International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 2 214 - 220

Measures to Evaluate the Superiority of a Search Engine

Soumya George[1], M. Sudheep Elayidom[2], T. Santhanakrishnan[3]
[1]Research Scholar, Department of Computer Applications, Cochin University of Science and Technology, Kochi-22, India
mariyam.george@gmail.com
[2]Associate Professor, Division of Computer Engineering, Cochin University of Science and Technology, Kochi-22, India
sudheepelayidom@hotmail.com
[3]Scientist, Govt. of India, Ministry of Defence, Naval Physical and Oceanographic Laboratory, Thrikkakkara, Kochi-21, India
santhannpol@gmail.com

Abstract— Main objective of a search engine is to return relevant results according to user query in less time. Evaluation metrics are used to measure the superiority of a search engine in terms of quality. This is a review paper presenting a summary of different metrics used for evaluation of a search engine in terms of effectiveness, efficiency and relevancy.

Keywords— *Evaluation, search engine, relevancy, metrics, effectiveness, efficiency*

*****

I. INTRODUCTION

Search Engines are Information retrieval systems that are used to mine information from the Big Data existing on the web. Most popular search engines include Google, Bing, and Yahoo etc. In general a search engine consists of 3 main tasks: web crawling, indexing and searching. Web crawling is an automated iterative process done by web crawlers, spiders, robots etc. that traverse through web to fetch relevant data for indexing using link traversal starting with a seed URL. Crawling techniques includes focused crawling and distributed crawling. Focused crawler will focus on documents on a particular topic or domain, say sports. Distributed crawlers crawl the documents in a distributed fashion using many processes so that it needs less time to complete.

The main component of a search engine is the inverted index and the posting lists which stores each unique term appearing in all web pages with a pointer to all the documents containing that term. Google uses full text indexing method that indexes all words. keyword indexing indexes only important words or phrases (e.g. Lycos).There is also another technique where humans find the key terms or phrases to index from all web pages to create human powered directories that organizes web pages by category or subject, which is followed by the Yahoo search engine[1].

Searching process uses Boolean methods or advanced user options to find the list of documents in which all keyword terms in the user queries appear and display results in the decreasing order of similarity.

II. EVALUATION OF A SEARCH ENGINE

Evaluation of a search engine is used to measure the quality of a search engine in terms of results and response time, mainly used to evaluate time, space, cost, and usability and retrieval performance of the retrieval system. Each phase of the search engine and the algorithms used like crawling algorithms, page ranking algorithms etc. affects the quality [1].The major goal of search engine is to return relevant results according to user's needs in less time. The measurable criteria for evaluating a search engine includes the no: of documents indexed per hour to measure the indexing speed, latency which measures the amount of time taken to search the user need and the ability and time taken to process complex queries[2].

Cleverdon [3] suggested six criteria for search engine evaluation. These criteria includes amount of documents in the collection called coverage, total time needed to process the query and to return results called the latency or time lag, how results are presented to the user, user's effort to search, recall which measures the exhaustiveness and precision used to measure accuracy.

Evaluation of search engine requires a corpus of documents, a set of queries and a collection of relevant documents to each query. A number of evaluation corpus or test collections are available to evaluate search engines. These collections are made by experts in the field. E.g., CRANFIELD, TREC, GOV2collections [4].

Evaluation metrics are the different measures used to measure the superiority of search engine in terms of quality. These measures include system centric measures and user centric measures as shown in Fig. 1[5] [6] [7][25].

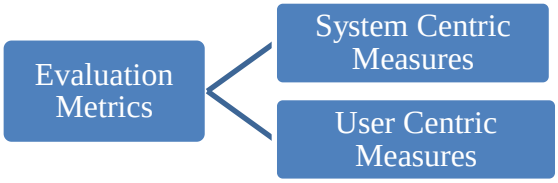

Fig. 1 Classification of evaluation metrics

III. SYSTEM CENTRIC MEASURES

System centric metrics include measures to evaluate the effectiveness and efficiency of search engines as shown in Fig. 2. These measures are needed to build better search engines. Effectiveness measures the ability of a search engine to find the relevant results and efficiency measures the time and space requirements of a search engine for the overall working[29].

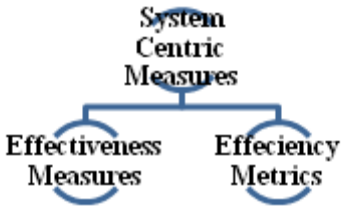

Fig. 2 System centric measures

A. *Effectiveness Measures*

Effectiveness measures are used to measure the effectiveness of the search engine in terms of user friendliness, ease of use, ability to find relevant results that satisfy user needs etc. [27]. Precision and recall are the main measures used to evaluate the superiority of a search engine.

1) *Recall / Precision and other related measures*: Precision and Recall are the most widely used basic measures to evaluate search engine. Fig. 3 shows the different precision or recall related measures used to evaluate search engine.

Recall / Precision & related measures

- Precision & Recall
- Interpolated Precison
- Recall-precision Graph
- Fallout
- AP, MAP & GMAP
- R-precision
- F-measure & E-measure
- Normalized Recall
- Utility Measure

Fig. 3 Recall / Precision & related measures

- *Precision and Recall*

Precision and recall are defined in terms of a set of retrieved documents and relevant documents[26]. Retrieved documents include the list of documents displayed by the search engine in response to user query and relevant documents include the list of all web documents that are relevant to user query. Superiority of a search engine depends on the number of relevant retrieved results relative to the total no: of relevant documents in the entire web

Precision measures the accuracy and is the proportion of relevant documents to the retrieved pages [8].

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Recall measures the exhaustiveness and is the proportion of retrieved documents to the relevant set of pages.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Precision and Recall are inversely proportional to each other. So when precision increase, recall decrease. So the desired high point of effectiveness is all relevant documents should be retrieved before the first non-relevant document in the collection.

The set of all relevant documents is represented by {relevant documents}. {Retrieved documents} represents the set of all retrieved documents. {Relevant} ∩ {retrieved} represents the set of all retrieved documents that are relevant. These relationships can be represented as Fig. 4 [9].

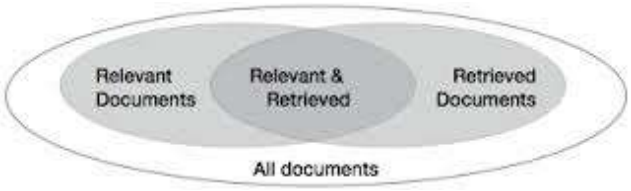

Fig. 4 Venn diagram showing {relevant}, {retrieved} and {relevant} ∩ retrieved} documents

Breakeven point is the point where precision becomes equal to recall.

- *Fall-out*

Fall-out is the ratio of number of non-relevant documents retrieved to the total number of non-relevant documents.

$$\text{fall-out} = \frac{|\{\text{non-relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{non-relevant documents}\}|}$$

In *Swets model*, precision, recall and fallout is calculated based on conditional probability.

The contingency table showing the relationship between relevant, non-relevant, retrieved and not retrieved is shown in the Fig. 5[10].

	RELEVANT	NON-RELEVANT	
RETRIEVED	$A \cap E$	$\bar{A} \cap E$	E
NOT RETRIEVED	$A \cap \bar{E}$	$\bar{A} \cap \bar{E}$	$\bar{E}$
	A	$\bar{A}$	N

(N = number of documents in the system)

Fig. 5 Contingency table

By using contingency table, precision, recall and fallout can be calculated as,

$$\text{PRECISION} = \frac{|A \cap E|}{|E|}$$

$$\text{RECALL} = \frac{|A \cap E|}{|A|}$$

$$\text{FALLOUT} = \frac{|\bar{A} \cap E|}{|\bar{A}|}$$

- *F-score / F-measure and E-measure*

F–measure is the weighted harmonic mean of Precision and Recall [11].

$$= 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

E-measure is a variant of F-measure used to give more weightage to precision or recall rather than giving equal importance.

$$= (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

- β is the parameter that gives the relative weightage to precision or recall.

- *Interpolated precision*

For each selected recall levels, interpolate the precision values by finding the maximum known precision value at any recall level above the current level. . The interpolated precision at the *j*-th standard recall:

$$P(r_j) = \max_{r_j \le r} P(r)$$

- *11-point interpolated average precision* - It is the average of the interpolated precision values calculated at 11 standard recall levels from 0 to 1.

- *Non – interpolated Average Precision (AP) & Mean Average Precision (MAP) & Geometric Mean of Average Precision (GMAP)*

Non-interpolated Average Precision, AP is a metric that finds average of the precision at each point where a relevant document is retrieved [12].

$$AP = (\sum_{i=1}^{R} \frac{i}{rank_i}) / R$$

Where R = Total no: of relevant documents for the query and $i/rank_i = 0$, if document i was not retrieved.

Mean Average Precision, MAP is the mean of the average precision values calculated for all set of topics [13]. MAP for a set of queries, Q can be calculated as,

$$MAP = (\sum_{i=1}^{Q} AP_i) / Q$$

Geometric Mean of Average Precision, GMAP uses geometric mean of the average precision values which is calculated as,

$$= \exp \frac{1}{n} \sum_{n} \log AP_n$$

Where n denoted the o: of queries and AP represents the average of precision values for query k.

- *R- Precision*

It is the calculation of precision value after R relevant documents are retrieved [14].

$$\text{R-precision} = \frac{1}{R} \sum_{i=1}^{R} d_i$$

Where d_i is the relevance level of i^{th} document in the ranked output.

- Average R-Precision-For all queries, take the average of the R-precisions.

- *Recall – Precision Graph*

It is the commonly used method for evaluating and comparing systems created by plotting recall values against precision. The same graph can be used to plot values of different trials and the curve which determines the ideal retrieval is the upper right hand curve where precision and recall value is maximum. Fig. 6 shows a Recall – Precision curve [15].

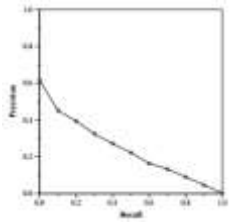

Fig. 6 Recall –Precision curve

- *Utility Measure*

Based on whether user is giving more importance to recall or precision, a utility measure is defined as,

$$\mathbf{A.}\ N_r + \beta.\bar{N}_r + \gamma\ .N_n + \delta.\bar{N}_n$$

Where N_r denotes the no: of retrieved documents which are relevant, $\bar{N}_r$ represents the no: of relevant documents that are not retrieved, N_n denotes the retrieved non-relevant documents and $\bar{N}_n$ represents the no: of documents not retrieved and are not relevant. α , β, γ and δ are the positive weights assigned according to user [16].

- *Normalized Recall*

Normalized Recall is calculated from the comparison between actual rank and ideal ranks calculated as,

$$R_{norm} = 1 - \frac{\sum_{i=1}^{n} r_i - \sum_{i=1}^{n} i}{n(N-n)}$$

Where N is the total number of retrieved documents and n represents the relevant no: of documents and r_i is the i^{th} relevant document [17].

2) *Other system centric measures:* A number of other system based measures are available as shown in Fig. 7.

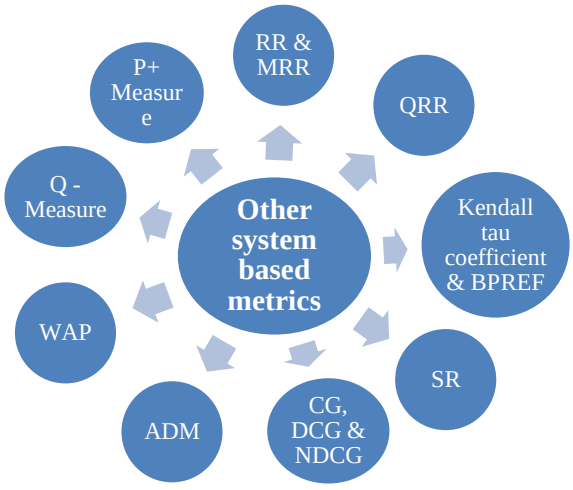

Fig. 7 System centric measures

- *Reciprocal Rank (RR) & Mean Reciprocal Rank (MRR)*

Reciprocal rank is calculated by taking the reciprocal of the rank of the first relevant result. RR is zero, if no relevant results obtained [7].

$$RR = \begin{cases} 0 & \text{if no relevant results} \\ \frac{1}{r} & \text{else} \end{cases}$$

Where r is the first rank when a relevant result is found.

Mean Reciprocal Rank is calculated from the mean of the RR value for each query.

$$MRR = \frac{1}{|Q|}\sum_{q \in Q} RR_q$$

Where Q is the set of queries and RR_q is the reciprocal rank measured for query q.

- *Quality of Result Ranking (QRR)*

QRR is a correlation based measure between ranking of search engine produced results and a new list created from the results ranked using user assigned relevance.

- *Kendall tau coefficient (τ) & Binary Preference (BPREF)*

Both the above measures are calculated using preferences, based on document relevancy. The Kendall tau coefficient *(τ)* for two rankings described using preferences where P is the no: of preferences that agree and Q is the number of preferences that disagree is

$$\tau = \frac{P - Q}{P + Q}$$

BPREF gives more preference to relevant documents of given topic using binary relevance judgments [18].

$$BPREF = \frac{1}{R}\sum_{d_r}(1 - \frac{N_{d_r}}{R})$$

Where d_r represents a relevant document and N_{dr} represents the number of non-relevant documents. BPREF can also be calculated as,

$$BPREF = \frac{P}{P+Q}$$

- *Sliding Ratio (SR)*

It is the ratio between search engine's actual output of ranked list and an ideal raking of same list of documents for every rank upto a limit or threshold value [19].

$$= \frac{\sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} d_{I(i)}}$$

Where d_i is the relevance score of each i^{th} ranked document in the list and $d_{I(i)}$ represents its ideal rank. A good ranking system should display documents in the decreasing order of relevance, i.e. if $d_i > d_j$, then document i should be displayed before document j. But SR is not sensitive to document relevance order. So, a modified sliding ratio, msr is proposed which is calculated as,

$$= \frac{\sum_{i=1}^{n} \frac{1}{i} d_i}{\sum_{i=1}^{n} \frac{1}{i} d_{I(i)}}$$

- *Cumulative Gain (CG), Discounted Cumulative Gain (DCG), Normalized Discounted Cumulative Gain (NDCG)*

Cumulative Gain, CG is the sum of the relevance score of all documents upto a particular rank [19]. CG is calculated as,

$$= \sum_{i=1}^{n} d_i$$

Where d_i is the relevance score of i^{th} ranked document.

Discounted Cumulative Gain, DCG is a graded relevance measure by giving importance to the position of the document to calculate the gain factor of the documents. Gain is for top rank documents upto a certain rank, P and discounted for documents with low rank. Normal discount value is 1/log (rank) [18].DCG_P is measured from the total gain accumulated at a particular rank P and is calculated as,

$$= rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

Or

$$= \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

Normalized Discounted Cumulative Gain, NDCG is the DCG value normalized using ideal DCG value, $IDCG_P$[18][27][28].

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

- *Average Distance Measure (ADM)*

For a query, Average Distance Measure, ADM is the average difference between system's relevance score and user assigned relevance score for the output list of documents, calculated as,

$$= 1 - \frac{\sum_{d \in D} |SRS(d) - URS(d)|}{|D|}$$

- *Weighted Average Precision (WAP)*

It is the extension of Average Precision, AP based on multigraded relevance score. It is calculated as,

$$= \frac{1}{R}\sum_{n=1}^{N} rel(d_n)\frac{cg}{cg_I}$$

Where R represents the no: of relevant documents, rel (d_n) represents the relevance score of nth ranked document, cg represents the cumulated gain and cg_I represents the cumulated gain of ideal ranking.

- *Q – Measure*

Q – Measure is a graded average precision value, graded AP. It is calculated as

$$= \frac{1}{R} \sum_{n=1}^{N} rel(d_n) \frac{cbg}{cg_I + n}$$

Where R represents the no: of relevant documents, rel (d_n) represents the relevance score of nth ranked document, cg_I represents the cumulated gain of ideal ranking and cbg represents the cumulated bonus gain value. Cbg is calculated as,

$$= \sum_{i=1}^{n} bg_i$$

Where $bg_i = d_{i+1}$ if $d_i > 0$, else 0.

- *P^+ Measure*

P+ Measure is a variant of Q – Measure [20]. It is calculated as

$$= \frac{1}{C(r_p)} \sum_{r=1}^{r_p} I(r) BR(r)$$

Where C (r_p) is the set of top relevant documents above or at rank r_p, I(r) = 1 if the document at rank r is relevant, otherwise 0. BR(r) represents the blended ratio calculated as,

$$= \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)}$$

Where β (>=0) is the user persistence parameter, and inherits both precision and normalized cumulative gain as precision(r) = C(r)/r and NCG = cg(r) / cg^*(r) respectively.

B. *Efficiency Metrics*

Efficiency measures the time and space requirements of a search engine for the overall working. It measures the amount of memory space, disk space, CPU time and other resources used for the overall working [6]. Some of the efficiency metrics used is listed in the Fig. 8. [21].

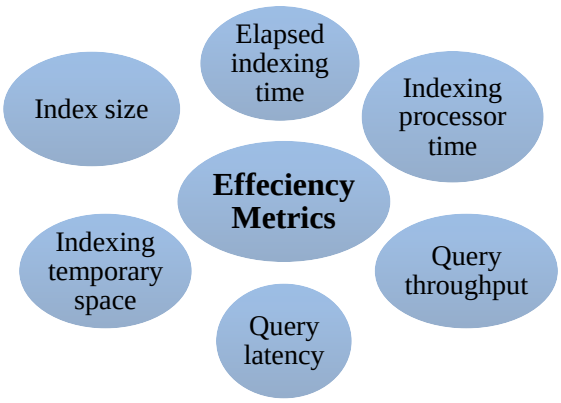

Fig. 8 Efficiency Metrics

1) Elapsed indexing time: It's the total time taken for indexing documents.
2) Indexing processor time: Total CPU time for indexing measured in seconds without considering the waiting time for I/O or speed gains from parallelism.
3) Query throughput: Total number of queries processed per second.
4) Query latency: Time for getting first response after issuing the query, measured in milliseconds.
5) Indexing temporary space: Total temporary disk space used for indexing.
6) Index size: Total space used for indexing files.

IV. USER CENTRIC EVALUATION

User centric measures are mainly used to measure user satisfaction. Relevance measure is different for different users and is multidimensional in nature[e].Various measures for user centric evaluation is given in Fig. 9[7].

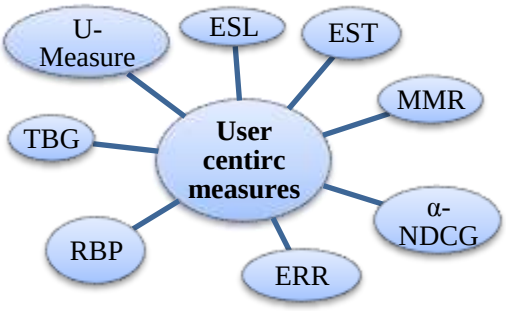

Fig. 9 User centric evaluation measures

A. *Expected Search Length (ESL)*

The number of non-relevant documents user has to visit in the hit list before finding the needed relevant ones is measured by using Expected Search Length, ESL.

B. *Expected Search Time (EST)*

It's the metric used to estimate the total time duration for user's search session to interact with search engine before finding the needed relevant results.

C. *Maximal Marginal Relevance (MMR)*

MMR uses a novelty factor to measure the relevance by finding similarity between query and retrieved documents.

D. *α–Normalized Discounted Cumulative Gain (α-NDCG)*

It uses novelty-biased gain measure which is based on the no: of information nuggets, which are the properties or part of documents that are relevant to the user [20].

$$= \sum_{i=1}^{m} li(r) (1-\alpha)^{reli(r-1)}$$

Where $\sum_{i=1}^{m} li(r)$ represents the graded relevance measured based on the no: of information nuggets, "m" presents in the document and li(r) denotes the relevance flag for each of them. $(1-\alpha)^{reli(r-1)}$ discounts the gain based on diminishing return where α denotes the probability that user "finds" a nonexistent nugget in document and reli(r) is the number of relevant documents for each nugget.

E. Expected Reciprocal Rank (ERR)

ERR measure is also based on diminishing return property which discounts the value of all new relevant documents after a relevant document is found. It is calculated as,

$$\Sigma_r \, dsat(r-1)\, Pr(r)\, (1/r)$$

Where dsat(r-1) Pr(r) represents the probability of user satisfaction at r and 1/r denotes the utility at r.

$$dsat(r) = \prod_{k=1}^{r}(1-Pr(k))$$

F. Rank-Biased precision(RBP)

RBP is a user based model that assumes the probability p for user visiting the next document after visiting a document at rank r or stop at probability (1-p).

$$= (1-p)\, \Sigma_r\, p^{r-1}\, g(r)/gain(H)$$

Where g(r) is the gain value at rank r and gain (H) is the gain value at highest relevance level, H.

G. Time-Biased Gain (TBG)

TBG is a binary relevance measure based on diminishing return property which uses the time taken by the user to reach rank r instead of document rank to discount the document value calculated for each relevant document as ,

$$= \sum_r g(r) \exp(-T(r)\frac{\ln 2}{h})$$

Where T(r) represents the time to reach rank r, h represents the discount function which is the half life period and g(r) represents the gain value. T(r) is calculated as,

$$= \sum_{k=1}^{r-1}(T_S + Pr_{click}(k) T_D(k))$$

Where TS represents the time in seconds to read the document of length, *L*(*r*) based on the number of words, *Prclick* (*r*) is a probability of click at r based on document relevancy and *TD*(*r)* denotes the estimated time to read the document at *r and calculated as*

$$= 0.018L(r) + 7.8$$

H. U- Measure

U-Measure can handle diversified search, multi-query sessions, summaries etc. Which is calculated based on trail text, a collection of strings which represents all texts that user read during searching. It is calculated as,

$$= \frac{1}{\mathcal{N}} \sum_{pos=1}^{|tt|} g(pos) D(pos)$$

Where N represents the normalization factor, g (pos) represents the position-based gain and D (pos) represents the position-based decay function for each string in trailtext, tt.

V. OTHER ADVANCED METRICS

A number of other advanced metrics are available like diversified search metrics, session metrics etc [20].

A. Diversified Search Metrics

Diversified search metrics are used when user queries are not specific or precise. So that search relevance will be based on intents or sub topics. Various measures include:

1) Subtopic / Intent Recall(I-rec):

This measure is used to find the no: of intents captured by the search engine calculated as,

$$= \frac{\sum_r newint(r)}{|\{i\}|}$$

Where "i" represents the no: of intents and newint(r) represents no: of intents covered at rank r calculated as,

$$newint(r) = \sum_i isnew_i(r) I_i(r)$$

Where $I_i(r)$ is 1 or 0 based on whether the document at rank r are relevant to intent i or not respectively. $isnew_i(r)$ be 1 I_i (k) = 0 for $1 \leq k \leq r-1$, and 0 otherwise

2) *Intent-Aware Metrics:*

A metric used for diversity evaluation proposed by Agarwal at al. [22]. Let pr (i/q) be the probability of each intents "i" for a given query "q" and M_i be the value computed for each intent using nDCG measure or any other metric. Then Intent-Aware metric can be computed as,

$$= \sum_i Pr(i|q) M_i$$

A number of other metrics are also available for diversity evaluation like D-Measures, Intent-Type-Sensitive metrics etc. D-Measure was proposed by Sakai and Song [23] that takes into account probability and graded relevance assessment for each intent for evaluation. Intent-Type-Sensitive metrics is sensitive to the navigational intent type or informational intent type labels.

B. Session Metrics

Session metrics are used to evaluate multiple lists of ranked documents during multi-query sessions.

1) Session DCG:

Session Discounted Cumulative gain measure is a session based metrics by concatenating the top "p" documents from the multiple ordered ranked lists. It's an extension of nDCG in which relevant document value will be discounted by its rank in the new list and no: of queries needed to access each. It is calculated as,

$$= \sum_{\mathbf{r}} \frac{g(\mathbf{r})}{\log_4(qnum(\mathbf{r})+3)\log_2(\mathbf{r}+1)}$$

Where g(r) is the relevance based or click-based gain at r and qnum(r) is the query number of document at r in the new list.

2) Click-Based U Measure:

It's a user based measure based on no: of user clicks involving multiple sessions or queries by constructing trailtext.This measure is based on assumption that user will click only relevant document.

VI. RELEVANCE METRICS

It's crucial to check whether the search engine return the most relevant results according to user's needs. Major features and criteria affecting relevancy include Search analytics, Content analysis, Geographic trends, Time based

trends, contextual searches, Social signals, personalized search [24].

A. Search Analytics

It calculates the frequency of search for each search term, total time spent by the users on a particular page, hit rate for Search result etc.

B. Content analysis

Relevancy of a results should not depend entirely on the result's title. There can occur mismatch between a document title and its contents. So proper content analysis should be done to check whether the title convey its contents accurately.

C. Geographic trends

User's search behavior and expectations upto some extent depends on the geographic region they belongs to and this constraint should also be taken into account to choose the hit list.

D. Time based trends

Time of the search and the searching parameter have a significant influence on the data user needs to capture. So time based trends is also an important parameter for measuring relevancy.

E. contextual searches

Its vital to consider the context in which the user is searching for, to capture user 's choice of relevant results.

F. Social signals

Social interactions is an important base factor nowadays in social networks or shopping sites etc and these social signals in the form of recommendations, suggestions and opinions also influence the search results.

G. Personalized search

Search results can be influenced by the user's search behavior by tracking past records, say purchase habits etc by using cookies.

VII. CONCLUSIONS

Superiority of a search engine is measured in terms of quality of results and response time. A no: of evaluation metrics are available to measure the superiority of a search engine in terms of efficiency and effectiveness. This paper presents a summary of a no: of evaluation metrics available to measure the quality of a search engine.

REFERENCES

[1] Monica Peshave and Kamyar Dezhgosha, How Search Engines Work and a Web Crawler Application, http://ebooks5.org/h/how-search-engines-work-and-a-web-crawler-application-book-w1245/
[2] Christopher Manning and Prabhakar Raghavan, Lecture 8: Evaluation, Introduction to Information Retrieval CS276 ,Information Retrieval and Web Search, http://www.cs.uvm.edu/~xwu/wie/CourseSlides/Evaluation.pdf
[3] Cleverdon, C.W., Mills, J., and Keen, E.M.., "An inquiry in testing of information retrieval systems", Cranfileld, U.K.: Aslib Cranfield Research Project, College of Aeronautics, 1966, pp. 230-232.
[4] http://nlp.stanford.edu/IR-book/pdf/08eval.pdf, Online edition (c) 2009 Cambridge UP
[5] Dietmar Wolfram," Applied Informetrics for Information Retrieval Research", New Directions in Information management, Number 36
[6] Alistair Moffat, Falk Scholer, Paul Thomas, "Models and Metrics: IR Evaluation as a User Process",
[7] Pavel Sirotkin, "On Search Engine Evaluation Metrics," Doctoral thesis, University of Dusseldorf, Germany, April 2012.
[8] Information Retrieval, Available: https://en.wikipedia.org/wiki/Information_retrieval
[9] Data Mining - Mining Text Data, Available: http://www.tutorialspoint.com/data_mining/dm_mining_text_data.htm
[10] Evaluation, Available: http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html
[11] Precision and recall, Available: https://en.wikipedia.org/wiki/Precision_and_recall
[12] Mark Sanderson, Test Collection Based Evaluation of Information Retrieval Systems, Foundations and Trends in Information Retrieval
[13] Evaluation and Relevance, Available: http://web.cecs.pdx.edu/~maier/cs510iri/IR_Lectures/CS_510iri_Lecture8RelevanceEvaluation-revised.pdf
[14] Bing Zhou and Yiyu Yao, "Evaluating Information Retrieval System Performance Based on User Preference",
[15] Common evaluation measures, Available: http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf
[16] Tanveer Siddiqui and U. S. Tiwary, Natural Language Processing and Information Retrieval, Oxford Higher Education
[17] Evaluation of IR, Available: web.simmons.edu/~benoit/lis466/slides/LIS531H-Evaluation.ppt
[18] J. Pei, Information Retrieval and Web Search – Evaluation, Available: L19 – Evaluation.pdf
[19] Bing Zhou and Yiyu Yao, "Evaluating Information Retrieval System Performance Based on User Preference"
[20] Tetsuya Sakai, Metrics, Statistics, Tests, February 6, 2013@PROMISE Winter School 2013 in Bressa none, Italy
[21] Evaluating Search Engines, Available: http://web.simmons.edu/~benoit/lis466/Evaluation-hap8.pdf
[22] R. Agrawal, S. Gollapudi, A. Halverson, and S. Leong. Diversifying search results. In Proceedings of ACM WSDM 2009, pages 5–14,2009.
[23] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C.-Y. Lin. Simple evaluation metrics for diversified search results. In Proceedings of EVIA 2010, pages 42–50, 2010.
[24] Available: http://www.infosys.com/manufacturing/resource-center/Documents/search-engine-evaluation.pdf
[25] Demartini, G., & Mizzaro, S., "A Classification of IR Effectiveness Metrics", In M. Lalmas (Ed.), European Conference on IR Research (pp. 488-491). London, UK: Springer, 2006.
[26] Ali Dasdan, Kostas Tsioutsiouliklis and Emre Velipasaoglu ,"Web Search Engine Metrics for Measuring User Satisfaction", 18th International World Wide Web Conference, 2009
[27] Jitendra Nath Singh and Dr. S.K. Dwivedi, " Evaluative Measures of Search Engines", IJECSE,Volume1,Number 2
[28] Ahmed Hassan, Yang Song and Li-wei He, "A Task Level Metric for Measuring Web Search Satisfaction and its Application on Improving Relevance Estimation", CIKM'11, Copyright 2011 ACM
[29] S.Sathya Bama, M.S.Irfan Ahmed and A.Saravanan, " A Survey On Performance Evaluation Measures For Information Retrieval System", (IRJET) Volume: 02 Issue: 02 | May-2015