Documents Clustering Based on Pearson's Correlation Coefficient

Dr. Kamlesh Malpani

Assistant Professor, Department of Computer Science, Shri Vaishnav Institute of Management, Indore, India. malpani_k1@rediffmail.com

ABSTRACT - Similarity should be about the same for mainly everyone. Two objects object A and B are similar if commonality is present. The more in common, are bigger similarities. The similarity of two objectsis related to their differences. The more they have in uncommon, the lesser their similarity is. Two Objects has maximum similarity when both are identical to each other, no matter how much commonality they share. Similarities are separated into different categories. These categories specify what kind of similarities is needed and for purpose like image similarity or sound similarity. Different types of algorithms specifying the object which similarity has to be found. Algorithms get filtered down to the important and related ones. Document similarity is a very basic task and can be used in many applications such as classification of documents, clustering and ranking of documents. Traditional approaches use different measures such as cosine, Jaccard, and dice to document representation and compute the document similarities using word present . Two documents are similar if they contain some of the same terms. Possible measures of similarity might take into consideration: The lengths of the documents, the number of terms in common, whether the terms are common or unusual, How many times each term appears. In this paper we proposed an efficient approach based on Pearson correlation coefficient to find similarity between two text documents.

Keywords - Documents. Similarity Tonimoto, Sorensen, Pearson.

I. INTRODUCTION

Documents processing plays an important role in information retrieval, data mining and web search. As the amount of digital documents has been increasing dramatically over the years as the internet grows, information management, search and retrieval has become an important problem. Similarity measure is a technique that organizes large number of objects into smaller coherent groups. Documents similarities aims at grouping similar documents in one class and separate this group as much as possible from the ones which contains information on entirely different topics. World Wide Web has huge applications of documents grouping such as clustering of results for users on search engines, grouping of comments to suggest products on online stores. Similarity measure plays an important role in deciding how similar or different the two documents are. A lot of similarity measures are in existence for computing the similarity between two documents. Euclidian distance is one of the well-known similarity metric taken from Euclidian geometry field. Cosine similarity is a measure taking the cosine of the angle between two vectors. The Jaccard co-efficient is a statistic used for comparing the similarity of two sample sets and is defined as size of intersection divided by size of union on sample data sets. An information-theoretic measure for document similarity is a phrase-based measure to compute the similarity based on suffix- Tree Document Model.

Similarity dynamically selects a number of features out of documents d1 and d2. Documents grouping require definition of a distance measure which assigns a numeric value to the extent of difference between two documents and which the algorithm uses for making different groups of a given dataset.

II. DOCUMENTS SIMILARITY MODEL

There are two most common models are used for Documents similarity

1. Vector space model
2. Distance based model

Similarity is generally used as a metric for measuring distance when the magnitude of the vectors does not matter. This happens for example when working with text data represented by word counts. We could assume that when a word (e.g. science) occurs more frequent in document 1 than it does in document 2, that document 1 is more related to the topic of science. However, it could also be the case that we are working with documents of uneven lengths. Text data is the most typical example for when to use this metric. However, you might also want to apply cosine similarity for other cases where some properties of the

DOI : 10.35291/2454-9150.2019.0479

instances make so that the weights might be larger without meaning anything different. Sensor values that were captured in various lengths (in time) between instances could be such an example.

The Tonimoto index (sometimes called the Tonimoto *coefficient*) compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations. Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations. Tonimoto measures the similarity between two nominal attributes by taking the intersection of both and divides it by their union. In terms of the above definitions this gives. The index is known by several other names, especially Dice index, index and Dice's coefficient. Other variations include the "similarity coefficient" or "index", such as Dice similarity coefficient (DSC). Sørensen's original formula was intended to be applied to binary data

III. PROBLEM STATEMENT

Many similarity measures specially Cosine, the most used, do not output similarity value of two documents in the percentage of common data to the total data. Many available measures such as Dice coefficient give proportional results if there are no repeating words in two documents. There is a problem how to handle the frequency of words. To weigh a word equal to its frequency is illogical and to ignore the frequency at all is also unrealistic. Algorithm is commonly used to weigh the frequent words. It gives relatively more weights to smaller numbers and vice versa. Measures such as Dice, Jaccard and Information Theoretic all use logarithms to give so called proper term weight. For these measures this weight can be considered proper when frequency difference of a word in two documents is a small number, say less than 10, but if this number is more than 100 then the logarithmic weight difference may not be proper. Cosine similarity produces unrealistic outputs in some situations. For example it gives similarity value "1" between two documents if they have some common words and no non-common words irrespective of their frequencies in both documents. Cosine also shows reverse trend of similarity on changing frequency of common words while remaining the non-common words and their frequencies unchanged. It increases the similarity values when the difference in the frequencies of common words between two documents goes on increasing while remaining the non-common words and their frequencies unchanged. Here we attempt to devise a measure to streamline the documents similarity giving appropriate weights to non-repeating, repeating, common and non-common words

IV. OBJECTIVES

There are several algorithms and methods have been proposed for documents grouping. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are-

1) Apply Tonimoto coefficient for Clustering documents based on similarity measure.

2) Apply Sorensen coefficient for clustering documents based on similarity measure.

3) Apply Pearson's correlation coefficient for Clustering documents based on similarity measure.

4) Compared all these three coefficients based on the similarity values and find out which similarity coefficient give more accurate values

V. LITERATURE SURVEY

In 2012 ChenguangWangy ,Yangqiu Song "KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks" They proposed a method to represent a document as a typed heterogeneous information network (HIN),where the entities and relations are annotated with types. Multiple documents can be linked by the words and entities in the inconsequently, we convert the document similarity problem to a graph distance problem. Intuitively, there could be multiple paths between a pair of documents [1].

In 2013 SapnaChauhan, PridhiArora ,PawanBhadana " Algorithm for Semantic Based Similarity Measure" This model combines phrases analysis as well as words analysis with the use of propbank notation as background knowledge to explore better ways of documents representation for clustering. The SBSM assigns semantic weights to both document words and phrases[2].

In 2014 Ming Che Lee, JiaWei Chang, and Tung Cheng Hsieh "A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences" This paper presents a grammar and semantic corpus based similarity algorithm for natural language sentences. Natural language, in opposition to "artificial language", such as computer programming languages, is the language used by the general public for daily communication [3].

In 2015 S. Mahalakshmi Challenging Issues and Similarity Measures for Web Document Clustering" Web itself contains a large amount of documents available in electronic form. The available documents are in various forms and the information in them is not in organized form. The lack of organization of materials in the WWW motivates people to automatically manage the huge amount of information [4].

DOI : 10.35291/2454-9150.2019.0479

In 2016 Christian Paul, AchimRettinger "Efficient Graph-Based Document Similarity" Assessing the relatedness of documents is at the core of many applications such as document retrieval and recommendation. Most similarity approaches operate on word-distribution-based document representations - fast to compute, but problematic when documents differ in language, vocabulary or type, and neglecting the rich relational knowledge available in Knowledge Graphs[5].

In 2016 Komal Maher, Madhuri S. Joshi "Effectiveness of Different Similarity Measures for Text Classification and Clustering" Clustering algorithms require a metric to quantify how different two given documents are. This difference is often measured by some distance measure such as Euclidean distance, Cosine similarity, Jaccard correlation, Similarity measure for text processing to name a few. In this research work, we experiment with Euclidean distance, Cosine similarity and Similarity measure for text processing distance measures[6].

In 2017 Muhammad Shoaib , Ali Daud and Malik Sikandar Hayat Khiyal An Improved Similarity Measure for Text Documents. In text mining applications such as clustering documents, citation matching and author name disambiguation (AND) similar documents are grouped together by estimating similarity among them in pair wise fashion. Most of similarity functions are relative measures and their output may not be the real picture of the similarity between the documents [7].

In 2018 Priya.V K. Umamaheswai Baskaran3, Paarkavi " An Effective Document Similarity Approach Using Grammatical Linkages In Semantic Graphs" Assessing the similarity of documents is at the core of many applications such as document retrieval and recommendation. Most similarity approaches operate on word-distribution based document representation-fast to compute, but problematic when documents differ in language, vocabulary or type, neglecting rich relational knowledge [8]

VI. PROPOSED APPROAH

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables (X , Y) {\displaystyle (X,Y)}the formula for ρ is:

Rearranging again given formula

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)}\,\sqrt{\sum y_i^2 - n\bar{y}^2}}$$

$$r_{xy} = \frac{\big(4 \times (2 \times 3 + 5 \times 5 + 3 \times 4 + 0 \times 1)\big) - ((10) \times (13))}{\sqrt{4 \times}\,[38] - (10)^2 \;\times \sqrt{4 \times [51] - (13)^2}}$$

VII. RESULT AND ANALYSIS

Documents present n the cluster

Table 1 documents present in the clusters

Documents	Pearson's Correlation coefficient	Tonimoto Coefficient	Sorensen Coefficient
C1	0	2	2
C2	0	2	2
C3	0	0	0
C4	0	0	0
C5	2	0	2
C6	2	2	0
C7	0	2	2
C8	2	4	4
C9	6	14	4
C10	12	4	8

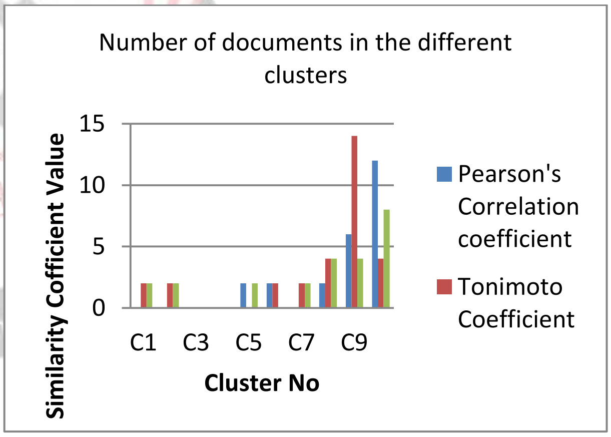

Figure 1 Documents present in the clusters

VIII. CONCLUSION

Documents clustering aims at grouping similar documents in one class and separate this group as much as possible from the ones which contains information on entirely different topics There are several method have been proposed in the last a few year. Several coefficients like are used to find the similarity between documents like Tonimoto, Sorensen, and Cosine. We proposed a new approach based on Pearson's correlation coefficient and compare with two matching coefficient Sorensen coefficient and Tonimoto coefficient. By the experimental analysis we showed that the Pearson's correlation coefficient gives better grouping as compared to extended Sorensen coefficient and extended Tonimoto coefficient.

DOI : 10.35291/2454-9150.2019.0479

REFERENCES

[1] ChenguangWangy , Yangqiu Song "KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks" School of EECS, Peking University Department of Computer Science, University of Illinois at Urbana-Champaign fwangchenguang, lihaoran 2012, mzhang csg@pku.edu.cn, fyqsong, hanjg@illinois.edu

[2] SapnaChauhan, PridhiArora ,PawanBhadana Algorithm for Semantic Based Similarity Measure International Journal of Engineering Science Invention ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 www.ijesi.org Volume 2 Issue 6 ǁ June. 2013 ǁ PP.75-78

[3] Ming Che Lee, JiaWei Chang and Tung Cheng Hsieh3 "A Grammar-Based Semantic Similarity Algorithm for Natural Language Sentences" Hindawi Publishing Corporation The Scientific World Journal Volume 2014, Article ID 437162, 17 pages http://dx.doi.org/10.1155/2014/437162

[4] S. Mahalakshmi "Challenging Issues and Similarity Measures for Web Document Clustering" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. IV (Jan – Feb. 2015), PP 55-59 www.iosrjournals.org

[5] Christian Paul, AchimRettinger "Efficient Graph-Based Document Similarity" Springer International Publishing Switzerland 2016 H. Sack et al. (Eds.):

[6] Komal Maher, Madhuri S. Joshi "Effectiveness of Different Similarity Measures for Text Classification and Clustering" Komal Maher et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (4) , 2016, 1715-1720ESWC 2016, LNCS 9678, pp. 334–349, 2016.DOI: 10.1007/978-3-319-34129-3 21

[7] Muhammad Shoaib , Ali Daud and Malik Sikandar Hayat Khiyal "An Improved Similarity Measure for Text Documents" Basic. Appl. Sci. Res., 4(6)215-223, 2014 ©2014, TextRoad Publication ISSN 2090-4304 Journal of Basic and Applied Scientific.

[8] Priya.V , K. Umamaheswai , Baskaran3, Paarkavi " An Effective Document Similarity Approach Using Grammatical Linkages In Semantic Graphs" International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 02 | Feb-2018 www.irjet.net p-ISSN: 2395-0072.

DOI : 10.35291/2454-9150.2019.0479