Document Clustering Using Sample Weighting

Chengzhi Zhang
[1] Dept. of Information Management,
Nanjing University of Sci. & Tech,
Nanjing 210093, China
[2] Institute of Sci. & Tech. Information of
China, Beijing 100038, China;
zhangchz@istic.ac.cn

Xinning Su, Dongmin Zhou
Dept. of Information Management,
Nanjing University,
Nanjing 210094, China
xnsu@nju.edu.cn
dylan_184@163.com

Abstract

Clustering algorithm based on Sample weighting has been noticed recently. In this paper, a novel sample weighting clustering algorithm is presented based on K-Means and fuzzy C-Means algorithm. The algorithm uses academic documents as the clustering objects. The PageRank value of each document is calculated according to the cited relationship among them, and it is used as the weight in the algorithm. Experiments show that the proposed algorithm is effective to improve performance of document clustering.

Keywords: document clustering, sample weighting clustering, text mining, Page-Rank

1 Introduction

Cluster analysis is a powerful technique in the field of data analysis. As a typical unsupervised learning technique, clustering method can be divided into several kinds, such as partitional clustering, hierarchical clustering, density-based clustering, grid-based clustering and model-based clustering (Han and Kamber, 2000).

Most applied clustering algorithms treat all samples or objects equally in the clustering process, such as the K-Means algorithm (MacQueen, 1967), fuzzy C-Means algorithm (Bezdek, 1981) and EM type clustering algorithm (Dempster et al, 1977). However, different samples or objects should play different roles in clustering process. Hence, it is very useful to give the appropriate sample weighting in cluster analysis. For this purpose, sample or object weighting clustering algorithm is proposed (Pedrycz, 1996; Rose, 1998; Nock and Nielsen, 2004; Nock and Nielsen, 2006; Li et al, 2005). As a noticed algorithm recently, there are still some unsolved problems of the sample weighting clustering algorithm. Whether the structure information among the clustering objects is helpful to sample weighting clustering or not? How to transform structure information into the weight of samples?

In this work, the authors use academic documents as the clustering objects and K-Means algorithm and fuzzy C-Means algorithm as the baseline. The PageRank value of each document is computed automatically according to the cited relationship among them, and it is used as the weight in the sample weighting clustering algorithm.

The rest of the paper is organized as follows: the next section reviews some related work on the sample weighting clustering. In section 3, a detailed description of the proposed approach is given. Section 4 details the methods of Sample weighting. Subsequently in section 5, the authors report experiments results that evaluate the proposed approach. We make concluding remarks in section 6.

2 Related Work

In the previous papers, there are only several algorithms considering sample weighting, such as conditional fuzzy c-means (Pedrycz, 1996), deterministic annealing for clustering (Rose, 1998), etc. But unfortunately, the application of the algorithms above is limited for that they need users or heuristic principle to weight samples. So, it is interesting to find an approach to automatically compute weighting of each sample.

Recently, Nock and Nisseslen proposed a formalized clustering framework motivated by boosting algorithm, which offers penalizing solutions via weights on the samples (Nock and Nielsen, 2004). They pointed out the significance of calculating the sample or object weight automatically during the process of clustering (Nock and Nielsen, 2006). Li and Gao, et al. have proposed a typical-sample-weighting clustering algorithm for large data sets. It can obtain original clustering samples using the atom-clustering algorithm. Then weight them according to the atom number of samples (Li et al, 2005).

3 Document Clustering Algorithm Using Sample Weighing

Just as indicated above, the underlying idea of our approach is as follows: different samples or objects should play different roles in clustering process. So, in the clustering process, different weights are assigned to different samples or objects. Sample weighting clustering could enhance the clustering effect (Nock and Frank, 2006). In this paper, the differences between the 'sample' and 'object' are ignored. The more sample deviates from the clustering center, the less its weight is.

3.1 K-Means Document Clustering Algorithm Based on Sample Weighting

Without regard to the weight of sample, the traditional K-Means algorithm ends clustering when the criterion function is convergent. The criterion function in document clustering process can be represented as the following formula:

$$J = \sum_{i=1}^{K} \sum_{j=1}^{m_i} Sim(\vec{d}_j, \vec{c}_i) \qquad (1)$$

Where J, which can be also called cohesion degree, can bed used to measure the performance of clustering. K is the total number of clusters and mi is the sum of all members in cluster i. $\vec{d}_j$ denotes the j-th member in cluster i. $\vec{c}_i$ is the center vector of cluster i, which can be computing according to formula (2). Sim($\vec{d}_j$, $\vec{c}_i$) is the similarity degree between cluster center vector $\vec{c}_i$ and document vector $\vec{d}_j$, which are both represented by vector space model after feature extraction and other steps such as feature weight computing respectively (Salton and Mcgill, 1983). One common measure of similarity is calculated by the cosine of the angle between the vectors (Salton and Mcgill, 1983).

$$\vec{c}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} \vec{d}_j \qquad (2)$$

In sample weighting clustering algorithm, the clustering criterion function is derived from the formula (3) just after weighting the clustering samples.

$$J' = \sum_{i=1}^{K} \sum_{j=1}^{m_i} (w_j \cdot Sim(\vec{d}_j, \vec{c}_i')) \qquad (3)$$

Where w_j denotes the weight of sample j with the constraint of $\sum_{j=1}^{m_i} w_j = 1$. $\vec{c}_i$ is the prototype of cluster i after clustering samples are weighted , and it can be computed according to the formula (4).

$$\vec{c}_i' = \sum_{j=1}^{m_i} (w_j \cdot \vec{d}_j) \qquad (4)$$

Obviously, the weight w_j plays an important role in adjusting the clustering prototype of a cluster. In the case of $w_j = \frac{1}{m_i}$, amely, the typicality of each sample is uniform on the clustering result, then formula (4) is converted into formula (2) and formula(3) is converted to formula (1) and results in the sample weighting clustering algorithm degrading to the traditional K-Means algorithm.

3.2 Fuzzy C-Means Document Clustering Algorithm Based on Sample Weighting

In 1973, Bezdek presented the fuzzy C-Means algorithm (Bezdek, 1981). The iterative formulas of fuzzy C-Means algorithm in document clustering process as follows.

$$u_{ij} = \frac{Sim^2(\vec{d}_j, \vec{c}_i)}{\sum_{k=1}^{c} Sim^2(\vec{d}_j, \vec{c}_k)} \qquad (5)$$

$$\vec{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^2 \vec{d}_j}{\sum_{j=1}^{n} u_{ij}^2} \qquad (6)$$

The criterion function in fuzzy C-Means clustering process can be represented as the following formula:

$$J = \sum_{i=1}^{c} \sum_{j}^{n} (u_{ij}^2 \cdot Sim^2(\vec{d}_j, \vec{c}_i)) \qquad (7)$$

In sample weighting clustering algorithm, the clustering criterion function is derived from the formula (7) just after weighting the clustering samples.

$$J' = \sum_{i=1}^{c} \sum_{j}^{n} (u_{ij}^2 \cdot Sim^2(\vec{d}_j, \vec{c}_i')) \qquad (8)$$

The center vector of cluster i, which can be computing according to formula (9).

$$\vec{c}_i' = \frac{\sum_{j=1}^{n} (w_j \cdot u_{ij}^2 \cdot \vec{d}_j)}{\sum_{j=1}^{n} (w_j \cdot u_{ij}^2)} \qquad (9)$$

4 Sample Weighting Methods

In the sample weighting clustering, one of the important works is automatic assignment weights to the clustering samples. In this work, the weights are assigned to the samples according to the importance of samples. In general, the citing relationship between academic documents could indicate the authority degree of a document approximately, which also denotes the structure information in document set. The authors calculate the simplified PageRank value of a document according to the citing relationship. In the comparison experiment on document weighting clustering, the authors use the cited frequency, simplified PageRank value of a document as its weight respectively. The cited frequency, simplified PageRank value of a document is defined as follows:

Definition 1: Document cited frequency is the total cited frequency of document $\vec{d}_j$ cited by the other documents in document set D, which is noted as Cited_Freq($\vec{d}_j$).

In this paper, the authors use references of the documents to compute the Cited_Freq($\vec{d}_j$). In general, if the documents' titles appeared in the references of $\vec{d}_j$, these documents must have been published before $\vec{d}_j$ publishing. Once a document A is cited by document B, it will never happen that B is cited by A at the same time[1]. So the link between documents is single-direction relationship shown in figure 1. It is completely different to the link relationship between web pages because two web pages may be linked to each other. According to the feature of documents, a simplified method to compute the PageRank value of a document with time attribute is presented.

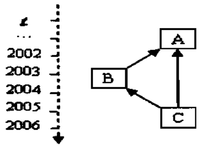


Figure 1. Citing relationship graph among the documents

Definition 2: The simplified document PageRank value is the authority degree of the document $\vec{d}_j$, which is noted as PageRank ($\vec{d}_j$). The initial value is 1, namely PageRank($\vec{d}_j$)=1.

The link relationship among in the documents can be regarded as a single-direction graph, which is different from relationship among in the web pages. PageRank($\vec{d}_j$) of document $\vec{d}_j$ is computed according the algorithm in paper (Brin and Page, 1998) and the formula in the computing process is as follow:

$$\textbf{PageRank}(\vec{d}_j) = r/N + (1-r) \sum_{i=1}^{N} (PR(\vec{d}_i)/C(\vec{d}_i)) \qquad (10)$$

Where PageRank ($\vec{d}_j$) denotes the rank of document $\vec{d}_j$, PageRank ($\vec{d}_j$) is the rank of document $\vec{d}_j$, and document $\vec{d}_j$ is linked to document $\vec{d}_j$. C($\vec{d}_j$) denotes the amount of out-links by document $\vec{d}_j$, namely the total number of the references in document $\vec{d}_j$. r is

[1] Document A can refer document B, and vice versa, if the respective authors are aware of each others' works at pre-print time. Because there aren't any pre-printed documents in the test data, the assumption of having acyclic graph for academic publication is presented in this paper.

damping factor which is an empirical value and r∈[0,1]. It is usually set at 0.15 and can be used to reduce the contribution of other documents for the ranking of document $\vec{d}_j$. N is the total number of documents in document set D.

The rank of a document is determined by other documents which citing it. However, the contribution of each citing document is different. The larger out-links in pi, the less contribution to document $\vec{d}_j$ it has. Meanwhile, the more cited frequency (i.e. in-link) of document $\vec{d}_j$, the higher its rank is. PageRank ($\vec{d}_j$) will be convergent after one iteration computing because that the link relationship among in the documents is a single-direction graph. The simplified PageRank value of the documents which hasn't been cited is set at 0.15.

After computing the Cited_Freq and PageRank value, the weight w_j of document $\vec{d}_j$ can be represented as $w_j^{Cited\_Freq}$ and $w_j^{PageRank}$, which can be computed according to formula (11) and formula (12) respectively.

$$w_j^{Cited\_Freq} = Cited\_Freq(\vec{d}_j) / \sum_{p=1}^{N} \text{Cited\_Freq}(\vec{d}_p) \quad (11)$$

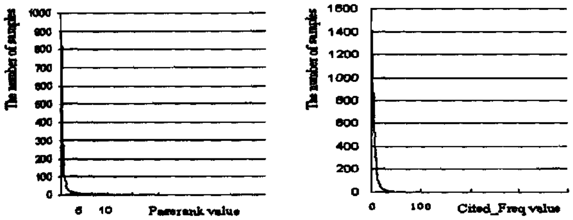
$$w_j^{PageRank} = PageRank(\vec{d}_j) / \sum_{p=1}^{N} \text{PageRank}(\vec{d}_p) \quad (12)$$

5 Experiments and Evaluation

5.1 Evaluation Data

The paper takes Chinese academic documents as clustering samples, each of which has a class tag. The class tags are used in the evaluation methods. The area of the academic documents is economics. The data size is 100000. The document cited frequency and simplified PageRank value of each paper are computed and the maximum values of which are 311, 18.49928 respectively. Figure 2 shows their distributions.

We use 100000 documents as clustering samples in the clustering experiment. In order to evaluate the document clustering algorithm based on sample weighting, this section presents groups of comparison experiments respectively on samples without weighting, which is noted as baseline, weighting by $w_j^{Cited\_Freq}$, which is noted as Cited_Freq, and weighting by $w_j^{PageRank}$.



(a) (b)

Figure 2. Distribution of PageRank value (a) and Cite_Freq value (b) of samples

This paper does three groups of experiments, in which the number of clustering prototype, i.e. K is set at 10, 15 and 20 respectively. With the different initial clustering prototypes, each group is repeated five times. The comparison experiments of baseline, Cited_Freq and PageRank are carried out each time. Lastly the average value is computed.

5.2 Evaluation Measure

The evaluation measures for clustering performance can be divided into two kinds approximately: supervised evaluation and unsupervised evaluation.

5.2.1 Supervised Evaluation

Evaluation measure based on supervised learning measure the extent to which the clustering structure discovered by a clustering algorithm matches some external structure (Pang et al, 2005). This measure method is also called the external evaluation method.

5.2.1.1 Entropy

It measures the degree to which each cluster consists of samples of a single class. Firstly, the entropy of each cluster is computed according to the following formula:

$$E_i = -\sum_{j=1}^{L} (p_{ij} \log_2 p_{ij}) \quad (13)$$

Where E_i denotes the entropy of cluster i. L is the total number of class labels. p_{ij} represents the probability that members of cluster i belong to class j, and can be calculated by formula (14) based on maximum likelihood principle.

$$p_{ij} = \frac{m_{ij}}{m_i} \quad (14)$$

Where m_i denotes the total number of members in cluster i, and m_{ij} is the total members of category j in clusters i. The total entropy of cluster set can be computed by weighting entropy of each cluster, shown in formula (15).

$$E=\sum_{i=1}^{K}\left(\frac{m_i}{m}E_i\right) \qquad (15)$$

Where E denotes the total entropy of cluster set with K as the total number of clusters, m is the total number of clustering samples.

The less the total entropy is, the better the clustering performance is.

5.2.1.2 Purity

It's another measure of the extent to which a cluster contains samples of a single class (Pang et al, 2005). The purity of cluster i can be computed by formula (16), then the total purity can be obtained by formula (17). The higher the total purity is, the better the clustering performance is.

$$\text{Purity}_i=\text{Max}\ (p_{ij}) \qquad (16)$$

$$\text{Purity}=\sum_{i=1}^{K}\left(\frac{m_i}{m}Purity_i\right) \qquad (17)$$

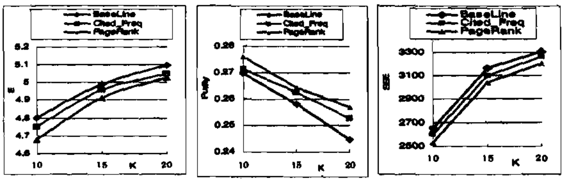
5.2.2 Unsupervised Evaluation

It is also called the interior evaluation method, which measures the goodness of a clustering structure without respect to external information. It is often further divided into two classes: measures of cluster cohesion, which determine how close related the samples in a cluster are, and measures of cluster separation, which determine how distinct or well-separated a cluster is from other clusters (Pang et al, 2005). It has been proved that the minimizing SSE is equivalent to maximizing SSB (Pang et al, 2005). Therefore, this paper chooses SSE only to evaluate the document clustering performance in the unsupervised evaluation. The cohesion degree can be computed according to formula (3). This paper evaluates the cohesion degree by $\text{SSE}=\sum_{i=1}^{K}\sum_{j=1}^{m_i}(1-Sim(\vec{d}_j,\vec{c}_i'))$. So the less SSE, the better clustering performance is.

5.3 Evaluating Results and Analysis

5.3.1 Evaluating Results

According to the experiment in section 5.1 and the evaluation measure in section 5.2, the measure results can be obtained.



(a) (b) (c)

Figure 3. Comparison of total entropy (a), purity (b), cohesion degree (c) of WKM

Based on the average result of each group, the comparison of total entropy, purity, cohesion degree of different clustering methods are shown in Figure 3(a), Figure 3(b), Figure 3(c), which is corresponding to 3 kinds of K value, i.e. K=10, K=15, K=20, respectively.

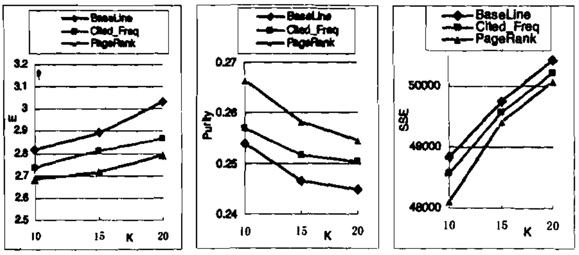
According to Figure 3(a), when K is set at 10, 15 and 20 respectively, the result of the total entropy (E) comparison is: BaseLine > Cited_Freq > PageRank. The less the total entropy is, the better the clustering performance is. So, taking into account of the total entropy, the weighting method based on PageRank has the best performance, and then is the weighting method based on Cited_Freq, the method without weighting is the worst.

According to Figure 3(b), when K is set at 10, 15 and 20 respectively, the result of the purity comparison is: PageRank > Cited_Freq > BaseLine. The performance of the weighting method based on Cited_Freq is close to the method without weighting. According to the purity of clustering results, the weighting method based on PageRank has the best performance.

According to Figure 3(c) , when K is set at 10, 15 and 20 respectively, the result of the cohesion degree（SSE）comparison is: BaseLine > Cited_Freq > PageRank. The less SSE is, the better the clustering performance is. So, from the point of view of SSE, the weighting method based on PageRank has the best performance, and then is the weighting method based on Cited_Freq. The worst is the method without weighting.

5.3.2 Evaluating Results and Analysis of WFCM

The comparison of total entropy, purity, cohesion degree of different clustering methods in WFCM are shown in Figure 4(a), Figure 4(b), Figure 4(c), which is corresponding to 3 kinds of K value, i.e. K=10, K=15, K=20, respectively.



(a) (b) (c)

Figure 4. Comparison of total entropy (a), purity (b), cohesion degree (c) of WFCM

To sum up, document clustering algorithm based on sample weighting can enhance the performance of document clustering, which can be proved by the results of supervised evaluation and unsupervised evaluation. To some extent, it also answers the questions proposed at the beginning of the paper, namely: The structure information between samples is helpful to sample weighting. Taking the advantage of citing relationship among documents to automatically computing the simplified PageRank value of each document, and then being used in the sample weighting clustering algorithm can enhance the performance of document clustering.

6 Conclusions and future work

In this paper, a novel sample weighting clustering algorithm is presented based on K-Means algorithm and fuzzy C-Means algorithm. The algorithm uses academic documents as the clustering samples. The PageRank value of each document is calculated according to the citing relationship among them, and it is used as the weight in the sample weighting clustering algorithm. The effectiveness has been verified by the experiment results.

As the future work, the authors plan to study the issues of (1) extending the K-Means and fuzzy C-Means algorithm to EM algorithm and some other partition algorithm, and (2) exploring the sample weighting clustering based on the Page-Rank value of web pages.

Acknowledgement

The work described in this paper has been supported in part by Supported by National Key Project of Scientific and Technical Supporting Programs funded by Ministry of Science & Technology of China (NO. 2006BAH03B03), Youth Research Support Fund funded by Nanjing University of Science & Technology (NO. JGQN0701), and Innovation Project of Graduate Education of Jiangsu Province (2006) funded by Jiangsu Education Commission.

References

J. C. Bezdek. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

S. Brin and L. Page. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: *Proceedings of the 7th ACM-WWW International Conference*, 107~117.

A. P. Dempster, N. M. Laird, D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statistical Soc. B*, Vol. 39, 1-38.

J. Han and M. Kamber. 2000. *Data Mining: Concepts and Techniques*, Morgan Kaufmann.

Jie Li, Xinbo Gao, and Licheng Jiao. 2005. A Novel Typical-Sample-Weighting Clustering Algorithm for Large Data Sets, *LNAI* 3801, 696–703.

J. MacQueen. 1967. Some Methods for Classification and Analysis of Multivariate Observation, In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

R. Nock and F. Nielsen. 2004. An Abstract Weighting Framework for Clustering Algorithms. In: *Proceedings of the Fourth International SIAM Conference on Data Mining*, 200-209.

R. Nock and F. Nielsen. 2006. On Weighting Exponent, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, NO. 8, 1223-1235.

W. Pedrycz. 1996. Conditional Fuzzy C-Means. *Pattern Recognition Letters*, Vol. 17, 625-632.

K. Rose. 1998. Deterministic Annealing for Clustering Compression, Classification, Regression, and Related Optimization Problems. In: *Proceedings of the IEEE*, 86(11), 2210-2239.

G. Salon and M. J. Mcgill. 1983. *Introduction to modern information retrieval*. McGraw-Hill Book Co., NewYork.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. 2005. *Introduction to Data Mining*, Pearson Education, Inc.