Evaluation of Clustering Results Web Pages Using DBSCAN

Waiting Guo
B9415023
CSIE, NTUST

June 1, 2009

Abstract

Nowadays, almost all people use Internet everyday. We must to use the search engine to find the answer we want while using Internet. While we key the search word into search engine, it will return a lot of web pages that are mixed and confused. It is hard for users to select one web page from these mixed and confused web pages.

To solve this problem, we need a clear way to present these searched web pages. We can present these searched pages in category. For this way, we need the clustering technique to cluster these pages. Therefore, I simulate an environment of clustering results using DBSCAN[1] with Euclidean distance to evaluate similarity of each clustering documents. I try to cluster these web pages into some clusters.

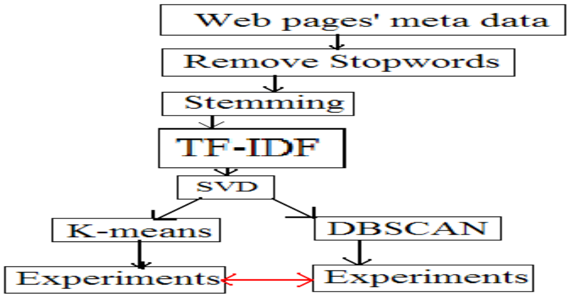
1 Introduction

In this project, I cluster some web pages using DBSCAN[1]. Moreover, I present the experiment result between DBSCAN[1] and K-Means[2]. There are three parts, dataset construction, clustering algorithm programming and result evaluation.

This paper is organized as follows. In section 2, I explain the whole system Finally, in section 3, I propose my conclusion of this project.

2 System Architecture

2.1 Whole Architecture



Above fiqure is the whole architecture of this project. First step, I need to get the "metadata" from web pages. Second, I need to remove the "stopwords", which are not important words. Third, I need to "stem" those non-stopwords. Fourth, I use "TF-IDF" method to decide which terms are the most important terms. Fifth, I use "SVD" to reduce the dimension of data. Sixth, I use "DBSCAN" and "K-Means" to cluster the experiment data, and I compare the clustering result.

The following subsections are the brief explanion of these steps.

2.2 Remove Stopwords

Stopwords are useless for clustering process, such as this, it, is, are, ain't, or was, etc. For example, "This is a Computer Structure book.", some words in this sentence are stopwords, "This", "is", "a". Therefore, I need to remove these stopwords before clustering.

2.3 Porter Stemming Algorithm

The Porter stemming algorithm[3] is a for removing the commoner morphological and inflectional endings from words in English.

2.4 TF-IDF

The "TF-IDF" weight (tem frequency-inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a documents relevance given a user query.

The term count in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents to give a measure of the importance of the term ti within the particular document dj.

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where ni,j is the number of occurrences of the considered term (ti) in document dj, and the denominator is the sum of number of occurrences of all terms in document dj.

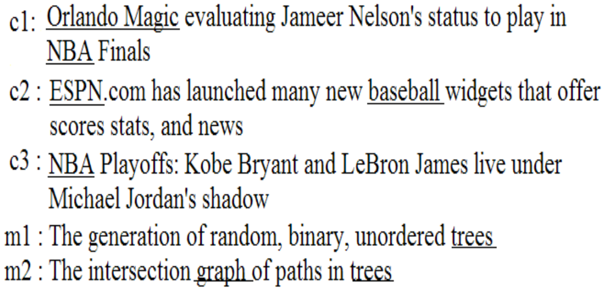
The inverse document frequency is a measure of the general importance of the term.

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

where $|D|$ is total number of documents in the corpus, $|d : ti \epsilon d|$ is number of documents where the term ti appears.

2.5 Term by Document Matrix

I will use an simple example to explain this process. If the document collection has following five documents, the first three documents are related to sport and the last two documents are related to graphs.



c1: Orlando Magic evaluating Jameer Nelson's status to play in NBA Finals

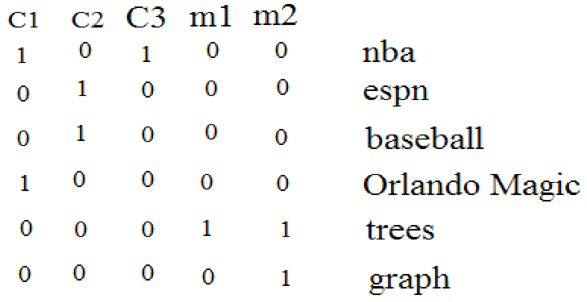
c2 : ESPN.com has launched many new baseball widgets that offer scores stats, and news

c3 : NBA Playoffs: Kobe Bryant and LeBron James live under Michael Jordan's shadow

m1 : The generation of random, binary, unordered trees

m2 : The intersection graph of paths in trees

After I consider which term is the important term, I can form the "Term by Document matrix". Here are the 6X5 matrix for above five sentences.



$$\begin{array}{ccccc|l}
\mathrm{C1} & \mathrm{C2} & \mathrm{C3} & \mathrm{m1} & \mathrm{m2} & \\
1 & 0 & 1 & 0 & 0 & \text{nba} \\
0 & 1 & 0 & 0 & 0 & \text{espn} \\
0 & 1 & 0 & 0 & 0 & \text{baseball} \\
1 & 0 & 0 & 0 & 0 & \text{Orlando Magic} \\
0 & 0 & 0 & 1 & 1 & \text{trees} \\
0 & 0 & 0 & 0 & 1 & \text{graph}
\end{array}$$

Furthermore, this term by document matrix stand for the above five sentences, so I cluster this matrix to get the clusters.

2.6 Singular Value Decomposition

I use SVD mechanics in analogy to eigenvalue/eigenvector mechanics. In linear algebra, the singular value decomposition(SVD) is an important factorization of a rectangular real or complex matrix. Applications which employ the SVD include computing the pseudoinverse, least squares fitting of data , matrix approximation.

Suppose M is an m-by-n matrix whose entries come from the field K, which is either the field of real numbers or the field of complex numbers. Then there exists a factorization of the form

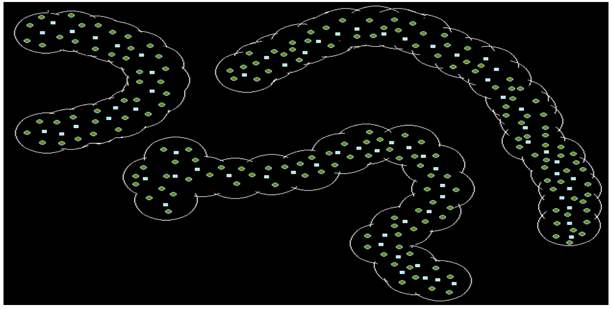
$$M = U\Sigma V^*,$$

where U is an m-by-m unitary matrix over K, the Σ is an m-by-n diagonal matrix whith nonnegative real numbers on the diagonal, and V* denotes the conjugate transpose of V, an n-by-n unitary matrix over K. Such a factorization is called a singular-value decomposition of M.

2.7 DBSCAN

DBSCAN[1](Density Based Spatial Clustering of Applications with Noise) is designed to discover the clusters and the noise in a spatial database.

The key idea in DBSCAN[1] is that for each point of a cluster the neighborhood of a given radius has to contain at least a minimum number of points. A naive approach could require for each point in a cluster that there are at least a minimum number(MinPts) of points in an Eps-neighborhood of that point. However, this approach fails because there are two kinds of points in a cluster, points inside of cluster(core points) and points on the border of the cluster (border points). In general, an Eps-neighborhood of a border point contains significantly less points than an Eps-neighborhood of a core point. Therefore, DBSCAN set the minimum number of points to a relatively low value in order to include all points belonging to the same cluster. The following fiqure is a simple sample of DBSCAN[1] .



2.8 K-Means

In statistics and machine learning, K-Means[2] clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with nearest mean. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data.

Given a set of observations (x1,x2,...,xn), where each observation is a d-dimensional real vector, then K-Means[2] clustering aims to partition this set into k partitions(k¡n) S=(S1,S2,...,Sk) so as to minimize the within-cluster sum of squares.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2$$

2.9 Euclidean Distance

I compute the distance between centers and data instances, so I must has a distance method. In this project, I use Euclidean Distance method to compute the distance.

In the plane, the distance between points (x1,y1) and (x2,y2) is given by

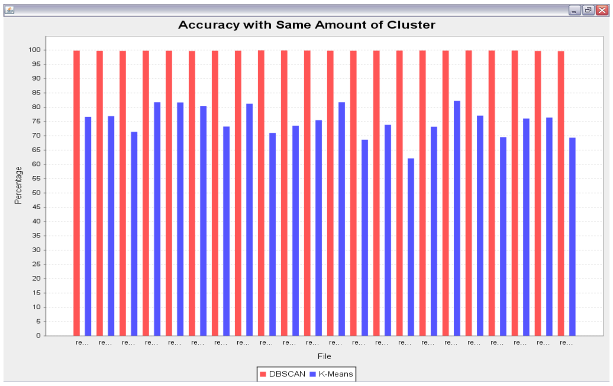
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

In general, the distance between two points, x and y, in a Euclidean space is defined as

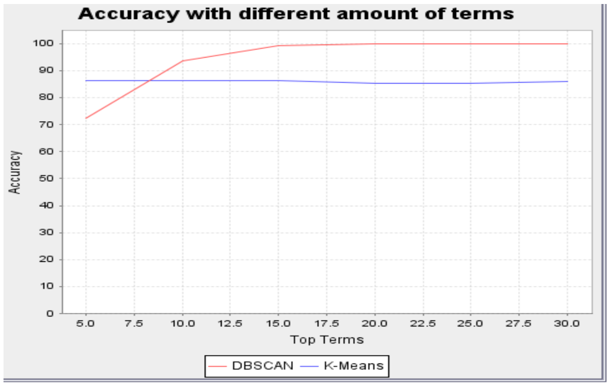
$$d = \sqrt{\sum_{i=1}^{n} |xi - y_i|^2}$$

3 Experiment

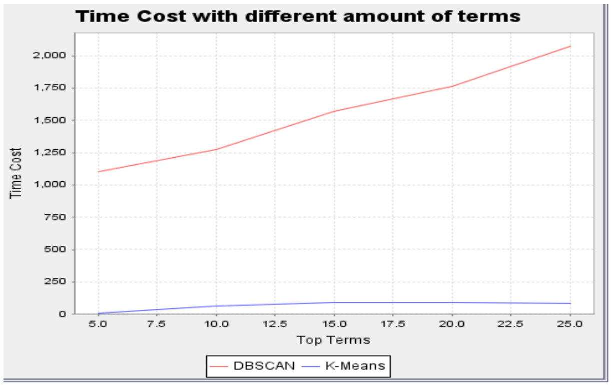
I use "Reuters-21578" data set to complete this experiment. There are 21578 documents in this data set, and these documents are separated into 22 files.



In this experiment, I try to compare the accuracy with the same amount of clusters. Therefore, this experiment shows that DBSCAN has better accuracy than K-Means.



We all know the same document show different face in term-by-document matrix with different top terms. Therefore, in this experiment, I try to compare the accuracy with different amounts of terms. This experiment shows that DBSCAN has better accuracy while the amount of term is bigger.



The time cost experiment is important experiment in data mining. Therefore, in this experiment, I try to compare the time cost with different amounts of terms. This experiment shows that DBSCAN spends more time than K-Means in the same amounts of terms. This is the disadvantage of DBSCAN.

4 Conclusion

- The top terms will affect the clustering result.
- DBSCAN spends more time than K-Means.
- However, DBSCAN give you more accurate clustering result.

According to above three personal points, we can refine the DBSCAN algorithm. Therefore, we can use efficient DBSCAN to provide more accurate clustering result don't cost much time.

References

[1] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A Desity-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In Proc. of KDD, 1996.

[2] MacQueen, J. B.(1967). "Some Methods for classification and Analysis of Multivariate Observations" in Proceedings of 5th Berkeley Symposium on Mathernatical Satistics and Probability.

[3] Porter Algorithm, http://tartarus.org/martin/PorterStemmer/

[4] Data Set, http://www.research.att.com/ lewis