RESEARCH ARTICLE OPEN ACCESS

EmpiricalAnalysis of Document Similarity Using Statistical Model

JyotiPhogat*, Atul Kumar**
*(Department of Computer Science, KIIT College of Engineering, Gurugram
**(Department of Computer Science,KIIT College of Engineering, Gurugram

ABTRACT
Information retrieval is great technology behind web search services. This paper presents the statistical method for content based information. Mainly three paradigms of models are used in retrieving information. These are Boolean, probabilistic and vector space model. This paper also presents empirical studies of document similarity and discusses the issue of information retrieval system using statistical model. Vector space model is classical and most used retrieval model. The operation of retrieving information is calculated by using the cosine similarity function of query vector and set of documents vector. Finally, we concludethe results with human score various type documents like sports, politics and short stories.

Keyword: Information Retrieval, Vector Space Model, Tf-idf, Dot Product, Document Similarity.

I. INTRODUCTION

In Information retrieval system information is organized as a collection of documents and documents are not structured, no schema**.** Semantic information retrieval is not applicable to navigational searches. In the document, all the words are not equal to represent semantics of the document. The words with high frequency are generally stopwords which do not provide any meaning to the content of documents. The words with less frequency are rarely meaning bearing words. Normally middle frequency words are meaning bearing words. These words provide meaning to content of the documents.Therefore, to determine the index term, some preprocessing of document is required. The field of information retrieval or document similarity attained peak popularity during last fifty years, number of researchers contributed through their efforts and achieved several remarkable milestones in order to facilitate the internet users with easiest and accurate searching in very small slots of time. In past years' performance of the search engine and their differentiation is the main issue. After lots of propose solutions satisfactory results are not achieved[1].

We have studied mainly three types of information retrieval model. Set theoretic or Boolean model represent documents as set of words or phrase. Boolean model can only give result for the exact match which was the greatest drawback of this model. It means that Boolean model has not given partial. Suppose one document has three term present of given four query terms [11]. This document is also most accurate. Boolean model cannot retrieve this document. Algebraic or vector space model uses vector, matrices or tuples to represent document and queries. The third model probabilistic model use probabilistic theorem like two or three level Baye's theorem in query processing[2]. Google has given result on the basis of Boolean retrieval.

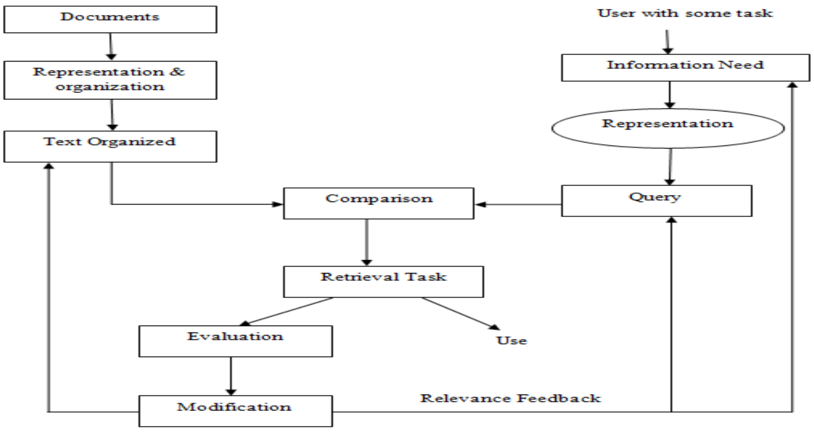

Figure 1: Semantic Information Retrieval Model

1. Vector Space Model

We use vector space model because query language is expressive as well as less complicated and we can get result with partial matching [6]. All inputs like document or queries are in form of vectors. Non-binary weights for index terms in queries and documents are used in the calculation of degree of similarity. This similarity of a document vector to a query vector is the cosine of angle between these vectors [8].

Document vector is defined as d_j= ($W_{1,,j}$,$W_{2,,j}$,…$W_{n,,j}$)
Query vector is defined as q= ($W_{1,q}$,$W_{2,q}$,…$W_{n,q}$)

Vector space model procedure is simply divided into three stages:

1.1 Document Indexing

For removal of non-significant words(non-function words like the, is) we use automatic indexing. This indexing can be based on term frequency, where terms that have both high and low frequency within a document are considered to be function words. In practice, term frequency has been difficult to implement in automatic indexing [3]. Instead the use of a stop list which holds common words to remove high frequency words (stop words), which makes the indexing method language dependent. In general, 40-50% of the total number of words in a document is removed with the help of a stop wordlist [4]. Recently, an automatic indexing method which uses serial clustering of words in text has been introduced. The value of such clustering is an indicator if the word is content bearing.

1.2 Term weighting

The term weighting for the vector space model has entirely been based on single term statistics. There are three main factors term weighting: collection frequency vector,term frequencyfactor and length normalization factor. These three factor are multiplied together to make the resulting term weight. [7, 9]

Term frequency factor (tf) means how well a term describes its document.

$$tf_{i,j} = \frac{f_{i,j}}{max_j f_{i,j}}$$

$$tf_{i,j} = 1 + \log f_{i,j}$$

$$tf_{i,j} = 0.5 + \frac{0.5 \times f_{i,j}}{max_j f_{i,j}} \qquad tf_{i,j} = k + \frac{(1-k) \times f_{i,j}}{max_j f_{i,j}}$$

Inverse document frequency (idf) measures the importance of a term in a document [10].

$$Idf_t = \log\left(1 + \frac{N}{n_t}\right)$$

$$idf_t = \log\left(\frac{N-n_t}{n_t}\right)$$

Where N= documents in coll ,n_t = documents containing term t

Weight of a term can be calculated by $W_{d,t}$. Where $W_{d,t}$= $tf_{d,t}$×$idf_{t.}$[9]

1.3 Similarity Coefficients

The similarity in vector space models is determined by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity. The inner product is usually normalized [5].

Cosine similarity is measured by dot product
Sim(d_i,q)= cos θ

(x.y) =|x||y|cosθ

$$= \frac{d_i.q}{|d_i||q|} = \frac{\sum_j w_{i,j} \times w_{i,q}}{\sqrt{\sum_j w^2{}_{i,j} \sum_j w^2{}_{i,q}}}$$

II. EXPERIMENTAL ANALYSIS

We need corpus to empirical analysis, query and indexing. In our research,we have prepared various Datasets. These datasets have taken from various website and electronic paper etc. These datasetsshow richness of Algorithm. These documents have evaluated many human beings and given score and take average of them. We will compute similarity score from different types of articles for example same news from different newspapers, or similarity scores from some similar documents like sport, science, health, political science etc. Those similarity scores are compared with human similarity scores and calculate the difference. Some of the comparisons are as follows:

2.1Analysis of Dataset 1

When we have collected articles from websites and e-paper which are completely different from each other and calculate similarity scores with respect to bbc.txt then the result we get is shown in table 1.0. As we compared all articles with bbc.txt file so the same file is completely similar to itself so the score is 100% others. The result shows how much similarity between other files and bbc.txt file:

File name	Similarity through VSM	Similarity Score(in%)	Human Similarity Score
bbc.txt	0.9999	100	100
Business.txt	0.683657	68	40
Cnn.txt	0.869827	87	85
Entertainment.txt	0.5805117	58	30
Gaurdiantimes.txt	0.886641	89	75
Health.txt	0.720498	72	65
Nytimes.txt	0.857249	86	88
Science.txt	0.729932	73	38
Sport.txt	0.7117429	71	40
Technology.txt	0.714766	71	50

Table 1(Similarity Scores for different stories with respect to bbc.txt)

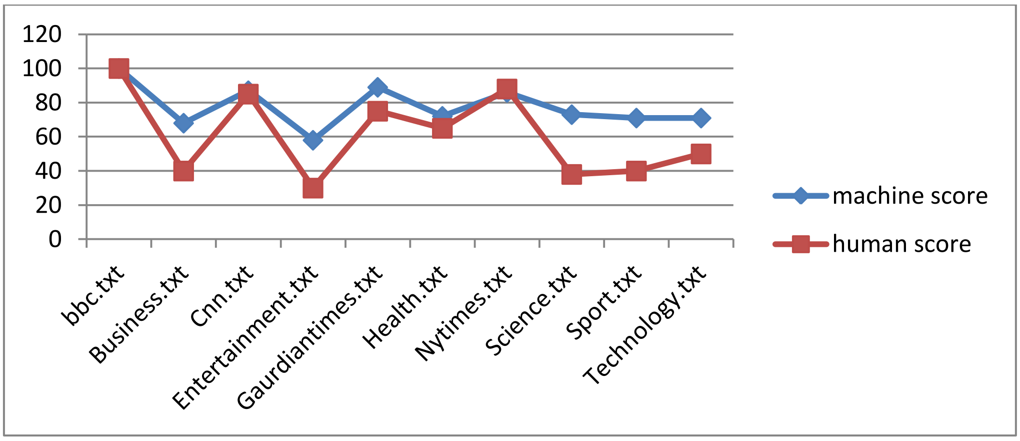

Figure 1: Machine Score Vs Human Score on Data Set1

2.2 Analysis of Dataset2

If we take some interrelated articles and compute similarity scores with respect to 18.txt is stored in table2. All files are compared with 18.txt so score of 18.txt is 100% and others are different. All files are related to same topic so similarity scores are high. In this dataset, we have found result similar to human beings. So, we can say that Vector Space model works close to human score in some cases.

File name	Similarity through VSM	Similarity score (in%)	Human Similarity Score
11.txt	0.801312	80	85
12.txt	0.7441142	74	80
13.txt	0.841190	84	80
14.txt	0.799902	80	82
15.txt	0.802550	80	82
16.txt	0.836700	84	85
17.txt	0.82357	82	80
18.txt	0.99999	100	100
19.txt	0.84036	84	72
sport.txt	0.81237	81	80

Table 2 (Similarity Score for sport Articles w.r.t 18.txt)

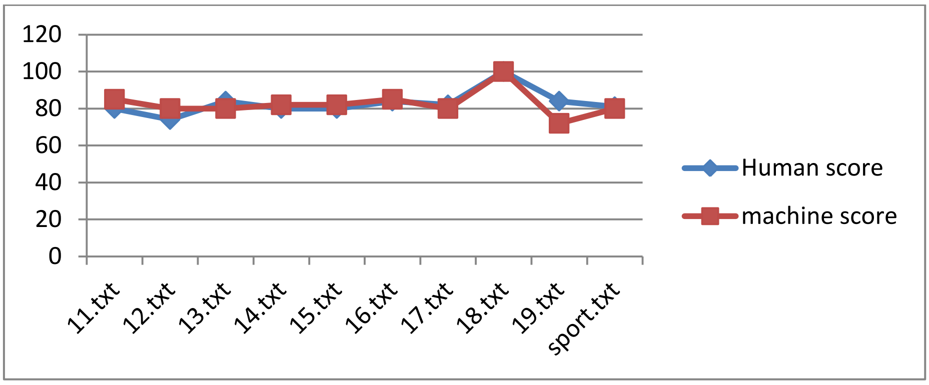

Figure 2: Machine Score Vs Human Score on Data Set2

2.3 Analysis of Dataset3

If we compute similarity scores for dataset3 science related documents with respect to 22.txt.The result is : stored in table3. Vector space model works almostlike human beings with dataset3

File name	Similarity through VSM	Similarity score (in%)	Human Similarity Score
21.txt	0.771003	77	70
22.txt	1.000	100	100
23.txt	0.67075	67	80
24.txt	0.6664725	67	70
25.txt	0.756248	76	75
26.txt	0.698015	70	70
27.txt	0.632688	63	60
28.txt	0.676790	68	75
29.txt	0.61570	62	60
30.txt	0.659377	66	70
Science.txt	0.687299	69	75

Table 3(Similarity score for science stories w.r.t 22.txt)

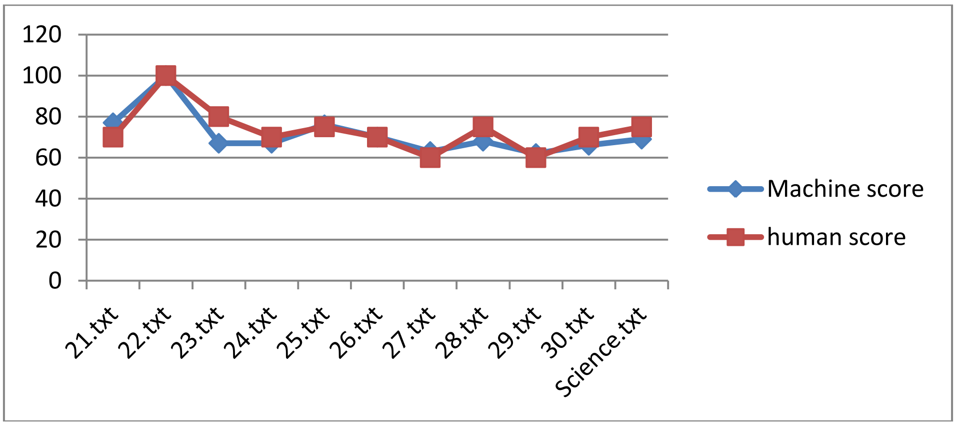

Figure 3: Machine Score Vs Human Score on Data Set3

III. RESULT AND DISCUSSION

On the basis of computation, we calculate that some interrelated documents similarity score of human and machine is not much different as in not related documents, for Table 1 scores of human and machine did not match, but for table 2 and 3 scores was more similar. For table 1 the scores of human and machine are totally different from each other. When we calculated similarity scores for the file with itself it should be 100%. For table 1.0 bbc.txt, for table 2 18.txt and for table 3 22.txt files having scores 100%. For table 1 and 2 score are almost same that means our system works perfectly with these types of documents.Our aimto design a tool that will enable users to retrieve information from the Internet more efficientlyand effectively.Finally, we have seen that

Science stories results have been more accurate result vs. human being. So, we can say further improvement will be increase by Natural Language processing technique.

IV. CONCLUSION AND FUTURE WORK

As we know that retrieving information from internet or from any large unstructured database is quite difficult and very time consuming. A lot of algorithms and techniques are developed in this field yet retrieving' information is problematic. In this research work we used vector space model for retrieving information and compare them with the human similarity scores. It gives partial matching of document. This method has many application like essay checking, theoreticalevaluation of answer sheet.Finally,we conclude that it is easier to retrieve data or information based on their similarity measures but for documents which are completely different it is slightly complicated. Table 1 result not close to human being.To overcome this problem, we will use NLP technique and probabilistic model. Using NLP and latency semantic analysis we can get better result for the same.

REFERENCES

[1] Information Retrieval Experiment, K. Sparck-Jones, ed., Butterworths, London, 1981.

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic retrieval"Journal of Information Processing and Management 24(5):513-523, 1988.

[3] Sanjay K. Dwivedi,Jitendra Nath Singh, Rajesh Gotam "Information Retrieval Evaluative Model" FTICT 2011: Proceedings of the 2011, International conference on "Future Trend in Information & Communication Technology, Ghaziabad, India, Feb -2011.

[4] D.L. Lee, "Document Ranking in BASISplus" Information Dimensions,Dublin, Ohio, 1993.

[5] Yi Shang Longzhuang Li, "Precision Evaluation of Search Engines" World Wide Web (2002).

[6] G. Salton and M. E. Lesk,"Computer evaluation of indexing and text processing"Journal of the ACM, 15(1):8-36, January 1968.

[7] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Journal of InformationProcessing and Management, Vol. 24, No. 5, 1988, pp. 513-523.

[8] Christopher D. Manning, PrabhakarRaghavan, and HinrichSchutze, "Introduction to Information Retrieval"CambridgeUniversity Press, New York, USA, 2008.

[9] Gerald Salton and Chris Buckley,"Term weighting approaches in automatictext retrieval. Journal of Information Processing and Management, 24(5):513-523, Issue5. 1988.

[10] JinbiaoHou: "Research on Design of an Automatic Evaluation System of Search Engine" .In proceeding of ETP International Conference on Future Computer and Communication .FCC/2009.

[11] Singhal, "Modern Information Retrieval: A Brief Overview," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, Issue 1, pp. 35-42, 2001.