Volume 4, Issue 11, November 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Novel Approach to Query Expansion Using Genetic Algorithm on Clustered Query Sessions for Effective Personalized Web Search

Dr. Suruchi Chawla

Assistant Professor, Shaheed Rajguru College of Applied Science for Women,
Vasundhra Enclave, University of Delhi, India

Abstract— Research has been done based on query expansion to bridge the gap between vocabulary of user search query and documents relevant to user in order to retrieve relevant documents early in search results. In this paper novel approach is proposed for recommending optimal set of terms for query expansion using genetic algorithm based on clustered query sessions. The genetic algorithm is applied on clustered query sessions for user input queries optimization in order to identify the optimal set of expansion terms relevant to the domain of cluster. The initial input query issued for search on the web is used to select the cluster for recommending the set of terms for query expansion. The process of retrieval of web search results using input query expanded with the selected terms along with the recommendations of next set of query terms for query expansion continues till the search is personalized according to the information need of the user. Experiment was conducted on the data set of query sessions captured on the web in three domain Academics, Entertainment and Sports and results show the significant improvement in the precision of search results using the proposed approach.

Keywords— Information Retrieval, Personalized Web Search, Genetic Algorithms, Search engines, Query Expansion.

I. INTRODUCTION

Information on the Web is huge and growing in size. User uses the keyword based input query to search the web using search engines. This keyword based input query is composed of very few terms and most of the times, the keywords used in the input query does not matches the words used in documents relevant to the query and hence these relevant documents cannot be fetched from the web. In this paper novel approach is proposed for automatic query expansion using genetic algorithm based on clustered query sessions. Work related to the proposed approach has been done in [28] for automatic expansion based on clustered query sessions where the user input query is used to select the cluster for related queries terms suggestions. This process of related queries terms recommendations continue till the query expansion with the selected terms personalizes the web search to the information need of the user. The related terms associated with the cluster are the terms which are present in the clicked documents of the cluster but there are some relevant documents which are not present in the cluster. These relevant documents could not be retrieved due to insufficient terms of the user input queries and hence cannot be clicked and contribute terms for query expansion. Therefore the terms suggested for query expansion may lack the significant terms for effective automatic query expansion.

In the proposed approach of automatic query expansion using genetic algorithm based on clustered query sessions, initial set of input queries associated with each cluster form the population of chromosomes in the execution of genetic algorithm. The terms of queries chromosomes are evolved using crossover and mutation operators. In crossover, two parent query chromosomes are selected for swapping the terms between each other to generate the child chromosomes. Mutated user queries are the queries in which some terms of the query is replaced with terms found in mutation pool specific to the domain of cluster. The terms in the mutation pool are extracted from the documents relevant to the domain of cluster which could not be retrieved due to small size input queries. Thus these queries chromosomes are evolved using mutation and crossover genetic operators by adding more and more terms used in documents relevant to queries at each generation till the terminating condition is reached. Upon termination, the queries obtained in the last generation contain sufficient significant terms to retrieve more and more relevant document early in search results. Thus the terms pool is generated for each cluster which collect the terms found in queries obtained in the last generation after optimization. This pool of terms associated with each cluster is recommended for input query expansion in order to retrieve more and more relevant documents early in search results using the proposed approach.

The algorithm is proposed for Personalized Web Search based on clusterwise terms suggestions for automatic query expansion using genetic algorithm. The entire processing of the algorithm is divided into two Phases: Phase I and Phase II. In Phase I, offline processing is performed, the query sessions containing the input query and the associated clicked URLs collected on the web are processed to generate the query sessions keyword vector. This query session keyword vectors are then clustered using k-means algorithm in order to group similar information need query session keyword vectors in clusters. The genetic algorithm is applied on each cluster in order to identify the optimal set of terms relevant to the topic associated with the cluster. In Phase II, online processing is performed. In online processing, the user input query issued for search on the web is used to retrieve the web search results and the cluster is selected which is

most similar to input query for related queries term recommendations. The terms selected by the user are added to the input query each with the weight 1. As the user request for next result page, this expanded input query is issued to the search engine to retrieve the search results and the user session keyword vector generated from the clicked URLs of the current user search query session is used to identify the most similar cluster for the recommendations of next set of queries terms. This process of retrieval of search results corresponding to the expanded input query and recommendations of related queries terms suggestion on each page continues till user search is personalized according to his information need.

Experiment was conducted on the data set of query sessions captured on the web in the Academics, Entertainment and Sports domain. The experimental results were compared with the PWS using automatic query expansion (without genetic algorithm) based on clustered query sessions in [28]. The experimental results confirm the effective improvement in average precision using the proposed approach.

II. RELATED WORK

Research has been done on query expansion using number of different existing approaches like work based on query log [35][21], general or domain-specific ontology [17][8][10], user profile and collaboratively filtering [15], selecting popular words in the results [3][9][31][11][19][20]. In order to generate new queries, methods based on TF.IDF(term frequency inverse document frequency) are proposed in [11] , probabilistic language model in [29],vector space model in [32] and only term frequency is considered in [31]. In [20] proximity to the original query keywords is considered when selecting words from results or corpus to compose new queries. In [19] products attributes are selected as expanded terms based on co-occurrence patterns, extreme rating and consistent rating. Query expansion is also related to the topics of faceted search in [14][1][4], cluster labeling/summarization in [5][18], result differentiation in [36]. In [28] a method is proposed for queries expansion 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 terms for query expansion in specific domain. This process of recommendations continues till the search using expanded query is personalized to the information need of the user. The effectiveness of this method is confirmed with the experimental results.

In [16] 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. Genetic Algorithm(GA) inspired from natural theory of evolution has the ability to work on many solutions in parallel. The solution space is explored in multiple direction at one time and has a better chance at finding the true global maximum of the system on first try.[2] Due to parallel nature of genetic algorithms it is much more efficient at navigating vast space than traditional algorithms. GA being insensitive to the initial condition imposed on them discards any solution that is not promising. Due to this feature of GA it is more flexible, robust and simple in design.[25] Thus GA almost always produce a relatively very good solution to the problem at hand.[2] In [30] a hybrid of Genetic Algorithms and Fuzzy Logic was applied for personalized search results where the Fuzzy set techniques were used for better document modeling and genetic algorithm was used for query optimization.

Work related to the approach proposed in this paper has been done in [28]. But it is found that in [28] the pool of terms specific to each cluster are derived from clustered clicked documents which are retrieved using user initial input queries issued on the web. However there are some relevant documents on the web which could not retrieved due to limited keywords used in the input query. Thus these relevant documents could not be clicked and hence not present in the cluster. Hence this pool of terms still misses some of the significant terms belonging to those relevant documents that are not present in the cluster. Therefore pool of terms lack some of the significant terms for effective query expansion.

The research in this paper is motivated to use the optimization like Genetic Algorithm to refine the initial set of user queries associated with each cluster by adding domain specific terms found in web documents relevant to the user queries using genetic operators like mutation and crossover. However these user queries evolve from one generation to another and evaluated on the basis of fitness function till the terminating condition is reached. The optimal set of queries obtained in the last generation for each cluster is used for generation of pool of terms for a given cluster. The generation of pool of terms from these optimal queries for query expansion are likely to retrieve more and more relevant documents early in the search results and hence satisfies the information need of the user effectively. Hence query expansion based on pool of terms generated using query optimization bridges the gap between the vocabulary of search queries and keywords used by the authors in the relevant documents found on the web.

Thus the method is proposed for Personalized Web Search with query expansion based on clusterwise query optimization using genetic algorithm. The web search results are adapted according to the user query expanded with terms selected from clusterwise pool of optimal terms recommended on each requested web page.

III. BACKGROUND

A. Genetic Algorithm

Genetic Algorithm is a search method based on the natural theory of evolution [12]. 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. [22]

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. [6]

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.

B. 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. [23][24]

1) Information Scent metric:

The Inferring User Need by Information Scent (IUNIS) algorithm is used to quantify the Information Scent s of the pages d clicked by the user in i[th] query session. [7][13]

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

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

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

$PF.IPF(P_i$: PF correspond to the page d normalized frequency f in a given query session i where n is the number of distinct clicked page in session i and IDF correspond to the ratio of total number of query sessions M in the whole data set to the number of query sessions m that contain the given page .

$Time(P_i$: It is the ratio of time spent on the page i in a given session i to the total duration of query session i. [27]

***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 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

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

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 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)* Clustering of Query session keyword vector:**

The k-means algorithm is used for clustering query sessions keyword vectors since its performance is good for document clustering. [26][33]

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:

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

where be a cluster found in a k-way clustering process (p∈1..k) , is the centroid of p^{th} cluster , is the vector representing some query session belonging to the cluster and is the total number of query sessions in all clusters as defined below .[34]

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

The centroid of the cluster is defined as below:

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

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

IV. PERSONALIZED WEB SEARCH WITH QUERY EXPANSION BASED ON CLUSTERWISE SEARCH QUERIES OPTIMIZATION USING GENETIC ALGORITHM.

In this paper an algorithm is proposed for Query expansion in which genetic algorithm is applied on clustered query sessions in order to generate the pool of optimal terms in a specific domain for recommendations. The genetic algorithm is applied on clustered query sessions keyword vectors where query session keyword vector is generated from user query session using Information Scent and content of clicked URLs present in the query session. These query session keyword vectors are clustered using k-means algorithm and a given cluster group similar information need query sessions. The genetic algorithm is applied on each cluster by representing the queries in a given cluster as the set of query chromosomes. The single point crossover and the single point mutation are the genetic operators which are applied on the query chromosomes. In single point crossover, two parent query chromosomes are selected and terms of the query chromosomes after the selected point are swapped with each other to generate the child chromosomes. In single point mutation, random term of parent chromosome is selected and replaced by the term found in the mutation pool and not already present in a given chromosome. Mutation pool is the set of terms extracted from the documents relevant to the domain of the cluster. These genetic operators are applied on the query chromosomes in the current generation to generate the next generation of chromosomes. At each generation, fitness function is calculated for each chromosome which measures the ability of the query chromosome to retrieve the maximum relevant document from the web. The chromosomes are evolved in such a way so that maximum fitness value of query chromosome is increased with each generation produced. This process of evolving the population of chromosomes using genetic operators continues from one generation to another till the terminating condition is reached. Upon the termination, the distinct terms of the query chromosomes in the last generation is collected and stores in the term set ET_j associated with a given cluster.

The algorithm is proposed for Personalized Web Search with query expansion using Genetic Algorithm on clustered query sessions. The entire processing of the algorithm is divided into two phases: Phase I, Phase II.

In Phase I, offline processing is performed during which query sessions keyword vector are generated and clustered. The genetic algorithm is applied on each cluster to generate the pool of terms associated with each cluster for further use in query expansion. The stepwise description of Phase I 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.
5. Each cluster j is associated with the mean keyword vector $clust\_mean_j$.
6. For each cluster j maintain the list of Queries in Q_j.
7. For each cluster j apply the algorithm **Genetic Algorithm for terms pool generation using clustered query sessions** on the List Q_j . associated with the cluster j to determine the pool of terms for queries expansion associated with each cluster and is represented by ET_j (Optimal Queries j).

Algorithm 2:

Genetic Algorithm for terms pool generation using clustered query sessions

Input: List Q_j, cluster mean keyword vector $clust\_mean_j$

Output: Optimal Term set , ET_j

1. The set of distinct queries associated with a given $cluster_j$ forms the population of chromosomes where each chromosome represents the user query associated with a given cluster. The number of genes in each chromosome is equal to the number of distinct terms associated with all user queries in a given cluster. Every term of the user query is represented by the non zero weight in the gene position allocated to that term in the chromosome.
2. Once the population is initialized with the chromosomes, the fitness value of the candidate solutions represented by chromosomes is evaluated. Each chromosome C representing the user query q is evaluated using the Fitness Function. Fitness function Fitness(q) is defined as follows

   $$Fitness(q) = \max_{d_i \in A_q} (\sigma(t, d_i))$$

   Where t is the mean keyword vector associated with a given $cluster_j$ and A_q is the set of top 10 URL retrieved from search engine using query q. σ is the similarity measure for a pair of documents. Only the snippets of d_i returned by the search engine are used for computing the similarity. The mutation pool is a set of terms that initially contains terms extracted from the description of the topic under analysis. As the system collects relevant content, the mutation pool is updated with new terms from the snippets of the relevant documents returned by the search engine. This procedure brings new terms to the scene, allowing a broader exploration of the search space.
3. Select those chromosomes which have the highest Fitness value using Tournament selection and also followed Elitism which copies the best chromosome (or a few best chromosomes) to new population without mutation and crossover.
4. Apply the single point crossover and single point mutation with mutation probability 0.25 and crossover rate of 0.8 on the selected chromosomes not included in Elitism.
5. Apply the steady-state-no-duplicates replacement policy to replace the population of parent chromosome with the reproduced offspring chromosomes obtained in step 3 and 4 in order to generate the next generation of population P.
6. Goto step 2 until the required number of n1 iterations or terminating conditions is satisfied where the difference between the optimal Fitness values of last 50 generation is less than the threshold value τ.
7. Upon termination, collect the terms of the query chromosomes in the last generation in the set ET_j associated with the cluster. The queries represented in the last generation of population have been evolved using crossover and mutation from one generation to next based on their ability to retrieve relevant results when presented to a search engine and is of high fitness value in comparison to the initial set of queries represented in first generation of population.

In Phase II, online processing is performed. During online processing, the input query issued for web search is used to retrieve the search results from the web and at the same time it is used to select the most similar cluster. The selected cluster is used to suggest pool of terms for the query expansion along with the retrieved search results. The user selected terms are added to the input query for query expansion and user's response to the search results is stored in the user's profile. As the user request for next result page, the expanded input query is issued on the web to retrieve the search results for the next page and at the same time user session keyword vector generated from user session profile is used to select the cluster for query terms suggestion on the next result page. Thus search result retrieval is personalized according to the expanded input query and query terms suggestion is also personalized according to the user access patterns on the retrieved web search results. This process of optimal term pool recommendation in a specific domain along with the search retrieval using expanded input query continues till user search is personalized according to his needs. The stepwise description of Phase II is given below.

Phase II

Online Processing.

1. The search query entered by the user is used to retrieve the search results on the web and at the same time select the j^{th} 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 Optimal query term set ET_j associated with the cluster j is selected.
3. The selected ET_j is presented to the user.
4. If the user request for the next result page
    a. The users clicks to the search results on the previous page have been tracked and current user input query with selected expansion terms are stored in user profile.
    b. The user expanded input query where each added term is given weight 1 in query vector is used to retrieve the search results from web for the next requested result page.
    c. 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 $current\_usersessionvector_t$.
    d. Select the j^{th} cluster which is most similar to the information need associated with the $current\_usersessionvector_t$
    e. The Optimal query term set ET_j associated with the cluster j is selected.
    f. The selected ET_j is presented to the user along with the search results obtained in step b.
    g. Goto step 4.

else

Current search session is terminated

V. EXPERIMENTAL STUDY

The experiment was conducted on a data set of user query sessions collected on the web. The architecture is developed to generate the data set of query sessions by capturing the URLs clicked by users in the Google search results. The user is required to enter the input query through a GUI based interface of the architecture in order to retrieve the Google search engine results. These search results are 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.

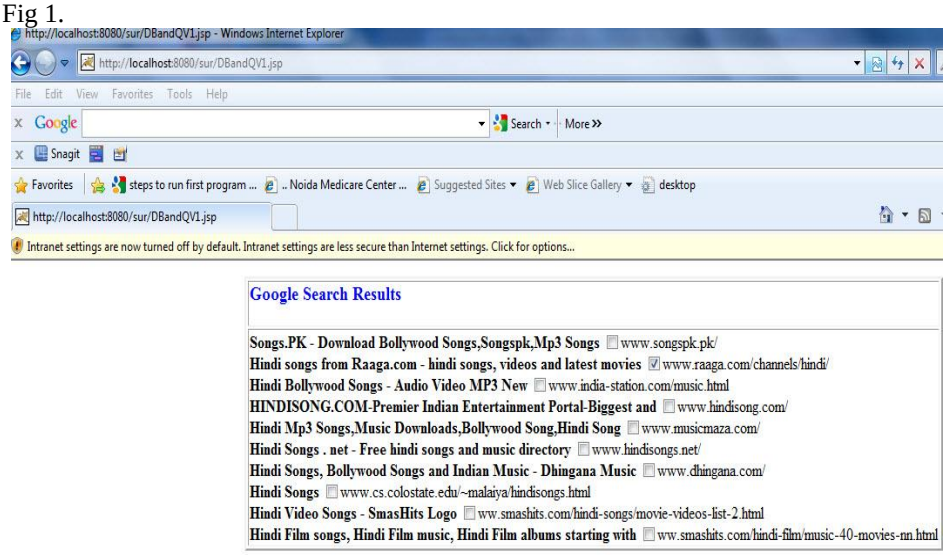

Fig.1. Screen SnapShot of architecture displaying Google Search results along with the checkboxes.

The experiment was performed on the 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 using query expansion based on genetic algorithm, 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 50 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(P) was m for each cluster where m is the number of user queries in set Q_j associated with each j^{th} 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 generations 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 of the proposed approach, the user clicks on the retrieved search results, are captured through the check boxes displayed on the GUI and stored in the database. 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 captured user query sessions on the web are processed further to find the query session keyword vector using Information Scent and content(tf.idf) of clicked URLs. The clustering agent developed in JADE is executed to generate the clusters of query session keyword vectors. The Genetic algorithm is performed on each cluster in order to get the term set associated with each 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, single point crossover, fitness function and output function are defined by the user in MATLAB. The output function is defined in MATLAB for storing ET_j set of terms associated with the given cluster in the database for the later retrieval for personalized web search.

The approach proposed for PWS using query expansion based on genetic algorithm was compared with approach used for improving information retrieval precision with query expansion(without genetic algorithm) based on same clustered query sessions in [28] in order to determine the effectiveness of PWS with query expansion using genetic algorithm 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 Query expansion(with/without genetic algorithm). In PWS with query expansion(with genetic algorithm), the input query is used to select the cluster most similar to the information need of the user. The set of query terms associated with the selected cluster are recommended along with the web search results retrieved for the current user input query displayed with checkboxes in order to capture the user's clicks.

Fig. 2. Shows the Personalized Web Search results with Optimal queries recommendations along with the search results are shown with CheckBoxes to capture the user clicks.

The Fig 2. above shows the search results for the current input query 'hindi song' along with the optimal terms set recommendations for query expansion. The user's clicks to the 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 for generating the next set of terms for query expansion along with search results retrieved using search query expanded with selected terms from the previous page. This process of recommendations of optimal term set for query expansion along with the personalized retrieval of web search results using expanded input query from the previous page continues till the user search is personalized to the need of the user.

The performance of PWS using query expansion based on genetic algorithm is evaluated from the average precision of Personalized Search Results and compared with average precision of Search Results with query expansion(without genetic algorithm) in each of the selected domains (Academics, Entertainment and Sports). The GUI of the Personalized Results with queries expansion(without genetic algorithm) is given below in Fig 3.where user's clicks are captured through checkbox ticks.

Fig. 3. Shows the Personalized Web Search results without Optimal Queries 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.

During online searching, the test queries were issued in each of the selected domain to the GUI based interface to retrieve the personalized search results with query expansion based on genetic algorithm/with query expansion(without genetic algorithm). The average precision is computed using the fraction of retrieved documents which are relevant in the personalized search results. The experimental results in Fig 4 shows the average precision of test queries computed in the domains of academics, entertainment and sports using PWS with query expansion(with/without Genetic Algorithm).

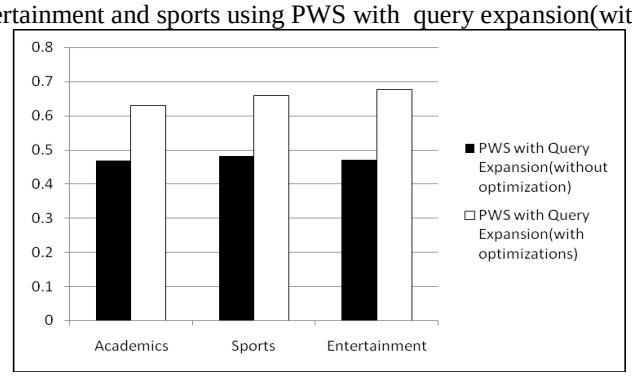

Fig.4. Shows the avgprecision of PWS with query expansion(with/without optimization(Genetic Algorithm)) in Academics, Sports and Entertainment.

The average precision is improved in each of the selected domains using personalized web search with query expansion based on genetic algorithm. The obtained results were analyzed using the statistical paired t-test for average precision of PWS with query expansion (with Genetic Algorithm/ without Genetic Algorithm) . The test was conducted on the 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 62.91038 for the combined sample. Value of t for paired difference of average precision was 43.46043 for academics, 61.59097 for entertainment and 55.29328 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 using query expansion(with genetic algorithm) in comparison to improvement in average precision of search results with query expansion (without genetic algorithm).

This proves that use of genetic algorithm in generating the term set for query expansion in PWS add some terms in the web queries which are used in web documents relevant to the user initial query and hence generates a higher number of relevant clicked URLs up in top ranked clicked documents and increases their probability of being clicked by the users. The increase in the ratio of relevant documents retrieved to the total documents retrieved is responsible for the improvement in the average precision in each of the selected domains. Thus the query expansion based on term set generated using genetic algorithm personalizes the web search more effectively with respect to the information need of the user. The experimental results which were also verified statistically confirm the significant improvement in precision when compared to PWS with query expansion (without Genetic Algorithm). Hence PWS using query expansion based on genetic algorithm are more likely to return more and more relevant documents early in search results and is responsible for the improvement in the average precision of search results using the proposed approach.

VI. CONCLUSIONS

In this paper novel approach is proposed for query expansion based on clusterwise optimal term set generated using Genetic Algorithm for effective personalization of web search. The performance of the Personalized Web Search based on query expansion using optimal set of terms algorithm was evaluated on the data set of query sessions captured in domain(Academics, Entertainment and Sports) in order to determine its effectiveness. The effectiveness of proposed algorithm is compared with Personalized Web Search with query expansion(without genetic algorithm). The experimental results verified statistically confirm the improvement in the precision of search results using the Query expansion with optimal set of terms.

REFERENCES

[1] A. Kashyap, V. Hristidis, and M. Petropoulos. *FACeTOR: Cost-Driven Exploration of Faceted Query Results*. In CIKM, pages 719–728, 2010.

[2] Adam Marczyk, *Genetic Algorithms and Evolutionary Computation* The Talk.Origins Archive, 2004. Retrieved December 4, 2004 from the World Wide Web: http://www.talkorigins.org/faqs/genalg /genalg.html

[3] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. *An Information-Theoretic Approach to Automatic Query Expansion*. ACM Trans. Inf. Syst., 2001, 19(1):1–27.

[4] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das. *Facetedpedia: Dynamic generation of query-dependent faceted interfaces for wikipedia*. In WWW, 2010, pp 651–660.

[5] D. Carmel, H. Roitman, and N. Zwerdling. *Enhancing Cluster Labeling Using Wikipedia*. In SIGIR, 2009, pp 139–146.

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

[7] E H. Chi, P. Pirolli, , K. Chen, & J. Pitkow, *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, 2001, pp. 490-497.

[8] F. A. Grootjen and T. P. van der Weide. *Conceptual Query Expansion*. Data Knowl. Eng., 2006, 56(2):174–193.

[9] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. *Selecting Good Expansion Terms for Pseudo-Relevance Feedback*. In SIGIR, 2008, pp. 243–250.

[10] G. Fu, C. B. Jones, and A. I. Abdelmoty. *Ontology-Based Spatial Query Expansion in Information Retrieval*. In OTM Conferences (2), 2005, pp 1466–1482.

[11] G. Koutrika, Z. M. Zadeh, and H. Garcia-Molina. *Data Clouds: Summarizing Keyword Search Results over Structured Data*. In EDBT, 2009, pp. 391–402.

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

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

[14] K. Chakrabarti, S. Chaudhuri, and S. won Hwang. *Automatic Categorization of Query Results*. In SIGMOD Conference, 2004, pp. 755–766.

[15] L. Fu, D. H.-L. Goh, and S. S.-B. Foo. *Evaluating the Effectiveness of a Collaborative Querying Environment*. In ICADL, 2005, pp. 342–351.

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

[17] M. Baziz, M. Boughanem, and N. Aussenac-Gilles. *Conceptual Indexing Based on Document Content Representation*. In CoLIS, 2005,pp. 171–186.

[18] M. Muhr, R. Kern, and M. Granitzer. *Analysis of Structural Relationships for Hierarchical Cluster Labeling*. In SIGIR, 2010, pp. 178–185.

[19] N. Sarkas, N. Bansal, G. Das, and N. Koudas. *Measure-driven keyword-query expansion*. PVLDB, 2009, 2(1):121–132.

[20] O. Vechtomova, S. E. Robertson, and S. Jones. *Query Expansion with Long-Span Collocates*. Inf. Retr., 2003, 6(2): pp.251–273.

[21] P.-A. Chirita, C. S. Firan, and W. Nejdl. *Personalized Query Expansion for the Web*. In SIGIR, 2007, pp. 7–14.

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

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

[24] P. Pirolli, *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.

[25] Richard Baker, *Genetic Algorithms in Search and Optimization*. Financial Engineering News, 1998. Retrieved December 4, 2004 from the World Wide Web: http://www.fenews.com/fen5/ga.html

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

[27] S. Chawla, & P. Bedi, *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), 2007, pp. 483-488.

[28] Suruchi Chawla and Punam Bedi. *Improving Information Retrieval Precision by Finding Related Queries with similar Information need using Information Scent*. Proc, ICETET'08 – The 1st International Conference on Emerging Trends in Engineering and Technology, (Proceedings published by IEEE Computer Society Press and Papers also available in IEEE Xplore, 2008, pp.486-491, July 16-18.

[29] S. E. Robertson. *On Term Selection for Query Expansion.* Journal of Documentation, , 1990, 46: pp 359–364.

[30] V. Snasel, A. Abraham, , S. Owais, , J Platos,., & P. Kromer, *Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System*, Foundations of Computational Intelligence, 2009,4, pp 299-324.

[31] Y. Tao and J. X. Yu. *Finding Frequent Co-occurring Terms in Relational Keyword Search*. In EDBT, , 2009, pp 839–850.

[32] Y. Xu, G. J. F. Jones, and B. Wang. *Query Dependent Pseudo-Relevance Feedback based on Wikipedia*. In SIGIR, , 2009,pp. 59–66.

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

[34] Y Zhao, & Y. Karypis, *Criterion functions for document clustering: Experiments and Analysis*. Technical report,University of Minnesota, Minneapolis, MN, 2002.

[35] Z. Bar-Yossef and M. Gurevich. *Mining Search Engine Query Logs via Suggestion Sampling*. PVLDB, 1(1): pp 54–65, 2008.

[36] Z. Liu, P. Sun, and Y. Chen. *Structured Search Result Differentiation*. PVLDB, 2009, 2(1): pp. 313–324.

