An Automated Framework based on Three Level Scoring Method to Choose Best Search Engine in a Particular Domain

Naresh Kumar
Asst. Professor, CSE Dept,
MSIT, New Delhi, India

Rajender Nath
Professor, DCSA,
K. U., K, Haryana, India

Pooja Kherwa
Asst. Professor, CSE Dept,
MSIT, New Delhi, India

ABSTRACT
Today, using search engine is the most widely used activity on the World Wide Web. But choosing the most suitable search engine that can provide the most relevant contents is really a tough job. One search engine is better in one domain may not be better in another domain. To address this problem of finding suitable search engine for a given domain, this paper presents an automated frame work based upon Three Level Scoring (TLS) methods for choosing the best search engine in a given domain. The proposed framework has been tested experimentally by implementing in C# programming language.

Keywords
WWW, search engine, precision, relevance, user efforts, technical domain, domain, medical domain, mixed domain, domain query, Criteria, ranking.

1. INTRODUCTION
According to [1], a lot of search engines with different ranking methods and different coverage area are available to search the information on the web. It is also observed that no single search engine (SE) can provide better results all the time [2]. So, to evaluate the better precision and coverage area various search engine experiments were carried out manually by taking some general queries of some specific domains [3], [4]. The main drawback of such evaluation system is time consumption and hence cannot be adopted in ever changing search engine technologies. So, the automatic evaluation of search engine is highly desirable in search scenarios.
The rest of the paper is organized as follows: The related work is discussed in section 2. Section 3 describes the problems with currently available search engine evaluation techniques. Section 4 describes the proposed framework. Section 5 describes Experimental results and their comparison and section 6 concludes the paper.

2. RELATED WORK
In [1], statistical evaluation of the search engine in terms of relevance and precision was carried out. The methods considered for this purpose were Okapi, CDR, VSM and TLS. They compared the *AltaVista, Fast, Google, Go, iWon*, and *NorthernLight* search engine individually in terms of relevance and precision. The results of evaluation showed that different search engine behaved differently for different queries and in overall performance Google was best performer for each type of query.

Fazli Can et. al. in [5], proposed an automatic Web search engine evaluation method.They calculated recall and precision at various document cut-off values and used them for statistical comparison. They evaluated AlltheWeb, AltaVista, HotBot, InfoSeek, Lycos, MSN, Netscape and Yahoo by taking 25 different queries and top 20 results. They also perform the calculation manually and they found a high level statistical significant consistency between the automatic and human-based assessments both in terms of effectiveness and also in terms of selecting the best and the worst performing search engine.

In [6], proposed a 'Ranked Precision' (RP) metric to evaluate the performance of search engine. The RP returns a number between 0 and 1, which showed the effectiveness of the search engine in terms of retrieved documents, ranked according to their relevance and present them in order of first n document(s).

Study in [7], carried out the work on manual basis for the user satisfaction measurement. In this work, different categories of queries were applied by the 35 under graduate students on Google, Bing and Blekko. The results of evaluation showed Google as a winner, Bing was closely behind, and Blekko still required some work to be carried out for better performance.

3. PROBLEM FORMULATION
Many evaluation techniques have been proposed by many authors [2], [3], [4], [8], [9]. The major concern of all these techniques is primarily related to automated and manual evaluation of precision. The main problems of these techniques are summarized below:

(a) These techniques cannot be applied to the dynamic web environment. (b) Manual evaluation is very time consuming process.

The motivation of this paper is the facts that identify the most effective web search engine satisfying the current information need is very important. So, to address the problems mentioned above an automated frame work for evaluating search engine is proposed in the next section.

4. PROPOSED WORK

4.1 Selection of Search Engine, Domain, Query and Criteria
To determine the suitability of the search engine for a particular domain, first the search engine evaluation criteria and domain query evaluation criteria are defined. To do this, a mathematical model is formulated as given below:

Let $S = \{S_A\}$, A=1, 2, …, a, be the set of 'a' no. of participant SEs.

Let $D = \{D^n\}$, n=1, 2, …, v be the set of 'v' domains covered by the SE.

Let $Q = \{\{Q^T\}, \{Q^M\}, \{Q^X\}\}$, be a set of three different sets of queries i.e. Q^T represents the technical query

set, Q^M represents the medical query set and Q^X represents the mixed query set.

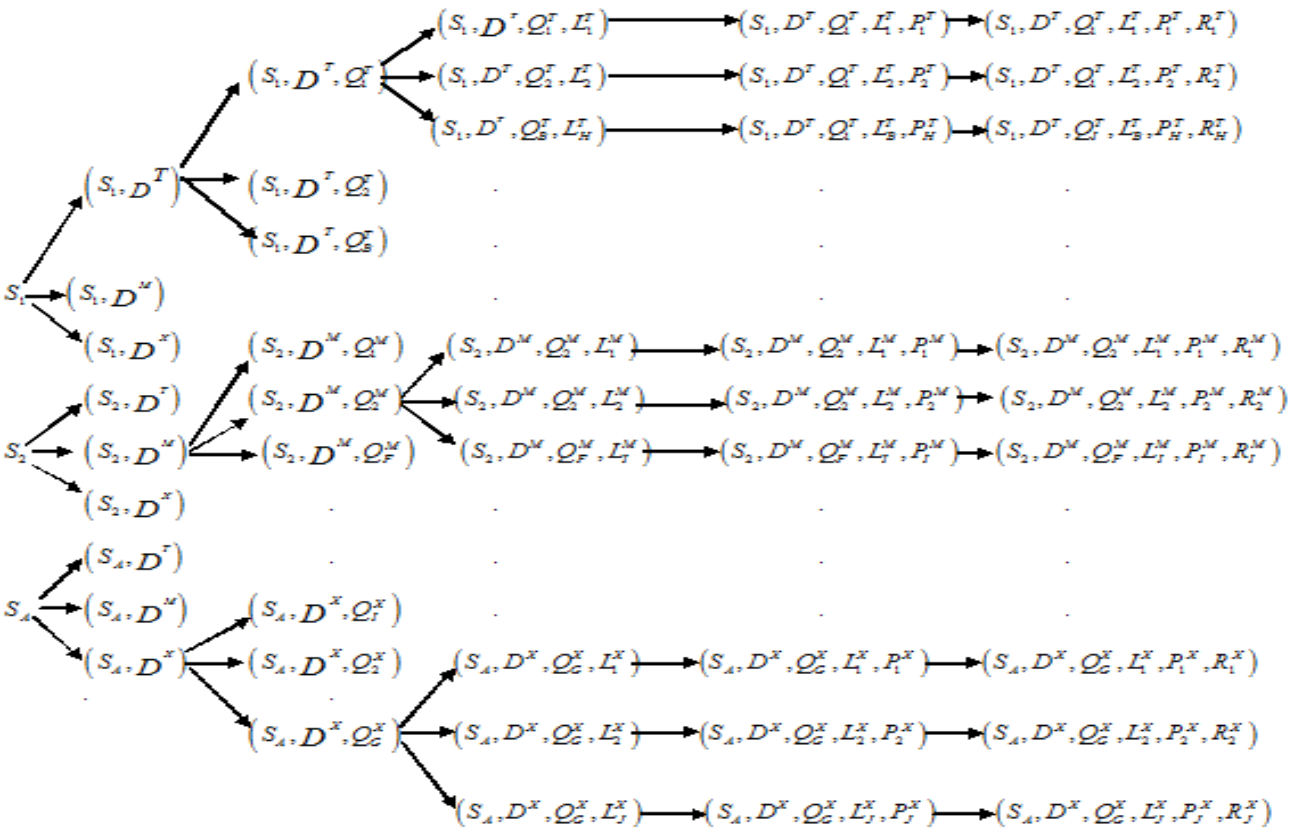

Figure 1. Graphical view of TLS

Algorithm: - Relevancy score and relevancy calculation for the web links.
Input: - Downloaded web pages.
Output: - Relevancy score to each downloaded and selected web link.
// starting of algorithm
Step 1: - Download the web page from the WWW for the given - SE, given domain and for given query.
Step 2: - Let the returned web link for the given query Q is L_l , terms of criteria is C_r, the web page for L_l is W_{pl} and terms of W_{pl} is E_z.
Step 3: - For each SE S do
For each domain D do
For each query Q do
For each link L_l do
For each term E_z of W_{pl} do
If(a term E_z = =a term C_r)
Count++ //counter
End
P_l = (count/n(C))*100 // count % of matched terms, n(C) no. Of terms in C
If(P_l >=70)
Assign score 2 to L_l
Else If(70 < P_l >30)
Assign score 1 to L_l
Else
Assign score 0 to L_l
End
End
End
End
Step 4: - For each P_l do
*Relevancy= (sum of assigned score to the relevant and partially relevant web pages / L*2)*100*
Step 5: - Return Relevancy of each link.
Step 6: - Stop.

Figure 2: - Algorithm for Relevancy calculation and score assignment

links, L^M represents the medical links and L^X represents the mixed links returned by the SE for the submitted query.

Let $L^T = \{L^T_H\}$, H = 1, 2, …, h be the set of 'h' number of links returned by the SEs.

Let $L^M = \{L^M_I\}$, I = 1, 2, …, i be the set of 'i' number of links returned by the SEs.

Let $L^X = \{L^X_J\}$, J = 1, 2, …, j be the set of 'j' number of links returned by the SEs.

Hence L can be written as

$$L = \bigcup_{H=1}^{h} L^T_H \bigcup_{I=1}^{i} L^M_I \bigcup_{J=1}^{j} L^X_J$$

Further let $C = \{\{C^T\}, \{C^M\}, \{C^X\}\}$, be a set of three different set of criteria i.e. C^T represents the technical criteria, C^M represents the medical criteria and C^X represents the mixed criteria for evaluation of SE(s).

Table 1 Domain, Query and Criteria of evaluation

Domain (D)	**Query (Q)**	**Criteria (C)**
Technical (Q^T)	Java	Download java, basic feature of Java, Java tutorials in ppt and pdf form, java developer, java run time environment, frequently asked questions on java
	Hibernate	Basic features of hibernate, platform requirements of hibernate, queries for hibernating and download, hibernate jobs, hibernate interview questions.
	DB2	Basic features, queries, tutorials in ppt and pdf form, test information, certificate information, DB2 commands, DB2 training, DB2 software.
	Vmware vsphere	Basic features, download Vmware vsphere, versions of Vmware vsphere, Vmware vsphere client, Vmware vsphere training.
	Eclipse	Download Eclipse, versions of Eclipse, Eclipse for Java, Eclipse plug-in, Eclipse packages.
Medical (Q^M)	Peptic Ulcer	Peptic Ulcer causes, Peptic Ulcer diseases, Peptic Ulcer diet, Peptic Ulcer treatment.
	Polio	Polio Symptoms, Diagnostic tests, treatment and vaccines available for prevention.
	Arthritis	Arthritis causes, Arthritis Symptoms, Arthritis causes, Arthritis treatment and preventions.
	Diabetes	Diabetes symptoms, Diabetes test, Diabetes Treatment, Diabetes diet, Diabetes causes, Diabetes types and machine used to check it.
	Cancer	Cancer types, Cancer Prevention, Cancer causes, Cancer tests and treatment.
Mixed (Q^X)	Mobile number portability	General overview of Mobile number portability, Number lookup services, Cost for portability in different area / regions / country.
	Hair straightening	Hair straightening cream, Hair straightening methods, Hair straightening machines, Images for straight hairs, Methods for Hair straightening, Side Effects of Hair straightening.
	Big boss	Big boss Videos, Big boss Participants and Big boss previous winners.
	Katrina Kaif	Images for Katrina Kaif, Videos of Katrina Kaif, movies of Katrina Kaif, forthcoming movies of Katrina Kaif.
	DPS Bangalore	Overview of DPS Trust ,foundation year, Images of DPS, reviews, locations of DPS, fees structure and admission procedure
	Firefly e-ventures Pvt. Ltd	Overview of Company Profile, CEO, list of key executives, Location and Products/Stats.
	Dinosaur	Dinosaur-Movie, Dinosaur Images, Dinosaur types.
	The Big Bang theory	Production-directed by, theme song, Actors, editors, number of episodes, language country.
	Ballistics	Introduction and history of Ballistics, Ballistics ppt, Ballistics tutorials, use of Ballistics, Gun ballistics, forensic ballistic.
	Books by Durjoy Dutta	Information, Books Rating and reviews, Release dates of the books and Price of the books

Let C^T = { C^T_K }, K = 1, 2, …, k be the set of 'k' number of terms used for evaluation of technical web pages.

Let C^M = { C^M_U }, U = 1, 2, …, u be the set of 'u' number of terms used for evaluation of medical web pages.

Let C^X = { C^X_P }, P = 1, 2, …, p be the set of 'p' number of terms used for evaluation of mixed web pages.

Hence C can be written as

$$C = \bigcup_{K=1}^{k} C^T_K \bigcup_{U=1}^{u} C^M_U \bigcup_{P=1}^{p} C^X_P$$

Furthermore, let E = {{ E^T }, { E^M }, { E^X }}, be a set of three different sets i.e. E^T represents the technical terms, E^M represents the medical terms and E^X represents the mixed terms available in resultant web pages of SE.

Let E^T = { E^T_R }, R = 1, 2, …, r be the set of 'r' number of terms available in the web page.

Let E^M = { E^M_V }, V = 1, 2, …, v be the set of 'v' number of terms available in the web page.

Let E^X = { E^X_W }, W = 1, 2, …, w be the set of 'w' number of terms available in the web page.

Hence E can be written as

$$E = \bigcup_{R=1}^{r} E^T_R \bigcup_{V=1}^{v} E^M_V \bigcup_{W=1}^{w} E^X_W$$

The final value of TLS is based on the Cartesian product (CP). CP of sets depends upon the number of sets and the number of elements in the each set.

To demonstrate the above mention mathematical model let us takes the following example. Let 5 number of sets are available i.e. Y = {S, D, Q, L, E}. If n represents the number of elements in corresponding set then each element of S appears n(D) * n(Q) * n(L) * n(E) times and elements of D appears n(Q) * n(L) * n(E) times and so on. But the elements of E appear once in each set. So, the numbers of combinations required are $\pi_{Z=1}^{z} |Y_Z|$ where 0< z < = 5. This whole process of TLS is described with the help of a tree structure as shown in Figure 1.

4.2 Relevancy Score and Relevancy calculation

The downloaded web pages corresponding to the given query are analyzed based on the terms available in the web page. The numbers of terms available in the webpage are compared with the previously defined criteria of SE evaluation. This comparison provides the relevancy of the web page with the given query as per the Three Level Scoring rules and these rules are defined below:

Relevance-: Relevance is the ratio of retrieved relevant documents to the total number of retrieved documents. TLS uses the following criteria to determine the relevancy of a web page with the given query:

a) If more than 70% terms of evaluation criteria are present in the web page then the web page is considered as 'relevant' and given a score 2.

b) If 30% to 70% terms of evaluation criteria are present in the web page then the web page is considered as 'partially relevant' and given a score of 1.

c) All other web pages that lies neither in the categories of 'relevant' document nor in the category of 'partially relevant' document are considered as 'irrelevant' and assigned a score 0. Irrelevant pages are the links that contains irrelevant information to the query and contain duplicate links, inactive links or error messages like file not found, forbidden errors etc. So,

*Relevancy= (sum of assigned score to the relevant and partially relevant web pages / L*2)*100.*

To compute relevancy an algorithm is designed as shown in Figure 2.

User efforts applied by the user to search the information and precision of the returned results of SE can be tested by using the below given formula.

User Effort (Search Cost) - It is the maximum effort that is applied by a user to find the relevant document from the retrieved documents. It can be measured on the basis of relevant and partially relevant documents.

*User Effort = (no. of links followed by user to find the desired information / L)*100*

If the user gets the required information in first web link out of 10 links then the user efforts is only 10%.

Precision- It is the ratio of retrieved relevant and partial relevant documents to the total number of retrieved documents.

*Precision = (total no. of relevant and partial relevant web pages / L)*100.*

Table 2(a) Results for Technical query set

Query	Google (%)	Bing (%)	AltaVista (%)
Java	60	80	80
Hibernate	70	80	60
Db2	90	80	70
Eclipse	60	50	40
Vmware vsphere	70	50	80
Average	**70**	**68**	**66**

Table 2(b) Results for Medical query set

Query	Google (%)	Bing (%)	AltaVista (%)
Peptic ulcer	90	70	70
Polio	90	70	80
Arthritis	70	90	70
Diabetes	60	70	70
Cancer	60	60	60
Average	**74**	**72**	**70**

5. EXPERIMENTAL RESULTS AND THEIR COMPARISON

To perform the experiment domains, queries and criteria are defined as shown in Table 1. The first column in Table 1 Specifies the name of domains, second column specifies the Queries available in those domains and third column specifies the criteria of evaluation. For the experiment the value of S = {Google, Bing, Altavista}, domain D = {Technical, Medical, Mixed}. And for each domain the query set was as: Q_T = {Java, Hibernate, Db2, Eclipse, Vmware vsphere}, Q_M = {Peptic ulcer, Polio, Arthritis, Diabetes, Cancer} and Q_X = {Mobile no. Portability, Hair straightening, Big Boss, Katrina Kaif, DPS Bangalore, Firefly e – venture Pvt. Ltd., Dinosaur – movie, Book by Durjoy Dutta, The Big Bang theory, Ballistics}. SE evaluation criteria have variation in number of terms. Top 10 results are considered most relevant. That's why top 10 links of selected search engine(s) were taken for experiment.

Table 2(c) Results for Mixed query set

Query	Google (%)	Bing (%)	AltaVista (%)
Mobile No. Portability	60	100	80
Hair Straightening	40	60	40
Big Boss	50	60	70
Katrina Kaif	50	60	60
DPS Bangalore	40	80	80
Firefly e-ventures Pvt. Ltd	50	80	80
Dinosaur-movie	80	50	70
Books by Durjoy Dutta	50	70	70
The Big Bang Theory	70	60	70
Ballistics	50	70	70
Average	**54**	**69**	**69**

Table 3(a) Results for Technical query set

Query	Google			Bing			Altavista		
	Relevance (%)	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**
Java	45	60	60	45	60	60	35	50	100
Hibernate	45	50	20	40	40	30	30	30	20
Db2	30	40	40	50	60	30	40	50	80
Vmware vsphere	80	90	30	50	60	100	70	80	100
Eclipse	30	30	30	30	30	40	30	30	70
Average	46	54	32	43	50	52	41	48	74

Table 3(b) Results for Medical query set

Query	Google			Bing			Altavista		
	Relevance (%)	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**
Peptic Ulcer	45	70	40	40	60	50	40	60	60
Polio	25	40	10	50	80	80	30	40	70
Arthritis	40	70	10	35	60	30	35	60	100
Diabetes	35	60	40	25	40	100	25	40	90
Cancer	35	40	30	30	40	30	30	60	70
Average	**36**	**56**	**26**	**36**	**56**	**64**	**32**	**52**	**78**

Table 3(c) Results for Mixed query set

Query	Google			Bing			Altavista		
	Relevance (%)	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**	**Relevance (%)**	**Precision (%)**	**User effort (%)**
Mobile No. Portability	40	50	50	45	60	70	50	70	50
Hair Straightening	50	70	60	50	60	70	50	60	40
Big Boss	35	50	20	15	20	100	45	50	20
Katrina Kaif	60	80	50	70	90	30	50	60	30
DPS Bangalore	65	80	50	60	70	100	65	80	90
Firefly e-ventures Pvt. Ltd	40	50	80	45	70	40	45	70	40
Dinosaur-movie	45	50	60	60	70	50	40	50	70
Books by Durjoy Dutta	80	100	20	55	70	70	55	70	70
The Big Bang Theory	20	20	70	30	30	30	30	30	20
Ballistics	40	40	40	65	80	80	55	70	90
Average	47.5	59	50	48.5	61	52	49.5	62	64

Three SE's were selected for evaluation by using the proposed framework. The Proposed framework to evaluate SE was implemented in C# programming language using the parameters (discussed in Section 4.2) and the results are obtained from the experiments are shown in Table 2(a), 2(b) and 2(c). To cross check the results of experimental results, the manual calculation was also done using the same queries that were given in Table 1. Furthermore, the efforts applied by the user and precision of results were also computed in manual calculations as parameters defined in Section 4.2 and are shown in Table 3(a), 3(b) and 3(c).

5.1 Discussion of Results

The authors of the paper compared the results of three SE's on three domain set viz. Technical, Medical and Mixed. For the technical domain, the experimental results showed that Google had the highest relevancy results i.e. 70% whereas AltaVista gave the results with lowest relevancy i.e. 66% (see Table 2(a). The results of manual calculation also showed that Google returns the results with highest relevancy i.e. 46% and AltaVista with lowest relevancy i.e. 41% (see Table 3(a)).

For the medical domain, the automated results showed that Google has the highest relevancy results i.e. 74%, whereas AltaVista gave the results with lowest relevancy i.e. 70% (see Table 2(b)). The results of manual calculation also showed that both Google & Bing returned the results with highest relevancy i.e. 36% as well as highest precision i.e. 56%, whereas AltaVista showed lowest relevancy i.e. 32% and lowest precision 52% (see Table 3(b)). However, in terms of user effort, Google edged out Bing with lowest effort required i.e. 26% compared to 64% and 78% required by Bing and AltaVista, respectively.

For mixed domain query set Bing and AltaVista showed highest relevancy i.e. 69% whereas Google showed the lowest relevancy i.e. 54% in experimental results. Whereas in manual calculation AltaVista showed a little better relevancy than Bing (49.5% of AltaVista compare to 48.5% of Bing) whereas Google showed the lowest relevancy of 47.5%.

6. CONCLUSION

This paper has presented a new framework to evaluate SE's in a particular domain based on TLS method. The experimental results have shown that the proposed framework is capable of differentiating SE's on the basis of their performance. The experimental results have tallied with the manual results indicating the validity of the proposed framework. The final results showed that for technical and medical query set Google returned the most relevant results with higher precision and less user efforts. Whereas for mixed queries Altavista returned most relevant results with higher precision. Hence, the proposed work is able to make the difference between different search engines.

7. REFERENCES

[1] Shang Yi and Longzhuang Li (2002), "Precision Evaluation of Search Engines", in WWW, ACM journal, Volume 5, Issue 2, pp. 159 – 173.

[2] Xiaoying Dong and Louise T. Su (1997),"Search engines on the world wide web and information retrieval from the internet: A review and evaluation", in Emerald – online information *Review*, pp. 67–81.

[3] Leighton H. Vernon and Srivastava Jaideep (1999), "First 20 precision among world wide web search services, in Journal of the American Society for Information Science, pp. 870–88.

[4] Wishard, Lisa (1998) Precision Among Internet Search Engines: An Earth Sciences Case Study. Issues in Science and Technology Librarianship. No. 18, Spring: http://www.library.ucsb.edu/istl/98-spring/article5.html.

[5] Fazli Can, Nuray Rabia, Sevdik A. B. (2003), "Automatic performance evaluation of Web search engines", in Information Processing and Management (Elsevier), pp. 495-514.

[6] Kumar Rakesh, Suri P.K., & Chauhan R.K. (2005), "Search Engines Evaluation", in DESIDOC Bulletin of Information Technology, Vol. 25, No. 2, pp. 3-10.

[7] Shang Y. and Longzhuang (2000), "A new method for automatic performance comparison of search engines" in World Wide Web (Springer), pp. 241 – 247.

[8] Fu Yiqun Liu, Yupeng, Zhang Min, Shaoping Ma, Liyun Ru (2007), "Automatic Search Engine Performance Evaluation with Click-through Data Analysis", In Proceedings of the 16th international conference on World Wide Web, pp. 1133-1134.

[9] Cormack G. V.,. Clarke C. L. A, Palmer C. R. (1999), "Fast Automatic Passage Ranking" in Proceedings of the Eighth Text Retrieval Conference pp. 735-742.