Selecting and Evaluating Hierarchical Cluster Representative

Yaser A. M. Hasan, Muhammad A. Hassan♣, and M. J. Ridley

Y.A.M.Hasan@bradford.ac.uk, mohdzita@zpu.edu.jo, M.J.Ridley@Bradford.ac.uk

University of Bradford, West Yorkshire, BD7 1DP, UK

♣ Zarqa Private University, P. O. Box 2000, Zarqa 13110, Jordan

***Abstract* – Cluster retrieval was proposed to improve retrieval efficiency, since user needs are compared with a cluster representative; or centroid, instead of all documents. It is important to select the centroid in a way that strongly represents the semantics of the cluster members. In this paper we proposed a method to form the centroid in case of hierarchical clustering is used, it depends on index terms of the parent documents in the hierarchy, combining these terms into a virtual document vector of entries composed of the accumulated weight of each index term. The centroids were evaluated by using two variables; distance to other centroids, and connectivity with the same cluster member, it proved efficiency even when using a subset of the top most important terms to represent the centroid; called top-n% of term.**

***Index Terms* — Information Retrieval, Cluster Retrieval, Cluster Representative**

I. INTRODUCTION

Clusters have been used for grouping documents of relevant semantics in order to enhance retrieval efficiency; either by returning documents that are relevant to the users needs, or present results in a way that the user can get quickly to his/her needs (browsing). Clustering proved efficient since documents that are belonging to a cluster most likely to be relevant to a user request, which was declared by Rijsbergen as the cluster hypothesis[1]. Some other researchers extend this hypothesis to the opposite direction; i.e. documents which are relevant to a request are most likely to belong to the same cluster. [2]

In cluster retrieval user requests, are compared to a representative of a cluster instead of comparing them to all documents in a collection.[1, 3, 4] So the representative will have a crucial effect on efficiency and should be more accurate and carefully selected to reflect the characteristics of the cluster it represents.

In this paper we proposed a method of selecting cluster representative for clusters having hierarchical structure, in which each cluster is represented by a node (or a root document) in the zero level of the hierarchy. And the other documents are represented by child nodes in higher levels. The relationship between the root document and documents in higher levels is either direct relation (having similarity greater than the initial threshold), or *incremental transitive* relation, defined in section II, where a document in level (L) has similarity to its parent equals the initial threshold (δ) plus some increment (ε) depends on L.

Clustering method has great influence on representative selection, next is a brief survey of what effect the clustering method have on representative selecting and creating.

In case of partitioning clustering (as in the case with k-mains and its variants) the centroid is calculated as the statistical average of the objects vectors, dependent on the number of clusters which is predefined, and the initial location selected for these centroids (or means).[5-7].

In the case of agglomerative clustering, there are two approaches depending on the goal of clustering: static or incremental [8].

Incremental methods[7, 9] use an arbitrary document vector as a seed (or representative) of a cluster, if a new coming document is not relevant to this seed, it will be considered as a new seed of a new cluster; such as in the case of STC (Suffix Text Clustering).[8, 10] Browsing retrieval results using scatter/gather[3], also, considered cluster representative as a vector of terms represent some topic as a vector of topical entries, users can browse using topical terms, and cluster summaries, or *tags* as in[11]. Clusters created by agglomerative methods were represented by a wide range of centroid choices:

- Maximally-*linked document*,[1] where the centroid is "that document which is linked to the maximum number of other documents in the cluster". The centroid is not unique. The centroid could be a vector of topics in some context or prototypical document[12, 13].
- The centroid as the Center of gravity[1]; divide the sum of all normalized vectors in a cluster over the number of vectors (documents) in that cluster. Or the average of the weights in all documents vectors[7], top-n ranked documents in a cluster could be used as a representative.[2, 14]
- It could be the maximal predictor of the cluster[1], where a term is considered as a member of the representative if it occurs in at least half of the documents belonging to the cluster. The ratio selected as a threshold will affect the widening or shortening of context represented by such a cluster.[12] For that it is recommended not to choose terms that have very low or very high document frequency for context definition.[13] but it is known from information theory that the information content "entropy" related to the logarithm of the document frequency, so high document frequency terms have higher entropy values[12], and so it is important to make some kind of

tradeoff when using document frequency as a criterion for term selection into a cluster centroid, in this paper we included both frequent terms, and high weighted terms.

We can conclude the following general principles of selecting cluster centroid:

- Generally, terms, topics, or categories used to represent both documents and centroids of clusters were driven from the same set of documents.
- The way in which clusters are created affects the centroid selection, which means that the centroid should be consistent with the clustering method.
- Cluster centroid should support the objective of clustering; if the objective is browsing then the centroid is built as a hierarchy of topical terms or summaries.
- The centroid is selected to distinguish one cluster from others, so the distance between the centroid and other objects in the cluster should be minimized, while the distance to other clusters should be maximized.
- The centroid should be selected in a way minimizes the effect of similarity measure, and/or the indexing method.
- In a peer-to-peer environment, clusters belonging to a node should reflect the topics it includes, and so it is more useful to include high frequency terms, as well as highly weighted terms, in the cluster representative, similar condition presented in [13] for text categorization.

Throughout the paper, cluster representative is referred to as: representative vector, centroid, or virtual document.

The rest of the paper is organized as follows; section II introduces the clustering method by which clusters were built, section III presents the proposed method for selecting a cluster centroid, section IV gives details about how to evaluate the selected centroid, and conclusions are in section V.

II. INCREMENTAL TRANSITIV CLUSTERING

Definition: Incremental transitive relevance♠:

Given a collection of normalized vectors od documents (D), and the relevance relation ($\Re$) defined on D, such that $\Re$={(x,y): sim(x, y) ≥ δ, ∀ documents x, y ∈ D} (δ) is the relevance threshold, and sim(x,y) is the similarity measure. Let d1, and d3∈,and sim(d1,d3)<δ, then d1 and d3 can have incremental transitive relevance if there exists d2∈ D such that sim(d1, d2) ≥ δ, and sim(d2, d3) ≥δ+ε, where ε is a positive real number, $\Re$ is the incremental transitive relevance relation.

Incremental transitivity proposes increments of threshold value to get a sequence: $\delta i=\delta°+i.\varepsilon$ (1)

As translating to higher levels of transitivity, where δ° is the initial relevance threshold.

Expanding this definition to d1, d2, … ,dn , of n-1 incremental transitive relevance levels, then any document vector di ∈ D, i = 1 ,…, n , that has direct relevance with di-1 can has transitive relevance with d1 if sim(d1, di-1) ≥$\delta°+(^{l}$ -1)×ε. Documents in higher levels do not have direct similarity to the root document but it has incremental transitive relevance through its parent node(s), and so the hierarchy is called incremental transitive hierarchy.

III. BUILDING THE CENTROID

Our centroid is consistent with the incremental transitivity clustering method, that it should reflect the contents of the transitive hierarchical structure; as in fig.1 Each parent node (or that has children) represents a document all of whose children are relevant documents with similarity greater than or equal to the threshold plus the product of the increment and the child level. Note however that those children don't have direct relevance to the zero-level node.

So the cluster representative should include those terms of the node in the zero-level which represent the topmost topic, and the terms belong to parents in higher levels where those terms do not belong to the top most parents in the zero level of the hierarchy.

Formally, let C_i is a cluster whose root document is d_i, and d_i is represented by the vector: $d_i=(w_{\Delta(1)}, w_{\Delta(2)}, \dots , w_{\Delta(n)})$, $\Delta(j), j=1, .. , n$ is the index of the jth none zero entry in the vector d_i, and $w_{\Delta(j)}$ is the weight of the index term $t_{\Delta(j)}$.

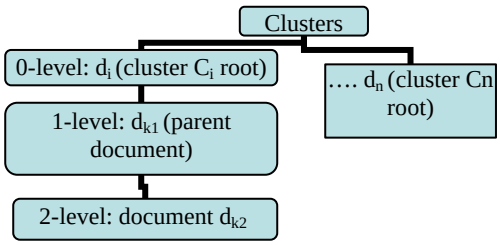

Fig.1. Incremental transitive cluster hierarchy

And d_k is a child node of d_i in the cluster hierarchy, d_k it self is a parent to an other set of relevant children that are decided to have relevance to d_i through incremental transitivity, so there must be some terms in d_k that do not exist in d_i; i.e. $\exists\, t_{\Delta(\upsilon)}$ entries in d_k, but do not exist in d_i. If $\Delta(\upsilon)$ is a set of indexes of terms that are elements of d_k, then the set of terms that form the centroid of the cluster C_i is the set union of all sets of terms that have non zero entries in the parent document nodes in all other levels. That is if there are k parent documents in the cluster C_i, with $\Delta(\upsilon_1), \Delta(\upsilon_2), \dots, \Delta(\upsilon_k)$ indexes, then the set of terms from which the centroid vector is created has the following form:

$$\pi(C_i) \subseteq \{t_{\Delta(1)}, \dots , t_{\Delta(n)}\} \cup \{t_{\Delta(\upsilon 1)}\} \cup \{t_{\Delta(\upsilon 2)}\} \cup \dots \cup \{t_{\Delta(\upsilon k)}\} \quad (1)$$

Centroid vector is formed from the *accumulated* weights of terms belonging to $\pi(C_i)$, where the accumulated weight for a term t_k is given by the equation:

♠ A new method proposed by the authors and under the process of publication.

$$w(t_k, C_d) = \sum_{j=1}^{n} w(t_k, d_j) \quad (2)$$

Where $w(t_k,d_j)$ is the weight given to the term t_k in a document d_j, and C_d is the centroid of the cluster whose root is the document d, and n is the number of documents in the C_d that contain the term t_i.

All the terms could be included, or the centroid space could be reduced to a subset of the original terms, in this paper all (100%), half (50%), and one third (33%) of the terms of the representative are tested. Term vectors of parent documents in a cluster hierarchy were combined with each other, adding weights of terms whenever a term was repeated in more documents, to give higher accumulated weights to more frequent terms.

IV. EVALUATING THE CENTROID

Centroids are being formed as *hypothetical*, *virtual* or *prototypical* documents [7, 10, 12, 13]. In this study clusters centroids are created as virtual documents of terms taken form parent documents, these terms are descended sorted according to the following two criteria:

1. Accumulated weight of a term is considered; higher accumulate weighted terms first.
2. Document frequency of that term in all parent documents that form the centroid, higher frequency terms first.

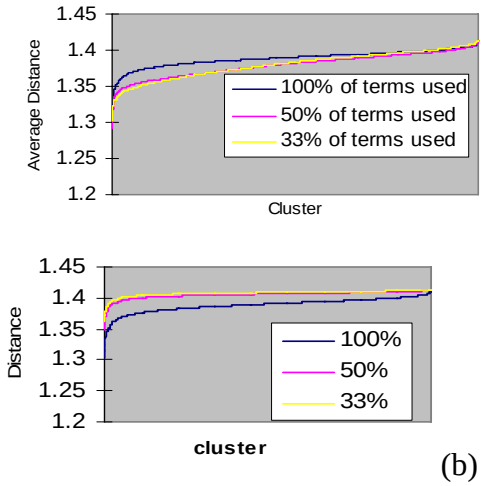

Fig.2. Average distance between centroids, when (a) sorted by frequency (b) sorted by accumulated weights

After descend sorting terms of a representative, top-n% percentage of terms are selected. To evaluate the centroid efficiency, two variables are to be examined:

- The average distance among all cluster centroids, or the *dissimilarity* among clusters centroids [1], and
- The average similarity between the centroid virtual documents and all documents belonging to the same cluster, or the *internal connectivity*.

The first variable is measuring the diversity among clusters, or how efficient is the clustering method (and so the adopted representative selection method) in partitioning the set of documents in a collection. The second variable is measuring the commonalty of a topic among the documents in the same cluster. Consequently, clusters (and their centroids) are better formed when the distance between them is maximized, and the internal connectivity among their documents is maximized.

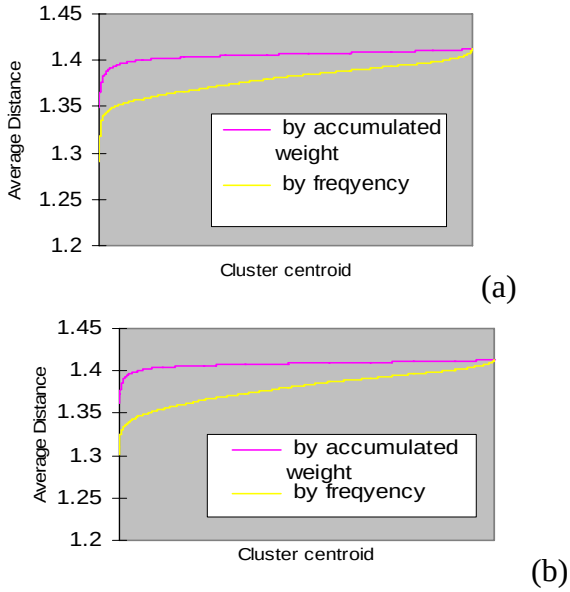

Fig.3. Average distance from cluster centroids, compared for the two sorting criteria. (a) 50% term contribution, (b) 33%

Clusters were formed using an implementation of the incremental transitive clustering algorithm, applied on 18650 documents derived from Reuters21578 collection; all documents are selected to have a title and body text.

A. Average distance: using top-n% terms:

After sorting terms using the above criteria, the experiment was repeated for three selections of top-terms; the first involves all terms (100%) in calculating the distance between centroids, the second involves the top 50% terms, and the third uses the top 33% terms; i.e. a percentage of terms is used instead of a fixed number of terms, since clusters centroid vectors are of different numbers of terms. Each vector is normalized and then the distance is calculated by using the Cartesian distance formula between two points in R^n:

$$d = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \ldots + (y_n - x_n)^2} \quad (3)$$

For each cluster centroid, the distances to all others are calculated, and then the average for each is calculated. In order to compare the results for the above two sorting criteria, and for each top-n percentage, average values are ascending sorted and plotted, as in fig.2.

When frequency is used as the sorting criterion, the distance is larger for 100% term contribution, and smaller

distance for both 50%, and 33%. But when accumulated weight is used as sorting criterion, the distance became larger as the percentage of contributing terms getting smaller. This result can be explained by noting that most frequent terms are used by more documents; i.e. have larger document frequency, and when elected among the top-n percentage terms, documents will share more common terms, consequently, more x_i, and y_i, in the distance equation are non-zeroes, and so $(x_i - y_i)^2$ is less than both x_i^2, and y_i^2, which makes the distance smaller. And so the distance will be greater when contributing higher percentage of terms, because the probability to have specific unshared terms in more cluster representatives becomes larger, so we can find more elements of equation 3 in which either x_i=0, or y_i=0, which means larger values of

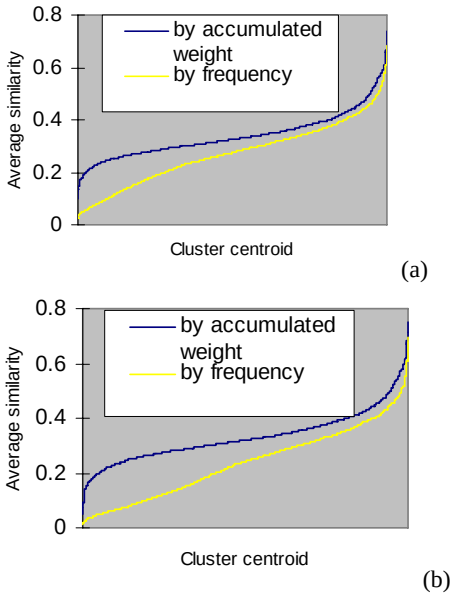

Fig.5. Compare average similarity for three term sorting criteria, when terms contribution is: (a) 50% (b) 33%

$(y_i-x_i)^2$ elements, see fig.2 a.

When using accumulated weight as sorting criterion the distance between centroids become larger as a smaller percentage of terms is involved, since terms that have high accumulated weights in a centroid are more specific to the topic of the cluster than other clusters, and so less probable to co-exist in other clusters, or at least will have very small accumulated weights in other cluster centroids. This result is relevant to the characteristics of specificity, where a "specific term indicates that the term is involved in a topical relationship with a document, with another term, or with a set of documents" [14], and didn't contradict Hideo Joho and Mark Sanderson's belief in [15], because they decided that document frequency is more accurate in determining term specificity in case of "the terms are very specific".

And so accumulated weight, criterion is better at specifying the relationship between terms and topics represented by clusters, this fits with Sebastiani's finding that "terms that have very low or high document frequency as not being informative" [13]. And fits –also- with what Rooney et. al. find that is: the context defined by terms that have high document frequency will be wider. [12]

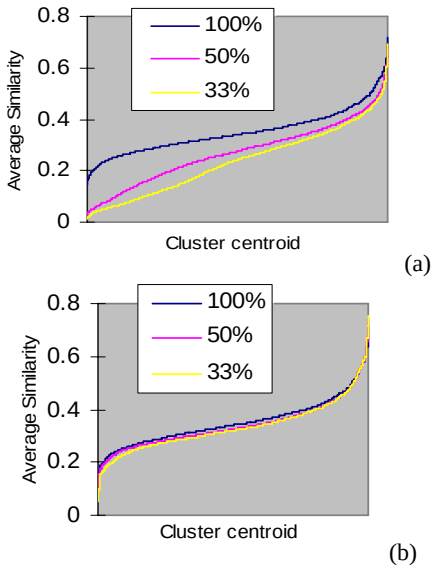

Fig.4. Average similarity between centroid and documents, when: (a) Sorted by Frequency (b) Sorted by Acc. weight

Fig.3 presents a comparison between the two criteria, and for a reduced term contribution percentages.

*B. Average similarity (connectivity) between documents and cluster centroid***:**

The same experiment is repeated for the second variable: the internal similarity (connectivity) between the centroid and the documents of a cluster, where the same sorting criteria of distance measurements are used. The effect of both term selection and space reduction (contributing only the top-n% of terms) on the internal connectivity, between the centroid virtual document vector and cluster documents, is examined by calculating the average cosine similarity between the centroid and the vectors of all documents in the cluster. Similarity values are sorted ascending and then plotted as in fig.4.

When term frequency is used as the sorting criterion, the average similarity gets smaller as the percentage of contributed terms falls, and so space reduction has a major effect on internal connectivity, see fig.4 a. But when using accumulated weight as the sorting criterion, space reduction has a minor effect on the internal connectivity among cluster members, as shown in fig.4 b.

These results can be explained as: when accumulated weight is used to sort terms, frequent terms with small original weights will gain more accumulated weight since weights of a term are added when a it is repeated in more than one document contributing to form the centroid, and so

both mostly frequent, and original high weighted terms are positioned at the top of the list, and so will be selected when reducing the centroid space, which results in minor changes to the average similarity. But when using the frequency as a sorting criterion, the less frequent but originally high weighted terms will not be selected in the top most terms in the list, so when reducing the percentage they will not included.

Fig.5 presents a comparison between the two sorting criteria's effects on similarity for the reduced term involvement percentages, since when using 100% of terms there is no difference because all terms will remain contributed regardless of the sorting criteria. It can be seen from fig.6 that the average similarity distribution is skewed toward the left; i.e. biased to smaller values when using frequency as a sorting criterion. Centroid virtual documents were defined as the combination of parent documents terms in a cluster, so the similarity values between it and other documents in the cluster is considered as an indicator of the representation power of a centroid. In the case using accumulated weight the distribution is slightly shifted to the left with a minor change of the distribution, as shown in fig.7.

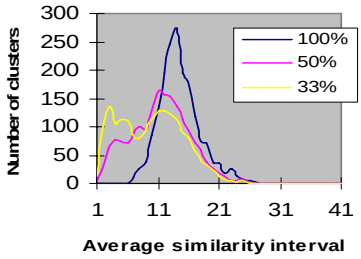

Fig.6. Similarity distribution of frequency as sorting criterion

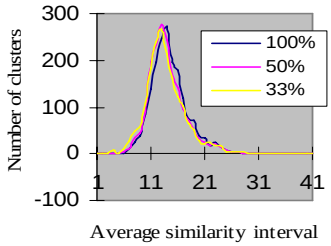

Fig.7. Similarity distribution of acc weight as sorting criterion for 100%, 50%, and 33% of term contribution percentage.

V. CONCLUSION

Terms selected to form the centroid are better selected from parent documents of a cluster hierarchy, since they are relevant to more other documents in the cluster with higher similarity values. Accumulated weight of terms is the preferred criterion when choosing the top-n% terms; i.e. when reducing the space of centroid virtual document. Because when using accumulated weight as a criterion to select contributing terms, both weights and document frequencies are involved which rise the importance of using both when selecting the terms of the centroid.

REFERENCES

[1] C. J. V. Rijsbergen, *Information Retrieval*, second ed., 1979.

[2] S.-H. Na, I.-S. Kang, J.-E. Roh, and J.-H. Lee, "An empirical study of query expansion and cluster-based retrieval in language modeling approach," *Information Processing and Management*, vol. 43, pp. 302-314, 2007.

[3] M. A. Hearst and J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," in *the Nineteenth Annual International ACM SIGIR Conference*, Zurich, June 1996.

[4] A. Gionis, P. Indyk, and R. Motwani, "Similarity Search in High Dimensions via Hashing," *The Very Large Data Bases Journal*, pp. 518-529, 1999.

[5] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967, pp. 281-297.

[6] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Fourth printing ed.: Cambridge University Press, 2005.

[7] J. Allan, A. Leouski, and R. Swan, "Interactive cluster visualization for information retrieval," Department of Computer Science, University of Massachusetts, Amherst, Technical Report IR-116, CIIR, 1996.

[8] E. Greengrass, "Information Retrieval a Survey," 2000.

[9] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental Clustering And Dynamic Information Retrieval," *Society for Industrial and Applied Mathematics, SIAM J. COMPUT.* , vol. 33, pp. 1417-1440, 2004.

[10] O. Zamir and O. Etzoni, "Web Document Clustering: A Feasibility demonstration," in *ACM/SIGIR*, Melbourne, Australia, 1998.

[11] J. Caverlee, L. Liu, and D. Buttler, "Probe, Cluster, and Discover: Focused Extraction of QA-Pagelets from the Deep Web," in *the 20th International Conference on Data Engineering (ICDE'04)*, 2004.

[12] N. Rooney, D. Patterson, M. Galushka, and V. Dobrynin, "A scaleable document clustering approach for large document corpora," *Information Processing and Management*, vol. 42, pp. 1163-1175, 2006.

[13] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, pp. 1-47, March 2002 2002.

[14] G. Kim, "Relationship between index term specificity and relevance judgment," *Information Processing and Management*, vol. 42, pp. 1218-1229, 2006.

[15] H. Joho and M. Sanderson, "Document frequency and term specificity," in *Recherche d'Information Assistée par Ordinateur Conference (RIAO)*, 2007.