퍼지 및 지능시스템학회 논문지 2001, Vol. 11, No. 7, pp. 579-583

# An Incremental Similarity Computation Method in Agglomerative Hierarchical Clustering

Sung-young Jung and Taek-soo Kim

Machine Intelligence Group, LG Electronics Institute of Technology
16 Woomyeon-Dong, Seocho-Gu, Seoul, 137-140, Korea

## Abstract

In the area of data clustering in high dimensional space, one of the difficulties is the time-consuming process for computing vector similarities. It becomes worse in the case of the agglomerative algorithm with the group-average link and mean centroid method, because the cluster similarity must be recomputed whenever the cluster center moves after the merging step. As a solution of this problem, we present an incremental method of similarity computation, which substitutes the scalar calculation for the time-consuming calculation of vector similarity with several measures such as the squared distance, inner product, cosine, and minimum variance. Experimental results show that it makes clustering speed significantly fast for very high dimensional data

**Key words** : Clustering, Agglomerative Hierarchical Clustering, Clustering for High Dimensional Data, Incremental Cluster Similarity Computation Method

## 1. Introduction

High dimensionality is one of the important aspects to be considered in clustering large data. When the data are represented on high dimensional space, the calculation of the cluster similarity requires high cost of computational resource due to the vector manipulation itself. As the dimension of the data space increases, the computational time for cluster similarity becomes a critical factor. There are several well-known methods for the dimensional reduction, such as the singular value decomposition [1], the feature selection [2] [3], and the projected clustering [4]. In many cases, however, these are not so practical since they require high computational cost, sometimes the cost of the reduction in these approaches overwhelms that of the clustering itself. Even these approaches become infeasible when the principal features vary over time.

To cope with the problem of high dimensionality, we propose a method of incremental similarity computation which drastically reduces the amount of computation for high dimensional input data by substituting scalar calculation for time-consuming vector calculation in computing the cluster similarity by an incremental manner.

In the incremental clustering algorithm, every input data is set as a new cluster until the number of clusters exceeds the predefined limit. When the limit is reached, the most similar cluster pairs are found and merged to keep the limit. The similarities of the cluster created by merging must be recalculated against all other clusters, which brings about a great deal of vector operations because it is heavily repeated for all merging steps.

In this paper, we introduce the way of computing vector similarities by means of scalar operations using geometrical properties to get such similarity measures as the squared distance, the cosine, and the inner product, even up to the minimum variance in the agglomerative hierarchical clustering with mean centroid method.

This paper is organized as follows: we describe how the incremental clustering method is applied to each similarity measures. Subsequently, we show the experimental evaluation of the proposed method. And we finally summarize the results as our conclusion.

## 2. Incremental Computation Methods for Cluster Similarity

The main idea is to devise an incremental method for vector similarity computation since many part of vector similarity is computed after cluster merging in order to estimate the similarity of the cluster created by the merging step. Here, the similarities between the created cluster and the others can be inferred from the previous similarities related with two clusters which are merged to create new cluster. The important thing here is that it can be done without vector computation such as inner product and vector cosine.

Now we will define general functional formation for the incremental computation of cluster similarity. As Figure 1 describes, the cluster **x** and **y** are going to be

Manuscript received september 15, 2001 ; revised Nobember 10. 2001.

merged to cluster **n**. The target which must be computed is the similarity between cluster **n** and the other cluster **w**. The available information sources given are the similarities between cluster **w** and **x**, cluster **x** and **y**, cluster **w** and **y**, and the merged cluster element ratio which determines the relative position of **n** on $\overline{xy}$ line. The ratio is equivalent to $\overline{xn}$ / $\overline{xy}$. Then the similarity between cluster **n** and **w** can be defined as equation (1). Figure 1 describes the functional formation on the vector space.

$$\mathrm{sim}(\mathbf{w},\mathbf{n}) = \mathrm{f}(\mathrm{sim}(\mathbf{w},\mathbf{x}).\mathrm{sim}(\mathbf{x},\mathbf{y}),\mathrm{sim}(\mathbf{w},\mathbf{y}),\overline{xn}/\overline{xy}) \quad (1)$$

This function should use only scalar calculation or triangular function without any vector operation. For the similarity of absolute or square of vector difference, the function can be easily derived based on geometrical property. We will describe about it next.

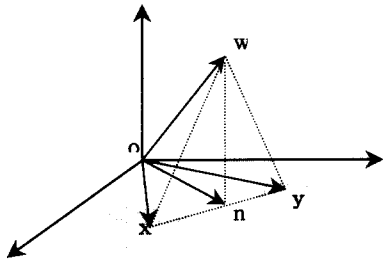



Fig. 1. Cluster merging step : **x**, **y** is the most similar clusters, **n** is the new cluster created by merging, **w** represents one of the other clusters

### 2.1 Incremental Vector Similarity Equation for Geometrical Distance

The similarity for squared geometrical distance is defined as equation (2)

$$\mathrm{sim}(\mathbf{p},\mathbf{q}) = |\mathbf{p}-\mathbf{q}|^2 \quad (2)$$

The similarity between w and n becomes the squared length of the line $\overline{wn}$. The formula for the incremental similarity computation becomes the equation (3).

$$\mathrm{sim}(\mathbf{w},\mathbf{n}) = \overline{wn}^2 = \mathrm{f}(\overline{wx}^2, \overline{xy}^2, \overline{wy}^2, a) \quad (3)$$

where $a = \overline{xn}/\overline{xy}$. It can be expressed with the edge line and the angle of the triangle $\Delta won$ as equation (4) by cosine theorem.

$$\overline{wn}^2 = \overline{xw}^2 + \overline{xn}^2 - 2\cdot\overline{xw}\cdot\overline{xn}\cdot\cos(\angle wxn) \quad (4)$$
$$\overline{xn} = a\cdot\overline{xy}$$

The $\cos(\angle wxn)$ value is converted with the line length of triangle $\Delta wxy$ like equation (5)

$$\cos(\angle wxn) = \cos(\angle wxy) = \frac{\overline{wx}^2 + \overline{xy}^2 - \overline{wy}^2}{2\cdot wx\cdot wy} \quad (5)$$

Finally the similarity for vector **w** and **n** is derived from previously computed values by applying equation (5) to (4).

$$\therefore \overline{wn}^2 = \overline{xw}^2 + (a\cdot\overline{xy})^2 - a\cdot\frac{\overline{xy}}{\overline{wy}}\cdot(\overline{wx}^2 + \overline{xy}^2 - \overline{wy}^2) \quad (6)$$

### 2.2 Incremental Vector Similarity Equation for Inner Product and Cosine measure

The similarity for inner product is defined as equation (7).

$$\mathrm{sim}(\mathbf{p},\mathbf{q}) = \mathbf{p}\cdot\mathbf{q} \quad (7)$$

The corresponding formula for incremental similarity is defined as follows.

$$\mathrm{sim}(\mathbf{w},\mathbf{n}) = \mathbf{w}\cdot\mathbf{n} = \mathrm{f}(\mathbf{w}\cdot\mathbf{x},\mathbf{x}\cdot\mathbf{y},\mathbf{w}\cdot\mathbf{y},a = \overline{xn}/\overline{xy}) \quad (8)$$

The inner product can be replaced by the line component of triangle $\Delta won$ by cosine theorem

$$\begin{aligned}\mathbf{w}\cdot\mathbf{n} &= |\mathbf{w}|\cdot|\mathbf{n}|\cdot\cos(\angle won)\\ &= |\mathbf{w}|\cdot|\mathbf{n}|\cdot\frac{|\mathbf{w}|^2+|\mathbf{n}|^2-\overline{wn}^2}{2\cdot|\mathbf{w}|\cdot|\mathbf{n}|}\\ &= \frac{1}{2}(|\mathbf{w}|^2+|\mathbf{n}|^2-\overline{wn}^2)\end{aligned} \quad (9)$$

$\overline{wn}^2$ term can be replaced by known values such as line length values and the input parameters in the function for incremental similarity (8)

$$\overline{wn}^2 = \overline{xw}^2 + \overline{xn}^2 - 2\cdot\overline{xw}\cdot\overline{xn}\cdot\cos(\angle wxn)$$
$$\begin{aligned}\overline{xw}^2 &= \overline{ox}^2 + \overline{ow}^2 - 2\cdot\overline{ox}\cdot\overline{ow}\cdot\cos(\angle wox)\\ &= |\boldsymbol{x}|^2 + |\boldsymbol{w}|^2 - 2\cdot\boldsymbol{x}\cdot\boldsymbol{w}\end{aligned}$$
$$\overline{xn} = a\cdot\overline{xy}$$
$$\overline{xy}^2 = |\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 - 2\cdot\boldsymbol{x}\cdot\boldsymbol{y}$$
$$\overline{wy}^2 = |\boldsymbol{w}|^2 + |\boldsymbol{y}|^2 - 2\cdot\boldsymbol{x}\cdot\boldsymbol{y}$$
$$\cos(\angle wxn) = \cos(\angle wxy) = \frac{\overline{xw}^2 + \overline{xy}^2 - \overline{wy}^2}{2\cdot\overline{xw}\cdot\overline{xy}} = \frac{(|\boldsymbol{x}|^2 - \boldsymbol{x}\cdot\boldsymbol{w} - \boldsymbol{x}\cdot\boldsymbol{y} + \boldsymbol{w}\cdot\boldsymbol{y})}{\overline{xw}\cdot\overline{xy}}$$
$$\therefore \overline{wn}^2 = \overline{xn}^2 + \overline{xn}^2 - 2\cdot\overline{xw}\cdot\overline{xn}\cdot\frac{(|\boldsymbol{x}|^2 - \boldsymbol{x}\cdot\boldsymbol{w} - \boldsymbol{x}\cdot\boldsymbol{y} + \boldsymbol{w}\cdot\boldsymbol{y})}{\overline{xw}\cdot\overline{xy}}$$
$$= |\boldsymbol{x}|^2 + |\boldsymbol{w}|^2 - 2\cdot\boldsymbol{x}\cdot\boldsymbol{w} + a^2\cdot(|\boldsymbol{x}|^2 + |\boldsymbol{y}|^2 - 2\cdot\boldsymbol{x}\cdot\boldsymbol{y}) - 2a\cdot(|\boldsymbol{x}|^2 - \boldsymbol{x}\cdot\boldsymbol{w} - \boldsymbol{x}\cdot\boldsymbol{y} + \boldsymbol{w}\cdot\boldsymbol{y}) \quad (10)$$

Consequently, the inner product similarity is deduced in the incremental similarity functional formation. Now the inner product similarity can be computed from the previous similarity values and squared absolute vector values in equation (11). Absolute value of the vector can be included in incremental method because its value is static for the given vector.

$$\therefore \mathbf{w}\cdot\mathbf{n} = \frac{1}{2}\{|\mathbf{n}|^2 - |\mathbf{x}|^2 + 2\cdot\mathbf{x}\cdot\mathbf{w} - a^2\cdot(|\mathbf{x}|^2 + |\mathbf{y}|^2 - 2\cdot\mathbf{x}\cdot\mathbf{y}) + 2a\cdot(|\mathbf{x}|^2 - \mathbf{x}\cdot\mathbf{w} - \mathbf{x}\cdot\mathbf{y} + \mathbf{w}\cdot\mathbf{y})\} \quad (11)$$

Now we can calculate the inner product similarity incrementally using equation (11). By the way, there is another way to get the inner product with equation (9). We already showed that the term $\overline{wn}^2$can be replaced by line component in equation (6). Therefore, by managing squared distance similarity matrix as described in the previous section, inner product value can be directly computed with equation (9) and (6).

The cosine similarity can be directly computed from the inner product similarity which is computed incrementally

$$\cos(\angle won) = \frac{\mathbf{w} \cdot \mathbf{n}}{|\mathbf{w}| \cdot |\mathbf{n}|} \tag{12}$$

### 2.3 Incremental Similarity Computation for Minimum Variance Clustering

The similarity value for minimum variance clustering [5] [6] can be defined by the amount of variance increase when two clusters are merged. Among all the cluster pairs, the pair which has minimum increment of variance is selected to be merged for each step.

$$\begin{aligned} sim(i,j) &= \Delta E_{ij} = E_{ij} - (E_i + E_j) \\ E_i &= \sum_{x_k \in c_i} |x_k - c_i|^2 = N_i \sigma_i^2 \end{aligned} \tag{13}$$

where $E_i$ is the total amount of variance in cluster $i$, $c_i$ is $i$-th cluster centroid, $N_i$ is the number of element in cluster $i$. Then the similarity is defined by the increment of total amount of variance, $\Delta E_{ij}$. $\sigma_i^2$. is the variance. $c_{ij}$ is the merged cluster of $c_i$ and $c_j$, and is determined by the center of gravity between two clusters.

$$\sigma_i^2 = \frac{1}{N_i} \sum_{x_k \in c_i} |x_k - c_i|^2 = \frac{1}{N_i} \sum_{x_k \in c_i} |x_k|^2 - |c_i|^2 \tag{14}$$

$$c_{ij} = \frac{N_i c_i + N_j c_j}{N_i + N_j}$$

In order to apply incremental computation method, the similarity equation can be derived with the terms of cluster and number of element as follows.

$$\begin{aligned} \Delta E_{ij} &= E_{ij} - (E_i + E_j) \\ &= N_{ij}\sigma_{ij}^2 - (N_i\sigma_i^2 + N_j\sigma_j^2) \\ &= \sum_{x_k \in c_{ij}} |x_k|^2 - N_{ij}|c_{ij}|^2 \\ &\quad - \sum_{x_k \in c_i} |x_k|^2 + N_j|c_i|^2 - \sum_{x_k \in c_j} |x_k|^2 + N_j|c_j|^2 \\ &= N_i|c_i|^2 + N_j|c_j|^2 - N_{ij}\left|\frac{N_i c_i + N_j c_j}{N_{ij}}\right|^2 \\ &= \frac{N_i N_j |c_i - c_j|^2}{N_i + N_j} \end{aligned} \tag{15}$$

The equation (15) shows that the increment of variance can be computed by the squared distance of two clusters $|c_i - c_j|^2$, and their number of elements $N_i$ and $N_j$. We showed before, the method of incremental computation for squared distance similarity. Consequently, the similarity of minimum variance clustering, $\Delta E_{ij}$ can be computed in an incremental way by managing incremental squared distance similarity method, by applying the square distance value to equation (15) to get $\Delta E_{ij}$ value. It makes clustering drastically fast for high dimensional data space.

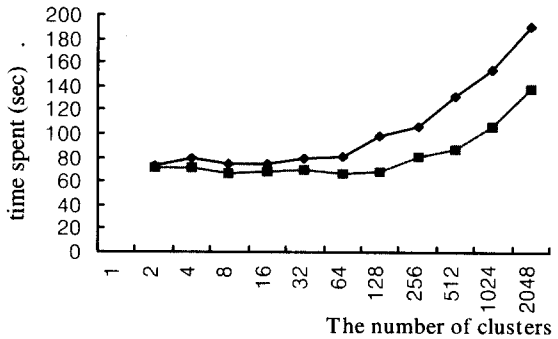
## 3. Experimental Results

We constructed the collaborative filtering model using Pearson correlation measure [8] [9] in order to compare clustering results. This system receives input data which is clustered by the clustering algorithms proposed in this paper. We regard given clustering result as better when the collaborative filtering result is more accurate. We adopted precision as an accuracy measure. The precision is defined by the number of correct answers divided by number of answers which the system gives.

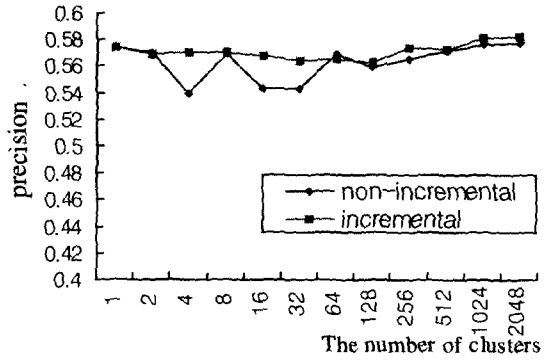
$$precision = \frac{\text{the number of correct answers}}{\text{the number of answers}} \tag{16}$$

We used EachMovie data on public domain [7]. It consists of 2,811,983 transactions in which 72,916 users rated for 1,628 items.

Fig. 2 shows the results for the incremental cluster similarity computation for inner product. The time spent for the inner product and the cosine similarity calculation is almost same since one of them can be directly computed from the other. It shows that the clustering time spent for incremental method is less than non-incremental way as in Figure 2(a). We can also see that the precision does not be changed significantly. It means there is no significant difference in computational results between incremental and non-incremental method. The slight difference can be made by the inaccuracy of triangular functions and the accumulated errors. Figure 3 shows the results for the squared distance similarity computation. The result of minimum variance can be directly inferred from squared distance similarity as equation (15), therefore there is no time difference between them. It shows that incremental method is significantly faster then non-incremental way Fig. 3 (a), without severe accuracy degradation Fig. 3 (b)

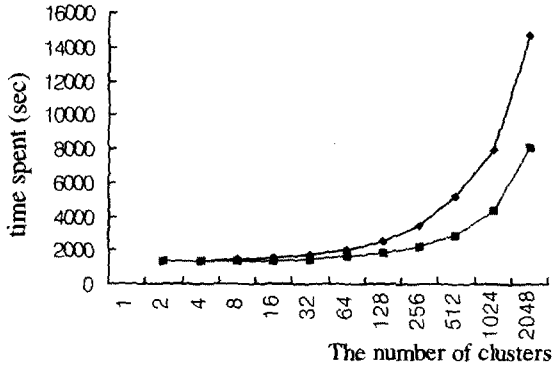



(a) The spent time

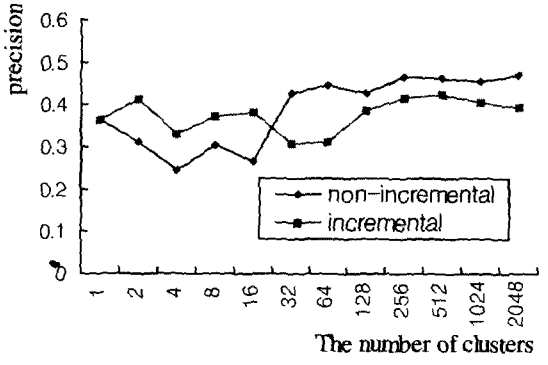



(b) Precision

Fig. 2. Incremental cluster similarity computation for inner product similarity. Training data size is 3000, test data size is 1000, and the number of recommendation : 10




(a) The spent time




(b) Precision

Fig. 3. Incremental cluster similarity computation for square distance similarity. Training data size is 50000, test data size is 5000, and the number of recommendation : 10

## 4. Conclusions

For dealing with the problem of high dimensionality in clustering large data, we proposed incremental similarity computation method for several similarity measures such as squared distance, inner product, cosine, and minimum variance in agglomerative hierarchical clustering. The incremental method replaces vector operation with scalar operation which results in computational speed up for high dimensional vector space. The experimental results show that incremental method is significantly faster than non-incremental way without sever accuracy degradation.

There are many approaches to speed up clustering algorithm process for large data. Most of them are based on approximation resulting in accuracy damaged. Our work does not bring about accuracy degradation because it is not designed using approximation mechanism. So it can contribute to make many other clustering algorithms more feasible for high dimensional data when combined with other approach.

## Reference

[1] Leon, S. J., "Linear Algebra with applications", third edition, Macmillan Publishing Company, 1990.

[2] Blum, A. L., and P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, 1997

[3] Koller, D., and M. Sahami, "Toward Optimal Feature Selection", Machine Learning Proc. of 13th International Conference, pp. 284-292, 1996

[4] Aggarwal, C. C., and C. Procopiuc, "Fast Algorithms for Projected Clustering", ACM SIGMOD, pp. 61-72, 1999, Philadelphia,

[5] Frakes, W. B., Baeza-Yates, R., "Information Retrieval: Data Structures & Algorithms", Prentice-hall, 1992.

[6] Ward, J. H., "Hierarchical Grouping to Optimize an Objective Function", *American Statistical Association*, 58 (301), pp. 235-244, 1963.

[7] McJones, P. "EachMovie collaborative filtering data set.", DEC Systems Research center. http://www.research. digital.com/SRC/eachmovie/.

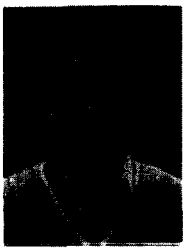
[8] Breese, J. S., D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence, 1998.

[9] Delgado, J., and N. Ishii, "Memory-Based Weighted-Majority Prediction for Recommender Systems," ACM SIGIR'99 Workshop on Recommender Systems, 1999.

[10] Boughettaya, A. "On-line Clustering", *IEEE Transaction on Knowledge and Data Engineering*, vol. 8, no. 2, April 1996.

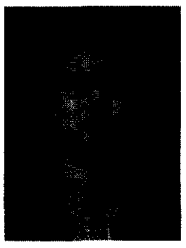
[11] Ward, J. H., and E. H. Marion, "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles", Educational and Physiological Measurement, 1963.

[12] Voorhees, E. M. "Implementing Agglomerative Hierarchic Clustering Algorithms for use in Document Retireval", *Information Processing & Management* vol. 22, no. 6, 1986



**Sung Young Jung**
received his BS and MS degrees in computer science from the Korea Advanced Institute of Science and Technology in 1994 and 1996. He is a assistant research engineer at Machine Intelligence Group in LG Electronics of Technology. His research interests include machine learning, Data Mining, and artificial intelligence.

E-mail : syjung@LG-Elite.com
Phone : +82-2-3497-8516



**Taek-Soo Kim**
received the BS degree and the MS and PhD degree in Electrical Engineering all from the University of Yonsei, Seoul Korea, in 1990, 1992, and 1996 respectively. He is a chief research engineer of Machine Intelligence group at LG Electronics Institute of Technology. His areas of research include data mining, biomedical engineering and artificial intelligence.

E-mail : tskim@LG-Elite.com
Phone : +82-2-3497-8517