IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10143

A study of Page Relevance in Information Retrieval

Kompal

Govt. College, Panchkula

Abstract: Search engines and web crawlers aim to provide users with the most relevant and useful information in response to their search queries.In web crawling, the relevance of a page refers to how well it matches a user's search query or a specific topic. Search engines uses various complex algorithms to determine the web page relevance to ensure that users receive the most accurate and useful results when they perform a search.

I. PAGE RELEVANCE

It is a important factor in determining search engine ranking, which determines the order in which pages are presented to users in response to a search query. Many techniques exist for calculating page relevance.

II. RELEVANCY CALCULATION TECHNIQUES

2.1 Weighted Page Rank

In this, Web page weight is calculated based on outgoing and incoming links and based on weight, the page importance is decided. The relevancy is less as ranking of the page is calculated based on Web page weight at indexing time.

2.2 HITS

HITS is search algorithm called Hyperlink-Induced Topic Search. The relevant pages authority values and hubs is computed by this algorithm. The important relevant page is result of the algorithm. When two pages having roughly the same number of citations are compared, the one receiving many citations from P1 and P2 (called prestigious or important pages) needs to be highly ranked.

2.3 Eigen Rumor Algorithm

It is a challenge to display quality blogs to users due to large number of blogs on the web for the service providers. The low rank scores is given to blogs and such scores cannot be used to decide the blog importance. To resolve this problem, Fujimura proposed an algorithm for blogs ranking. This Eigen Rumor Algorithm provides to each blog a rank score by weighting the hub score and the bloggers authority based on the eigen vector calculation.

2.4 Hawk Algorithm

The HAWK algorithm predicts and select the relevant URL based on web page content, and the URL priority was determined from the crawled queue.

The advantage of this algorithm is that the web page content was not only used for improvement of relevance of the webpage, but the link structure of the webpage is used for improvement of specific topic coverage.

2.5 Page Rank Algorithm

The web page importance is determined by counting page citations or back links.

2.6 Ontology based focused crawling

In this system, the processes interact with each other. The crawling cycle and ontology cycle are two main processes. In the ontology cycle, the ontology define the crawling target and the relevant documents for the ontology enrichment are returned to the user. The documents on the web are retrieved by the crawling cycle and the documents relevance is determined by interacting with the ontology

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10143

2.7 Based on Classifier

It is a approach based on learning and to improve the unvisited URLs relevance prediction without downloading and many pages are visited irrelevant in nature. In this technique, the unvisited URLs classification is done based on visited URLs attribute score, i.e. Anchor text relevancy, parent page relevancy, cohesive text relevancy, URL relevancy. The vector space model is used in calculation of Relevancy score and supervised or unsupervised classifier is used for classification.

2.8 Based on Page Weight

The bandwidth is used for downloading the page. The used bandwidth can be better utilized and more can be get out of it. The downloaded pages bandwidth can be used and the same bandwidth can be used to get the body, title and the outgoing links on that page particularly. The page weight can be calculated based on occurrence of keywords in various locations of page and weightage given to title, body and links on that page. The higher the page weight, higher will be page relevance.

2.9 Based on Topic specific Weight table

The page relevancy can be calculated by using Topics. Topic is the set of keywords and its associated weights. The following equation defines the topic vector and can be written as given in equation (1).

Topic = {(k1, w1), (k2, w2)…..(kn ,wn)}…… (1)

where k_j represent j^{th} keyword of topic T.

w_j is the j^{th} keyword weight and represent j^{th} keyword importance in the topic T.

The different sources provides the topic by asking from the user to specify them.

The Google web search engine had been given a topic and first few results are retrieved based on the topic. The assigned weight wi = tf * df,where df and tf is document frequency and term frequency respectively. The weights had been normalized after assigning words weight by using the following equation

$W_{i+1} = W_i / W_{max}$

Where W_{max} is the maximum weight of any keyword and W_{i+1} is the new weight assigned to any keyword. The topic specific weight table helps in determination of page relevancy and by calculating page weight for each keyword in topic specific weight table. The higher the page weight, higher will be relevancy.

2.10 Rank Aggregation Approaches

The rank aggregation is used to acquire a “better” ordering by merging many distinct rank orderings on the similar set of candidates, or alternatives. It is used comprehensively in the background of social choice theory in which various “voting paradoxes” have been revealed. The conflict also rises in most other situations such as meta-search and mutual filtering in machine learning, merging results from numerous databases in database middleware and concept of correlation in statistics. The approach of rank combination can combine various ranked terms used for query expansion into a single list of terms. A few of terms scored higher selected from the generated terms list are taken as QE terms for the query of the user. A few methods for rank combination based on rank positions are given below:

2.10.1 Borda Ranking Approach

According to this approach, the list of candidates preferred by each voter has its own. For each voter, m points obtains by top candidate, m-1 points obtains by the next to top candidate, and *m-2* points obtains by the candidate having position top third and so on. The addition of points obtained by each voter gives the final points and provides to each candidate. Some candidates are also unranked. So the balance points are equally divided among the unranked candidates. The highest points attaining candidate wins. A few candidate terms highly ranked selected by Borda scheme are used for expanding the user query. This type of query expansion is called Borda Based Query Expansion (BBQE).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 1, January 2021

DOI 10.17148/IJARCCE.2021.10143

2.10.2 Condorcet Ranking Approach

This algorithm of ranking aggregation is based on majority concept. In this algorithm, every other candidate beat the winner candidate in the pair-wise comparison. The candidate which was not ranked by voter lose score than other ranked candidates. If unranked candidates are more than one then all unranked combine with each other. Some of the high ranked candidate terms which are chosen by Condorcet scheme are utilized for extending the user query.

2.10.3 Reciprocal Ranking Approach

In Reciprocal ranking approach, top first candidate term for each voter obtain score one, and the candidate term on second top position obtain score one-half and the candidate term on position top third obtain score one-third and so on. A unranked candidate term by a voter is not used for the computation.

At last, depending on their final scores, all the candidate terms are ranked. Some of the high ranked candidate terms selected by the reciprocal approach are utilized for extending the user query.

2.10.4 SumScore Ranking Approach

The addition of the similarities scores from all QE terms selection techniques produces the joined value of similarity score of each candidate term. Before joining, the similarity score of candidate terms acquired from distinct QE terms selection techniques is normalized. At last, some of the high ranked candidate terms chosen by SumScore technique are utilized for extending the user query.

III. CONCLUSION

Search engines use complex algorithms to analyze page relevance factors and determine the page relevance to specific queries. Web crawlers continuously index and re-index pages to ensure that search results remain as relevant as possible to users search intentions. The most valuable and pertinent information in response to uses search queries is provided to users.

REFERENCES

[1]. S. Lawrence and C. L. Giles, .Searching the World Wide Web, “Science, vol. 280, no. 5360, pp. 98-100,1998

[2]. Pal, D.S. Tomar, S.C. Shrivastava, “Effective focused crawling based on content and link structure analysis”, International Journal of Computer Science and Information Security (IJCSIS), Vol. 2, No. 1, pp. 1-5, 2009

[3]. S.M. Pavalam, S.K. Raja, M. Jawahar, F.K. Akorli, “Web crawler in mobile systems”, International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp.531-534, 2012.

[4]. S.N. Jain, P. Rawat, “A study of focused web crawlers for semantic web”, International Journal of Computer Science and Information Technologies, Vol. 4, No. 2, pp. 398-402, 2013.