Fuzzy Post-clustering Algorithm for Web Search Engine

Younghee Im, Jiyoung Song, and Daihee Park

Dept. of computer & Information Science, Korea Univ., Korea
{yheeim, songjy, dhpark}@korea.ac.kr

Abstract. We propose a new clustering algorithm satisfying requirements for the post-clustering algorithms as many as possible. The proposed "Fuzzy Concept ART" is the form of combining the concept vector having some advantages in document clustering with Fuzzy ART known as real-time clustering algorithms.

1 Introduction

Web-document clustering methods could be divided into pre-clustering methods and post-clustering methods; the former are off-line clustering of the entire document collection, and the latter are on-line clustering of the retrieved document set by Web search engines [1] [2] [3]. The post-clustering algorithms have different requirements from both conventional clustering algorithms and pre-clustering algorithms. Zamir et al. [4] have identified some key requirements for the post-clustering algorithms as follows: 1) Relevance; 2) Browsable Summaries; 3) Overlap; 4) Snippet-tolerance; 5) Speed; 6) Incrementaility.

We intend to devise a new clustering algorithm satisfying requirements for the post-clustering algorithms as many as possible. To devise a new post-clustering algorithm, we borrow two important concepts such as a concept vector [5] and Fuzzy ART [6]. The proposed one, which is named by "Fuzzy Concept ART(FCART)", is the form of combining the concept vector that have some advantages in document clustering with Fuzzy ART known as real-time clustering algorithms. FCART is satisfied with all of requirements for the post-clustering algorithms. Besides, we expect that it may be the alternative model to circumvent some drawbacks of Fuzzy ART such as sensitivity to the order of input sequence, time-complexity, and true meaning of fuzzy set theory [8].

2 Document Representations and Concept Vector

2.1 Document Representations

In the vector space model, each document is represented as the weighted term-frequency vector. According to the relevant researches [4], composing document vector with the snippets returned by Web search engine can reduce the search space significantly keeping the precision of clustering. Also, since the title of

G.G. Lee et al. (Eds.): AIRS 2005, LNCS 3689, pp. 709–714, 2005.

document stands for the whole content of document, we use only both snippets and title in constructing document vector instead of using the entire document to improve the speed of clustering. In addition, design of loading the clustering tool in the client machines may take an advantage of reducing overload of Web search engine's server.

2.2 Concept Vector

The concept vector [5] is the normalized centroid of the cluster to have unit Euclidean norm. The concept vector of a certain cluster is guaranteed to be closest in cosine similarity (in an average sense) to all document vectors in the corresponding cluster. In particular, the concept vectors are sparse and localized in the word space. The sparsity of concept vectors simplifies the computation of cosine similarity and cluster's coherence [5]. Hence the computational complexity of post-clustering can be remarkably decreased.

Furthermore, the locality of concept vectors is extremely useful in labeling the latent concepts for clusters. If it is provided users with well-represented labels of cluster, they can select the cluster containing the information that they want, by seeing the labels alone.

The keywords of cluster π_j are represented as a word cluster $\mathbf{Word_j}$. it is defined as follows [5]

$$\mathbf{Word_j} = \{k\text{th word} : \mathit{1} \leq k \leq d,\ \mathbf{c_{k,j}} \geq \mathbf{c_{k,m}},\ \mathit{1} \leq m \leq c,\ m \neq j\} \tag{1}$$

where d is total number of terms. Since the concept vectors are local to each word cluster, the word cluster $\mathbf{Word_j}$ provides good keywords for the corresponding cluster. Also among the document vectors in the cluster, the summary of cluster may be thought of as the document vector that is closet in cosine similarity to the concept vector. It is possible to understand the contents of cluster intuitively. So, in FCART, the cluster's summary of cluster π_j, $\mathbf{Summary_j}$ is defined as follows

$$\mathbf{Summary_j} = \arg \max_{\mathbf{x} \in \pi_j} \{\cos(\theta(\mathbf{x},\ \mathbf{c_j}))\} \tag{2}$$

3 FCART(Fuzzy Concept ART)

The basic idea of FCART is that the weight vector of cluster unit becomes concept vector of the corresponding cluster. FCART performs fuzzy clustering in the true sense of the word by applying the fuzzy set theory: it represents the degree of input pattern's membership for each cluster by relative fuzzy membership values(context-sensitive) and determines which of the input pattern is noise or outlier by absolute fuzzy membership value(context-insensitive). Also it updates not the weights of the cluster that is the most similar to input pattern(WTA strategy), but also the weights of every cluster according to the relative fuzzy membership values(soft-competitive learning). Therefore, the document which contains many topics can belong to many clusters. Now, Fuzzy Concept ART are now presented in details:

Initialization. The number of cluster,c, is initialized to be one. Input patterns are normalized to have unit L_2 norm. And initial weight vector is initialized to be the first input pattern:

$$\mathbf{w_1^{(0)}} = \mathbf{x_1} \tag{3}$$

Since the matching degree between input pattern and weight vector is measured by cosine similarity, FCART guarantees that the first pattern is assigned the first category without regard to the value of vigilance variable.

Activation Function(AF). The activation function is defined as the relative fuzzy membership function:

$$AF(\mathbf{w_j^{(t)}}, \mathbf{x_i}) = R_{ij} = \frac{A_{ij}}{\sum_{h=1}^{c} A_{ih}} \tag{4}$$

where the absolute fuzzy membership function, A_{ij}, is defined as the cosine similarity between the input pattern and the weight vector:

$$A_{ij} = \cos(\theta(\mathbf{w_j^{(t)}}, (\mathbf{x_i}))) = \mathbf{x_i} \cdot \frac{\mathbf{w_j^{(t)}}}{||\mathbf{w_j^{(t)}}||} \tag{5}$$

The weight vector of cluster π_j, $\mathbf{w_j^{(t)}}$, is defined as sum of the input patterns which are classified to cluster π_j:

$$\mathbf{w_j^{(t)}} = \sum_{x_i \in \pi_j} \mathbf{x_i} \tag{6}$$

Then the concept vector of cluster π_j, $\mathbf{c_j^{(t)}}$, is defined as follows

$$\mathbf{c_j^{(t)}} = \frac{\mathbf{m_j^{(t)}}}{||\mathbf{m_j^{(t)}}||} = \frac{\mathbf{w_j^{(t)}}}{||\mathbf{w_j^{(t)}}||} \tag{7}$$

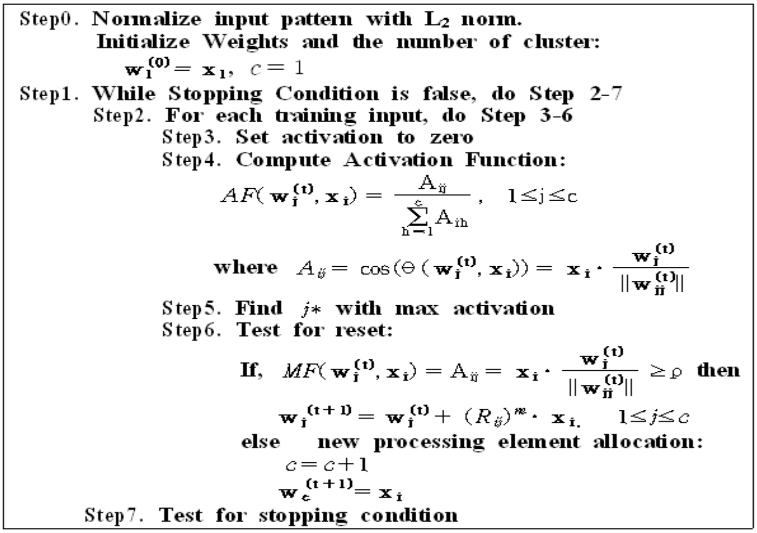
where the mean vectors, m_j, contained in the cluster π_j is

$$\mathbf{m_j} = \frac{1}{n_j} \sum_{x \in \pi_j} \mathbf{x} \tag{8}$$

where n_j is the number of document vectors in π_j. Note that the mean vector $\mathbf{m_j}$ need not have a unit norm. In FCART, the concept vectors are computed by normalizing the corresponding weight vectors to have unit norm without computing the mean vectors.

Matching Function(MF). The matching function which is applied to vigilance test is defined as the absolute fuzzy membership function:

$$MF(\mathbf{w_j^{(t)}}, \mathbf{x_i}) = A_{ij} \tag{9}$$



```
Step0. Normalize input pattern with L2 norm.
       Initialize Weights and the number of cluster:
          $\mathbf{w}_1^{(0)} = \mathbf{x}_1, \; c = 1$
Step1. While Stopping Condition is false, do Step 2-7
       Step2. For each training input, do Step 3-6
              Step3. Set activation to zero
              Step4. Compute Activation Function:
                     $AF(\mathbf{w}_j^{(t)}, \mathbf{x}_i) = \frac{A_{ij}}{\sum_{h=1}^{c} A_{ih}}, \quad 1 \leq j \leq c$
                     where $A_{ij} = \cos(\Theta(\mathbf{w}_j^{(t)}, \mathbf{x}_i)) = \mathbf{x}_i \cdot \frac{\mathbf{w}_j^{(t)}}{\|\mathbf{w}_{ji}^{(t)}\|}$
              Step5. Find $j*$ with max activation
              Step6. Test for reset:
                     If, $MF(\mathbf{w}_j^{(t)}, \mathbf{x}_i) = A_{ij} = \mathbf{x}_i \cdot \frac{\mathbf{w}_j^{(t)}}{\|\mathbf{w}_{ji}^{(t)}\|} \geq \rho$ then
                        $\mathbf{w}_j^{(t+1)} = \mathbf{w}_j^{(t)} + (R_{ij})^m \cdot \mathbf{x}_i. \quad 1 \leq j \leq c$
                     else new processing element allocation:
                        $c = c + 1$
                        $\mathbf{w}_c^{(t+1)} = \mathbf{x}_i$
       Step7. Test for stopping condition
```

Fig. 1. FCART Algorithm

From the above definitions, the activation function and the matching function are satisfied with the following condition [7].

$$MF(\mathbf{w_1}, \mathbf{x_i}) > MF(\mathbf{w_2}, \mathbf{x_i}) \Leftrightarrow AF(\mathbf{w_1}, \mathbf{x_i}) > AF(\mathbf{w_2}, \mathbf{x_i}) \tag{10}$$

That is, when best-matching template $\mathbf{w}_{\mathbf{j}*}^{(\mathbf{t})}$, selected according to $AF(\mathbf{w}_{\mathbf{j}}^{(\mathbf{t})}, \mathbf{x_i})$, does not satisfy the vigilance criterion, a new processing unit can be immediately allocated to match the input pattern $\mathbf{x_i}$. And the corresponding input pattern is assigned to weight vector of the new cluster. This means that no mismatch reset condition and search process are required to detect the resonance domain. Hence, in speed, FCART has an additional advantage for post-clustering.

Detection of resonance unit. To select resonance unit, the vigilance test is

$$MF(\mathbf{w}_{\mathbf{j}*}^{(\mathbf{t})}, \mathbf{x_i}) = \rho \tag{11}$$

where the best-matching cluster $j*$ is $\arg\max_{j=1,\cdots,c}\{AF(\mathbf{w}_{\mathbf{j}}^{(\mathbf{t})}, \mathbf{x_i})\}$.

That is, the value of activation function means the degree of the current input pattern's credit for the corresponding cluster and the value of matching function determines whether the input pattern is an outlier for the cluster or not.

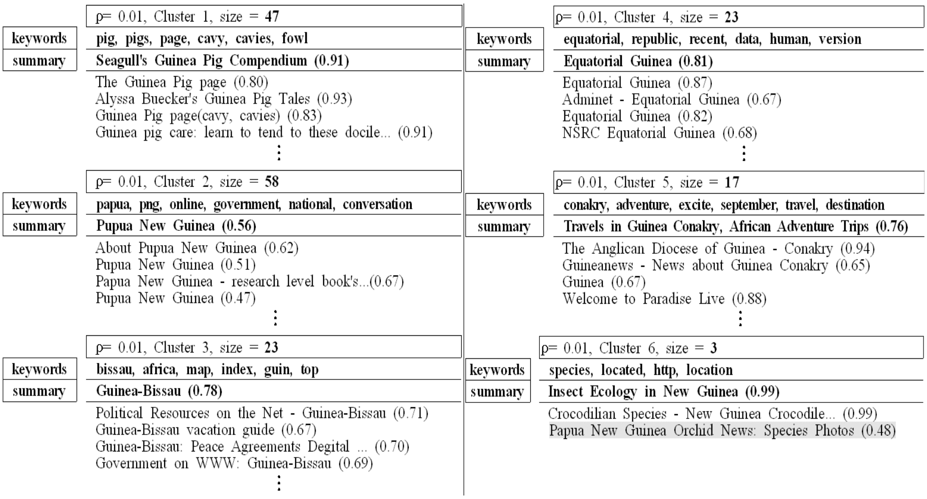
Updating Weights. In FCART, input patterns have the degree of the membership for each cluster by equation (4). So, when the weights are updated, input patterns have effect on the weight vectors of each cluster according to the relative membership value

$$\mathbf{w}_{\mathbf{j}}^{(\mathbf{t+1})} = \mathbf{w}_{\mathbf{j}}^{(\mathbf{t})} + (R_{ij})^m \cdot \mathbf{x_i}, \quad 1 \leq j \leq c \tag{12}$$

where $m \in (1, \infty)$ is weighting exponent for the degree of membership. By equation (12), FCART updates not only the weight of the best-matching cluster but also the weights of any other clusters. The algorithm of the FCART is summarized at fig.1

4 Experimental Results

We sampled the title and snippet of 185 high-ranked documents which are returned by Web search engine, Google, for a query "guinea". First, we removed



ρ= 0.01, Cluster 1, size = **47**

keywords: **pig, pigs, page, cavy, cavies, fowl**

summary: **Seagull's Guinea Pig Compendium (0.91)**

The Guinea Pig page (0.80)
Alyssa Buecker's Guinea Pig Tales (0.93)
Guinea Pig page(cavy, cavies) (0.83)
Guinea pig care: learn to tend to these docile... (0.91)
⋮

ρ= 0.01, Cluster 2, size = **58**

keywords: **papua, png, online, government, national, conversation**

summary: **Pupua New Guinea (0.56)**

About Pupua New Guinea (0.62)
Pupua New Guinea (0.51)
Papua New Guinea - research level book's...(0.67)
Pupua New Guinea (0.47)
⋮

ρ= 0.01, Cluster 3, size = **23**

keywords: **bissau, africa, map, index, guin, top**

summary: **Guinea-Bissau (0.78)**

Political Resources on the Net - Guinea-Bissau (0.71)
Guinea-Bissau vacation guide (0.67)
Guinea-Bissau: Peace Agreements Degital ... (0.70)
Government on WWW: Guinea-Bissau (0.69)
⋮

ρ= 0.01, Cluster 4, size = **23**

keywords: **equatorial, republic, recent, data, human, version**

summary: **Equatorial Guinea (0.81)**

Equatorial Guinea (0.87)
Adminet - Equatorial Guinea (0.67)
Equatorial Guinea (0.82)
NSRC Equatorial Guinea (0.68)
⋮

ρ= 0.01, Cluster 5, size = **17**

keywords: **conakry, adventure, excite, september, travel, destination**

summary: **Travels in Guinea Conakry, African Adventure Trips (0.76)**

The Anglican Diocese of Guinea - Conakry (0.94)
Guineanews - News about Guinea Conakry (0.65)
Guinea (0.67)
Welcome to Paradise Live (0.88)
⋮

ρ= 0.01, Cluster 6, size = **3**

keywords: **species, located, http, location**

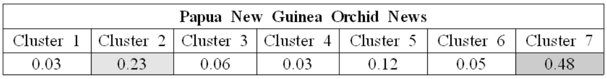
summary: **Insect Ecology in New Guinea (0.99)**

Crocodilian Species - New Guinea Crocodile... (0.99)
Papua New Guinea Orchid News: Species Photos (0.48)

Fig. 2. Clustering result of FCART(= 0.01)

HTML tags, then we eliminated non-content-bearing stopwords and terms which occurred in less than 2 documents. The documents set, GUINEA consists of 159 dimensional document vectors which are sparse (96 % sparsity). To validate the performance of FCART, we cluster the GUINEA and fig.2 gives the result of clustering. In this experiment, we set vigilance parameter to $0.01(\rho = 0.01)$ and weighting exponent to $2(\ m = 2\ )$. The value lying next to the document's title means the relative fuzzy membership value for the corresponding cluster.

As far as the incrementality is concerned, whenever a new pattern comes in, FCART can learn the pattern without relearning the entire system. Especially FCART can perform fuzzy clustering. For example, the document, "Papua New Guinea Orchid News" which is member of the Cluster 6 in fig.2, belongs to not only Cluster 6 (for species) but also Cluster 2 (for Papua New Guinea). fig.3 gives the relative fuzzy membership values of that document for each cluster.



Papua New Guinea Orchid News

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
0.03	0.23	0.06	0.03	0.12	0.05	0.48

Fig. 3. Fuzzy membership for document “Papua New Guinea Orchid News”

5 Conclusions

In this paper, we have proposed a novel clustering algorithm satisfying all of requirements for the post-clustering algorithms. Particularly, the proposed FCART is the form of combining the concept vector that have some advantages in document clustering with Fuzzy ART known as real-time clustering algorithms. Besides, we expect that it may be the alternative model to circumvent some drawbacks of Fuzzy ART such as sensitivity to the order of input sequence, time-complexity, and true meaning of fuzzy set theory [8].

References

1. A. Leouski and W. B. Croft, “An Evaluation of Techniques for Clustering Search Results”, Technical Report IR-76, University of Massachusetts at Amherst, 1996.
2. S. Osinski, J. Stefanowski, and D. Weiss, “Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition”, avalilable at www.cs.put.poznan.pl/dweiss/site/publications/slides/iipwm2004-dweiss-lingo.pdf, 2004.
3. M. A. Hearst and J. O. Pedersen, “Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results”, Proceedings of ACM SIGIR ’96, pp. 76-84, 1996.
4. O. Zamir and O. Etzioni, “Web Document Clustering: A Feasibility Demonstration”, Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 46-54, 1998.
5. I. S.Dhillon and D. S. Modha, “Concept Decomposition for Large Sparse Text Data using Clustering”, Technical Report RJ 10147(9502), IBM Almaden Research Center, 1999.
6. G. A. Carpenter, S. Grossburg, and D. B. Rosen, “Fuzzy ART: An Adaptive Resonance Algorithm for Rapid, Stable Classification of Analog Patterns”, Proceedings of International Conference on Neural Networks, Vol. II, pp. 411-416, 1991.
7. A. Baraldi and E. Alpaydin, “Simplified ART: A New Class of ART Algorithms”, International Computer Science Institute, TR 98-004, 1998.
8. Y. H, Im, “Fuzzy Concept ART: Post-Clustering Algorithm for Web information Retrieval”, M.Sc. Thesis, University of Korea, Korea, 2001.