IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-12 E-ISSN: 2347-2693

An Analysis of the Effectiveness of Various Similarity Measures for Web Page Clustering

J.Usharani[1], Dr.K.Iyakutti[2]

[1] Assistant Professor Department of Computer Science Madurai Kamaraj University College Madurai

[2] Professor Department of Physics and Nanotechnology SRM University Chennai India.

Received: Nov /22 /2014 Revised: Nov/30/2014 Accepted: Dec/12/2014 Published: Dec/31/ 2014

Abstract- One of the prominent challenges encountered with regard to web search engines is the large number of documents retrieved by the user in response to their queries. In this regard Various solutions have been proposed in the literature .One approach is to use clustering of web documents. In this paper we propose a genetic algorithm approach for clustering of web documents and study the effectiveness of using various similarity measures in this context. This paper proposes various similarities have been employed and the cosine similarity yields better results when compared to other similarity measures.

Keywords:- Web Page Clustering, vector space model, Genetic Algorithm

1. INTRODUCTION

With booming of the Internet, the number of the web pages is incremented in an explosive rate. Efficiently retrieval of information has become a challenge. Clustering is a process of grouping objects with similar properties. Web page clustering is one of the major and most important preprocessing steps in web mining analysis. Web page clustering puts to gather web pages in groups, based on similarity, or other relationship measures.

1.1Clustering

Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. A good clustering method will produce high quality clusters with high inter-cluster similarity and low intra-cluster similarity. Clustering is an unsupervised classification technique. Clustering is a data analysis technique that, when applied to a set of heterogeneous items, identifies homogeneous subgroups as defined by a given model or measure of similarity.

2. WEB PAGE RETRIEVAL FOR THE QUERY

The user gives query as input, based on that query the documents are retrieved from the search engine. Most of the documents are retrieved from the search engine may or may not be relevant to the user query.

2.1Preprocessing of Web pages

Tokenization

A document is treated as string, and then partitioned into a list of tokens. The process of breaking stream of text into words, pharses, symbols, or other meaningful elements called tokens.

Removing stop words

Stop words are frequently occurring, insignificant words. There are words that are non descriptive for the topic of a document such as a,and.etc.

Stemming

The process of conflating tokens to their root form(connection->connect).

2.2 Document Representation

The VSM (vector space model) is one of the most widespread models for representing documents to be clustered. The web document objects can be represented using the vector space model (VSM).The three characteristics of VSM is (Frey, 2012)

a. Each document is characterized by a vector of word frequencies.b.A distance measure is defined as a function of those document vector, in order to measure the similarity between or distance between any pair of documents.c.A clustering algorithm utilizes this distance measure to set related documents into clusters.

3. SIMILARITY MEASURES

A similarity measure is a function used to measure the degree of similarity between query and documents. It measures how much the query and document is similar with each other. It gives a value which decides the degree of similarity. In order to find similarity first query and document are converted in to vector form.

3.1Types of Similarity Measures

Euclidean Similarity

Euclidean distance determine the root of square difference between the coordinates of a pair of object.For vector x and y distance d(x,y) is given by

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Pearson similarity

Pearson Similarity is correlation Coefficient is standardized angular separation by centering the coordinates to its mean value.

$$\frac{\sum_{i=1}^{n}(A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\left(\sum_{i=1}^{n}(A_i - \overline{A})^2\right)\left(\sum_{i=1}^{n}(B_i - \overline{B})^2\right)}}$$

Jaccard similarity

Measuring the jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all divided by number of properties.

$$\Gamma(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Cosine Similarity

Cosine similarity is a measure of similarity between two high dimensional document vectors

$$\cos\theta = \frac{\sum_{i=1}^{t} x_i * y_i}{\sqrt{\sum_{i=1}^{t} x_i^2}\sqrt{\sum_{i=1}^{t} y_i^2}}$$

4. GENETIC ALGORITHM

The genetic algorithm is a heuristic search algorithm which is used for optimization of difficult problem. It is based on Darwin principle of natural selection.

4.1Components of GA

Chromosome Representation

All the document s and query are first converted in to chromosome. This is given an input to the genetic algorithm.

Fitness Function

It gives a value which is used to calculate the similarity between query and document based on this value chromosome is selected for selection mechanism.

Selection

Is the process in which chromosome is selected for next step in generation based on fitness value of chromosomes.

Crossover

In crossover two or more parent chromosome is selected and pair of genes are interchanging with each other.

Mutation

Is a process in which gene of the chromosome is changed. If one point mutation if gene is 0 then change it into 1 and if gene is 1 then change it into 0

5. PROPOSED METHODOLOGY

The main goal of the proposed system is apply genetic algorithm based clustering with different similarity measures.

Process

Extract all the words from documents

Remove stop words

Both query and documents are encoded into chromosome

Encoded chromosome are given as input to genetic algorithm

Run genetic algorithm until stopping criteria

5.1Evaluation of Clustering Quality

The following two measures are commonly used to judge cluster quality

Purity

The purity measure evaluates the coherence of a cluster. Given particular cluster C_i of size n_i ,the purity of C_i is formally defined as

$$P(C_i) = \frac{1}{n_i}\max_h(n_i^h)$$

Where $\max_h(n_i^h)$ is the number of documents that are from the dominant category in cluster C_i and n_i^h represents the number of documents from cluster C_i assigned to category h

Entropy

The entropy measure evaluates the distribution of categories in a given cluster. The entropy of a cluster C_i with size n_i is defined to be

$$E(C_{II}) = -\frac{1}{\log c}\sum_{h=1}^{k}\frac{n_i^h}{n_i}\log(\frac{n_i^h}{n_i})$$

. Where c is the total number of categories in the data set and n_i^h is the number of documents from the hth class that were assigned to cluster C_i

We need maximize the purity measure and minimized the entropy of cluster in order to accomplish high quality clustering results.

Clustering genetic algorithm

1) Randomly generate initial population
2) Evaluate the fitness of all individuals in the population.
3) Cluster the population according to fitness value.The fitness value is calculated based on the above mentioned similarity measures.
   i) Randomly generate cluster
   ii) Calculate distance between the centers and other chromosome in the population.
4) Select the parents from each cluster
5) Calculate fitness of individuals in the population
6) Apply genetic operator like crossover and mutation.
7) Repeat step 3 to 6 until the terminated condition

6. RESULT

The experimental results on different similarity measures on the five selected datasets are presented in the Table 1 and Table 2. The Table 1 shows the purity of clusters produced by using different similarity measures.

The Table 2 shows the entropy of clusters produced by using different similarity measures.

Table 1 Purity with different Similarity Measures

DATA	Euclidean	Cosine	Jaccard	Pearson
Java	0.51	0.73	0.52	0.59
Computer	0.62	0.82	0.59	0.62
Mouse	0.64	0.87	0.73	0.69
Mining	0.58	0.64	0.54	0.52
Cookies	0.69	0.89	0.74	0.71

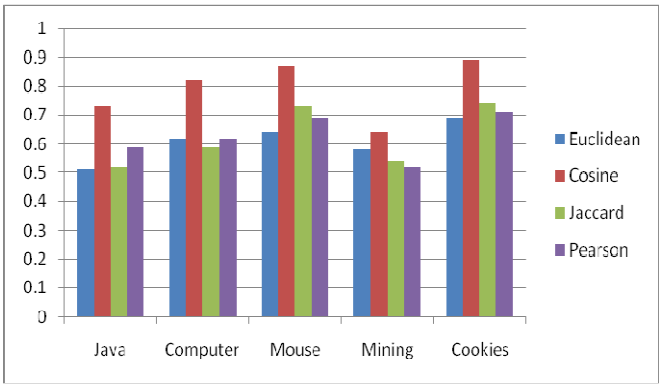

Fig 1 Purity with different Similarity Measures

Table 2 Entropy with different Similarity Measures

DATA	Euclidean	Cosine	Jaccard	Pearson
Java	0.61	0.45	0.54	0.62
Computer	0.57	0.38	0.59	0.54
Mouse	0.49	0.33	0.47	0.51
Mining	0.63	0.43	0.65	0.67
Cookies	0.65	0.47	0.66	0.69

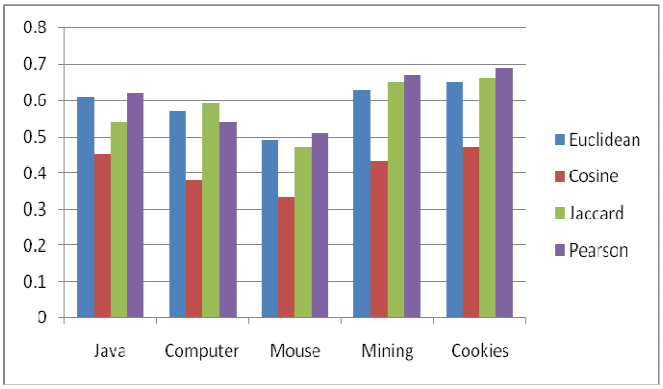

Fig 2 Entropy with different Similarity Measures

In this experiment we found that cosine similarity gives the best result in terms of purity and entropy.

7. CONCLUSION

In this paper we propose a genetic algorithm based clustering of web documents. We study the effectiveness of various similarity measures and found that cosine similarity gives better result.

References

[1]A.Huang "Similarity measures for text document clustering" NZCSRS(2008)

[2]A.Strehl,J.Ghoesh "Impact of similarity measures"

[3]N.Oikonomakon,M.vazirginnn "A review of web document Approaches"

[4]R.kala,A.Shukla and R.Tiwang " A novel Approach to clustering using genetic algorithm"International journal of engineering research 2010.

[5] U.Maulik,S.Bandyopadhyay "Genetic algorithm based clustering"