INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

COMPARATIVE ANALYSIS OF JACCARD COEFFICIENT AND COSINE SIMILARITY FOR WEB DOCUMENT SIMILARITY MEASURE

Neha Agarwal [1], Mukesh Rawat[2], Vijay Maheshwari [3]

[1] Dept CSE, MIET, Meerut
neha.agarwal2308@gmail.com
[2] Dept CSE, MIET, Meerut
m_rawat1976@yahoo.com
[3]FECI, Shobhit University, Meerut
maheshwarivijay@yahoo.com

Abstract: *Documents are available in form of unstructured, semi structured and structured information. Document clustering will be applicable for World Wide Web, electronic book site, online forums, electronic mails, online blogs, digital libraries and online government repositories. So it is a very tedious job to classify the web documents according to their domain. To solve this problem various algorithms are evolved for proper domain specific clustering of web documents, documents belonging to the same cluster having text similarity between them. Jaccard coefficient and cosine similarity are two techniques to test the similarity between the two documents. In this paper there is a comparative study of these two techniques with respect to the time complexity and relevant result according to the search query.*

Keywords: *clusters, clustering, Jaccard Coefficient , Cosine Similarity.*

1. INTRODUCTION

Clustering [11] means grouping of documents which are similar to each other into one cluster. Document clustering (or Text clustering) is automatic document organization, topic extraction [2] and fast information retrieval or filtering. It is closely related to data clustering.

Components of a Clustering Task

Typical pattern clustering activity involves the following steps [Jain and Dubes 1988]:
(1) Pattern representation [2] (optionally including feature extraction and/or selection),
(2) Definition of a pattern proximity measure appropriate to the data domain,
(3) Clustering or grouping,
(4) Data abstraction (if needed), and
(5) Assessment of output (if needed).

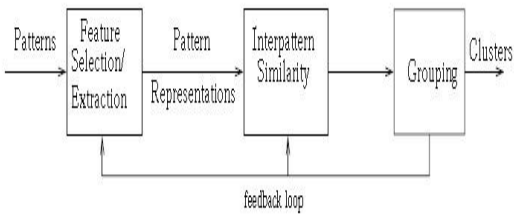

Figure 1-Pattern Clustering

The main uses of clustering of documents are –

- If a collection is well clustered, we can search only the cluster that will contain relevant documents.
- Searching a smaller collection should improve effectiveness and efficiency.

2. LITERATURE REVIEW

2.1 Jaccrad Cofficient

The Jaccard index[11], also known as the Jaccard similarity coefficient (originally coined coefficient de communaute by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:
Let D={D1, D2,……Dn) be the collection of N textual documents being crawled to which consecutive integers document identifiers 1…n are assigned. Each document Di can be represented by a corresponding set Si such that Si is a set of all the terms contained in Di. Let us denote that set by D* such that D*={S1,S2,…….. Sn}. The similarity of any two documents Si and Sj can be computed using the similarity measure
Similarity_measure(S_i,S_j)=$|S_i \cap S_j|$/ $|S_i \cup S_j|$

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

2.2 Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positivee space, where the outcome is neatly bounded in [0,1].The standard way of quantifying the similarity between two documents d_1 and d_2 is to compute the cosine similarity of their vector representations $\vec{V}(d_1)$ and $\vec{V}(d_2)$

$$\mathbf{sim}(d1, d2) = \frac{\vec{V}(\mathbf{d_1}).\vec{V}(\mathbf{d_2})}{\left|\vec{V}(\mathbf{d_1})\right|\left|\vec{V}(\mathbf{d_2})\right|}$$

where the numerator represents the dot product (also known as the inner product) of the vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$, while the denominator is the product of their Euclidean lengths.

3. COMPARATIVE ANALYSIS OF JACCARD COEFFICIENT AND COSINE SIMILARITY FOR CLUSTER GENERATION

The comparative analysis of Jaccard Coefficient and Cosine Similarity for Cluster Generation is done on the basis of three parameters-Time complexity-measure and relevancy of result. These three parameters are described below

3.1 Time Complexity between Jaccard Coefficient and Cosine Similarity for document similarity measure: The table 1.1 shows the time required by Jaccard Coefficient and Cosine Similarity for document similarity measure to retrieve the relevant result against a query term.

Note – These documents are taken for the computer science domain

No. of docs	No. of Cluster	Query term	Jaccard coefficient Time(in ms)	Cosine Similarity Time(in ms)
5	2	Deadlock	9	8
10	5	Process	11	10
15	11	Fuzzy	25	23
20	12	Network	30	28
25	14	Data Mining	41	37
30	16	Kernel	50	46

Table – 1.1

3.2 F-measure between Jaccard Coefficient and Cosine Similarity for document similarity measure

The performance of identifying correct cluster has been measured via three metrics: precision, recall, and F-measure. Precision is the percentage of correctly identified cluster for the document over all the documents in the repository, while Recall is the percentage of correctly identified cluster for the document over all the correctly identified cluster for the document and unidentified cluster for the document. Suppose the number of correctly identified cluster for the document is C, the number of wrongly identified cluster for the document is W and the number of unidentified cluster for the document is M, then the precision of the approach is given by the expression given below

$$P = C/ (C + W) \quad (1)$$

and the recall, R, of the approach is

$$R = C/ (C + M) \quad (2)$$

F-measure incorporates both precision and recall. F-measure is given by

$$F = 2PR/ (P + R) \quad (3)$$

Where: precision P and recall R are equally weighted.

No of documents	Cluster size	Query term	C	W	M	P	R	F
5	2	Deadlock	2	1	2	0.66666667	0.5	0.57142857
10	5	Process	5	2	3	0.71428571	0.625	0.66666667
15	11	Fuzzy	10	2	3	0.83333333	0.76923	0.8
20	12	Networks	12	4	4	0.75	0.75	0.75
25	14	Data Mining	15	5	5	0.75	0.75	0.75
30	16	Kernel	16	7	7	0.69565217	0.69565	0.69565217

Table – 1.2

The above table 1.2 describes the value of F-measure by Jaccard Coefficient for different number of documents.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS.....

No of documents	Cluster size	Query term	C	W	M	P	R	F
5	2	Deadlock	2	1	2	0.66666667	0.5	0.57142857
10	5	Process	6	2	2	0.75	0.75	0.75
15	11	Fuzzy	12	2	1	0.85714286	0.923	0.88888889
20	12	Networks	15	3	2	0.83333333	0.882	0.85714286
25	14	Data Mining	17	4	4	0.80952381	0.81	0.80952381
30	16	Kernel	20	6	4	0.76923077	0.833	0.8

Table 1.3

The above table 1.3 describes the value of F-measure by Cosine Similarity for different number of documents.

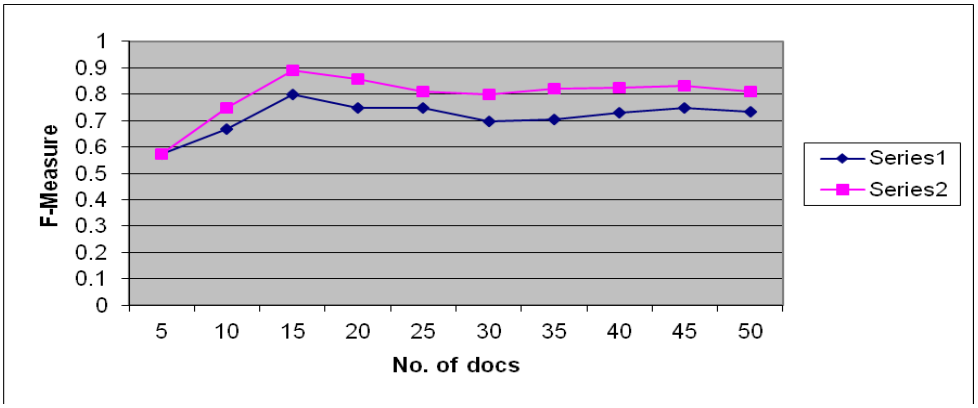

Figure- 2

Above figure 2 represents the comparison of F-measure for document clustering by Jaccard coefficient and cosine similarity in which series 1 represents the F-measure for Jaccard coefficient and series 2 represents the F-measure for cosine similarity.

3.3 Relevancy of results between the Jaccard Coefficient and Cosine Similarity for document similarity measure: The result obtained against a search query entered is more relevant for the clusters created by Cosine Similarity measure as compared to the cluster generated created by Jaccard Coefficient. The number of relevant result specified by the parameter 'C' i.e. correctly identified documents as shown in the above tables.

4. CONCLUSION

Jaccard coefficient and cosine similarity are two of best known techniques for finding the similarity between two documents, time required for cluster generation by using Cosine Similarity measure takes less amount of time as compare to Jaccard Coefficient measure because of using mathematical formula for calculating the similarity measure between the documents. On the other hand Jaccard Coefficient between the two documents by matching all the terms of one document to another which take much more amount of time. By implementing the model for both Jaccard Coefficient and Cosine Similarity cluster generated by Cosine Similarity gives more accurate and relevant result as compare to Jaccard Coefficient.

REFERENCES

[1] C. Aggarwal, S. Gates and P. Yu. On the merits of building categorization systems by Supervised clustering. In Proceedings of (KDD) 99, 5th (ACM) International Conference on Knowledge Discovery and Data Mining, pages 352– 356, San Diego, US, 1999. ACM Press, New York, US.

[2] R. Agrawal, C. Aggarwal, and V. V. V. Prasad. Depth-first generation of large item sets for association rules. Technical Report RC21538, IBM Technical Report, October 1999.

[3] R. Agrawal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. Journal of Parallel and Distributed Computing, 61(3):350–371, 2001.

[4] Deepti Gupta, Komal Kumar Bhatia, A.K.

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY

WINGS TO YOUR THOUGHTS.....

Sharma, A Novel Indexing Technique for Web Documents using Hierarchical Clustering, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.9, September 2009.

[5] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD)'2002, Edmonton, Alberta, Canada, 2002. http://www.cs.sfu.ca/˜ ester/publications.html.

[6] S. Chakrabarti. Data mining for hypertext: A tutorial survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM,1:1–11, 2000.

[7] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In Proceedings of the 29th Symposium on Theory of Computing STOC 1997, pages 626–635, 1997.

[8] P. Domingos and G. Hulten. Mining high-speed data streams. In Knowledge Discovery and Data Mining, pages 71–80, 2000.

[9] R. C. Dubes and A. K. Jain. Algorithms for Clustering Data. Prentice Hall College Div,Englewood Cli®s, NJ, March 1998.

[10] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In IEEE Symposium on Foundations of Computer Science, pages359–366,2000.

[11] Christopher D. Manning, Prabhakar Raghavan," An Introduction to information retrieval", Cambridge University Press Cambridge, England, Online edition (c) 2009.