Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7641924

EFFICIENCY ASSESSMENT OF SEARCH ENGINES WITH IMPROVED VSM & ENTROPY BASED LINK OPTIMIZATION ALGORITHM

Siddharth Ghansela
Assistant Professor, MCA Department, GBPIET Pauri Garhwal Uttarakhand

Abstract: Vector space model is a mathematical model for evaluating the similarities between large data set and a query in increasing order so that a user can find the best matching document among all. It calculates similarity value by using their cosine function. The cosine function evaluates the similarity value by using a weighting scheme. The available factors for weighting schemes are TF(Term-frequency) & IDF(Inverse document frequency). There are various stop words are used when we are writing a query, but only main query terms are important for us for finding best match. It is found that sometimes the results of vector space model are slightly different from other due to the separation of the stop words during similarity analysis. So here we are using some value for stop word so that they can also improve the rank of a document. Also, we are working with entropy-based link optimization algorithm for ranking document, so that we can compare the improved version of vector space model with the entropy-based link optimization algorithm.

Keywords: Optimization, Entropy, Data Mining

Background- In the preferences of search engine scenarios, the increasing and the large availability of datasets create a very huge concern that how we optimize and stored these vast and large amounts of data sets, generally storage and optimization both plays a very bigger role in search engine optimization. But we focused on the optimization techniques of these data sets that at any time 24 * 7 data availability are present here for the user , but we know data is in very vast form such as sky survey cataloguing data, Genomics data, scientific data, we trained data using various fundamental Data mining Strategies but a one major concern the availability of these data for the user at very short instant of time with very fast and enormous speed but how to achieve this, now the realistic answer of this is to a very strong reliable and sophisticated and adequate optimization techniques.

Entropy Based Link Optimization- The term entropy is related with the field of thermodynamics [1]. The entropy was developed by Shannon as part of his theory of communication, in which a data communication system is composed of three elements: a source of data, a communication channel, and a receiver. In Shannon's theory, the "fundamental problem of communication" – as expressed by Shannon – is for the receiver to be able to identify what data was generated by the source, based on the signal it receives through the channel. Shannon considered various ways to encode [2][3], compress, and transmit messages from a data source, and proved in his famous source coding theorem that the entropy represents an absolute mathematical limit on how well data from the source can be lossless compressed onto a perfectly noiseless channel. Shannon strengthened this result considerably for noisy channels in his noisy-channel coding theorem. Entropy in information theory is directly analogous to the entropy in statistical thermodynamics.

Entropy has relevance to other areas of mathematics such as combinatorics. The definition can be derived from a set of axioms establishing that entropy should be a measure of how "surprising" the average outcome of a variable is. For a continuous random variable, differential entropy is analogous to entropy.

Entropy based Link Optimization Algorithm: This algorithm worked with the different links like incoming links also called as inbound links as well as outgoing links also called as outbound links. The following points are related with the working as well as properties of link optimization algorithm.

- Worked with the pattern of links as inbound links or outbound links.
- Very efficient for search engines with respect to processing and computation.
- Using probabilistic techniques and finding the entropy of each page separately.
- At the end all page's entropy summation and counts the whole web application (websites entropy).
- And a website which have a highest entropy rank it first.

The following algorithm is used for ranking the n- number of documents generally available on a website. It works with incoming and outgoing links. Incoming links are those links which are coming to our website from another website whereas outgoing link are those links which are going to another site from our site. Incoming link is known as back links and outgoing links are known as outbound links [4].

when we have millenaries of links indicating to our website but sometimes search engine castigate us for it, if there are low quality and inappropriate backlinks indicating to our website. The reason behind castigate is that we have no authority over the inbound links which others post for us. Search engine only admire the links when appropriate links point to our website. There is no control over the inbound links. If the material presented in the web page is rich with unique information, then probability of the incoming links to our website is high else low. Search engine optimization is a continuous process that cannot stop at any time. The increment in incoming link depends upon the SEO process.

An external link, also called an outbound link, is a link from our website to another website. An external link for our webpage is an inbound link for someone another website. Users have multiple thoughts about outbound links and many users said that it is against the Search engine optimization regulations and search engines can castigate us for this work. Bloggers are of the view that you lose traffic when you put outbound links in your posts but if you share useful resources then chances are that our viewers will stick to your blog and will come back in search of more useful resources. But if we keep outcoming links in our content then make assured that they are indicating to appropriate, authentic and informative material".

ELOA (N, INLINK, TOTLINK, OUTLINK)

1. Start.
2. Take any websites and count its pages, such as N pages contains any websites.
3. Initial Population of Websites n.
4. Take first page of A1 website from n pages
5. 1 page ← A1(N).

6. Counting the total links which contains any websites such as TOTLINK.
7. Counting the separately total links for any webpage such as TOTLINKPAGE1.
8. Counting 1 page inbound or incoming link such as INLINK.
9. Find the probability such as P(INLINK) = INLINK / TOTLINKPAGE1.
10. Now Counting the entropy of 1st page such as, ENTROPY OF INLINK (1st page) = LOG2(P(INLINK)).
11. Similarly, Counting 1 page outbound or outgoing link such as OUTLINK.
12. Find the probability such as P(INLINK) = OUTLINK / TOTLINKPAGE1.
13. Now Counting the entropy of 1st page such as, ENTROPY OF OUTLINK (1st page) = LOG2(P(OUTLINK)).
14. Now combined entropy of 1st page such as: (ENTROPY OF INLINK + ENTROPY OF OUTLINK)/ENTROPY OF TOTLINK.
15. Now similarly count N pages entropy of website 1st.
16. At last summation such as Total Website Combined Entropy (WCE) = Entropy (Page1) + Entropy (Page2) +……+ Entropy (Page N).
17. A Website which has highest WCE selected for ranking and placed at top rank.
18. Exit.

Experimental Results:

Query: Nuclear Power Plants in America

A data set is given as in table no.1.:

Links			
Pages	**Incoming links**	**Outgoing links**	**Total Links**
1st	12	16	28
2nd	13	14	27
3rd	10	17	27
4th	16	10	26
Total	51	57	108

Table no.-1 Data Set of links

Computational Results:

Page 1st = P(INLINK) = 12/28=0.4285.

Page 1st = P(OUTLINK) = 16/28 = 0.5714.

ENTROPY(INLINK) = LOG2(0.4285) = -1.22246.

ENTROPY(OUTLINK) = LOG2(0.5714) = -0.807222.

ENTROPY(Page 1st) = ENTROPY(INLINK)+ENTROPY(OUTLINK)/ENTROPY(TOTLINK).

ENTROPY(TOTLINK) = LOG2(108) = 6.75488.

ENTROPY(Page 1st) = (-1.22246+(-0.807222))/6.75488
ENTROPY(Page 1st) = -0.30047.
Similarly we do for remaining 3 pages,
The entropy of the links is given in table no.2.:

Pages	**P(INLINK)**	**P(OUTLINK)**	**ENTROY (INLINK)**	**ENTROPY (OUTLINK)**	**ENTROPY(PAGES)**
1st	0.4285	0.5714	-1.2224	-0.807222	-0.30047
2nd	2.33314	1.2345	2.345	1.346	1.22223
3rd	1.4567	2.3456	1.2345	1.89655	3.4666
4th	1.2345	2.3456	1.456	1.5678	2.3456

Table no.-2 Entropy of links

Website Combined Entropy (WCE) = -0.30047+1.22223+3.4666+2.3456 = 6.73396.

Enhanced Vector Space Model: The Vector space model is a mathematical similarity analysis model uses very different method where information is fetched as limited matching [5]. The vector space method presents natural language document and queries as vectors in a multidimensional space. Generally, Vector space model is divided into three stages: Data set presentation/indication as terms, weighting the indicated terms to improve retrieval of documents helpful to the end user and grading the entire set of the documents according to the matching measure [6]. In the improved or enhanced version of VSM, there is a slightly change in the inverse document frequency where we have to add a value 1 for optimizing the rank of a document. Having a lot of inspections on the calculations of similarity outcomes using the original vector space model, we further investigate research articles to evaluate some improved methods for calculation of Inverse document frequency (IDF) [7,8,9]. The IDF has a measure aspect in term weight computation in various documents. The term weight has a very important role in similarity value calculations. We used following method for the calculation of IDF [10].

$$idf = log\left(\frac{D+1}{df_j}\right) \qquad (1)$$

The above equation is an improvement in the previous IDF method. With this improved version in IDF, weight of terms present in the document is calculated as given below:

$$w_{i,j} = TF \times log\left(\frac{D+1}{df_j}\right) \qquad (2)$$

Whereas *TF i*s term frequency. In this condition the IDF is calculated through Eq. (1) and the weight of the terms are evaluated through Eq. (2). For making the similarity calculation easy, our proposed new method of a cosine similarity function given by:

$$sim(Q, D_i) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\left(length\ \ of\ \ document_j - \ number\ of\ stop\ words\ \right)}} \quad (3)$$

Where the length of the document is number of unique terms present in document. Since IDF formula of Eq. (1) which is used in our proposed method cannot remove the stop words from the documents, it is removed using our new cosine function as given by Eq. (3). Similarity score is computed for each query. It is calculated, as a median around the number of available links.

Experimental Results of Enhanced VSM-

The TREC query selected is as follows:

QUERY: nuclear power plants in America

and the results of top three search engines are given as follows:

Google

D1- Nuclear power in the United States

D2- Nuclear Power in the USA

D3- Nuclear power plants

And the weights of documents based on Google results are given in table no. 3.

Terms	**Term in Q**	**Count tf_i**			df_i	$\frac{D+1}{df_i}$	$log(\frac{D+1}{df_i})$	**Weights,$W_i = tf_i \times IDF_i$**			
		D1	**D2**	**D3**				**Q**	**D1**	**D2**	**D3**
Nuclear	1	1	1	1	3	1.34	0.30	0.30	0.30	0.30	0.30
Power	1	1	1	1	3	1.34	0.30	0.30	0.30	0.30	0.30
Plants	1	0	0	1	1	4	1.39	1.39	0	0	1.39
America	1	0	0	0	0	0	0	0	0	0	0
United States	0	1	0	0	1	4	1.39	0	1.39	0	0
USA	0	0	1	0	1	4	1.39	0	0	1.39	0

Table no.-3 Weights of documents based on Google results

Yahoo

Query: nuclear power plants in America

and the results Yahoo search engine are given as follows:

D1- Nuclear power in the United States

D2- The biggest nuclear power plants in the USA

D3- Map of USA Nuclear Plants

And the weights of documents based on Yahoo results are given in table no. 4.

Terms	**Term in Q**	**Count tf_i**			df_i	$\frac{D+1}{df_i}$	$log(\frac{D+1}{df_i})$	**Weights,$W_i = tf_i \times IDF_i$**			
		D1	**D2**	**D3**				**Q**	**D1**	**D2**	**D3**
Nuclear	1	1	1	1	3	1.34	0.30	0.30	0.30	0.30	0.30
Power	1	1	1	0	2	2	0.70	0.70	0.70	0.70	0
Plants	1	0	1	1	2	2	0.70	0.70	0	0.70	0.70
America	1	0	0	0	0	0	0	0	0	0	0
United States	0	1	0	0	1	4	1.39	0	1.39	0	0
Biggest	0	0	1	0	1	4	1.39	0	0	1.39	0
USA	0	0	1	1	2	2	0.70	0	0	0.70	0.70
Map	0	0	0	1	1	4	1.39	0	0	0	1.39

Table no.-4 Weights of documents based on Yahoo results

Bing:

Query: nuclear power plants in America

and the results of Bing search engine are given as follows:

D1- Nuclear power in the United States

D2- The biggest nuclear power plants in the US

D3- Map of US Nuclear Plants

And the weights of documents based on Bing results are given in table no. 5.

Terms	**Term in Q**	**Count tf_i**			df_i	$\frac{D+1}{df_i}$	$log(\frac{D+1}{df_i})$	**Weights,$W_i = tf_i \times IDF_i$**			
		D1	**D2**	**D3**				**Q**	**D1**	**D2**	**D3**
Nuclear	1	1	1	1	3	1.34	0.30	0.30	0.30	0.30	0.30
Power	1	1	1	0	2	2	0.70	0.70	0.70	0.70	0
Plants	1	0	1	1	2	2	0.70	0.70	0	0.70	0.70
America	1	0	0	0	0	0	0	0	0	0	0
United States	0	1	1	1	3	1.34	0.30	0	0.30	1.34	0.30
Biggest	0	0	1	0	1	4	1.39	0	0	1.39	0
Map	0	0	0	1	1	4	1.39	0	0	0	1.39

Table no.-5 Weights of documents based on Bing results

1. **Similarity Analysis of search engines:** The similarity analysis is the final measurement of the matching between the documents and the TREC query. It is represented by the similarity angle between the documents and the query [11,12].

So the similarity analysis of all the three documents are given as follows:

Google:

$Cos\,Q\,.D1 = 0.09$
$Cos\,Q\,.D2 = 0.09$
$Cos\,Q\,.D3 = 1.03$

Yahoo:

$Cos\,Q\,.D1 = 0.04$
$Cos\,Q\,.D2 = 0.55$
$Cos\,Q\,.D3 = 0.33$

Bing:

$Cos\,Q\,.D1 = 0.05$
$Cos\,Q\,.D2 = 0.47$
$Cos\,Q\,.D3 = 1.66$

1. **Results Analysis:**
The enhanced VSM is advance version of Vector space model by which we are able to optimize the rank of a web page. After computing the similarity function by using the changed IDF method we have the following analysis of the three search engines in table no.6. By using the enhanced vector space model, we are able to achieve the optimized rank of a web page as shown in Table number 6.

Document	**Rank Based On Enhanced VSM**		
	Google	**Yahoo**	**Yahoo**
D1	2	3	3
D2	2	1	2
D3	1	2	1

Table no.-6 Rank of documents based on three engines

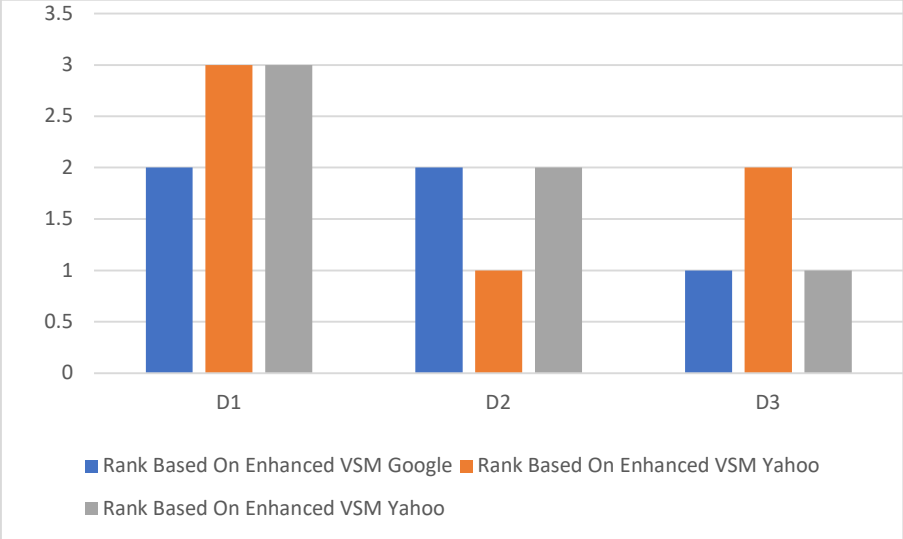

Figure no.-1 Rank Comparison

In the above results there is a good similarity found between the search engines rank when we are using the modified Inverse document frequency. A good match is found between the two search engines for second and third document.

For the same query we have ELOA (Entropy based Link Optimization Algorithm) results along with the actual results came from Google search engine. We found that the entropy-based algorithm provides a good match with the Google results. The below table no. 7 provides the analysis between them.

Query: nuclear power plants in America			
Documents	**ELOA Result**	**ELOA Rank**	**Google Rank**
D1	0.42617	1	1
D2	0.3895	2	2
D3	0.35196	3	3

Table no.-7 Rank Comparison

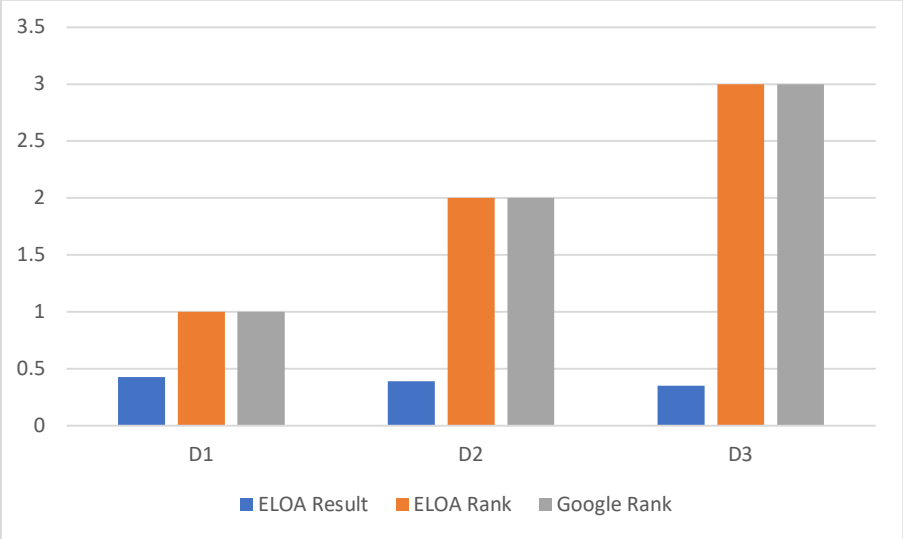

Figure no.-2 Result Analysis

Conclusion:

The both models we have used in our research can help in the improvement in ranking documents. All the experiments are based on TREC query. The ELOA analysis is a good technique for ranking where the incoming and outgoing links are not changing frequently. For some better results we have to monitor them for long time for getting accuracy in ranking. The improvement version of Vector Space Model is very much useful for the enhancement in ranking documents. The future scope of both methods is good for ranking, tracking and other evaluation of webpages. In future we can add some more queries and mathematical model for the improvement in the results.

References:

1. Clausius, R. On the Motive Power of Heat, and on the Laws which Can be Deduced from it for the Theory of Heat. In Annalen der Physik; Dover: Mineola, NY, USA, 1960.
2. Shannon, Claude E. *(July 1948).* "A Mathematical Theory of Communication". Bell System Technical Journal. *27 (3): 379–423.* doi:10.1002/j.1538-7305.1948.tb01338.x
3. Shannon, Claude E. *(October 1948).* "A Mathematical Theory of Communication". Bell System Technical Journal. *27 (4): 623–656.* doi:10.1002/j.1538-7305.1948.tb00917.x.
4. Brin, Sergey; Page, Lawrence (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". Stanford University. Archived from the original on February 11, 2012. Retrieved May 15, 2013.
5. Salton, G., Wong A., and Yang C. S,[1975], "A vector space model for information retrieval", Communications of the ACM,18(11):613–620.
6. G. Salton, E. A. Fox. and H. Wu,[1983]," Extended Boolean Information Retrieval," Communications of the ACM,sVol. 26, No. 11, pp. 1022-1036.
7. Gerald Salton and Chris Buckley, "*Term weighting approaches in automatic text retrieval",* Information Processing and Management, 24(5): issue 5. 1988.

8. J. Ramos, "*Using tf-idf to determine word relevance in document queries*. In First International Conference on Machine Learning", New Brunswick: NJ, USA, 2003.
9. S .Takao, J. Ogata, Y. Ariki, "*Study on New Term Weighting Method and New Vector space model", based on Word Space in Spoken Document Retrieval",* RIAO00, Volume I, pp. 116-131,2000-04.
10. Chris Buckley, "*The importance of proper weighting methods"*, In M. Bates, editor, Human Language Technology. Morgan Kaufman: 1993.
11. Singh, J.N. and S.K. Dwivedi, [2013]," A comparative study on approaches of vector space model in information retrieval", Proceedings on International Conference on Reliability, Infocom Technologies.
12. Singh, J.N. and S.K. Dwivedi, [2015] "Performance Evaluation of Search Engines Using Enhanced Vector Space Model", Journal of Computer Science 2015, DOI: 10.3844/jcssp.2015.692.698.