ISSN: 2454-9940

# INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

ISSN 2454-9940
INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
www.ijasem.org
Vol 18, Issue 1, 2024

# Enhancing Search Engine Performance Through Query Recommendation Techniques

*Mr.Rachamalla Shyambabu*
*Assistant Professor, Department of CSE,*
*Malla Reddy College of Engineering for Women.,*
*Maisammaguda., Medchal., TS, India*

## Abstract—

***Recently, search engines like google like Google and yahoo emerge as extra vital for locating facts over the World Wide Web in which internet content developing fast, the person's delight of seek engine consequences is decreased. This paper proposes a technique for suggesting a listing of queries which can be associated with the person enter question. The associated queries are primarily based totally on formerly issued queries via way of means of the customers. The proposed technique is primarily based totally on clustering technique wherein businesses of semantically comparable queries are detected. This facility provides a few queries which might be associated with the queries submitted via way of means of customers in order direct them closer to their required facts. This technique now no longer simplest determined the associated queries however additionally rank them in step with a similarity measure. Finally the technique has been evaluated the usage of actual statistics units from the hunt engine question log.***

## I. INTRODUCTION

With the boom of length and reputation of the World Wide Web, many customers locate it is tough to get the preferred records, even though they use maximum green search engines like Google like Google and yahoo (e.g. Google, yahoo).Actually theses search engines like Google like Google and yahoo permit customers to specify queries absolutely as lists of key phrases, following the technique of conventional records systems [1]. But this listing of key phrases is now no longer continually an awesome descriptor of the wished records, consequently it changed into essential to gain person's stratification of seek engine consequences and make it smooth to Retrieve the specified records. The hassle of enhancing seek engine consequences and acquiring the preferred records from this big quantity of web contents has been processed with the aid of using specific approaches along with clustering the quest engine consequences in particular subjects so the person can locate the specified consequences in decided on class of seek consequences [2]. Although, the person does not use the right seek phrases or seek question even as looking so this ends in a trouble of having un-required outcomes and the person should be acquainted with specific terminology in a know-how domain [3]. This isn't continually the case of many customers; they have got simplest a little history approximately the records they may be looking and unluckily they failed to get the specified outcomes. In order to conquer this trouble, it is now no longer sufficient to apply clustering seek outcomes approach due to the fact the trouble isn't in acquiring the huge outcomes however it is within side the key phrases utilized in looking aren't strongly Associated [3]. Query advice indicates associated queries for seek engine customers while they may be now no longer glad with the outcomes of an preliminary enter question, for this reason supporting customers in enhancing seek quality. Conventional tactics to question advice were targeted on increasing a question through phrases extracted from diverse records reasserts along with a word list like WorldNet, the pinnacle ranked files and so on [5]. The preceding queries saved in question logs may be a supply of extra proof to assist destiny customers. A question advice device primarily based totally on large-scale Web get right of entry to logs and net web page archive, and compare 3 question advice techniques primarily based totally on exceptional function spaces (i.e., noun, URL, and Web community) has been presented [5]. The counseled Method aimed to assist seek engine customers in locating their required consequences without difficulty and quickly, this technique indicates associated queries beside the enter question even as the consumer searches so he can construct a proper seek question with the understanding area terminology which is essential for seek engine to get the associated consequences. Also the extra time for enhancing the consequences needs to be unnoticeable via way of means of the consumer.

## II. RELATED WORK

Yates, R.B [6] has carried out a survey in to expose the different upgrades of seek engine aspects. Jawed [7] offered a set of rules for clustering seeks engine queries consistent with four notions consistent with: first, the context of the question; second, not unusual

place clicked URLs among queries; 1/3, Similar strings among the queries and fourth, the gap of the clicked files in a few pre-described hierarchies. Heffernan and Berger [8] cautioned a method for question clustering primarily based totally on the 1/3 notion. Fonseca [9] offered a brand new technique to discover the associated queries which can be primarily based totally on affiliation rules .The queries constitute gadgets in lifestyle affiliation rules. The question log report is taken into consideration as a group of transactions which constitute a consultation wherein the consumer post all associated queries in a particular time .The technique confirmed accurate results ,nevertheless bobbing up of problems The first trouble is the issue of figuring out which classes of those queries which are belong to the equal seek manner. The 2nd trouble the associated queries that are submitted through distinctive customers can not be discovered .This is due to the fact the aid of a rule will increase handiest of its queries are in the equal consultation and that they ought to through submitted through the equal user. M.Hosseini and H.Abolhassni have defined a technique for recommending related queries in line with clustering manner over internet queries from engines like google question log [4]. Zaiane and Streets [10] supplied a technique for recommending queries in line with seven elements of question similarity, Three of them are moderated versions of the primary and 2nd notions .In addition our technique advise the associated queries to the input question however my look for distinctive problems just like the previous statistics from question log file. There is another technique to advise associated queries through question growth. The researchers display that common question phrases are close to [11]. So maximum of the time, queries are ambiguous. One viable answer for this trouble is to make bigger a question with new phrases. Query clustering facilitates to discover applicable phrases for this growth that can be carried out in ways: 1- Query growth in phrases of comparable queries. 2- Expansion in phrases of decided on pages of comparable queries

## III. DESCRIPTION OF THE METHOD

In order to compute the similarity among queries, first we constructed a term-weighted vector for every question. All preceding queries saved within side the question document are taken into consideration with their clicked URLs to suggest the customers with the associated queries to the enter question. We constitute the Queries with the clicked URLs in a bipartite graph with the intention to be clean which can be the URLs clicked with the question submitted with the aid of using the customers [4]. We characterize the graph with the aid of using G (V, E) wherein V is the set of all vertices of queries (Q) and URLs (U), E is the set of edges among Q and U. The vertex units of the graph Q and U at the same time as $Q \cup U = V$ and $Q \cap U = \varphi$, every edge connects among the question issued with the aid of using consumer and the clicked URLs. This illustration of queries and hyperlinks as a bipartite graph makes it clean to discover the similarity among queries as confirmed in Fig.1.

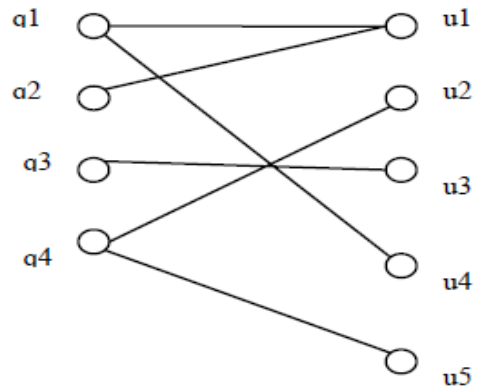


Fig 1. Query – URL representation as a bipartite graph it's vital to set a weight for each aspect among Q and U to reveal the significance of this aspect and distinguish among different edges. We taken into consideration weighted bipartite graph as G (V, E, W) where W is variety of clicked at the hyperlink U while the question Q is submitted via way of means of the user. Queries together with the clicked URLs extracted from Query log are clustered .This is a preprocessing section earlier than making use of question advice set of rules which queries are comparable and additionally to decide that is the maximum comparable cluster to the enter question. We compute clusters via way of means of k-mean set of rules due to its easy and greater respect for document clustering [12] as compared with different algorithms for document clustering.

A. QUERY RECOMMENDING ALGORITHM In our studies work, we implemented a set of rules to recommend associated queries to a question submitted through the person. The clustering technique objectives to categorize all associated queries into companies primarily based totally on all data within side the question log record. When the person put up a question, the set of rules unearths the right organization of associated queries and ranks

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

them consistent with it is relevance to the person enter question and sooner or later it shows all preceding associated queries to the person. The Query Recommendation Algorithm works as the subsequent steps: 1. Queries and Their clicked URLs extracted from the search engine question log record are clustered through k-suggest set of rules. 2. Even as the person put up an enter question, the set of rules unearths the comparable cluster to the enter question; how it is near the centralized of which cluster 3. Query factorization: wherein every question is represented as a vector in which jet detail constitute among the question and URL j. a question vector is proven in (1)

$$\vec{q}_i = [r_1, r_2, r_3, \ldots . r_j] \qquad (1)$$

Where the relation value between is and URL j, it's Computed as shown in (2)

$$r_j = \frac{w_{i_j}}{\sum_{k=1}^{n} w_{kj}} \times \log\left(\frac{|L|}{\sum_{k=1}^{m} connect\ (q_i, l_j)}\right) \qquad (2)$$

Where n is the entire wide variety of precise queries and m is the entire wide variety of wonderful URLs. The first component of (2) is the ration among wig (wide variety of click on times) and total wide variety of wig for all queries with the URL j. The second component is the logarithm of the ration among |L| (total wide variety of wonderful URLs) to the wide variety of URLs which connect (ilk) is a Boolean characteristic as proven in (3)

$$connect\ (q_i, l_j) = \begin{cases} 1 & ; w_{ik} = 0 \\ 0 & ; w_{ik} \geq 0 \end{cases} \qquad (3)$$

4. Queries similarity: in order to compute similarity between queries, we use Animator coefficient similarity measure to show how two queries are similar together as shown in (4)

$$T(q_i, q_j) = \frac{\vec{q}_i . \vec{q}_j}{\left|\vec{q}_i\right|^2 + \left|\vec{q}_j\right|^2 - \vec{q}_i . \vec{q}_j} \qquad (4)$$

5. Support of the question: that is a degree of ways a great deal this question belongs to its cluster. The help of the question is measured because the ratio of the wide variety of clicked URLs for the question |Li| to the whole wide variety of the URLs; for all queries Σ jack I | L | within side the cluster C as proven in (5) Finally, the queries within side the decided on cluster are ranked base on their similarity and their help; the rank rating is measured as proven in (6) wherein the similarity among all queries in the cluster quid and the enter question q. The time period constants α and β used for normalization [4].

$$Rank(q_i) = \alpha \times T(q_i, q) + \beta \times Sup(q_i) \qquad (6)$$

## IV. EXPERIMENTAL RESULTS

In our experiments, a question log of the AOL seek engine changed into used for amassing click on via data. A file on this question log represents the go to to an end result for a question or the submission of a question (if no end result is visited) [13]. Each file store: • A nameless ID that lets in to institution queries from the equal person without revealing the AOL person's nickname. • Query submitted via way of means of the person. • Date and time of the submission of the question • Rank function of the end result visited via way of means of the person on each file. Examples of this file may be discovered in Table.1 and a full description of the AOL question log is illustrated in [14]

***TABLE I. EXAMPLE OF AOL LOG QUERIES***

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 18, Issue 1, 2024

| User ID | Query | Time | Visited URLs |
|---|---|---|---|
| 123 | Computer virus | 2006-03-08 12:21:23 | http://www.microsoft.com/security/antivirus/whatis.aspx<br>http://www.howstuffworks.com/virus.htm<br>http://www.snopes.com/computer/virus/virus.asp<br>..... |
| 154 | Funny videos | 2006-05-12 .02:12:32 | http://www.break.com/<br>http://www.funnyordie.com/<br>http://www.dailyhaha.com/<br>..... |
| 217 | Albert Einstein | 2006-08-17 05:19:05 | http://www.westegg.com/einstein/<br>http://www.albert-einstein.org/ |

***TABLE2: RECOMMENDED QUERIES FOR "SCHOLARSHIP"***

| Query | Recommended Query | Similarity by Tanimoto | Similarity by Cosine | Rank Score by Tanimoto | Rank Score by Cosine |
|---|---|---|---|---|---|
| Q1 | International scholarship | 0.958 | 0.98 | 0.791 | 0.808 |
| Q2 | Grants and fellowships | 0.921 | 0.989 | 0.765 | 0.82 |
| Q3 | Fulbright Scholarship | 0.651 | 0.987 | 0.549 | 0.818 |
| Q4 | Study abroad chances | 0.117 | 0.257 | 0.122 | 0.234 |
| Q5 | Scholarship programs | 0.62 | 0.794 | 0.525 | 0.664 |
| Q6 | Full funded scholarship | 0.422 | 0.743 | 0.366 | 0.623 |

In order to assess our similarity measure, we compared it with the Cosine similarity which has been utilized by Mehdi and Hassan [8]. We extracted 10,000 queries from AOL dataset for clustering. After building the clusters we choose ten queries of the clustered dataset: (1) weather; (2) newspaper; (3) laptop Software; (4) airline flights; (5) scholarships; (6) Nobel award; (7) 5 supermegacelebrity hotels; (eight) human improvement books; (9) MIT universities; (10) migrations .All the chosen queries are dispatched to the recommendation set of rules so as to signify beneficial queries for the user. In addition to that, the endorse queries are ranked in accordance to rank rating that's calculated via way of means of Eq.6 with parameters $\alpha$ =0.eight and $\beta$ =0.2. Table 2, indicates the pleasant of rating primarily based totally on two exclusive similarity degree: Cosine similarity utilized by Mehdi and Hassan [4] and our similarity degree (Animator coefficient degree).In this example, the algorithm suggest 6 associated queries to the enter question "scholarship" In Fig.2, we in comparison the rating of effects primarily based totally on cosine similarity degree [4], and Animator coefficient similarity degree to expose the performance of rating. Cosine similarity has overestimate measurements of question similarity as we see question 2 "Grants and fellowships" is extra comparable to consumer questioning "Scholarships" as opposed to query1 "International scholarships". Although question 2 is semantically same to the consumer question, however question 1 is a good deal higher as it has the most comparable string and imply to the consumer question. Therefore this degree have an effect on on end result rating of seek engine via way of means of using Animator coefficient in similarity size has extra precision and accuracy in rating the quest engine effects.

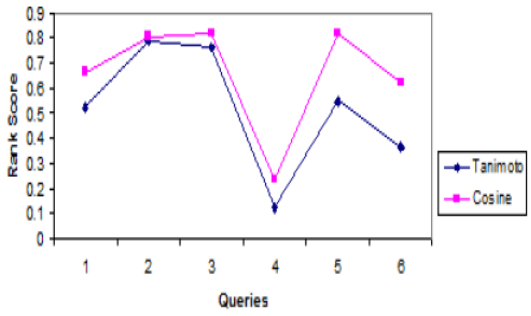


Fig. 2 ranking efficiency between Animator and Cosine Similarity

## V. CONCLUSION AND FUTURE WORK

We have provided a way for recommending the associated queries to the enter primarily based totally on clustering method over the net queries extracted from a seek engine question log. We are doing the experimentations with large logs greater than used and thinking about greater queries to enhance the opinions of our approach. In addition, we are attempting to make bigger the queries the usage of the key phrases associated with the cluster Also we recall the development of Similarity by thinking about the clicks in question solution to files that are much like the enter question As destiny work, we recall to enhance the perception of interest of the cautioned queries and to make bigger other notions of hobby for the advice algorithm. For instance locating the queries which percentage phrases however now no longer have not unusual place clicked URLs, this could contain the equal phrases however have extraordinary meanings if the textual content of the URLs is also now no longer percentage. Hence we are able to understand polysomic phrases.

## REFERENCES

*[1] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, chapter 3, pages 75–79. Addison-Wesley, 1999. [2] Caramia, G. Felici and A.*

*Pezzoli, "Improving search results with data mining in a thematic search engine," Computer & Operations Research 31, pp. 2387-2404, ( 2004) Elsevier*

*[3] R. Baeza-Yates, C. Hurtado and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," LNCS 3268, pp. 588-596, (2004), Springer-Verlag Berlin Heidelberg , 2004.*

*[4] M. Hosseini and H. Abolhassani, "Clustering search engines log for query recommendation," CSICC, CCIS 6, pp. 380-387, (2008), Springer-Verlag Berlin Heidelberg 2008*

*[5] L. Li, S. Otsuka, and M. Kitsuregawa "Query Recommendation Using Large-Scale Web Access Logs and Web Page Archive," LNCS 5181, pp. 134–141, (2008), Springer-Verlag Berlin Heidelberg 2008*

*[6] R. Yates, "Query usage mining in search engines," in Scime, A. (ed.) Web Mining: Applications and Techniques. Idea Group (2004)*

*[7] J. Wen, J. Nie, H. Zhang, "Clustering user queries of a search engine," in 10th International World Wide Web Conference. W3C, pp. 162–168 (2001)*

*[8] D. Beeferman, and A. Berger, "Agglomerative clustering of a search engine query log," in KDD, Boston, MA USA, pp. 407–416 (2000)*

*[9] B. Fonseca, P. Golgher, E. De Moura, and N. Ziviani, "Using association rules to discovery search engines related queries," in First Latin American Web Congress (LAWEB 2003), Santiago, Chile (November 2003)*

*[10] O. Zaiane, A. Strilets, "Finding similar queries to satisfy searches based on query traces," in Proceedings of the International Workshop on Efficient Web-Based Information Systems (EWIS), Montpellier, France (September 2002) .*

*[11] D. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: a study of user queries on the web," ACM SIGIR Forum 32(1), 5–17 (1998)*

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