Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 7, July-2017

Advance Image Search Based On Extended RMTF

[1]Mr. Durgaprasad B. Kadam, [2]Dr. Arti Mohanpurkar

[1]*Department of Computer Science, Dr. DY Patil School of engg and tech lohgaon. Maharashtra, India*

[2]*Department of Computer Science, Dr. DY Patil School of engg and tech lohgaon. Maharashtra, India*

Abstract — Information on internet is increasing every day at exponential rate. Utilization of ERP, social networking web sites, online shopping, advertisement causes tremendous hike in the data. To deliver exact information out of such huge amount of data is not at all an easy task. Various data, mining, indexing, filtration of data are looking ordinary to meet the requirement specified by the user. So there is need to have a model by considering user at center. Prime objectives of such a system need to be 1) to separate unwanted data 2) to deliver correct data intended by users. Many of times search engine delivering user specific results end up with mere customization of results. There could be several perceptions of input provided or there could be different interpretation of given input for search. This causes a gap between expectation of user and results delivered by search engines. There must be a mechanism to understand the specific requirement given by user and to deliver the solution accordingly. For carrying out searching general or random search approach is not mandatory because every user got some interest or requirements before that user proceeds for the search. To provide a solution as per the requirement mentioned by user novel approach is proposed ; keeping pace with time and to deliver the result without any influence of marketing strategies like Search Engine Optimization (SEO), Search Engine Marketing (SEM), false hits to improve page ranking is needed. Mechanism to calculate Interest of specific user and to identify the exact requirements based on interest of individual need to be the only parameter to decide index of web pages. System implemented performs search of query entered by the user by considering interest of the user collected with the help of set of images. This system delivers solution for query with single and multiple words. Concept of provision of set of images from various categories

Keywords- Knowledge Discovery and Data Mining, requirement centric results, interest calculation, data intensive computing, ranking of URL according to the interest of user

I. INTRODUCTION

In this era everything is getting automated causing huge amount of data on an internet. Since almost everything and information associated with it could be collected by using an internet just on a click. Huge numbers of users are there to avail this internet facility. Even after providing great computational power and high performance providing facility, it is difficult to deliver adequate information needed to a user. This is proposal is an attempt to address issues faced by user while searching, by providing solution in the form of most relevant result according to Jitao Sang et.al.[1], D. Lu. et. al. [13] and Y. Cai. et. al. [16].Even if we have great computing facility of hardware and optimum performance delivering algorithms, Probability of meeting the expectation specified by user, could not be as per the expectation. Ever increasing traffic on a web and huge amount of data could act as a hurdle while surfing across a net. Introducing a functionality which acts as interface between the user and rest of the web makes a sense. This will help in filtering the contents to be searched more than that we can reject the contents with less relevance as suggested by F. Liu et. al.[7] before performing the actual search. If we adopt this particular approach then probability of getting the desired data out available data gets increases due to considerable reduction in the sample space. Example discussed by Jitao Sang et.al. [1], word jaguar is searched by search engines. It is difficult to select parameters to decide ranking of data according to the interest of user, because jaguar is an animal, there are various organization and products with same name. In this paper there is discussion of various existing approaches and comment on the result that we retrieve.

II. EXISTING SYSTEM

Manual Categorization based approach [1], Popular according to people [5] , Recommendation based approach [8] , Semantic or Ontology Based Approach [19] , History, cookies and Log Based Approach [6], Clustering Based Approach [11] , Session Based Approach [8], Context Based Approach [12] are some the existing techniques to search results. These are few techniques which may give the appropriate result in peculiar situation, for some specific data. Completely generic approach to collect interest of individual and deliver the related data accordingly is required.

Many of the approaches cause diversion of traffic as mentioned by Borzsonyi et.al. [3] a site containing malicious code or a site which may not have high relevance with the requirement of user. Technique superior than this or adequate combination with suitable enhancement is required. Searching mechanism is designed, according to Dou. Z. et. al.[6] designed should first sort out the data which need not to be searched regarding particular user on a specific topic. Secondly once sample space is reduced, extend this filtration to further extent by considering interest of user. Main thirst in clustering is to improve recall process. In this case, sample space in which search is to be made is reduced. Cluster

All Rights Reserved, @IJAREST-2017

encourages forming group of important keywords and entire search process orient around that keyword. There are many more phases of this cluster, scatter and gather is one of the important phase in the process of clustering. In this approach several categories are formed, this phase is called as scattering. In next phase that is gather super categories are formed by considering the groups and categories made in scattering. Example of this approach is news displayed by the yahoo. In this case it is not influenced by any specific user; rather inference of user is not required to call something like breaking news.

Many organizations are targeting this domain to extend their business. After performing an analysis of various search engines and approaches that they adopt to perform the search, in most of the cases there is no real personalization of result is taken care of. Customization of graphical interface to display the result, separation of result completely managed by user without any automation, providing specific functionality with very limited scope is not.

So for designing personalized search engine interest of user has to be captured by using combination of long term technique like History, cookies and Log Based Approach and for change in the interest we can refer session based approach we can evaluate the performance of a system by collecting feedback to whom we are displaying the result.

.

III. ARCHITECTURE OF SYSTEM

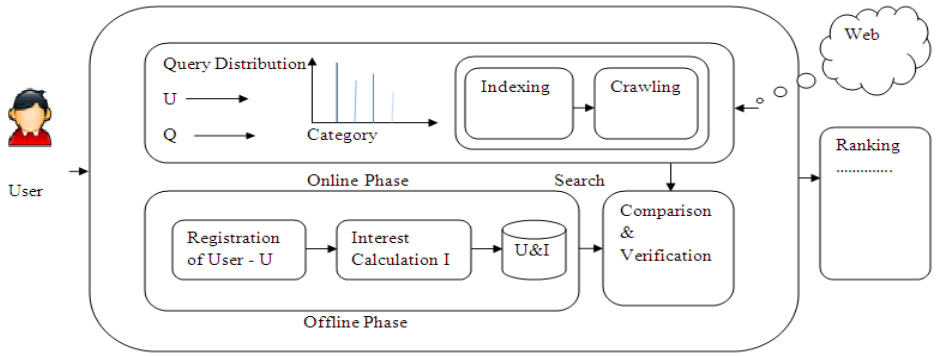

Fig 1: - Architecture of the System

3.1 Architecture of the system

Users will provide their respective areas of interest. According to the requirement provided by the user, system is trained from time to time. Once user provides requirements Re - ranking of displayed result based on interest of user, should take place. There is need to adopt a technique which will give the result for which, user is not supposed to navigate through the set of result again. Unique combination meeting this requirement is proposed

3.2 Offline Phase: Training Module

This is the core part of architecture. This component is responsible for Re- Ranking the result, which are received from the searching module. Intelligent is applied for making the decision to prioritize the data which is less relevant to the particular user. Based on the comparison best approach is selected to proceed with

3.2.1 User Registration:

User Specifies his interest through registration and continuous monitoring and filtration technique helps to know preferences. Least botheration is to be carried out by the user during the registration, so that user will not mind to register.

3.2.2 Interest Calculation:

To decide interest of user and to avoid the web pages and data retrieved from the resources where user is not interested; this technique is of real use [12] [19] . The set of questions, images [4], puzzles are designed to know interest of user [9].

3.2.3 Storage

This is where the combination takes place to deliver the optimum solution. Information associated with each user regarding his/her interest is stored in a database and it is use full when online functions complete its function and refer database for deciding final ranking of URLs.

All Rights Reserved, @IJAREST-2017

3.3. **Online Phase: Searching Module:**

This is to search a context on a web and deliver to indexing module to form indexing at level 1. Instead of going with available search engine API, development and continuation of separate indexing and searching technique is encouraged.

Crawling: Crawling Effective functioning of search engine can be, search engine should not start searching after the request sent by the user. Search engine is required to have knowledge of almost all web site register along with the data residing at that site. Practically it is difficult for an administrator of search engine to visit every website and keep monitoring the changes in a respective website. Here the concept of crawler comes into the picture, which is automatic process of collecting data from almost every website and monitors the changes after regular interval. Many other cases website administrator wants there site to be one of the top web sites displayed, so they approach to a searching site to deliver the changes made in a web site and main keywords associated with website. This time to time collected information is conveyed to the respective indexing technique to form index over the data collected after crawling. Crawling is the process of forming the index of URLs. This process always keeps running at the background. Role of this process is to hit URL search a data and form a base for ultimate indexing. To search the desired data across the web is equally important so as to address exact requirement of the user. Process of collecting information about various web sites before user actually demands for some information, saves the time required for searching.

1. Input User provides input in the form of user details and query to be searched.

Comparison and Verification After successful evaluation of interest of user and collecting general index comparison of these URLs is made with the Interest Calculation Factor. If these URLs contain the data which is of same interest shown by user, these URLs are considered and ranked at the top position.

2. Ranking

Interest Calculation Factor is considered as a factor to shuffle position of URLs across the pages. If these URLs contain the data which is of same interest shown by user, these URLs are considered and ranked at the top position.

3.3.1. Indexing:

Once the crawling process completes, results are delivered by crawling process to indexing [10] for the betterment of data storage. By forming the index over the collected data time required to search a data inside in minimized up to considerable extent. In this the unbiased results after search are expected, meaning it must not carry any influence of page ranking of recommendation which will help to avoid irrelevant data and thereby to achieve the result up to considerable extent.

3.3.2. Data Collection:

Effective functioning of search engine can be, search engine should not start searching after the request sent by the user. Search engine is required to have knowledge of almost every web site along with the data residing at that site. Practically it is difficult for an administrator of search engine to visit every website and keep monitoring the changes in a respective website. Here the concept of crawler comes into the picture, which is automatic process of collecting data from almost every website and monitors the changes after regular interval. Many other cases website administrator wants there site to be one of the top web sites displayed, so they approach to a searching site to deliver the changes made in a web site and main keywords associated with website. This time to time collected information is conveyed to the respective indexing technique to form index over the data collected after crawling. Crawling is the process of forming the index of images. This process always keeps running at the background. Role of this process is to hit image search a data and form a base for ultimate indexing. To search the desired data across the web is equally important so as to address exact requirement of the user.

Process of collecting information about various web sites before user actually demands for some information, saves the time required for searching.

3.3.3. Input

Input from the system user is complex multiple words-based queries.

3.3. Comparison and Verification

After calculating interest of user and collecting general index comparison of these images is made with the Interest Calculation Factor. If these images contain the data which is of same interest shown by user, these images are considered and ranked at the top position by using ranking algorithm.

3.4. Mathematical Model

Formation of domain for Relevant Search

$$I(u,q,c) = \begin{cases} 1 \; of \; I \in Y \\ 0 \; Otherwise \end{cases} \quad ---(1)$$

Where,

Y∈ | U | × | Q | × | C |

All Rights Reserved, @IJAREST-2017

C → Set of Category, U → Set of Unified Resource Locater, Q → Set of Query entered by user, c → Category, u → Unified Resource Locater, q → Query entered by user According to Bays' Rule Probability of getting relevant URL can be expressed by

P (Ru/U) =P (Ru) * P (u/Ru) /P (u) ……..(2)

Probability of getting relevant URL, which is not relevant, can be expressed by Ranking of a page

P(NRu/u)=P (NRu) *P (u/NRu)/P (u)…… (3)

Increment in a Rank

$Re+ = \{(u|(c, u, q) \in Y) \wedge (I(u, c, q) = 1)\} \cdot \cdot \cdot$ (4)

Re+ → Increment in ranking,
Ru → Relevant URLs to the user.
u → Total number of URLs displayed to the user,
NRu → Not

P(Ruu) + P(NRuu) = 1 ……. (5)

Decrement in a Rank

$R- = \{(u|(c, u, q) \in Y) \wedge (I (u, c, q) \neq 1)\} \cdot \cdot \cdot$ (6)

R⟶ Decrement in ranking
By applying Bays' Optimal Decision rule
url is usefull iff

P(Ruu) ≥ P(NRuu) …….. (7)
Precision = (Ru ∩ ui) / u …… (8)
Recall = (Ru ∩ ui) / Ru …….. (9)

Interest Calculation Factor ICF can be expressed by applying Harmonic equation

ICF = {β* Precision(TN) *Recall(TN)} / {Precision(TN) + Recall(TN) }……..(10)

β → Balancing factor to increase the precision value, which is 2 TN → Top N results displayed after indexing
While calculating interest of an user, interest category is divided into several sub categories and then interest is stored inside a table along with associated user, while considering a ranking, every parameter related with user is considered before displaying rank of particular URL. In this way divide and conquer approach is used in this architecture Implementation of this system requires memorization of data associated with user, which is gathered during the registration and topic or category wise interest of user in a website which is likely to be changed from time to time is collected from indexing phase and stored in terms of Interest Calculation Factor in a database.

IV. EXPERIMENT

To implement the architecture mentioned above project has been implemented. In this experiment API is configured with project for crawling to collect data from various websites.

1.Data set used for experiment

Total number of sites from which data is collected	250
Total number of categories	50
Total number of images	50

All Rights Reserved, @IJAREST-2017

Table 1: Data set used for experiment

Difference between the actual and visited site is due to the secured protocol which do not allow the crawling of respective site. After collecting this data, it is stored in secondary storage device.

2. Indexing

Data collected is categorized and saved

3. User Interest Registration

By offering wide range of images from various categories interest is evaluated. Total numbers of categories considered are 50 and images are 50. Total number of users registered is 100.

Ranking is done by –

Score=ICF+R …… (11)

Score- value deciding rank of URL, R – Relevance of URL So final rank of URL is decided with the help of ICF and relevance of contents in query.

4. RESULTS

Experiment were conducted on various types of queries for specific user by notifying their interests 1) for some ambiguous words containing query 2) Two word query. Experiment has given below table 2 and table 3 results

Keyword Results for Top	12 links
Crawler	0.8333
Jaguar	0.6666
Ticket Booking	0.833
International Journal	0.75
software download	0.6666

Table 2: Results of Accuracy for given sample queries using non – personalized approach for Top 12 links

In table 2 unbiased or non personalized results are displayed. In the non personalized mode accuracy is close to 1 but not the ideal.

Keyword Results for Top	12 links
Crawler	1
Jaguar	1
Ticket Booking	1
International Journal	1
software download	1

Table 3: Results of Accuracy for given sample queries using personalized approach for Top 12 links

In table 3 personalized results are displayed. In the personalized mode accuracy is exactly 1. Experiment is showing ideal results in the personalized mode because of consideration of category.

All Rights Reserved, @IJAREST-2017

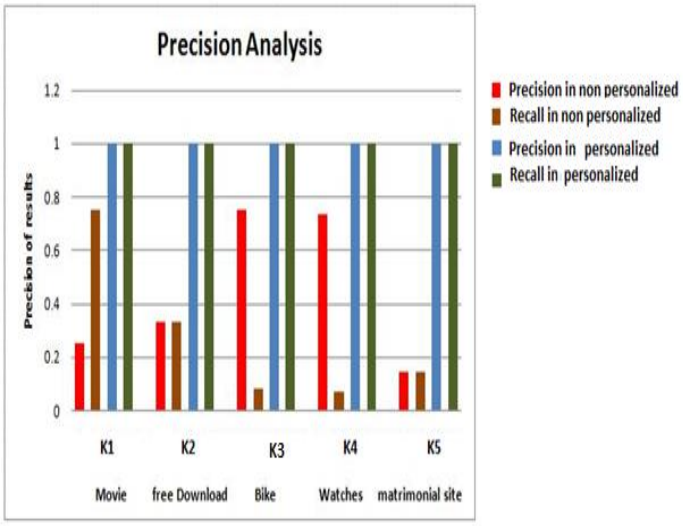

Figure 1: Comparison of precision for various keywords in personalized and non-personalized approach

Figure 1 is showing the results displayed for personalized and non -personalized approach for different keyword. This is showing comparison between personalized approach and non personalized approach is mentioned in a figure. Result is compared by using keywords Movie, Free download, Bike, Watches, Matrimonial site Thus the precision value is increased by distributing load over cluster. As cluster is formed by considering interest of user only, hence accuracy of results is 100% trough out. This also carries impact on the other elements.

V. CONCLUSION

Experiments were conducted to address the problem stated in abstract and introduction section and to support future trends. In advanced web technology has prominent feature that every application along with the data will reside in some server and machine of user will act as thin client. So customized application would be in great demand for performing search of particular data of user and also data across the web, which is relevant to a user. Advanced Image Search using Extended RMTF has delivered following outcomes.

1. Registered users can notify their interest for given set of images.
2. System effectively delivers results for query containing multiple words.
3. No marketing strategy or promotional activity to promote the result up in the order, so user can get user centric results.
4. Personal information of user like cookies or in any other format is never used to make business but for the betterment of experience during the search. Definitely there is room to enhance this model –

1. Where along with images various objects like videos, documents, software could be also used to collect more and more information about the user
2. Utilizing this information services and solutions could be offered instead of offering advertisements.

VI. ACKNOWLEDGMENT

I thank Head of the Department of Computer, for her guidance on the paper. I also express thanks to Post Graduate Coordinator, Project Guide and other senior staff members for their comments, suggestions on this paper and helpful discussions.

VII.REFERENCES

[1] Jitao Sang, Changsheng Xu, Senior Member, IEEE, Dongyuan Lu," Learn to Personalized Image Search fromthe Photo Sharing Websites", 2012 IEEE Transaction on Multimedia , Volume : PP, Issue:99.
[2] Brown, P., Bovey, J., Chen, X. (1997), "Context-aware applications: From the laboratory to the marketplace". IEEE Personal Communications, Vol. 4, No. 5.
[3] Borzsonyi, S., kossmann, D., Stocker, K. (2001). "The skyline operator. Intl. Conf. on Data Engineering (ICDE)", IEEE Computer Society, pages 421-432.

All Rights Reserved, @IJAREST-2017

[4] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu, "Exploring folksonomy for personalized search," in SIGIR, 2008, pp. 155–162.
[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. "Learning object categories from google's image search" In IEEE International Conference on Computer Vision, 2005
[6] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos, "A Unfied Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis", IEEE Transaction on knowledge and data Engg., VOL. 22, NO. 2, FEB 2010.
[7] Dou, Z., Song, R., Wen, J.R. and Yuan, X. 2009." Evaluating the Effectiveness of Personalized Web Search". IEEE Transactions on Knowledge and Data Engineering, 21, 1178-1190
[8] F. Liu, C. Yu, and W. Meng. "Personalized web search for improving retrieval effectiveness " . IEEE Trans. On Knowl. and Data Eng, 16(1) 28,40, 2004.
[9] Kevyn Collins-Thompson Paul N. Bennett, Susan T. Dumais "Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic", WSDM 2012.
[10] J. Teevan, S. Dumais, and E. Horvitz. "Potential for personalization ".ACM Trans. Comput.Hum. Interact. 17(1):31 March 2010.
[11] K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, D. Sontag. "Personalizing web search results by reading level ". In Proceedings of CIKM 2011. ACM, New York, USA.
[12] Alessandro Micarelli,Fabio Gasparetti, Filippo Sciarrone, and Susan Gauch 2, "Personalized Search on the World Wide Web "The Adaptive Web, LNCS 4321, pp. 195 – 230, 2007. c, Springer-Verlag Berlin Heidelberg 2007
[13] D. Lu and Q. Li , " Personalized search on flickr based on searcher preference prediction " in WWW (Companion Volume), 2011, pp. 81. 82.
[14] J. E. Pitkow, H. Schutze, T. A. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. M. Breuel, " Personalized search " Commun. ACM, vol. 45, no. 9, pp. 50.55, 2002.
[15] J. Teevan, M. R. Morris, and S. Bush, "Discovering and using groups to improve personalized search "in WSDM, 2009, pp. 15.24.
[16] Y. Cai and Q. Li, "Personalized search by tag-based user profile and resource profile in collaborative tagging systems " in CIKM, 2010, pp. 969.978
[17] B. Smyth, " A community-based approach to personalizing web search " Computer, vol. 40, no. 8, pp. 42 50, 2007.
[18] J. Teevan, S. T. Dumais, and E. Horvitz, "Personalizing search via automated analysis of interests and activities " in SIGIR, 2005, pp. 449.s 456.
[19] Semantic Web Personalization: A Survey, Information and Knowledge Management ,iiste ISSN 2224-5758 (Paper) ISSN 2224-896X (Online) Vol 2, No.6, 2012,
[20] Ayesha Ameen ,Khaleel Ur Rahman Khan B.Padmaja Rani J. Teevan, S. T. Dumais, and D. J. Liebling, " To personalize or not to personalize: modeling queries with variation in user intent " in SIGIR, 2008

All Rights Reserved, @IJAREST-2017