applying the genetic algorithm on the clustered data set using Algorithm 1. 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 refined set of clusters in the database for the later retrieval for personalized web search.

The approach proposed for PWS(with cluster optimization) was compared with PWS (without cluster optimization) in [6] in order to determine the effectiveness of PWS with cluster optimization in better satisfying the information need of the user .

During online processing, the input query is issued to GUI based interface designed for both PWS (with/ without cluster optimization). The input query is used to select the cluster(with/without optimization) most similar to the information need of the user. The set of clicked URLS associated with the selected cluster are recommended and displayed with checkboxes in the GUI Interface for capturing the user's clicks.

=:

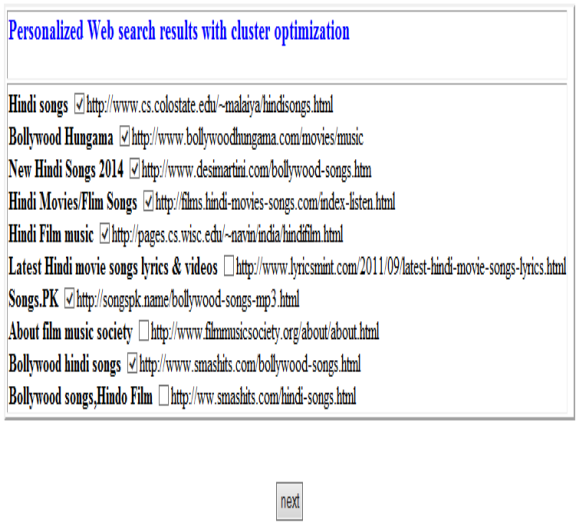

Fig. 2: Shows the Personalized Web Search results(with cluster optimization) are shown with CheckBoxes to capture the user clicks.

The Fig 2. And Fig 3. shows the Personalized Results (with/without cluster optimization). 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 cluster similar to the information need of the current user profile. The selected cluster is used to recommend the set of clicked URLs for the next requested result page. This process of clicked URLs recommendations for personalized web search continues till the search is personalized to the need of the user.

The performance of PWS (with cluster optimization) is compared with Personalized Search Results(without cluster optimization) in each of the selected domains (Academics, Entertainment and Sports).

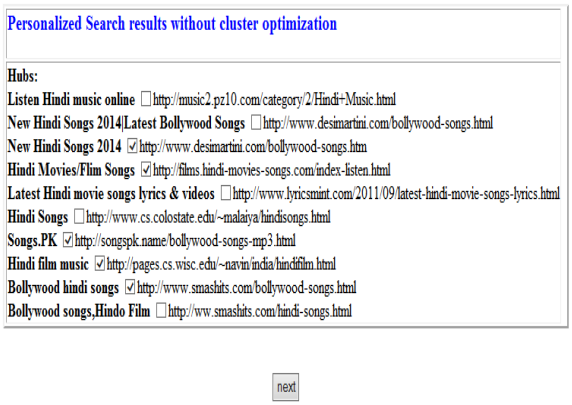

Fig.3: Shows the Personalized Web Search results without using cluster optimization are shown with CheckBoxes to capture the user clicks.

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 cluster optimization) . 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 cluster optimization) are shown in Fig 4.

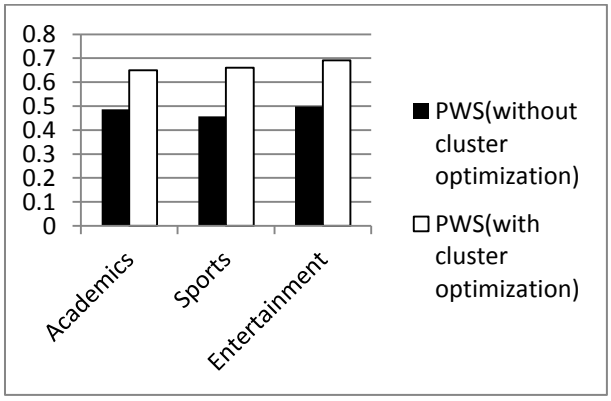

Fig.4: Shows the avgprecision of PWS (without/with cluster optimization) in Academics, Sports and Entertainment.

The average precision is improved in each of the selected domains using personalized web search (with cluster optimization). The obtained results were analyzed using the statistical paired t-test for average precision of PWS (with / without cluster optimization). 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 65.59 for the combined sample. Value of t for paired difference of

average precision was 54.37 for academics, 51.00 for entertainment and 50.73 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 when personalized web search (with cluster optimization) in comparison to improvement in average precision of Personalized search results (without cluster optimization).

This proves that Personalized Web Search based on cluster optimization personalizes the web search more effectively by overcoming the problem of misclassified data to clusters. Thus during online web search more and more relevant documents associated with selected cluster are retrieved and is responsible for the improvement in the average precision of test queries in each of the selected domains. The experimental results which were also verified statistically confirm the significant improvement in precision when compared to PWS (without cluster optimization). Hence PWS based on cluster optimization recommend more and more relevant documents early in search results and is responsible for the improvement in the average precision of search results.

6. CONCLUSION

In this paper genetic algorithm is used for cluster optimization in order to increase the quality of clusters for effective Personalization of Web Search using these optimal set of clusters of query sessions. The experiment was conducted on the data set captured in three domains entertainment, sports and academics to compare the performance of PWS (with/without cluster optimization). The average precision of personalized search results is improved with cluster optimization and effectively satisfies the information need of the user.

7. REFERENCES

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

[2] Arzanian, B., Akhlaghian, F., and Moradi, P. 2010. 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.

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

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

[5] Bremermann, H. J. 1958. 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.

[6] Chawla, S., and Bedi P.2007. 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.

[7] Chawla, S., and Bedi, P.2008. Improving information retrieval precision by finding related queries with similar information need using information scent. In First International Conference on Emerging Trends in Engineering and Technology, ICETET'08. (pp. 486-491). IEEE.

[8] Chawla, S. 2012a. Trust in Personalized Web Search based on Clustered Query Sessions. International Journal of Computer Applications, 59(7), pp. 36-44.

[9] Chawla, S.2012b. Semantic Query Expansion using Cluster Based Domain Ontologies. International Journal of Information Retrieval Research (IJIRR), 2(2), pp. 13-28.

[10] Chawla, S.2013. Personalised web search using ACO with information scent. International Journal of Knowledge and Web Intelligence, 4(2), pp. 238-259.

[11] Chawla, S. 2014a. Personalized Web Search using Trust based Hubs and Authorities. International Journal of Engineering Research and Applications, 4(7), pp. 157-170.

[12] Chawla, S.2014b. Novel Approach to Query Expansion using Genetic Algoirthm on Clustered Query Sessions for Effective Personalized Web Search . International Journal of Advanced Research in Computer Science and Software Engineering, 4(11), pp. 73-81.

[13] Chi, E H., Pirolli, P., Chen, K., and Pitkow, J. 2001. 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.

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

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

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

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

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

[19] Kanungo, T., Mount, D. M. , Netanyahu, Nathan S., Silverman, Christine D. Piatko, Ruth and Wu, A. Y.2002. "An Efficient k- Means Clustering Algorithm:Analysis and Implementation.", IEEE

Transactions on pattern analysis and machine intelligence, 24(7), 881-892.

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

[21] Kim, H., Lee, S., Lee, B., and Kang, S.2010. Building Concept Network-Based User Profile for Personalized Web Search, 9th International Conference on Computer and Information Science ,pp. 567 – 572.

[22] Leung, K.W.-T. , Ng, W. , and Lee, D.L.2008. Personalized Concept-Based Clustering of Search Engine Queries, Journal IEEE Transactions on Knowledge and Data Engineering, 20(11), pp. 1505 – 1518.

[23] Liu, F., Yu,, C., and Meng, W.2004. Personalized Web search for improving retrieval effectiveness", Journal IEEE Transactions on Knowledge and Data Engineering, 16(1), pp. 28 – 40.

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

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

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

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

[28] Peng, Wen-Chih, and Lin, Yu-Chin.2006. 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.

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

[30] Pirolli, P.2004. 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.

[31] Rekha, C., Sujatha, N. , and Iyakutti, K.. 2011 ,Algorithm to Improve the Cluster Quality using Genetic Algorithm, Research Journal of Computer Systems Engineering-RJCSE, 2(4).

[32] Smith, M.P, and, Smith, M.1997. The use of genetic programming to build Boolean queries for text retrieval through relevance feedback", Journal of Information Science, 23 (6), pp. 423–431.

[33] Tamine, L., Chrisment, C., and Boughanem, M.2003. Multiple query evaluation based on an enhanced genetic algorithm, Information Processing and Management , 39(2), pp. 215–231.

[34] Vrajitoru, D. 1998. Crossover improvement for the genetic algorithm in information retrieval, Information Processing& Management, 34(4), pp. 405–415.

[35] Wen, J. R., Nie, J. Y, and Zhang, H. J.2002. Query Clustering Using User Logs", Journal ACM Transactions on Information Systems, 20(1), pp. 59-81.

[36] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B. , Yu, P.S. , Zhou, Z.-H., Steinbach, M. , Hand D.J.,and Steinberg, D.2007 ―Top 10 Algorithms in Data Mining", ―Knowledge Information Systems, 14(1) pp. 1-37.

[37] Yang, J., Korfhage , R.R., and Rasmussen, E.1992. 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.

[38] Zhao, Y and Karypis, G. 2001. Criterion functions for document clustering: Experiments and Analysis. Technical report,University of Minnesota, Minneapolis, MN.

[39] Zhao, Y. and Karypis, G. 2002. Comparison of agglomerative and partitional document clustering algorithms, SIAM Workshop on Clustering High-dimensional Data and its Applications.

[40] Zhu, Z , Xu, J., Ren, X., Tian, Y. and Li, L.2007. Query Expansion Based on a Personalized Web Search Model, Third International Conference on Semantics, Knowledge and Grid, pp. 128 – 133.