Improving the Performance of User Search Goals Using Noval Approach

J.S.HARILAKSHMANRAJ,

PG SCHOLAR

M.E.(Software Engineering)

SRI RAMAKRISHNA ENGINEERING COLLEGE,

COIMBATORE, INDIA

hariproffessor@gmail.com

ABSTRACT

World Wide Web acts as storage repository and it used for an information retrieval. while, fetching information through user queries, a search engine in a large and un-effective collection of document. To avoid these conflict and to improve the search engine, I propose ***Novel based approach***. This approach states two strategies: ***clustering and prioritation***. Based on the user search goals the information is been restructured dynamically and it is been clustered and based on ranking algorithm, it is been prioritized.

Keyword: Clustering, ranking-algorithm, prioritization process, World Wide Web.

1. INTRODUCTION

The World Wide Web (WWW) is an large informatics center. It is been an storage repository for an millions of web pages is interlinked with each other [5].It contain information like text, audio, video ,image, e.t.c.., It is estimated that WWW has expanded by about 2000 % since its initial growth was doubling to its size of every six to ten months [1].Due to rapid growth of information sources available on the WWW and growing needs of users ,it is becoming difficult to manage the information on the web and satisfy the user needs. To represent the user search, queries may not exactly represent users' specific information. for an example, if an user search for an newspaper "THE SUN" United kingdom newspaper ,but user will get the result regarding Solar-system "SUN", it is been an irrelevant result. Using web restricting method, the result relevant to the user search is been restructured at time-bound manner [2],[3],[4]. Based on the restructured data-set ,clustering of relevant data-sets are been done. Clustering is used as a data processing technique in many different areas, including artificial intelligence, bioinformatics, biology, computer vision, city planning, data mining, data compression, earth-quake studies, image analysis, image segmentation, information retrieval, machine learning, marketing, medicine, object recognition, pattern recognition, spatial database analysis, statistics and web mining.

In this paper, I propose novel approach to predict the user search query, based on the user search ,queries flaws from multiple source are been retrieved and re-structing strategies are been applied and clusters are been formed. After clustering ranking algorithm is been used to prioritize the resultant data-set. Finally result is been stored to appropriate data source.

2. CONSTRUCTION OF NOVAL BASED FRAMEWORK

This framework consist of three process:-Re-structing, Clustering and prioritation. This three strategies used to improve the user search goals and improve the performance of search engine .

2.1 RE-STRUCTING METHOD AND THEIR RULES

Some rules to be consider for an re-structing . There are two rules to be considered **Rule-1** and **Rule-2,**which was stated below.

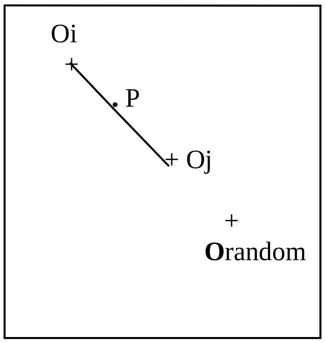


Fig 1.(a) Initial structure of an Data object (On)

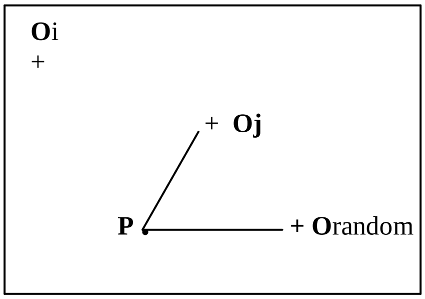


fig 1.(b) Reassigned to Orandom

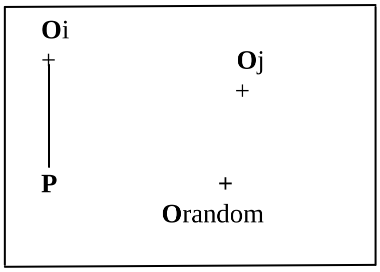


Fig 1.(c) Progressive search

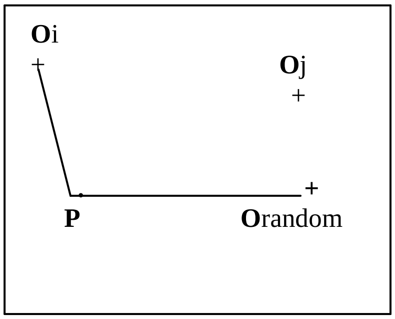


Fig 1.(d) Re-assigned to Orandom.

These re-structing process consider each and every data-set in database as *object*. if an user request for an query, based on the search result, appropriate object are fetched and assign the structure with **p,(*pivot pointer)***.pivot pointer extend its structure to other related object sets. It will set the match (or) not match constraint. based these structure is been formed.

RULE-1 :- Each and every object should be present in any one of the cluster.

RULE-2 :- Each cluster should have an single object.

So ,by using these above ***rules*** ,based on subject and information ,objects are been related and it acts an vital path to form an clusters.

3 . THE PROPOSED ALGORITHM FOR CLUSTERING AND RANKING

An clustering is an process is done using orclus [7] is an extended version of the algorithm Proclus[8].this algorithm is divided into three distinct phases: assign cluster, sub-space determination and finally merging. During the assign phase, the algorithm iteratively assigns data points to the nearest cluster centers. The distance between two points is defined in a sub-space dimension E, where E is a set of orthonormal vectors in some (fig.2.,)dimensional space. Subspace determination redefines the subspace associated with each clusters by calculating the co-variance matrix for a cluster and selecting the orthonormal vectors with the least spread (smallest Eigen's Values).

Clusters that are near each other and have similar directions of least spread are merged together, during the merge phase. The number of clusters and the size of the subspace dimensionality must be specified.

A statistical measure called the cluster sparse coefficient is provided which can be inspected after clustering to evaluate the choice of subspace dimensionality. The algorithm is computationally intensive due mainly to the computation of co-variance matrices. ORCLUS uses random sampling to improve speed and scalability and as a result may miss smaller clusters. The output clusters are defined by the cluster center and an associated partition of the data-set, with possible outliers.

An ranking of data-set or web oriented documents, an page ranking algorithm is been used. This algorithm has three phases: page initialization, page calculation, page_prioritation.

Page initialization: After clustering of each data-set, an web document is been constructed.

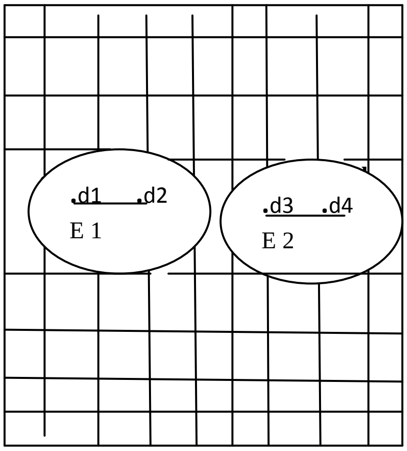


Fig 2. SUBSPACE DIMENSION E, WHERE E IS A SET OF ORTHONORMAL VECTORS

ALGORITHM

Step1: Create a Text-List (by links).
Step 2: Take query as a text: a String.
Step 3: For each Text in Text-List do:
(a) Construct Text-Vector-Space.
(b) Construct Domain-Dictionary of words.
(c) Using Statistical-Model() and Domain-Dictionary, Calculate relevance-value of Text with respect to Query.
(d) Construct Domain-Ontology of the Text.
(e) Calculate Domain-Similarity of Text value with Domain data-sets.
(f) Determine the maximum of Domain-Similarity value and relevance-value and call it Relevance-Score.
Step 4: Goto step 3 until no text is left in Text-List or no more texts are to be considered.
Step 5: Arrange the text (links) according to decreasing order of relevance-score and assign them ranks.
Step 6: Display the texts according to their ranks.
Text-Vector-Space: Consists of text words their weight-age.
Domain-Dictionary: Consists of text-words (nouns, pronouns, synonyms and their weight-age).
Domain-Ontology: A graph containing concepts as nodes and relations as edges.
Domain-Similarity is calculated for the Text with respect to Domain-Data-sets.
Statistical-Model is used to calculate the relevance score of text with respect to domain-Dictionary.

FIG 3. ALGORITHM FOR PAGE RANKING.

Page calculation: In this process, for each and every count of user search, an count for every single search is been done. In this counting strategy, an particular weight is been assigned for clusters and if any data-set wasn't unique and it was interlinked among several links(hyper-link),that kind of data-set was partitioned using web text process.

Page prioritation: in this paper, I have proposed an real time strategy for prioritation. If an user selects for an several links and it was been recorded by repository

USERS	LINKS	LINKS	LINKS
User 1	1	3	6
User 2	1	4	7
User 3	1	2	6

TABLE 1.1 Selection of web content done by user search.

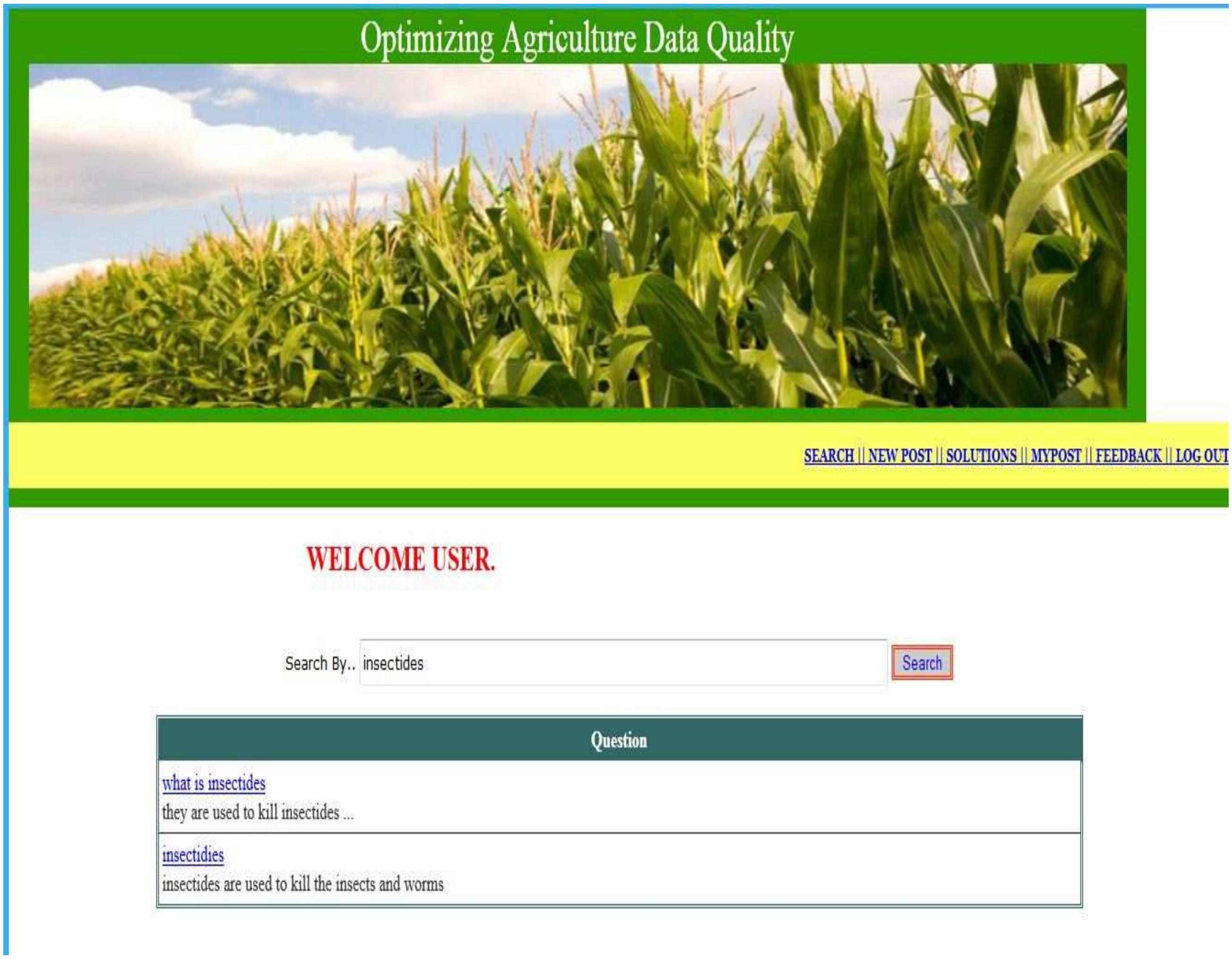
For an example, if an user-1,2,3 selects an link for same query search, if an new user starts search for an same query, based on the prioritation and page ranking algorithm, end-user will get the links in an prioritized manner, that have shown in TABLE.1.2

user	links	links	Links	links	links	Links
User-4	1	6	3	4	2	7

TABLE 1.2 Prioritized web contents

IMPLEMENTATION:

We have **implemented Ant based clustering algorithm** in our project **with visual studio as the IDE and C# as the programming language**. ***CLUSTERING*** is one of the process involved in our algorithm, it is an extends of other two process namely, prioritation and optimization. Screen-shot of clustering process is been picturized at below.


SEARCH || NEW POST || SOLUTIONS || MYPOST || FEEDBACK || LOG OUT
Search By..
Search

Optimizing Agriculture Data Quality

WELCOME USER.

Search By.. insectides

Question
what is insectides they are used to kill insectides ...
insectidies insectides are used to kill the insects and worms

FIG 4. CLUSTERING PROCESS

4. CONCLUSION

The noval based approach of clustering and prioritation provides several ways for improving search strategies and thus fetching relevant web pages efficiently. An paging algorithm improves the ranking by retrieving relevant web pages in the result-set produced by the search engine.

The novel ranking model presented in the paper takes the concepts. It makes relationship between the concept which exists both in the document and the user query to improve the retrieval of relevant documents in the result-set and finally it was produced by the search engine.

4. FUTURE WORK

My future efforts would be to design more meaningful and exhaustive ranking strategy by using the web pages and by deeply statistical analysis relevance of documents, so that the semantic search engine can evaluate more precisely relevance and also the similarity between the web page and the user query. The ranking can even be done by any ontology already created or automatically creating a new ontology for the documents and the user query and then comparing them for the relevance score. We will also try to make our approach scalable for the semantic web.

REFERENCES

[1] Naresh Barsagade, “Web usage mining and pattern discovery: A survey paper”. CSE 8331, Dec, 2003.

[2] H. Chen and S. Dumais, “Bringing Order to the Web: Automatically Categorizing Search Results,” Proc . SIGCHI Conf. Human Factors in Computing Systems (SIGCHI ’00), pp. 145-152, 2000.

[3] X. Wang and C.-X Zhai, “Learn from Web Search Organize Search Results,” Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development in information retrieval . (SIGIR ’07), pp. 87-94, 2007.

[4] J. HANDL,J. KNOWLES and M. DORIGO, “On the Performance of ant-based clustering, Lecture Notes in Computer Science Vol. 2977, 2004, pp. 90-104.

[5] Page L., S. Brin, R. Motwani, and T. Winograd, “The Page Rank Citation Ranking: Bringing Order to the Web”, Stanford ,Digital Library Technologies Project, 1998.

[6] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, “Learning to Cluster Web Search Results,” Proc. 27th Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’04), pp. 210-217, 2004.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In Advances in Knowledge Discovery and Data Mining., pages 307–328. AAAI/MIT press, 1996.

[8] K. Beyer, J. Goldstein, R. Ramakrishnan, and u. Shaft. When is nearest neighbor meaningful? In ICDT Conference, 1999 .

[9] Chahal, P. ; YMCAUST, Faridabad, India ; Singh, M. ; Kumar, S.”Ranking of web Document using semantic similarity” IEEE Conference Information Systems and Computer Networks (ISCON), 2013