Journal of Computer Science

Original Research Paper

Performance Evaluation of Search Engines Using Enhanced Vector Space Model

Jitendra Nath Singh and Sanjay K. Dwivedi

Department of Computer Science, Babasaheb Bhimrao Ambedkar University Lucknow, India

Article history
Received: 01-01-2014
Revised: 18-06-2014
Accepted: 15-07-2015

Corresponding Author:
Jitendra Nath Singh
Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India
Email: singhjn2000@gmail.com

Abstract: Vector space model allows computing a continuous degree of similarity between queries and retrieved documents and then ranks the documents in increasing order of cosine (similarity) value. It computes cosine or similarity value using their cosine function. The cosine function computes the similarity value by computing the weight of each term in the documents using a weighting scheme but it is a complex process to compute the weight of each term in the documents. It is also found that sometimes it fails to compute a similarity score, Firstly if there is only one document in the corpus and query terms match with the document and secondly, if the number of documents containing query terms and total number of documents retrieved are equal. To address this problem in order to improve the performance, we proposed an enhanced approach for computation of cosine or similarity value by enhancing the vector space model. Our work intends to analyze and implement our proposed method in performance evaluation of three search engines Google, Yahoo and MSN. To verify our method, we compared our proposed method with a manually computed relevance score and found that our evaluations match with manual method.

Keywords: Information Retrieval, Term Frequency, Cosine Value, IDF, Vector Space Model

Introduction

The search engine is an information retrieval system that helps users to find useful information from the web whereas the web is a system of interlinked documents. The information retrieved is usually key words or phrases that are possible indicators of what is contained on the web page as a whole, the URL of the page, the code that makes up the page and links into and out of the page. It has a user interface where users enter a search term, a word or phrase in an attempt to find specific information using search engines. It is important to us to choose the most appropriate search engine for a query and retrieved best information of interest to the user. Hence, performance evaluation of search engine is a great challenge. Many performance measures can be used to evaluate the performance of search engine. They are precision, recall, coverage, response time and interface etc. In this study, we focus on precision of search engine. Precision is commonly defined as the ratio of retrieved documents that are judged relevant. Performance evaluation of search engine also done manually based on precision (Chu and Rosenthal, 1996; Leighton and Srivastava, 1999). The major benefit of manual precision evaluation used in the existing methods is the high accuracy and drawback is that it is time consuming. Now automatic evaluation of search engine performance is most preferred due to fast changing nature of both the web and search engine. In evaluating the precision performance of search engines, automatic relevance evaluation is critical. So it uses a similarity measure for relevance evaluation of web documents which is generally used in Information Retrieval (IR). The commonly used similarity computation measures are Vector Space Model (VSM) (Hiemstra, 2009; Salton, 1989), Okapi similarity measurement (Okapi) (Robertson and Walker, 1999) and Cover Density Rankin (CDR) (Cormack *et al*., 1999) whereas Vector Space model and Okapi similarity measurement face certain problems in using them on the web because some of the parameters required by these measures, such as total number of documents on the web and the number of documents that contain the query terms are unknown. The VSM, however can be used be analysing only a fixed the number of hits for a query on the web. In a previous study (Singh and Dwivedi, 2012), we analyzed different approaches of Vector Space Model and various derivations of its weighting scheme and observed few problems. To improve this model (Vector Space Model), we present a new method for evaluating the performance of search engines on the web.

Search engines are evaluated in two steps based on sample queries: (a) Computing relevance scores of hits from each search engine and (b) ranking the search engines based on a statistical comparison of relevance score. Statistical metrics, including the Probability of win may be used in the performance comparison of search engines. In our experiment, the proposed new method has been applied to three popular search engines, Google, Yahoo and MSN, based on TREC pattern queries. The accuracy of our method was compared to an existing VSM and a manual method.

Classical Method of VSM

The vector space model (classical method) (Singh and Dwivedi, 2012; 2013), computes similarity score using following formula. We consider this method as base of our research in computation of similarity scores. The similarity is computed using the cosine function (Lee *et al*., 1997) given by:

$$sim(Q,D_i)=\frac{\sum_{j=1}^{V} w_{Q,j}\times w_{i,j}}{\sqrt{\sum_{j=1}^{v} w_{Q,j}{}^2}\times\sqrt{\sum_{j=1}^{v} w_{i,j}{}^2}} \quad (1)$$

where, w_{ij} is the weight of term j in the document *i* and w_{Qj} is the weight of term *j* in the query *Q*. The denominators in this equation, called the normalization factor, discard the effect of document length on document scores.

The weight of a term is computed by TF-IDF method (Buckley, 1993; Takao *et al*., 2000; Stephen, 2004; Jung *et al*., 2000) as given by Equation 2:

$$w_j = TF\times IDF \quad (2)$$

TF is the term frequency (number of occurrences of a query term in a document) and *IDF* is the inverse document frequency (global information). The simple method for computation of IDF (Salton and Buckley, 1988; Polettini, 2004; Papineni, 2001) is given by Equation 3:

$$idf = \log\left(\frac{D}{df_j}\right) \quad (3)$$

D is the number of documents in the document collection and df_j a number of documents containing the query term.

Issues in Similarity Value Computation

After analysis of classical vector space model we derive following observation.

If there is only one document in the corpus and query terms match with the document. Institution shows that cosine similarity would be one, but IDF will be zero by using an existing IDF method. So that similarity value becomes zero in such condition.

If all query terms present in the all documents, the IDF value computed by using the existing IDF method becomes zero, so it fails to compute similarity value of such corpus.

It favours for long documents but it is very difficult to compute a similarity score for long documents, due to high dimensionality.

Computation of weight of each term in the document is very difficult and requires large processing time.

Existence of stop words (a, an, the etc.) in the documents also affects computation similarity score.

Proposed Method

Having certain observations on the computations of similarity values using the classical vector space model, we further explored literatures to analyse some prominent methods for computations of IDF (Salton and Buckley, 1988; Ramos, 2003; Takao *et al*., 2000) as the IDF has a key role in term weight computation. The term weight has an influence in similarity value computation. We used following method for computation of IDF (Buckley, 1993):

$$idf = \log\left(\frac{D+1}{df_j}\right) \quad (4)$$

With this variation in inverse document frequency, weight of terms is computed as given by:

$$w_{i,j} = TF\times\log\left(\frac{D+1}{df_j}\right) \quad (5)$$

where, as *TF i*s term frequency. In this situation IDF is computed using Equation 4 and the weight of the terms using Equation 5. To make the similarity computation easy, our proposed new simpler method of a cosine similarity function given by:

$$sim(Q,D_i) = \frac{\Sigma_{j=1}^{V} w_{Q,j}\times w_{i,j}}{\sqrt{(length\ \ of\ \ document_j - number\ \ of\ \ stop\ words\ )}} \quad (6)$$

where, length of the document is number of unique term in document *j*. Since IDF formula of Equation 4 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 Equation 6. Similarity score is computed for each query. It is computed, as an average across the number links considered.

Similarity Values Computation Using Proposed and Classical Method of VSM

In the process of similarity computation, we have applied our proposed method and classical method of VSM to compute the similarity scores between documents and queries. These experiments were based on an accepted number of TREC pattern short queries.

These queries contain 2, 3 terms. The set of 50 queries are given in Table 1. There are various search modes discussed before, but we have applied only keyword based or defaults search mode and considered only top 10 documents from several documents retrieved. This is a mode that most users use in their searching because the vector space model mostly supports keyword based searching. Only lower case queries were used because different search engine treats capitalized queries differently. We applied these queries on three search engine Google, Yahoo and MSN and computed their similarity score.

We have computed similarity for top 10 retrievals of each query using classical method of VSM using Equation 1 and proposed method using Equation 6 for selected queries. Table 2 shows the average similarity values using three search engines obtained by our proposed method and classical method of VSM on queries listed in Table 1.

Comparison Between Proposed Method and Classical Method

We compared the similarity values computed by our proposed method and similarity values computed by the classical method of the VSM on three search engines Google, Yahoo and MSN. The comparison is shown in figures as given by.

The Figs. 1, 2, 3 show the comparisons of similarity scores between the two methods for three search engines. The observation clear-similarity scores computed by proposed method provided higher values in comparison to the classical method of VSM in all the cases. Since a document with higher similarity score is assumed more relevant to the user and always maintains the high ranking to such documents, we can say that our method has a better chance of evaluation and ranking the documents for the queries.

Based on the similarity values and figures, we have been able to establish that our proposed method provides a strong correlation between document and the query terms for each of the three search engines as compared with the classical method, hence is more effective in the evaluation of performance of search engines.

Manual Scoring Method

To check the accuracy of our proposed method, we have also given these queries to ten students (whom we selected for this task and have been carefully guided how to perform the task) to inspect the similarity score of search engine. The manual scoring method we have used extends the existing methods (Leighton and Srivastava, 1999). Following criteria has been used for manual scoring.

- The documents that are related to the information need of a query which may be useful to the given query are termed relevant. They get a score of 2
- Documents that are slightly related to the query or contain some short description relevant to the query are termed as *slightly* relevant. They get a score of 1
- *Duplicate links* are the pages that appear in the returned links with the same URL more than once. They are given a score of zero
- *Inactive links* are those which give an error message, like file not found (404) or server not responding (603) errors. They are also given a score of zero
- *Ir*relevant *links* are the links that contain irrelevant information. They also get a score of zero
- Based on the above criteria, the score of search engine is computed as the average of score per page per query

Probability of WIN

Performances of search engines have been compared based on similarity value computed in Table 2 and 3. The statistical metric probability of win (P_{win}) (Li and Shang, 2000a; 2000b; Shang and Li, 2002; Ieumwananonthachai and Wah, 1996) measure statistically how much better (or worse) sample mean of one hypothesis, μ_1 is as compared to other, μ_2. In hypothesis testing hypothesis $\{H: \mu_1>\mu_2\}$ is specified without alternative hypothesis and it is evaluated based on sample values. P_{win} is computed based on the mean and the variance of the performance data. First, we compute the difference of performance value (similarity value) between two search engines, that of $P1$ and $P2$ are the performance values of two search engines under considerations respectively, we compute ($P1$-$P2$) for n sample queries. Then compute μ as the sample mean of $P1$-$P2$, followed by sample variance σ^2. Now P_{win} is defined as Equation 7:

$$Pwin = Ft(n-1, \frac{\mu}{\sqrt{\sigma^2/n}}) \quad (7)$$

where, $Ft(v, x)$ is the cumulative distribution function of student's t-distribution with v degree of freedom. To compare a pair of search engines (say $S1$ and $S2$), if the P_{win} value is larger than 0.5, then $S1$ is better than $S2$, else $S2$ is better than $S1$.

Similarity Values Computation Using Manual Method

To check the accuracy of our method of relevance, we provided these queries to ten students as discussed in the manual scoring method. The similarity score is based on the manual method. For each of the three search engines, scores have been assigned by each student manually (as per the criteria of manual method) for each query. The final similarity score for a query has been obtained as the average of scores given by ten students as shown in Table 3.

Table 1. Selected TREC queries

1: Iodine in blood	11: Job safety analysis	21: Food services	31: New orleans	41: Radon inspector
2: Student jobs	12: Adobe Indian houses	22: Wright brothers	32: Optional form 306	42: Local civil rule 83.3
3: Weight of mail	13: Arizona game and fish	23: School bus safety	33: Chester an arthur	43: Storium 90
4: Global warming	14: Feta cheese preservatives	24: Nuclear commission	34: Action plan	44: Symptoms of heart attack
5: Loan proposal	15: Credit report	25: Listeria infection	35: Attorney for senior	45: Weather strip
6: Surface area evaporation	16: Quit smoking	26: Signature of first ladies	36: Eta form 9089 dl	46: Check my status
7: Corn price	17: Black history	27: Online coloring books	37: Family education rights	47: Civil right movement
8: Energy from coal	18: Computer programming	28: Capital hill massacre	38: Unique rare coins	48: Credit report
9: Weather radar	19: Sore throat	29: Earthquake in california	39: Diarrhea pregnancy	49: Internet phone service
10: March health awareness	20: Survey maps	30: Gangster disciples	40: Hand washing gel	50: Brooks brothers clearance

Table 2. Similarity score of three search engines: Google, Yahoo and MSN using proposed and classical methods

	Proposed method			Classical method		
Query ID	Google	Yahoo	MSN	Google	Yahoo	MSN
1	0.04121	0.03532	0.03243	0.00000	0.00000	0.00000
2	0.04330	0.04210	0.03915	0.00000	0.00000	0.03815
3	0.03893	0.03789	0.03153	0.03793	0.03689	0.03453
4	0.04311	0.03409	0.03312	0.00000	0.00000	0.00000
5	0.04150	0.04010	0.03985	0.00000	0.00000	0.00000
6	0.04476	0.03765	0.03676	0.00000	0.00000	0.00000
7	0.04332	0.03967	0.03845	0.00000	0.00000	0.00000
8	0.04127	0.04086	0.03992	0.00000	0.00000	0.00000
9	0.38760	0.03678	0.03552	0.00000	0.00000	0.00000
10	0.03567	0.03442	0.03334	0.00000	0.00000	0.00000
11	0.04132	0.03921	0.03832	0.04032	0.03999	0.03832
12	0.04345	0.03835	0.32030	0.04245	0.03935	0.03830
13	0.03970	0.03856	0.03675	0.00000	0.00000	0.00000
14	0.04376	0.04265	0.04164	0.04398	0.04265	0.04164
15	0.03987	0.03678	0.03402	0.03887	0.03778	0.03502
16	0.04327	0.03286	0.03192	0.04027	0.04186	0.03992
17	0.38760	0.03678	0.03352	0.37760	0.03698	0.03552
18	0.03567	0.03142	0.03034	0.00000	0.00000	0.00000
19	0.04132	0.03921	0.03732	0.00000	0.00000	0.00000
20	0.04245	0.03835	0.36030	0.04145	0.03935	0.38030
21	0.03970	0.03856	0.03675	0.03870	0.03756	0.03675
22	0.04376	0.04265	0.04164	0.04276	0.04165	0.04064
23	0.03987	0.03678	0.03402	0.03887	0.03778	0.03502
24	0.04345	0.03835	0.32030	0.04345	0.03835	0.32030
25	0.03870	0.03851	0.03674	0.00000	0.00000	0.00000
26	0.03887	0.03678	0.03402	0.03787	0.03699	0.03502
27	0.04127	0.03286	0.03192	0.04027	0.03986	0.03892
28	0.37760	0.03678	0.03352	0.00000	0.00000	0.00000
29	0.03667	0.03442	0.03234	0.03567	0.03442	0.03234
30	0.04032	0.03921	0.03732	0.03932	0.03821	0.03732
31	0.04145	0.03835	0.32030	0.04045	0.03935	0.36030
32	0.03870	0.03856	0.03675	0.00000	0.00000	0.00000
33	0.04176	0.04065	0.03964	0.04076	0.04099	0.03986
34	0.03887	0.03678	0.03402	0.03787	0.03778	0.03602
35	0.04145	0.03835	0.32030	0.04045	0.03935	0.36030
36	0.03870	0.03856	0.03675	0.03770	0.03656	0.03575
37	0.04021	0.03532	0.03243	0.00000	0.00000	0.00000
38	0.04230	0.04210	0.02015	0.04130	0.04019	0.04015
39	0.03993	0.03789	0.03153	0.03901	0.03701	0.03653
40	0.04011	0.03409	0.03312	0.04001	0.03909	0.03812
41	0.04150	0.04010	0.03985	0.04050	0.03910	0.03885
42	0.04176	0.03965	0.03876	0.00000	0.00000	0.00000
43	0.03632	0.03367	0.03245	0.03632	0.03367	0.03245
44	0.03627	0.03486	0.03392	0.03527	0.03486	0.03392
45	0.37760	0.03678	0.03452	0.36760	0.03678	0.03452
46	0.03567	0.03442	0.03334	0.00000	0.00000	0.00000
47	0.04032	0.03921	0.03732	0.03932	0.03821	0.03732
48	0.03845	0.03835	0.32030	0.03845	0.03835	0.32030
49	0.03970	0.03856	0.03675	0.00000	0.00000	0.00000
50	0.04176	0.04065	0.03964	0.03976	0.03865	0.03764

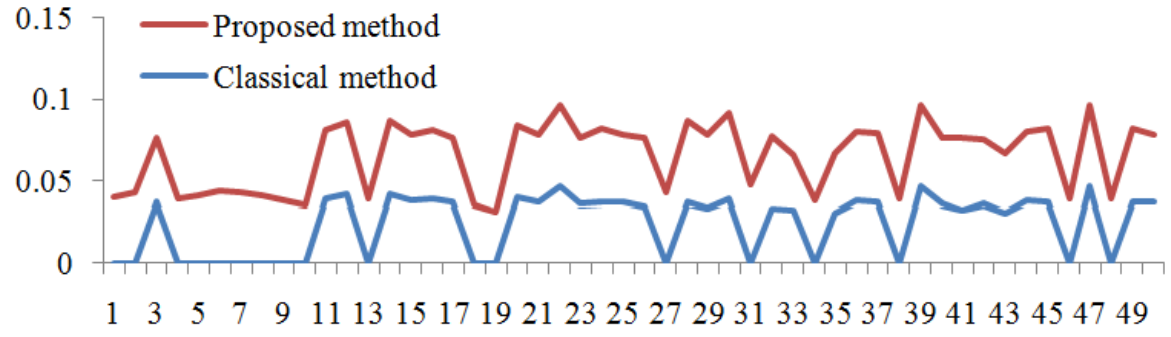

Fig. 1. Comparison of two methods based on average similarity score for Google

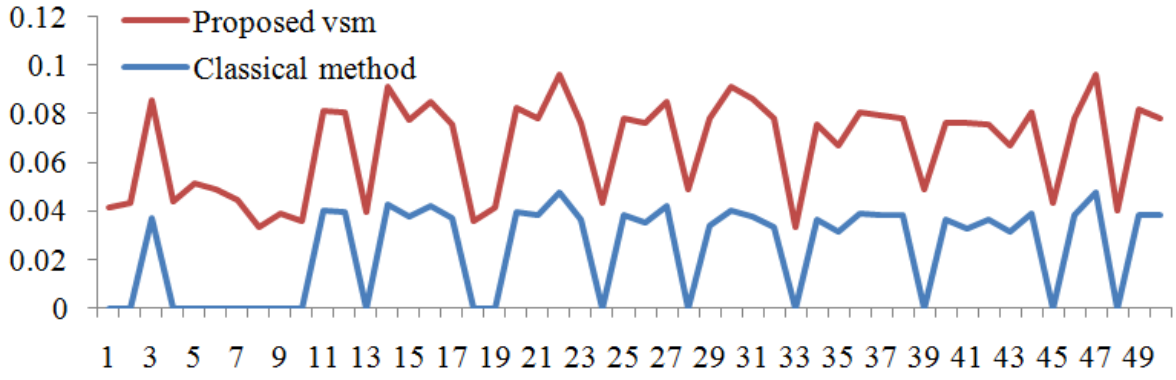

Fig. 2. Comparison of two methods based on average similarity score for Yahoo

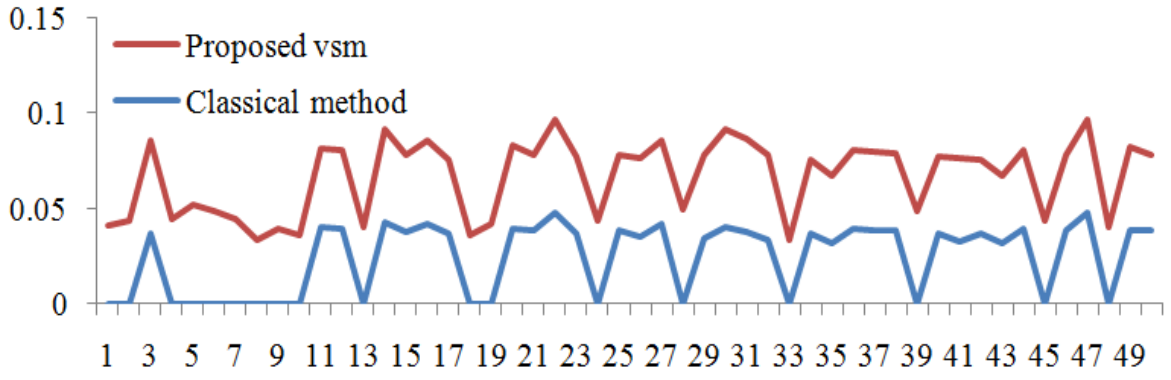

Fig. 3. Comparison of two methods based on average similarity score for MSN

Performance Comparision of Search Engines Using Proposed and Manual Method

Performances of search engines have been compared based on similarity value computed in Table 2 and 3 and using statistical metric probability of win discussed above section. The P_{win} values shown in the Table 4 have been computed for a pair of search engines. The values are similar for both our proposed method and for the manual method. For example, between Google and Yahoo, both the methods have give values greater than 0.5, which means performance of Google is better when compared with Yahoo. Similarity, the performance of Yahoo is found to be better than MSN as the P_{win} value computed greater than 0.5 with the proposed as well as manual method. Both methods arrive at similar comparison results: Google outperformed other two search engines. Yahoo took the second spot while the MSN got the third place. These results show that our method of computation of similarity values is accurate as the same is also justified by manual scoring method.

Table 3. Manual similarity score of three search engines: Google, Yahoo and MSN

Query ID	Google	Yahoo	MSN	Query ID	Google	Yahoo	MSN
1	1.53	1.34	1.250	26	1.53	1.34	1.250
2	1.43	1.25	1.240	27	1.43	1.25	1.240
3	1.62	1.43	1.370	28	1.62	1.43	1.370
4	1.16	1.14	0.950	29	1.16	1.14	0.950
5	1.58	1.55	1.240	30	1.58	1.55	1.240
6	1.42	1.35	1.230	31	1.42	1.35	1.230
7	1.10	0.93	0.920	32	1.10	0.93	0.920
8	1.35	1.32	1.280	33	1.35	1.32	1.280
9	1.42	1.43	1.390	34	1.42	1.43	1.390
10	1.68	1.67	1.610	35	1.68	1.67	1.610
11	1.74	1.73	1.730	36	1.74	1.73	1.730
12	1.54	1.42	1.350	37	1.54	1.42	1.350
13	1.43	1.40	1.390	38	1.43	1.40	1.390
14	1.33	1.32	1.310	39	1.33	1.32	1.310
15	1.52	1.51	1.500	40	1.52	1.51	1.500
16	1.66	1.62	1.610	41	1.66	1.62	1.610
17	1.87	1.84	1.830	42	1.87	1.84	1.830
18	1.56	1.51	1.490	43	1.56	1.51	1.490
19	1.33	1.32	1.310	44	1.33	1.32	1.310
20	1.52	1.51	1.500	45	1.52	1.51	1.500
21	1.66	1.62	1.610	46	1.66	1.62	1.610
22	1.87	1.84	1.830	47	1.87	1.84	1.830
23	1.56	1.51	1.490	48	1.56	1.51	1.490
24	1.67	1.65	1.600	49	1.67	1.65	1.600
25	1.52	1.51	1.500	50	1.52	1.51	1.500

Table 4. Comparison of Google (G), Yahoo (Ya) and MSN (M) based on similarity score computed

	Probability of win		
Relevance method	G > YA	Ya > M	G > MSN
Proposed method	0.876	0.678	0.786
Manual method	0.778	0.701	0.715

Conclusion

In this study, we have proposed an enhancement to the existing vector space model to compare the performance of search engines. Using an acceptable number of TREC pattern queries, we computed similarity values for our proposed method and classical method of VSM. The similarity values computed by our proposed method have been better as compared to the classical method of VSM which is shown in the figures. We also compared our proposed method with the manual method using the same query set for top 10 hits of each query. Both manual and our methods obtained similar results in which Google outperformed others two search engines, whereas Yahoo and MSN obtained the second and third spot respectively.

Funding Information

The authors have no support or funding to report.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Buckley, C., 1993. The importance of proper weighting methods. Proceedings of the workshop on Human Language Technology, (HLT' 93), Association for Computational Linguistics Stroudsburg, PA, pp: 349-352. DOI: 10.3115/1075671.1075753

Chu, H. and M. Rosenthal, 1996. Search engines for the World Wide Web: A comparative study and evaluation methodology. Proceedings of the Annual Conference for the American Society for Information Science, (SIS' 96), pp: 127-135.

Cormack, G.V., C.L.A. Clarke, C.R. Palmer, D.I.E. Kisman and C.R. Palmer, 1999. Fast automatic passage ranking multitext experiments for trec-8. TREC, BibSonomy.

Hiemstra, D., 2009. Information Retrieval Models. In: Information Retrieval: Searching in the 21st Century, Goker, A. and J. Davies (Eds.), John Wiley and Sons, Chichester, ISBN-10: 0470033630, pp: 1-18.

Ieumwananonthachai, A. and B.W. Wah, 1996. Statistical generalization of performance-related heuristics for knowledge-lean applications. Int. J. Artificial Intell. Tools, 5: 61-79. DOI: 10.1142/S0218213096000055

Jung, Y., H. Park and D. Du, 2000. An effective term-weighting scheme for information retrieval. University of Minnesota.

Lee, D.L., H. Chuang and K. Seamons, 1997. Document ranking and the vector space model. IEEE Trans. Software, 14: 67-75.

Leighton, H.V. and J. Srivastava, 1999. First 20 Precision among world wide web search services (search engines). J. Am. Soc. Inform. Sci., 50: 870-881. DOI: 10.1002/(SICI)1097-4571(1999)50:10<870::AID-ASI4>3.0.CO;2-G

Li, L. and Y. Shang, 2000a. A new statistical method for performance evaluation of search engines. Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence, Nov. 13-15, IEEE Xplore Press, Vancouver, BC., pp: 208-215. DOI: 10.1109/TAI.2000.889872

Li, L. and Y. Shang, 2000b. A new method for automatic performance comparison of search engines. World Wide Web, 3: 241-247. DOI: 10.1023/A:1018790907285

Papineni, K., 2001. Why inverse document frequency? Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, (LLT' 01), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 1-8. DOI: 10.3115/1073336.1073340

Polettini, N., 2004. The vector space model in information retrieval-term weighting problem. University of Trento.

Ramos, J., 2003. Using tf-idf to determine word relevance in document queries. Proceedings of the 1st International Conference on Machine Learning, (CML' 03), New Brunswick: NJ, USA.

Robertson, S.E. and S. Walker, 1999. Okapi/keenbow at trec-8. Proceedings of the TREC-8.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manage., 24: 513-523. DOI: 10.1016/0306-4573(88)90021-0

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. 1st Edn., Addison-Wesley Series in Computer Science, Reading, Mass, ISBN-10: 0201122278, pp: 530.

Shang, Y. and L. Li, 2002. Precision evaluation of search engines. World Wide Web, 5: 159-173. DOI: 10.1023/A:1019679624079

Singh, J.N. and S.K. Dwivedi, 2012. Analysis of vector space model in information retrieval. Proceedings of the IJCA National Conference on Communication Technologies and its impact on Next Generation Computing, (NGC' 12), pp: 14-18.

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 and Optimization, (ITO' 13), pp: 14-18.

Stephen, R., 2004. Understanding inverse document frequency: On theoretical arguments of IDF. J. Document., 60: 503-520. DOI: 10.1108/00220410410560582

Takao, S., J. Ogata and Y. Ariki, 2000. Study on new term weighting method and new vector space model based on word space in spoken document retrieval. Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval, Apr. 12-14, College de France, France, pp: 116-131.