Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Usage of Similarity Measures to Cluster Related Web Links

A.Haritha*, P.V.S.Lakshmi
Dept of Information Technology
PVP Siddhartha Institute of Technology, India

Abstract— The explosive amount of information present on the internet attracted many users. Due to the unstructured property of the data present on the internet the users are unable to retrieve the needed information in efficient manner. We concentrated on providing related pages which are of current interest to the user. To make this happen we collected different hyperlinks, transformed them to documents and used the numerical measures like Euclidian distance and Cosine similarity to measure the orientation of the websites to each other. Then applied the clustering algorithm to find out which of them are more associated to each other and are likely to form a cluster.

Keywords— Similarity Measures, Clustering, Euclidian Distance, Cosine Similarity, Inverse Document Frequency.

I. INTRODUCTION

The World Wide Web has lured users to seek and retrieve information from the Internet due to the explosive amount of wealth available on it. However, lacking the integrated structure or schema and this plethora often creates its own set of problems with users being unable to retrieve useful and relevant information in an efficient manner. One of the latent methods to deal with this problem is to consider navigational patterns of users interacting with one or more web sites.
Various search engines are using various algorithms to find similar pages. Similar sites will improves users browsing experience and one can get complete data on the topic which is searched on. Customization involves knowledge acquisition done by analysis of user's navigational behaviour. A user when goes online would like to get the links which suits his requirements or usage in the website he visits. The next business requirement in the online industry will be personalizing/customizing the web page fulfilling for each individuals requirement[1]. The personalization of the web page will involve clustering of different web pages having common usage pattern. In this paper we applied an effective technique by using two numerical measures like Cosine similarity and Euclidian distance on a set of different web links to find out the pages which are more inclined to each other. Then a clustering methodology has been used to cluster them.

II. TRANSFORMING DATA TO DOCUMENTS

A. Document Pre processing

Though the web links are just the hyperlink, they contain lot of information. First the hyperlinks are converted into documents. It comprises of the following tasks.

- Tokenization:
  It is the first step in preparing text. The document which is the outcome from each hyperlink is treated as a string (or bag of words), and then partitioned into a list of tokens.
- Removing the stop words:
  The frequently occurring and insignificant words like prepositions, pronouns are treated as stop words. This step is done to eliminate the stop words.
- Word Stemming:
  The technique of reduction of words into their root words is stemming. This step is the process of conflating tokens to their root form.

B. Document Representation

The index terms are created by generating N-distinct words from the corpora . The document collection is then represented as a N-dimensional vector in term space.

C. Calculation of Term Weights

The goal of term weighting is to assign to each term found a specific score that measures the importance, with respect to a certain goal, of the information represented by the term
Term Frequency. Inverse Document Frequency. Compute the TF-IDF weighting.

D. Analysis of TFIDF

By taking into account the term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically.

TFIDF = TF * IDF where:

TF = C / T

where C = The number of occurrences of a given word in a document
T = total number of words in a document.

IDF = D / DF

where D = total number of documents in a corpus
DF = total number of documents containing a given word [2]

III. DISTANCE/SIMILARITY MEASURE

Similarity between two objects is the numerical measure of the degree to which two objects are alike. Similarity are higher for pair of objects that are more alike. It lies between 0 if it is not similarity and 1 if there is complete similarity. Dissimilarity is the numerical measure of the degree to which the two objects are different. Dissimilarities are lower for more similar objects. Dissimilarity also ranges between 0 and 1.Similarity/Distance measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. [3]

Before clustering, a similarity/distance measure must be determined. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, the closeness is quantified as the distance/similarity value, we can see that large number of distance/similarity computations are required for finding dense areas and assigning new object to that cluster.[4] Here we can take different websites and apply similarity measures to find the closeness of the web links by using cosine similarity and Euclidian distance. On passing the metrics to the clustering algorithm each website is thrown to the respective cluster.

A. Cosine Similarity

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications [5] and clustering too [6]. It is the most popular vector based similarity measure in text mining and information retrieval. The attributes (words, in the case of the documents) is used as a vector to find the normalized dot product of the two documents. By determining the cosine similarity, the user is effectively trying to find cosine of the angle between the two objects. Cosine similarity is 1, the angle between x and y is 0 degree. Document x and document y are the same. If cosine similarity is 0, then the angle between x and y is 90 degree. The documents does not share any terms or words.

Expressed as a mathematical equation

$$similarity(x, y) = cos(\theta) = \frac{x \cdot y}{||x|| * ||y||}$$

This metric is a measurement of orientation and not magnitude, We are not only taking into account the magnitude of each word(TF-IDF) of each document, but the angle between the documents to find how close the documents are to each other. Cosine Similarity will generate a metric that says how related are two documents by looking at the angle instead of magnitude. The Vector Space Model of documents modelled as vectors and also have a formula to calculate the similarity between different documents in this space[7]

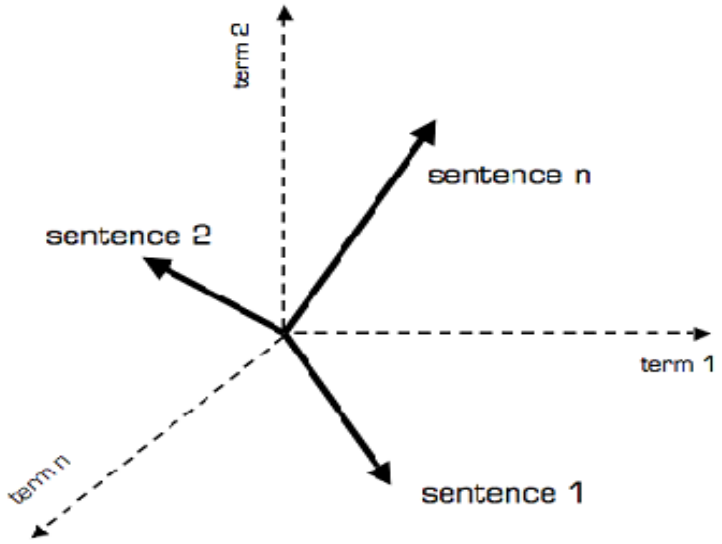

Fig 1.Vector Space Model

B. Euclidian Distance

Euclidean distance is a standard metric for geometrical problems. It is the ordinary distance between two points . Euclidean distance is mostly used for clustering text. It is the common distance between two. It is also the default distance measure used with the K-means algorithm.[8]

Two points, i and j, with coordinates (X1i ; X2i) and (X1j ;X2j), respectively are considered.

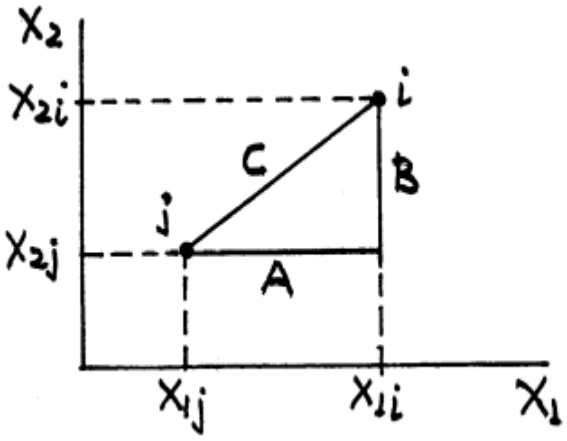

Fig 2. Euclidian Distance

The Euclidian distance between two points is the hypotenuse of triangle ABC.

$$D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$

IV. ALGORITHM

Given a set of data objects D and a pre-specified number of clusters k, k data objects are randomly selected to initialize k clusters, each one being the centroid of a cluster.

The remaining objects are then assigned to the cluster represented by the nearest or most similar centroid. Next, new centroid are re-computed for each cluster and in turn all documents are re-assigned based on the new centroids. This step iterates until a converged and fixed solution is reached, where all data objects remain in the same cluster after an update of centroids.[9]

A. Clustering by Refinement

1. Choose a number of clusters k and determine an initial assignment of terms.

2. Calculate the centroids for the cluster as μ_1 to μ_k .

1.Pick k data points and set cluster centers to these terms.

2. Or assign terms to clusters randomly and take mean of the clusters.

3. For each term, Compute the cluster center it is closest to using(Similarity/Distance measure).

4. Reassign each term to class whose centroid is most similar.

5. Stop when there are no new reassignments.

V. RESULTS AND DISCUSSIONS

A. Step 1

The bunch of websites taken for analysis

1. http://www.jetking.com
2. http://www.makemytrip.com
3. http://www.yatra.com
4. http://www.jetairways.com
5. http://aptechnpower.com
6. http://www.cleartrip.com

B. Step 2

Now we have to convert that web pages crawl the links reads the links and convert it into documents. Using Data to Documents operator[11].

C. Step 3

Once the web links are converted into documents we have to process the text by tokenizing, transform the cases and Filter Stop words. This task results in a word list indicating the total occurrences of the words and document occurrences.

D. Step 4

How similar are the websites to each other? We use Data to Similarity
and we use grammar called as numerical measures. The numerical measures used are Euclidean Distance and Cosine Similarity and the closeness/distance between the web links is computed.

E. Results

Based on the metrics the formation of clusters is shown.

Euclidian Distance and Cosine Similarity

First	Second	Distance
1.0	2.0	1.409
1.0	3.0	1.407
1.0	4.0	1.409
1.0	5.0	1.319
1.0	6.0	1.403
2.0	3.0	1.380
2.0	4.0	1.412
2.0	5.0	1.408
2.0	6.0	1.403
3.0	4.0	1.394
3.0	5.0	1.408
3.0	6.0	1.370
4.0	5.0	1.401
4.0	6.0	1.406
5.0	6.0	1.407

Least

Fig 3. Distance

First	Second	Similarity
1.0	2.0	0.008
1.0	3.0	0.010
1.0	4.0	0.008
1.0	5.0	0.130
1.0	6.0	0.015
2.0	3.0	0.048
2.0	4.0	0.003
2.0	5.0	0.008
2.0	6.0	0.016
3.0	4.0	0.028
3.0	5.0	0.008
3.0	6.0	0.062
4.0	5.0	0.018
4.0	6.0	0.011
5.0	6.0	0.010

Highest

Fig 4.Similarity

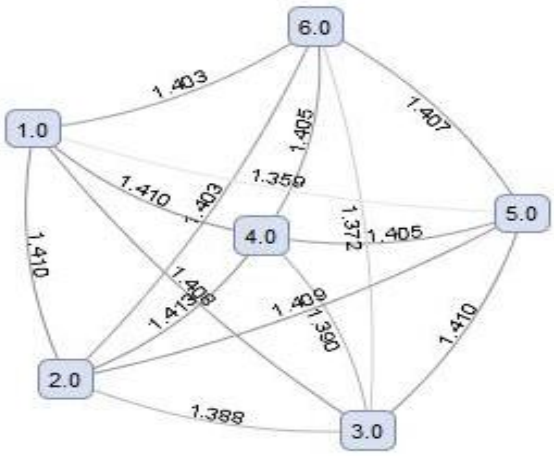

Fig 5.Euclidian Distance-graph

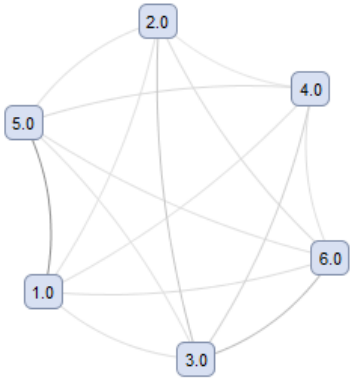

Fig 6.Cosine Similarity -Graph

It shows that Links 1 and Link 5 are less distant from one another using Euclidian
It shows that link 1 and Link 5 are more similar.
On applying clustering the resultant grouping of links into clusters is shown

A. Euclidian Distance *B.Cosine Similarity*

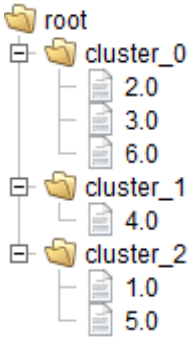

Fig 7.Euclidian Distance-Cluster-Folder View

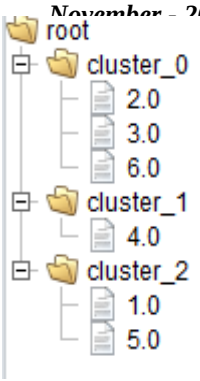

Fig8.Cosine Similarity-Cluster-Folder View

VI. CONCLUSIONS

We presented a comparative study of the numerical measures(Similarity/Distance) and these are considered as the metric for the clustering by refinement algorithm which was discussed earlier. Our experiment conducted on the web links using in cosine similarity measure resulted in 0.130 which is nearing to 1 for sites 1 and 5. By applying Euclidian distance 1(jetking.com) and 5(aptechnpower.com) are far away. So we could say that links 1 and 5 are much oriented towards each other. And we applied clustering by refinement algorithm we found that 1and 5 fall into one cluster. So on comparing two metrics we are able to group the most related links. The effectiveness of the refinement algorithm is shown by the optimized number of clusters.

REFERENCES

[1] Cooley, R., B. Mobasher and J. Srivatsava, 1997. Web mining: Information and pattern discovery on the world wide web. Proceeding of the 9th IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA., pp: 558-567. DOI: 10.1109/TAI.1997.632303.

[2] M.F.M Firdhous,Automating Legal Research through Data Mining, International Journal of Advanced Computer Science and Applications, Vol. 1, No. 6, December 2010.

[3] Venkata Gopala Rao S. Bhanu Prasad A., Space and Cosine Similarity measures for Text Document Clustering, International Journal of Engineering Research & Technology Vol. 2 Issue 2, 2013.

[4] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In AAAI-2000: Workshop on Artificial Intelligence for WebSearch, July 2000.

[5] R. B. Yates and B. R. Neto. Modern Information Retrieval. ADDISON-WESLEY, New York, 1999.

[6] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[7] Machine Learning:Cosine Similarity for Vector Space Models http://pyevolve.sourceforge.net/wordpress/?p=2497

[8] Cluster Analysis- http://www.yorku.ca/ptryfos/f1500.pdf

[9] K.Aruna Prabha, R.Saranya. Refinement of K-means clustering using genetic algorithm. Journal of Computer Applications , 2011.

[10] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[11] RapidMiner- http://www.neuralmarkettrends.com/

