Document Clustering Method Based on Visual Features

Yucong Liu, Bofeng Zhang, Kun Xing, Bo Zhou
School of Computer Engineering & Science
Shanghai University
Shanghai, China
e-mail: liuyucong@163.com, bfzhang@shu.edu.cn

***Abstract*—There are two important problems worth conducting research in the fields of personalized information services based on user model. One is how to get and describe user personal information, i.e. building user model, the other is how to organize the information resources, i.e. document clustering. It is difficult to find out the desired information without a proper clustering algorithm. Several new ideas have been proposed in recent years. But most of them only took into account the text information, but some other useful information may have more contributions for documents clustering, such as the text size, font and other appearance characteristics, so called visual features. This paper proposes a method to cluster the scientific documents based on visual features, so called VF-Clustering algorithm. Five kinds of visual features of documents are defined, including body, abstract, subtitle, keyword and title. The thought of crossover and mutation in genetic algorithm is used to adjust the value of *k* and cluster center in the k-means algorithm dynamically. Experimental result supports our approach as better concept. In the five visual features, the clustering accuracy and steadiness of subtitle are only less than that of body, but the efficiency is much better than body because the subtitle size is much less than body size. The accuracy of clustering by combining subtitle and keyword is better than each of them individually, but is a little less than that by combining subtitle, keyword and body. If the efficiency is an essential factor, clustering by combining subtitle and keyword can be an optimal choice.**

Keywods-document clustering; k-means; visual features; genetic algorithm

I. INTRODUCTION

In recent years, personalized information services play an important role in people's life. There are two important problems worth researching in the fields. One is how to get and describe user personal information, i.e. building user model, the other is how to organize the information resources, i.e. document clustering. Personal information is described exactly only if user behavior and the resource what they look for or search have been accurately analyzed. The effectiveness of a personalized service depends on completeness and accuracy of user model. The basic operation is organizing the information resources. In this paper we focus on document clustering.

At present, as millions of scientific documents available on the Web. Indexing or searching millions of documents and retrieving the desired information has become an increasing challenge and opportunity with the rapid growth of scientific documents. Clustering plays an important role in analysis of user interests in user model. So high-quality scientific document clustering plays a more and more important role in the real word applications such as personalized service and recommendation systems.

Clustering is a classical method in data mining research. Scientific document clustering [6][8][9] is a technique which puts related papers into a same group. The documents within each group should exhibit a large degree of similarity while the similarity among different clusters should be minimized.

In general, there are lots of algorithms about clustering [1][5][10][13], including partitioning methods[5] (k-means, k-medoids etc), hierarchical methods [16] (BIRCH, CURE, etc), density-based methods (DBSCAN, OPTICS, etc), grid-based methods (STING, CLIQUE, etc) and model-based methods, etc.

In 1967, MacQueen first put forward the k-means [2][3][4][7] clustering algorithm.The k-means method has shown to be effective in producing good clustering results for many practical applications. However it suffers from some major drawbacks that make it inappropriate for some applications. One major disadvantage is that the number of cluster k must be specified prior to application. And another is the sensitivity to initialization. The two drawbacks of k-means not only affect the efficiency of the algorithm but also influence clustering accuracy.

There are many existing document representation approaches [11], including Boolen Approach, Vector Space Model (VSM), Probabilistic Retrieval Model and Language Model. At present the most popular document representation is Vector Space Model (VSM). In the 1960's, G. Salton and other people proposed VSM. VSM is an algebraic model for representing text documents as vectors of identifiers. Documents are represented as vectors, such as $d_i = (w_{i1}, w_{i2}, \dots, w_{ij}, \dots, w_{in})$. The main advantages of this

978-0-7695-4580-6/11 $26.00 © 2011 IEEE
DOI 10.1109/iThings/CPSCom.2011.69

representation are its conceptual simplicity and its efficiency of similarity computation. Its main disadvantage is the fact that it loses important information about the original document.The classic vector space model proposed by Salton, Wong and Yang. The model is known as term frequency-inverse document frequency model (TF-IDF) [12]. In document clustering, the document size is not taken into account when calculating the weight of clustering keywords using TF-IDF.

The most important goal of this paper is to develop a technique which will guide the user to get desired information with proper clustering of scientific documents in web or information retrieval systems. In this paper we propose a high performance document clustering algorithm (called VF-Clustering) based on document's visual features, including body, abstract, subtitle, keyword and title. We integrate several visual features to represent documents. We also use the thought of crossover and mutation in genetic algorithm [14][15] to improve the k-means algorithm. We merge and add cluster centers during the process of clustering to adjust the value of k and cluster center dynamically. Experimental result shows that our approach is better in terms of clustering performance of the scientific documents.

The paper is organized as follows. Section 2 expresses the key steps of document clustering. Section 3 presents the document clustering algorithm based on visual features. Section 4 shows the implementation of VF-Clustering in Chinese scientific document clustering. Section 5 concludes the paper.

II. KEY STEPS OF DOCUMENT CLUSTERING

A. Document Segmentation

As it is necessary to segment document into words before document feature extraction, in our research, we use lexicon-based Word segmentation tools of the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). However, its lexicon version is too low so that we add a large amount of new words into this lexicon and remove stop words from the result set of words segmentation.

B. Document Representation and Feature-Words Selection

As we know, Vector Space Model (VSM) is widely used in document clustering, in which each n-dimensional vector represents a document. In this paper, VSM can be represented as (1).

$$\mathrm{d_i} = ((\mathrm{t_{i1}}, \mathrm{w_{i1}}), (t_{i2}, w_{i2}), \dots, (t_{ij}, w_{ij}), \dots (t_{in}, w_{in})) \quad (1)$$

where $\mathrm{d_i}$ means the i-th document, t_{ij} expresses the j-th keyword words of the i-th document, and w_{ij} represents the weight of the j-th keyword in the i-th document.

This paper adopts classical TF-IDF as the clustering keywords weight calculation method because it has an advantage in considering words occurrence frequency not only in a document but also in the whole date set. Furthermore, in this paper the size of each document is also taken into account, and the parameter weight is defined by (2).

$$w_{ij} = \frac{tf(i, t\_ij) * log\left(\frac{N}{m} + 0.01\right) * \frac{\sum_{k=1}^{k=N} size(k)}{N}}{size(i)} \quad (2)$$

where *size* (*i*) means the number of effective characters of the *i-th* document, $\frac{\sum_{k=1}^{k=N} size(k)}{N}$ shows the average size of all the document in date set, $tf(i, t\_ij)$ expresses the words occurrence frequency of keyword t_ij appeared in document d_i, m is the number of documents containing t_ij, and N is the total number of document contained in a document set.

C. Similarity Measurement

After the document representation using VSM, a document can be represented by a point in n-dimensional space, while the similarity measurement between different documents was represented by the distance between corresponding points. The closer the distance between the two points is in n-dimensional, the more similar the documents represented by the two points is, and vice versa. To calculate the distance, there are many different methods, such as Mahalanobis distance and Euclidean distance, etc. The more similar two documents is, the more similar coefficient close to 1, conversely, the similar coefficient is close to 0. In this paper, documents' similarity here is presented by cosine similarity which is defined by (3).

$$cos(i, j) = \frac{d_i \cdot d_j}{||d_i|| \cdot ||d_j||} \quad (3)$$

For example，

$d_i = (1,2,2,1,0)$， $d_j = (0,1,2,1,1)$

$d_i \cdot d_j = 1 * 0 + 2 * 1 + 2 * 2 + 1 * 1 + 0 * 1 = 7$

$||d_i|| = sqrt(1 * 1 + 2 * 2 + 2 * 2 + 1 * 1 + 0 * 0)$

$||d_j|| = sqrt(0 * 0 + 1 * 1 + 2 * 2 + 1 * 1 + 1 * 1)$

$$cos(i, j) = \frac{7}{\sqrt{10} * \sqrt{7}} = \sqrt{0.7}$$

III. DOCUMENT CLUSTERING ALGORITHM BASED ON VISUAL FEATURES

The main characteristics of document clustering algorithm based on visual features, so called VF-Clustering are as follows:

1) Five kinds of visual features are defined according to the analysis of content and structure of scientific document, including body (B), abstract (A), subtitle (S), keyword (K) and title (T). And the importance of these features to scientific document clustering will be compared through experiments.

2) In view of the two drawbacks of k-means algorithm, the thought of crossover and mutation in genetic algorithm is used to improve the k-means algorithm. Adjust the values of k and cluster center dynamically by merging and adding cluster centers in the process of clustering.

The implementation of clustering algorithm introduces below.

A. Document Presentation Based on Visual Features

As the most widely used document presentation method, the mentioned model VSM represents document in two ways. In one way we can segment words and select clustering keywords according to words' frequency by mainly analyzing the body of the document, or put clustering keywords selected in the first time into selection from the whole document, and according to the clustering keywords' position, their weight shall be adjusted if they occurrences in title or abstract. In the other way, only title and abstract are analyzed to retrieve clustering keywords and do further clustering, though effective, the result obtained in this way is not accurate enough.

In this paper, a document representation based on visual features is defined with a full consideration of the importance of each visual feature in the whole document. Therefore, we segment words on the basis of every visual feature independently and retrieve clustering keywords from each part with features extraction method introduced above. And according to the importance of every visual feature, it shall be adjusted for the clustering keywords' weight (w'_{ij}) of comprehensive document representation, with w'_{ij} be obtained by (4) and comprehensive document presentation shown in (5).

$$w'_{ij} = \frac{a*B(w_{ij})+b*S(w_{ij})+c*A(w_{ij})+d*K(w_{ij})+e*T(w_{ij})}{a+b+c+d+e} \quad (4)$$

$$d'_i = \left((t_{i1}, w'_{i1}), (t_{i2}, w'_{i2}), \dots, (t_{ij}, w'_{ij}), \dots, (t_{in}, w'_{in})\right) \quad (5)$$

where $B(w_{ij})$ means the weight of clustering keyword i in body part, and the values of *a, b, c, d, e* must be either part equal to 0 or all no less than 1. In our experiment we set the values of *a* and *b* equal to 2, others equal to 1.

B. K-means Algorithm Optimization Based on Crossover and Mutation

Take advantage of the idea of crossover and mutation in genetic algorithm during the process of clustering, this algorithm dynamically adjusts the values of k as well as cluster center by means of mergence and addition, to achieve k-means algorithm optimization.

Optimized clustering algorithm process is as follows:

Input: The initial number of cluster center k, Similarity threshold λ. In our experiment we set the value of k equal to 4.

Output: The clustering clusters formed finally (the number of clusters not necessarily equals k).

Step 1 Initialize cluster centers. First of all, it is necessary to check whether the newly selected cluster center is the existed one. If it is, the cluster center can be reproduced. Or else calculate the similarity between the current centers and selected one and compare this similarity with λ. If the similarity is bigger, reselect a document as a new center and go back to execute step 1 once more until the number of cluster center equal to k.

Step 2 Calculate the similarity between each data and each cluster center, and then compare the biggest similarity with a given threshold λ. On one hand, if the similarity is bigger, the data shall be put into a cluster with its similarity biggest.On the other hand, the thought of mutation in the genetic algorithm is used in here, the data should be added into cluster center as a new one which can cause cluster center number change.

Step 3 Recalculate the center of each cluster which is defined as the arithmetic average value of all data in this cluster. For example, it is assumed that there are 3 documents in the first cluster, which are

$$d_1 = \big((A, 2), (B, 3)\big)$$
$$d_2 = \big((A, 3), (B, 3)\big)$$
$$d_3 = ((A, 4), (B, 3), (C, 1))$$

Their new cluster center should be:

$$center = \left(\left(A, \frac{2+3+4}{3}\right), \left(B, \frac{3+3+3}{3}\right), \left(C, \frac{0+0+1}{3}\right)\right)$$
$$= ((A, 3), (B, 3), (C, 0.33))$$

Step 4 Calculate the similarity for every pair of new cluster centers obtained in step 3. The thought of crossover in genetic algorithm is used in here. Two clusters have to be merged if the similarity between them is bigger than λ. For example, there are 2 cluster centers: center1= ((A, 3), (B, 4), (C, 2)), center2= ((A, 2), (B, 4), (C, 3)), and the two merged into one cluster center, that is,

$$center = \left(\left(A, \frac{3+2}{2}\right), \left(B, \frac{4+4}{2}\right), \left(C, \frac{2+3}{2}\right)\right)$$
$$= ((A, 2.5), (B, 4), (C, 2.5))$$

Step 5 Execute step 2, step 3 and step 4 once more, and finish this process if cluster center reaches a stable value or maximize iteration times, or else return to step 2 and continue to execute this process.

IV. Implementation of VF-Clustering in Chinese Scientific Document Clustering

A. Evaluation of Clustering Results

There is still no uniform standard for the evaluation of document clustering results, however, precision rate and recall rate which reflect two different aspects of quality clustering must be taken into account together. Since *F*1 test value combines the two precisely, we use the most commonly evaluation, precision rate, recall rate and *F*1 test value to evaluate the effect of the document clustering.

Each artificial labeled theme T_i in data set corresponds to a clustering result set C_{ij} in clustering result. Now we define recall rate, precision rate and *F*1 as follows:

$$R(T_i, C_{ij}) = \frac{|T_i \cap C_{ij}|}{T_i} \quad (6)$$

$$P(T_i, C_{ij}) = \frac{|T_i \cap Ci_j|}{C_{ij}} \quad (7)$$

$$F1 = \frac{2*R(T_i,C_{ij})*P(T_i,C_{ij})}{R(T_i,C_{ij})+P(T_i,C_{ij})} \quad (8)$$

B. Experiment and Result Analysis

Text data sets are from 195 articles of Chinese scientific and technical document in CNKI, including 47 articles of Clustering Algorithm (CA), 58 articles of Data Mining (DM) 43 articles of Cloud Computing (CC) and 47 articles of Genetic Algorithm (GA). We pre-treatment the data set, we separately extract five visual features of each document to a save to the database table. The part of the experimental original data is shown in Fig. 1.

表 - dbo.webpage | 表 - dbo.titlekeyword | 表 - dbo.summarykeyword | 表 - dbo.subtitlekeyword | 表 - dbo.keyword | 表 - dbo.keykeyword | 表 - dbo.bodykeyword | 摘要

pageid	title	summary	keyword	body	visual	size	label	nl0	nl1	nl2	nl3	nl4	nl5	nl6	n7	nl8	nl9
1	一种抑制早熟…	遗传算法在许…	遗传算法; 早熟…	引言遗传算法…	基本遗传算法…	3349	4	4	4	4	4	4	4	4	4	4	4
2	新技术思想与…	文章通过探讨…	数字图书馆云…	经济社会环境…	新技术思想原…	5350	3	3	3	3	2	3	3	3	3	3	3
3	遗传算法在背…	索朝着计算机…	1背包问题; 简…	引 言背包问…	简单遗传算法…	3304	4	1	1	1	4	4	4	4	4	4	4
4	遗传算法在车…	在车间作业调…	遗传算法;作业…	引言车间作业…	问题的描述基…	2938	4	4	4	4	4	4	4	4	4	4	4
5	注入式的遗传…	该文提出了一…	遗传算法; 知识…	遗传算法是模…	算法的思想与…	1873	4	4	4	1	4	4	4	4	1	4	1
6	云计算在图书…	云计算技术以…	云计算 云计算…	1．1云计算的…	云计算的概念…	5235	3	3	3	3	3	3	3	3	3	3	3
7	空间数据挖掘…	随着现代科学…	空间数据挖掘;…	引 言随着卫…	空间数据挖掘…	4226	2	2	2	2	2	2	2	2	2	2	2
8	云计算给图书…	与云计算相关…	云计算 图书…	引言云计算(Clo…	云计算与图书…	7297	3	3	3	3	3	3	3	3	3	3	3
9	时空数据挖掘…	本文系统阐述…	数据挖掘; 空间…	引言空间数据…	空间数据挖掘…	6600	2	2	2	2	2	2	2	2	2	2	2
10	自适应遗传算…	利用水头实测…	参数反演; 渗透…	由于岩土工程…	遗传算法及其…	3046	4	1	4	4	4	4	4	4	1	4	4
11	遗传算法在建…	遗传算法易搜…	离散变量; 结构…	在解决复杂设…	离散变量结构…	3087	4	4	4	1	4	4	4	4	4	4	4
12	基于领域本体…	数据挖掘已成…	领域本体 语义…	伴随着信息时…	领域本体和语…	3361	2	2	2	2	2	2	2	2	2	2	2
13	可视化数据挖…	通过对可视化…	可视化数据挖…	引言可视化数…	依据可视化数…	3773	2	2	2	2	2	2	2	2	2	2	2
14	一种实现全父…	该文提出了一…	遗传算法 神经…	引言作为一种…	全父辈交叉遗…	4537	4	4	4	1	4	4	4	4	1	4	4
15	国外先进数据…	近年来，国外…	数据挖掘; 知…	概述数据挖掘…	数据挖掘工具…	3953	2	2	2	2	2	2	2	2	2	2	2
16	遗传算法 (原理…	遗传算法(Gene…	自然选择 进化论	自然选择（Nat…	自然选择遗传…	1454	4	2	4	1	4	4	4	4	4	4	4
17	基于XML的WEB…	互联网的广泛…	数据挖掘	随着互联网的…	Web数据挖掘…	2289	2	4	2	2	2	2	2	2	2	2	2
18	虚拟企业合作…	虚拟企业是新…	虚拟企业合作…	引言世纪知识…	虚拟企业合作…	4217	1	3	1	1	1	4	1	1	1	1	1

Figure 1. The part of the experiment original data

where the *label* is artificial classified marks, while *nl* shows the clustering result.

The first step of the experiment: firstly, make word segment for five visual features independently, remove stop words and extract clustering keywords; then, make a clustering for each visual feature that represents documents independently. The experimental result is shown in TABLE I. Where the k-means shows the basic clustering algorithm and make the body representing the documents, all others adopt the improved algorithm.

TABLE I. RESULTS OF CLUSTERING BY FIVE VISUAL FEATURES

		k-means(%)	B (%)	A (%)	S (%)	K (%)	T (%)
CA	R	76.60	76.60	74.47	76.60	55.32	53.19
	P	83.72	83.72	53.03	85.71	92.86	48.93
	*F*1	80.00	80.00	61.95	80.90	69.33	50.97
DM	R	93.10	93.10	91.38	93.10	86.25	87.93
	P	90.00	90.00	82.81	88.52	84.85	82.26
	*F*1	91.53	91.53	86.89	90.76	85.54	85.00
CC	R	93.02	93.02	90.69	90.70	95.35	93.02
	P	90.69	90.69	88.38	86.04	97.62	78.43
	*F*1	91.84	91.84	89.50	88.30	96.47	85.11
GA	R	93.62	93.62	40.43	89.36	100.00	91.79
	P	84.62	84.62	95.00	86.27	79.66	58.11
	*F*1	88.89	88.89	56.72	87.79	88.68	71.07

Through the analysis of the first step of the experimental results, we could conclude as follows:

1) Because the value of the *k* is equal to 4, so the basic algorithm and the improved algorithm to clustering have the same results when make the body representing document independently. But the clustering running time are reduced when use the improved algorithm.

2) The clustering performance by visual features body and subtitle are best in representing documents independently, and good steady is exhibited in these types of data sets. What's more, the visual feature body has slightly better clustering results than subtitle.

3) The visual feature keyword is better than abstract and title in clustering effect, moreover, abstract and title are poor in the stability of the clustering result by representing document independently. Among these three visual features, clustering has a good effect in a new subject or a subject with fewer applications. However, it has a relatively poor effect in subject with extensive applications.

4) The visual features title has poor clustering results in subject with extensive applications.

Under the first step of the experimental results, we make an analysis of clustering results obtained through different visual features representing the document independently. We make different combinations of visual features to represent the document and clustering. The result is shown in TABLE II.

Summarize the analysis of the results of the second step of experiment as follows:

1) From the whole analysis of the two results in TABLE I and TABLE II, it's obviously draw that the clustering result of the comprehensive visual features is better than any single visual feature in representing documents.

TABLE II. RESULTS OF CLUSTERING BY DIFFERENT COMBINATION

		S, K (%)	B, S (%)	B, S, K (%)	B, S, K, A (%)	B, S, A, K, T (%)
CA	R	80.85	85.11	95.74	95.74	95.74
	P	90.48	93.02	91.84	90.00	93.75
	*F*1	85.39	88.89	93.75	92.78	94.74
DM	*R*	94.83	98.28	93.10	93.10	94.83
	P	96.49	90.48	100.00	96.43	98.21
	*F*1	95.65	94.21	96.43	94.74	96.49
CC	R	97.67	100.00	100.00	97.67	97.67
	P	97.67	100.00	100.00	100.00	100.00
	*F*1	97.67	100.00	100.00	98.82	98.82
GA	R	95.74	93.62	100.00	95.74	100.00
	P	84.91	95.65	95.92	95.74	95.92
	*F*1	90.00	94.62	97.92	95.74	97.92

2) Although the clustering results of visual features that consist of subtitle and keyword are slightly better than the visual feature body representing documents independently, the effective number of characters of subtitle and keyword is less than the body's, so it greatly enhances the efficiency of feature words selection when making words segment. This way could be used to meet the high requirements of the clustering results and efficiency.

3) The integrated independent visual feature includes body, subtitle and keyword, in which each one has the best clustering results to express text, and its clustering results is almost the same as the one that integrate five visual features to represent a document. Moreover, the clustering results of these two combinations are the best, although the efficiency is not so good. In order to meet the higher requirement of the clustering results, we could combine body, subtitle and keyword together to represent a document.

V. CONCLUSION

This paper implements a method to cluster the scientific documents based on visual features (VF-Clustering). And through the deep analysis of these clustering results we find some useful information as follows:

1) In the five visual features, body representing documents independently to cluster have the best accuracy and steadiness, and subtitle is next. However, the clustering effect of abstract, keyword and title are not very good, especially in the widely applied field of knowledge clustering.

2) The accuracy of clustering by combining subtitle and keyword is better than each of them individually. Moreover, operation time can be saved greatly for the less effective characters in the two parts. If the efficiency is an essential factor, clustering by combining subtitle and keyword can be an optimal choice.

3) If the higher accuracy is demanded, clustering combining body, subtitle and keyword is a better choice.

This paper also uses the thought of crossover and mutation in genetic algorithm to improve the k-means algorithm and heightens the efficiency greatly by adjusting the values of *k* and cluster center dynamically in the process of clustering.

ACKNOWLEDGMENT

This work is supported by Shanghai Leading Academic Discipline Project (J50103) and Innovation Program of Shanghai Municipal Education Commission (11ZZ85).

REFERENCES

[1] S. Guha, R. Rastogi, and K. Shim, "An efficient clustering algorithm for large databases," ACM SIGMOD international conference on Management of data, Volume 27 Issue 2, June 1998.

[2] A. Likasa, and N. Vlassisb, "Verbeekb. The global k-means clustering algorithm. Pattern Recognition," 2003, pp. 451 – 461.

[3] J. A. Hartigan, and M. A. Wong, "A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 28, No. 1,1979, pp.100-108.

[4] K. Wagsta, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-means Clustering with Background Knowledge," Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 577-584.

[5] R. Dutta, I. Ghosh, A. Kundu, and D. Mukhopadhyay, "An Advanced Partitioning Approach of Web Page Clustering utilizing Content & Link Structure," Journal of Convergence Information Technology Volume 4, Number 3, 2009.

[6] J. L. Neto, A. D. Santos, and C. A. A. Kaestner, "Alex A. Freitas, Document Clustering and Text Summarization," Information Processing and Management, 2000.

[7] J.M. Pena , J.A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the K-Means algorithm," Pattern Recognition Letters, 1999, pp.1027-1040.

[8] L. Yanjun, M. Chung , and D. Holt, "Text document clustering based on frequent word meaning sequences," Data & Knowledge Engineering, 2008, pp. 381–404.

[9] E Rasmussen, P. Hall, and E. Cliffs, "Clustering algorithms," Information Retrieval, 1992, pp.419-442.

[10] A. K. Jain, and M. N. Murty, "Data Clustering: A Review," ACM Computing Surveys (CSUR), 1999, pp.264–323.

[11] W. B. Cavnar, "Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model," Proc. of TREC-3 (Third Text REtrieval Conference), Gaithersburg, 1994.

[12] G. Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing and Management 24, 513-523. 1988. Reprinted in: Sparck Jones, K. and Willet, P. Eds. Readings in Information Retrieval, 1997, pp.323-328.

[13] N. Grira, Crucianu, and M. Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," 7th ACM SIGMM international workshop on Multimedia information retrieval, 2005, pp.9-16.

[14] U. Maulik, and S. Bandyopadhyay, "Genetic algorithm-based clustering technique, Pattern Recognition," 2000, pp.1455-1465.

[15] K. Krishna, and M. Narasimha Murty, "Genetic K-Means Algorithm, Item Identifier S," 1999, pp.1083-4419.

[16] J. F. Navarro, C. S. Frenk, and S. D. M. White, "A universal density profile from hierarchical clustering," The astrophysical jouranl, 1997, pp.490-493.