A New Test of Cluster Hypothesis Using a Scalable Similarity-Based Agglomerative Hierarchical Clustering Framework

Xinyu Wang, Julien Ah-Pine, Jerome Darmont

A New Test of Cluster Hypothesis Using a Scalable Similarity-Based Agglomerative Hierarchical Clustering Framework

Xinyu Wang* — **Julien Ah-Pine*** — **Jérôme Darmont***

* *Université de Lyon, Lyon 2, ERIC, EA3083*

RÉSUMÉ. *L'hypothèse de cluster est l'hypothèse fondamentale de l'utilisation du clustering dans la recherche d'information. Elle indique que les documents semblables ont tendance à être pertinents pour la même requête. Des travaux passés testent intensivement cette hypothèse avec les méthodes de la classification ascendante hiérarchique (CAH). Mais leurs conclusions ne sont pas cohérentes en termes d'efficacité de la recherche. La limite principale dans ces travaux est le problème de passage à l'échelle lié a là CAH. Dans cet article, nous étendons nos travail précédent à un nouveau test de l'hypothèse de cluster en appliquant un système extensible de CAH basé sur la similarité. Principalement, la matrice de similarité cosinus est sparsifiée par des seuils pour réduire l'occupation mémoire et le temps de calcul. Nos résultats montrent que même quand la matrice est largement sparsifiée, l'efficacité de la recherche est maintenue pour toutes les méthodes, dont le complete et l'average ne dominent pas toujours les autres.*

ABSTRACT. *The Cluster Hypothesis is the fundamental assumption of using clustering in Information Retrieval. It states that similar documents tend to be relevant to the same query. Past research works extensively test this hypothesis using agglomerative hierarchical clustering (AHC) methods. However, their conclusions are not consistent concerning retrieval effectiveness for a given clustering method. The main limit of these works is the scalability issue of AHC. In this paper, we extend our previous work to a new test of the cluster hypothesis by applying a scalable similarity-based AHC framework. Principally, the input pairwise cosine similarity matrix is sparsified by given threshold values to reduce memory usage and running time. Our experiments show that even when the similarity matrix is largely sparsified, retrieval effectiveness is retained for all tested methods. Moreover, for two clustering methods, complete link and average link, they do not always dominate the other methods as reported in past works.*

MOTS-CLÉS : *hypothèse de cluster$_1$, classification ascendante hiérarchique$_2$, efficacité$_3$.*

KEYWORDS: *cluster hypothesis$_1$, agglomerative hierarchical clustering$_2$, effectiveness$_3$.*

1. Introduction

Introduced in (Jardine et van Rijsbergen, 1971), the Cluster Hypothesis states that "the association between documents conveys information about the relevance of documents to requests", i.e., similar documents are expected to be relevant to the same query. In previous research works, this hypothesis has been extensively tested using four agglomerative hierarchical clustering (AHC) methods, they are single link, complete link, average link and Ward's method. However, in terms of retrieval effectiveness, results of these tests are not consistent concerning a certain clustering method. For example, (Griffiths *et al.*, 1984) claims that "average link gave the best results" in his test, while (Willett, 1988) concludes that "compete link is probably the most effective method", and (Griffiths *et al.*, 1997) states that "Ward's method was found to give the best overall results". Besides, computing efficiency is usually out of scope in these works, as complexity of conventional AHC is up to $O(N^3)$, efficiency is an important factor to consider in practice.

In this paper, we extend our previous work (Ah-Pine et Wang, 2016) to a new test of cluster hypothesis by applying a scalable similarity-based AHC framework. Different from our previous work, here we focus on the retrieval effectiveness and efficiency using AHC methods in the context of cluster hypothesis. To serve this objective, our experimental setting and evaluation measure vary. The contributions of this work include : (I) optimal cluster search is applied to compare the retrieval effectiveness of seven AHC methods ; (II) computing efficiency is addressed, and the impact of improving efficiency to effectiveness is discussed. To our knowledge, our work is the first test on cluster hypothesis that addresses the efficiency issue and extends the scope to seven AHC methods.

The outline of this paper is organized as follows : Section 2 introduces the state of the art. The scalable similarity-based AHC framework that we proposed in (Ah-Pine et Wang, 2016) is described in Section 3. Section 4 details our experiment and presents obtained results. Conclusion and future work is in Section 5.

2. State of the Art

2.1. *Cluster Hypothesis Tests and Optimal Cluster Search*

Research works on cluster hypothesis tests are largely inspired by the *overlap test* (Jardine et van Rijsbergen, 1971), the *nearest neighbor test* (Voorhees, 1985) and the *density test* (El-Hamdouchi et Willett, 1987). The objective of these tests is to verify whether or not the cluster hypothesis characterizes a document collection. (Jardine et van Rijsbergen, 1971) conduct the test by measuring the overlap between distributions of similarities of RR pairs and RN pairs of documents (R :relevant, N :non-relevant). A collection with a low overlap value is believed to cluster strongly for a set of queries, and thus better separated from the non-relevant documents. Conclusion of this test claims that cluster search has potential to outperform the inverted file system. Ho-

wever, this test is only experimented on one dataset. Besides, as in a collection most documents are non-relevant, RN pairs dominate RR pairs, result based on the number of RN and RR pairs is biased. (Voorhees, 1985) questions the overlap test and provides an alternative. This work assumes that if a document is relevant to a query, then the document's k nearest neighbors contain relevant documents as well. With k set between $[0, 5]$ and four tested datasets, this test concludes that cluster hypothesis holds for one collection but not for another. Limitation of this test is that it is dependent on k, setting k to different values may alter the conclusion. (El-Hamdouchi et Willett, 1987) apply the density test to eliminate impact of external parameter on the experiment. This test is empirically demonstrated to be more correlated than the overlap test and the nearest-neighbor test, in proving that cluster-based retrieval outperforms document-based retrieval.

Apart from verifying cluster hypothesis, other tests are carried out to compare retrieval effectiveness using different clustering techniques, among which, four hierarchical clustering methods, single link, complete link, average link and Ward's method are widely experimented (Griffiths *et al.*, 1984) (Willett, 1988) (Griffiths *et al.*, 1997). As stated in Section 1, these works result in different conclusions. (Tombros *et al.*, 2002) agrees with the conclusion in (Griffiths *et al.*, 1984) that average link is found to be the most effective, however unlike other works, it computes retrieval effectiveness with partial but not the entire dendrogam of document clusters.

Another branch of tests focus on comparison of cluster-based searching strategies using hierarchical clustering, such as bottom-up search and top-down search (Van Rijsbergen et Croft, 1975) (Willett, 1988). Both searching strategies require to compute the similarity between a query and a cluster representative, mostly cluster centroid is used as the representative of a cluster. These works suggest that bottom-up search results in higher retrieval effectiveness than top-down search. Different from the two searching strategies, **optimal cluster search** (Jardine et van Rijsbergen, 1971) does not involve actual matching between query and cluster centroids. It searches for the optimal cluster for a given query by scanning the whole hierarchies of the dendrogram. The advantage of this searching strategy is that, it directly concerns the internal connections of hierarchies when computing the retrieval effectiveness for a query, and it eliminates any bias brought from external sources to the document hierarchy.

2.2. *Agglomerative Hierarchical Clustering and Lance-Williams Formula*

Compared to flat clustering, hierarchical clustering is more informative as it thoroughly reveals internal connections among clusters. This property is favorable in many text mining tasks. Hierarchical clustering family has two principle members, the agglomerative (AHC) and the divisive (DHC). DHC is a top-bottom process that begins with splitting a given dataset and results in clusters of which each is a data instance. AHC works in the opposite way. For DHC, if the dataset is large enough, finding an optimal split can be NP-hard. However, the worst complexity of AHC is $O(N^3)$.

There are many works that improve the performance of AHC (Xu *et al.*, 2005). In this paper, our scope is the conventional AHC and its methods.

Given a dataset of N instances, the general procedure of conventional AHC is shown as in Algorithm 1, the output is a binary tree-like structure, named dendrogram.

Algorithm 1: General procedure of AHC

Data: pairwise distance matrix $\mathbb{D}$ of a dataset
Initialize a dendrogram of N leaves with null height values;
while *number of clusters* > 1 **do**
1. $(C_i, C_j) = argmin D(C_x, C_y)$, i.e. search for the minimal distance in $\mathbb{D}$ and the pair of clusters (C_i, C_j);
2. Merge C_i and C_j into C_{ij} and add a corresponding parent node in the dendrogram with height value $D(C_i, C_j)$;
3. Compute distances between C_{ij} and other clusters C_k, and update $\mathbb{D}$ accordingly.

end
Result: a dendrogram of $2N - 1$ nodes

In the Vector Space Model and when TF-IDF weighting system is applied, documents can be projected into the feature space where the projection is sphere-like. In this case Euclidean distance is a suitable choice to determine the dissimilarity of two document vectors, $D(x, y)$. In AHC, an essential step is to compute $D(C_{ij}, C_k)$ after cluster C_i and cluster C_j are merged. There are seven methods to determine $D(C_{ij}, C_k)$ by using either clustering centroids or graphic representations of clusters.

(Lance et Williams, 1967) formulate the seven conventional clustering methods in a unified framework and name it the Lance-Williams (LW) formula. Equation 1 and Table 1 display the formula and its parameter values. Thanks to their work, one can simply switch parameter values to compute $D(C_{ij}, C_k)$ using a certain method.

$$\begin{aligned} D(C_{ij}, C_k) =& \alpha_i D(C_i, C_k) + \alpha_j D(C_j, C_k) + \beta D(C_i, C_j) \\ &+ \gamma |D(C_i, C_k) - D(C_j, C_k)| \end{aligned} \quad [1]$$

3. A Scalable Similarity-based AHC Framework

In (Ah-Pine et Wang, 2016) we establish a new expression of the LW formula by replacing distances with similarities, we use “Sim-AHC" to refer this approach. Sim-AHC is mathematically and experimentally proved to be equivalent to its counterpart. Input of Sim-AHC is the pairwise similarity matrix $\mathbb{S}$ of a normalized dataset, for example, $\mathbb{S}$ can be a cosine similarity matrix. Having all values in $\mathbb{S}$ between $[0, 1]$ permits us to sparsify $\mathbb{S}$ by a threshold value, so that less memory is required to store $\mathbb{S}$ and less time is required to compute on it. This is an advantageous property of Sim-

Methods	α_i	α_j	β	γ
Single	1/2	1/2	0	-1/2
Complete	1/2	1/2	0	1/2
Average	$\frac{\|C_i\|}{\|C_i\|+\|C_j\|}$	$\frac{\|C_j\|}{\|C_i\|+\|C_j\|}$	0	0
Mcquitty	1/2	1/2	0	0
Centroid	$\frac{\|C_i\|}{\|C_i\|+\|C_j\|}$	$\frac{\|C_j\|}{\|C_i\|+\|C_j\|}$	$-\frac{\|C_i\|\|C_j\|}{(\|C_i\|+\|C_j\|)^2}$	0
Median	1/2	1/2	-1/4	0
Ward	$\frac{\|C_i\|+\|C_k\|}{\|C_i\|+\|C_j\|+\|C_k\|}$	$\frac{\|C_j\|+\|C_k\|}{\|C_i\|+\|C_j\|+\|C_k\|}$	$-\frac{\|C_k\|}{\|C_i\|+\|C_j\|+\|C_k\|}$	0

Table 1. *Lance-Williams formula : methods and parameter values*

AHC. However, this property cannot be found in the conventional AHC that uses distances, because in this setting, zero and close-to-zero values signify high similarity of two data instances, which is of the most interest in clustering, thus these values have to be stored for computation instead of being ignored. This is a major difference between Sim-AHC and the conventional AHC. More importantly, Sim-AHC is capable to return the same or better clustering results even when $\mathbb{S}$ is largely sparsified. This characteristic offers Sim-AHC computing efficiency of substantially reduced memory usage and running time. To this extend, Sim-AHC is superior to the conventional AHC. To our knowledge, Sim-AHC is the first framework that complements the conventional AHC methods formulated by the LW formula, with advantageous characteristics.

As we know, the squared form of the Euclidean distance can be expressed by a linear combination of inner products, $D^2_{Euclidean} =< x, x > + < y, y > -2 < x, y >$, where x and y are two data vectors and $<, >$ represents the inner product. With this connection, we assume that the similarity between x and y is defined by the inner product of their normalized form, shown as Equation 2. As $S(x, x) = S(y, y) = 1$ holds, the corresponding dissimilarity $D(x, y)$ can be expressed as in Equation 3. Sim-AHC amounts to working with $-\frac{1}{2}D(x, y)$ instead of $D(x, y)$. Due to the length limitation of this paper, we do not detail the mathematical deduction of Sim-AHC here. Interested readers can refer to (Ah-Pine et Wang, 2016) for more details. Algorithm 2 illustrates the computing procedure of Sim-AHC.

$$S(x,y) =< \frac{x}{||x||}, \frac{y}{||y||} > \quad [2]$$

$$\begin{aligned} D(x,y) &= ||\frac{x}{||x||} - \frac{y}{||y||}||^2 \\ &= S(x,x) + S(y,y) - 2S(x,y) \\ &= 2[1 - S(x,y)] \end{aligned} \quad [3]$$

Algorithm 2: Computing procedure of Sim-AHC

Data: pairwise similarity matrix $\mathbb{S}$ of a normalized dataset with N instances
Initialize a dendrogram of N leaves with null height values;
while *number of clusters* > 1 **do**

1. $(C_i, C_j) = argmax[S(C_x, C_y) - \frac{1}{2}(S(C_x, C_x) + S(C_y, C_y))]$, i.e. search for the maximal similarity in $\mathbb{S}$ and the pair of clusters (C_i, C_j);
2. Merge C_i and C_j into C_{ij} and add a corresponding parent node in the dendrogram with height value $[S(C_i, C_j) - \frac{1}{2}(S(C_i, C_i) + S(C_j, C_j))]$;
3. Update $\mathbb{S}$ by computing similarity of $S(C_{ij}, C_k)$ and the self similarity $S(C_{ij}, C_{ij})$:

$$\begin{aligned} S(C_{ij}, C_k) &= \alpha_i S(C_i, C_k) + \alpha_j S(C_j, C_k) + \beta S(C_i, C_j) \\ &\quad - \gamma |S(C_i, C_k) - S(C_j, C_k)| \\ S(C_{ij}, C_{ij}) &= \delta_i S(C_i, C_i) + \delta_j S(C_j, C_j) \end{aligned}$$

end
Result: a dendrogram of $2N - 1$ nodes

Sim-AHC keeps the values of parameters α_i, α_j, β and γ unchanged as in Table 1. To guarantee the equivalence for each individual clustering method, the values of newly added parameters δ_i and δ_j are set differently : $\delta_i = \delta_j = \frac{1}{4}$ for median method ; $\delta_i = \frac{|C_i|^2}{(|C_i|+|C_j|)^2}$, $\delta_j = \frac{|C_j|^2}{(|C_i|+|C_j|)^2}$ for centroid method ; and for the other five methods, values of δ_i and δ_j can be determined freely as long as their sum is 1.

4. Experiments and Results

Our experiments apply Sim-AHC to verify : (I) if retrieval effectiveness comprises when efficiency is improved by sparsifying the similarity matrix, and (II) which clustering method outperforms the others in this setting. Four classic medium-sized datasets described in Table 2 are used in our tests. These datasets have been experimented in previous tests (Voorhees, 1985) (Tombros *et al.*, 2002), they contain a complete query set and a relevance judgment file, which allow us to evaluate our results. In preprocessing, each dataset is converted into a document-term matrix using the TF-IDF weighting scheme. Stop words are removed and terms are stemmed by Porter Stemmer. A rough feature selection is applied by removing terms that appear in less than 0.2% and more than 95% documents in each dataset, and each document vector is normalized by L^2-norm.

We apply E-measure (Jardine et van Rijsbergen, 1971) to evaluate the retrieval effectiveness. Expressed as $E = \frac{(\beta^2+1)PR}{\beta^2 P+R}$, smaller E value indicates better retrieval effectiveness. P and R represent precision and recall, respectively. β is the parameter that balances the importance between precision and recall, it takes values of 0.5, 1.0 and 2.0. Given a clustering method, Sim-AHC takes a preprocessed document-term

Datasets	MED	CISI	CACM	LISA
Num of docs	1033	1460	3204	6004
Num of terms after preprocessing	4389	2495	1463	2052
Num of queries	30	76	52	35
Ave num of docs per query	23.2	41.0	15.3	10.8

Table 2. *Description of datasets*

matrix as input, computes its pairwise cosine similarity matrix $\mathbb{S}$ and outputs a dendrogram. In order to test the impact of improving efficiency on retrieval effectiveness, we choose five threshold values at 10th, 25th, 50th, 75th and 90th percentiles of the distribution of similarities in $\mathbb{S}$. Table 3 shows the threshold values at these percentiles for tested datasets.

Dataset/ Percentile	10%	25%	50%	75%	90%
MED	0.0047	0.0094	0.0188	0.0351	0.0618
CACM	0.0111	0.0215	0.0396	0.0692	0.1240
CISI	0.0100	0.0200	0.0381	0.0659	0.1024
LISA	0.0092	0.0180	0.0353	0.0626	0.0981

Table 3. *Threshold values at 10th, 25th, 50th, 75th and 90th percentiles*

In each dataset, for each clustering method, we apply Sim-AHC and sparsify the cosine similarity matrix by the threshold values. After a dendrogram is obtained, we flatten it and apply optimal cluster search for each query. Given a query Q, E value at each level of the flattened dendrogram is computed. The optimal cluster is the one that has the minimal E value, and this E value is recorded for Q. In order to reflect the overall retrieval effectiveness of a clustering method, optimal E values for all queries in a dataset are averaged.

Comparison of seven clustering methods. Table 4 displays the averaged optimal E values for each clustering method using three β values obtained from the full-sized dendrograms of each dataset, no filtering strategy is applied to sparsify the cosine similarity matrices. Values highlighted in bold are column-wise minimums, indicating the best retrieval effectiveness among seven clustering methods. This result partially assents to previous finding that complete link and average link outperform single link and Ward's method. However, this only holds true for CACM and CISI datasets. It is obvious that Ward's method performs the best for LISA dataset. Besides, Mcquitty outperforms the other methods in some cases. This experiment demonstrates that complete link and average link do not always dominate the other methods as reported in previous works.

Dataset	MED			CACM		
Method	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
Single	0.791	0.876	0.869	0.743	0.819	0.846
Complete	**0.755**	**0.849**	0.857	**0.714**	**0.793**	**0.807**
Average	0.771	0.855	0.851	0.745	0.816	0.830
Mcquitty	0.768	0.851	**0.850**	0.742	0.812	0.828
Centroid	0.847	0.913	0.885	0.759	0.837	0.864
Median	0.813	0.894	0.881	0.756	0.834	0.863
Ward	0.772	0.857	0.855	0.730	0.805	0.822

Dataset	CISI			LISA		
Method	$\beta = 0.5$	$\beta = 1$	$\beta = 2$	$\beta = 0.5$	$\beta = 1$	$\beta = 2$
Single	0.821	0.881	0.849	0.795	0.859	0.880
Complete	0.801	0.856	0.819	0.725	0.776	0.776
Average	**0.799**	**0.853**	**0.817**	0.735	0.795	0.799
Mcquitty	**0.799**	0.857	0.825	0.732	0.792	0.796
Centroid	0.901	0.922	0.873	0.871	0.916	0.934
Median	0.857	0.897	0.857	0.842	0.891	0.906
Ward	0.801	0.858	0.823	**0.710**	**0.765**	**0.764**

Table 4. *Comparison of retrieval effectiveness for seven clustering methods*

Impact of sparsifying similarity matrix on retrieval effectiveness. We are interested to test if cluster retrieval effectiveness comprises when the computing efficiency of Sim-AHC is improved by sparsifying the similarity matrix $\mathbb{S}$. In this experiment, our baseline is the absolute running time (in seconds) and memory usage when a full-sized similarity matrix $\mathbb{S}$ is used as input. We then record the relative running time and relative memory usage of Sim-AHC when $\mathbb{S}$ is sparsified by a certain threshold value. Results of this experiment are displayed in Figure 1, where x-axis corresponds to percentiles. The dotted line with circle signs and the solid line with triangle signs respectively represent the relative memory usage and the relative running time. The dashed line with plus sign, and dashed line with cross signs and the dashed line with square signs plot the E values at $\beta = 0.5, 1.0\ and\ 2.0$, respectively. A surprising finding is that retrieval effectiveness does not comprise when $\mathbb{S}$ becomes more and more sparsified, as we can observe in each sub-figure, that the optimal E values are mostly invariant to the changes of threshold values. This demonstrates that even when $\mathbb{S}$ is largely sparsified, when memory usage and running time are significantly reduced, retrieval effectiveness can be guaranteed. This discovery also implies that Sim-AHC is a scalable algorithm, and it is favorable in tasks where efficiency is demanded.

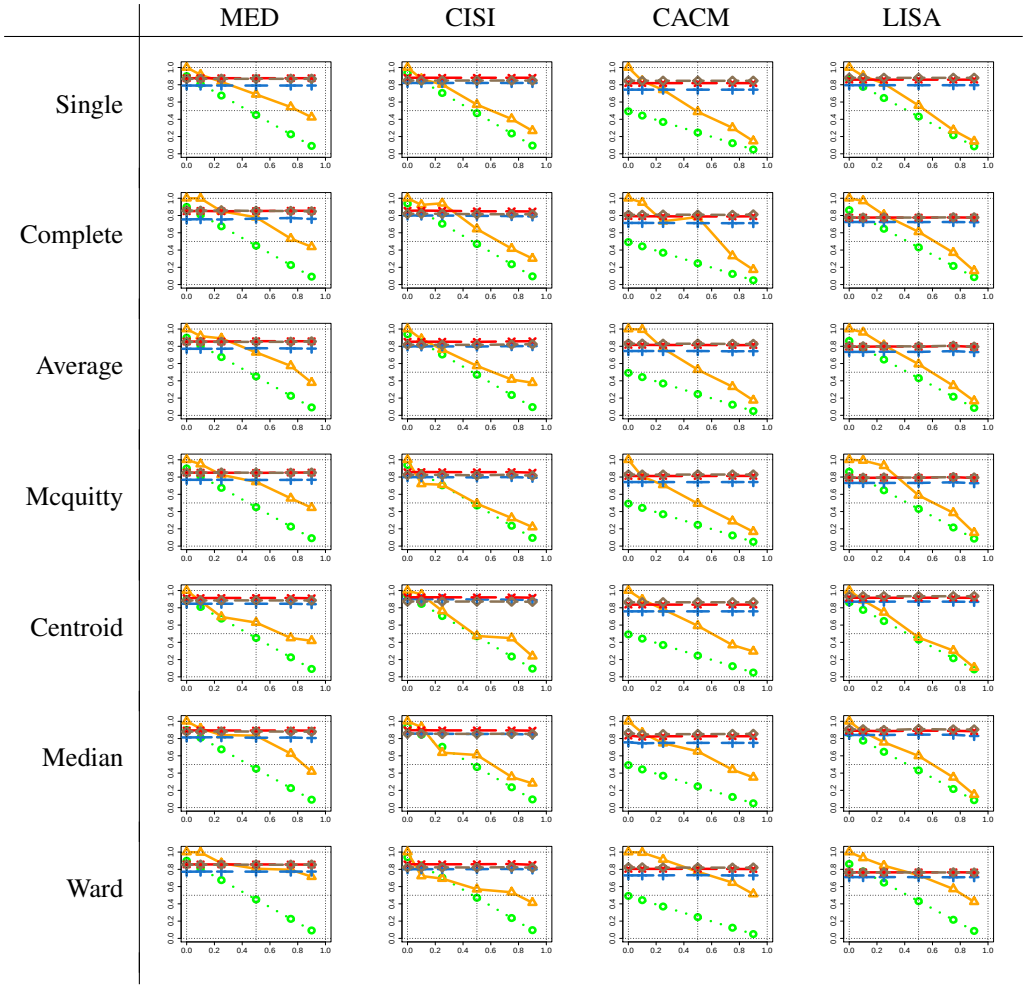

Figure 1. *Restuls of retrieval effectiveness, relative memory usage and relative running time when sparsifying similarity matrix*

5. Conclusion and Future Work

In this paper, we conduct a new cluster hypothesis test. The innovative features of this test are that (I) seven AHC methods unified by the LW formula are experimented, (II) a scalable similarity-based AHC framework is applied, the impact of improving efficiency on retrieval effectiveness is discussed, and (III) optimal cluster search is employed, the retrieval effectiveness is addressed based on the optimal E values. Our results reveal that complete link and average link are not always the best-performing methods in terms of retrieval effectiveness. We also discover that when applying filte-

ring strategy in Sim-AHC, retrieval effectiveness does not comprise when the similarity matrix becomes more and more sparse, that is, computing efficiency can be achieved when retrieval effectiveness is guaranteed. This property inspires us to conduct future tests of Sim-AHC on larger datasets. Currently, we are working on an implementation of a distributed version of Sim-AHC supported by Apache Spark engine. We anticipate that future tests using this implementation would help us reveal more interesting facts of cluster hypothesis in the scale of big data environment.

Acknowledgment

This work is supported by the French national project REQUEST PIA/FSN.

6. Bibliographie

Ah-Pine J., Wang X., « Similarity Based Hierarchical Clustering with an Application to Text Collections », *International Symposium on Intelligent Data Analysis*, Springer, p. 320-331, 2016.

El-Hamdouchi A., Willett P., « Techniques for the measurement of clustering tendency in document retrieval systems », *Journal of Information Science*, vol. 13, n° 6, p. 361-365, 1987.

Griffiths A., Luckhurst H. C., Willett P., « Using interdocument similarity information in document retrieval systems », *Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, CA*p. 365-373, 1997.

Griffiths A., Robinson L. A., Willett P., « Hierarchic agglomerative clustering methods for automatic document classification », *Journal of Documentation*, vol. 40, n° 3, p. 175-205, 1984.

Jardine N., van Rijsbergen C. J., « The use of hierarchic clustering in information retrieval », *Information storage and retrieval*, vol. 7, n° 5, p. 217-240, 1971.

Lance G. N., Williams W. T., « A general theory of classificatory sorting strategies II. Clustering systems », *The computer journal*, vol. 10, n° 3, p. 271-277, 1967.

Tombros A., Villa R., Van Rijsbergen C. J., « The effectiveness of query-specific hierarchic clustering in information retrieval », *Information processing & management*, vol. 38, n° 4, p. 559-582, 2002.

Van Rijsbergen C. J., Croft W. B., « Document clustering : An evaluation of some experiments with the Cranfield 1400 collection », *Information Processing & Management*, vol. 11, n° 5, p. 171-182, 1975.

Voorhees E. M., « The cluster hypothesis revisited », *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, p. 188-196, 1985.

Willett P., « Recent trends in hierarchic document clustering : a critical review », *Information Processing & Management*, vol. 24, n° 5, p. 577-597, 1988.

Xu R., Wunsch D. *et al.*, « Survey of clustering algorithms », *Neural Networks, IEEE Transactions on*, vol. 16, n° 3, p. 645-678, 2005.