Distance Dimension Reduction on QR Factorization for Efficient Clustering Semantic XML Document Using the QR Fuzzy C-Mean (QR-FCM)

Hsu-Kuang Chang[1,2] and I-Chang Jou[1]

[1] Institute of Engineering Science and Technology,
National Kaohsiung First University of Science and Technology, Kaohsiung, Taiwan
[2] Department of Information Engineering, I-Shou University, Kaohsiung, Taiwan
hkchang@isu.edu.tw, icjou@ccms.nkfust.edu.tw

Abstract. The rapid growth of XML adoption has urged for the need of a proper representation for semi-structured documents, where the document semantic structural information has to be taken into account so as to support more precise document analysis. In order to analyze the information represented in XML documents efficiently, researches on XML document clustering are actively in progress. The key issue is how to devise the similarity measure between XML documents to be used for clustering. Since XML documents have hierarchical structure, it is not appropriate to cluster them by using a general document similarity measure. Dimension reduction plays an important role in handling the massive quantity of high dimensional data such as mass semantic structural documents. In this paper, we introduce distance dimension reduction (DDR) based on the QR factorization (DDR/QR) or the Cholesky factorization (DDR/C). DDR generates lower dimensional representations of the high-dimensional XML document, which can exactly preserve Euclidean distances and cosine similarities between any pair of XML documents in the original dimensional space. After projecting XML documents to the lower dimensional space obtained from DDR, our proposed method QR fuzzy c-mean to execute the document-analysis clustering algorithms (we called the QR-FCM). DDR can substantially reduce the computing time and/or memory requirement of a given document-analysis clustering algorithm, especially when we need to run the document analysis algorithm many times for estimating parameters or searching for a better solution.

Keywords: QR factorization, singular value decomposition, distance dimension reduction, PEWF, PEIDF, PESSW, fuzzy C-mean, QR-FCM.

1 Introduction

XML document which is a semi-structured data, have hierarchical structure. Therefore, rather than using the similarity measure of the general document clustering techniques as is, a new similarity measure that considers the semantic and structural information of XML document must be investigated. However, some XML clustering methods used

P. Mueller, J.-N. Cao, and C.-L. Wang (Eds.): Infoscale 2009, LNICST 18, pp. 272–287, 2009.

the similarity measure that takes only the structural information of XML documents into account. Hwang proposed a clustering method that extracts a typical structure of the maximum frequent pattern using *PrefixSpan* algorithm [1] on XML documents [2, 3]. However, since such a typical structure extracted from XML document is not only the structure that represents the XML document itself, but also it cannot be the representative of the whole documents corpus, there is an accuracy issue of similarity. Lian summarized XML documents into *S-graph* which is a structural graph and proposed the calculation method of distance between *S-graphs* to be used for clustering [4]. However, they have no consideration for semantic information on XML documents as they focus on structural information only. Since dimension reduction is one of the fundamental methods for data analysis, there have been a lot of studies on effective and efficient dimension reduction algorithms. There are linear dimension reduction algorithms including principal component analysis (PCA) [5] and multidimensional scaling (MDS) [6]. There are also nonlinear dimensional reduction algorithms (NLDR) including Isomap [7], locally linear embedding (LLE) [8], [9], Hessian LLE [10], Laplacian eigenmaps [11], local tangent space alignment (LTSA) [12] and distance preserving dimension reduction based on the singular value decomposition (DPDR/SVD) [13]. These dimension a variety of areas such as biomedical image recognition, biomedical text data mining, and biological data analysis.

In this paper, we introduce distance dimension reduction (DDR) based on the QR factorization or the Cholesky factorization. DDR can produce t dimensional representations where t is the rank of the original XML data set, which exactly preserve Euclidean L_2-norm distances as well as cosine similarities between any pair of XML documents in the original m-dimensional space when m is much larger than the number of XML documents n, i.e. $m > n$. It projects the original XML data set into a much lower dimensional space without any loss of the pair-wise Euclidean distance information. Then, other XML documents analysis algorithms can be executed so that we can substantially reduce their computing times and memory requirements without any quality loss of their results.

The rest of this paper is organized as follows. In Section 2, we introduce the preparation XML documents on vector space model. In Section 3, we introduce DDR based on the QR factorization and DDR based on the Cholesky factorization after showing theorems of distance and cosine similarity preserving properties. Section 4 presents experimental results illustrating properties of the proposed DDR methods. Summary is given in Section 5.

2 Preparation for Semantic-Based XML Documents

In this section, we first introduce pre-processing steps for the incorporation of hierarchical information in encoding the XML tree's paths. It is based on the preorder tree representation (PTR) [14] and will be introduced after a brief review of how to generate an XML tree from an XML document. To do so, we have to first go through

Table 1. Preprocessing steps for XML document

Step 1. **Conversion**	Convert the XML document to tree Format. The values of the elements in the tree are not considered here, and only the structural information will be passed on to the subsequent steps.
Step 2. **Path Extraction**	Traverse the elements from the root to each leaf node of the tree. Record the sequence and hierarchical information for each path.
Step 3. **Similar Element Identification and Transformation**	Rename the terms of the paths with unique identifiers. Replace every element with a unique identifier in increasing order.
Step 4. **Nested and Duplicated Path Removal**	Remove any nested and duplicated path. The nested and duplicated paths in the tree are not considered here, and only the unique ones will be passed on to the next step.
Step 5. **Path Elements Encoding**	Based on the structural summary by Step 4, the path elements are encoded.

the following five preprocessing steps for XML documents. They are illustrated in Table 1.

From five steps preprocess, now XML document is modeled as a XML tree $T=(V,E)$. T is connected tree with $V=\{ v_1, v_2, ...\}$ as a set of vertices and $v_1 \in V$, $v_2 \in V$, $(v_1, v_2) \in E$ as a set of edges. One distinguished vertex $r \in V$ is designated the root, and for all $v \in V$, there is a unique path from r to v. As an example, Figure 1 depicts a sample XML tree containing some information about collection of books. The *book* consists of *intro* tags, each comprising *title*, *author* and *date* tags. Each *author* contains *fname* and *lname*, each *date* includes *year* and *month* tags. Figure 1 left shows only the first letter of each tag for the simplicity.

XML document has a hierarchical structure and this structure is organized with tag paths. Each tag path represents document characteristics that can predict the contents of XML document. Strictly speaking, it shows the semantic structural characteristics of XML document. In this paper, we propose a method for calculating the similarity using all tag paths of XML tree representing the semantic structural information of XML document.

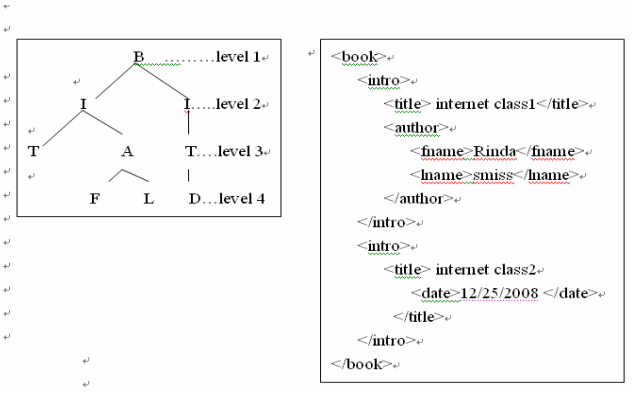


Fig. 1. An example of XML Document

3 Path Element Vector Space Model (PEVSM)

Vector model represents a document as a vector whose elements are the weights of the path elements within a document. In calculating the weight of each path element within a document, Term Frequency and IDF (Inverse Document Frequency) method is used [15]. We define PESSW (Path Element Structural Semantic Weight) that calculate the weight of a path element in a XML document. The PESSW is PEWF (Path Element Weighted Frequency) multiplied by PEIDF (Path Element Inverse Document Frequency). $PESSW_{ij}$ of i^{th} path element in the j^{th} document is shown in equation (1).

$$PESSW_{ij} = PEWF_{ij} \times PEIDF_{ij} \tag{1}$$

$PEWF_{ij}$ is shown in equation(2).

$$PEWF_{ij} = freq_{ij} \times \frac{1}{x^n} \tag{2}$$

$PEIDF_{ij}$ is shown in equation (3),

$$PEIDF_{ij} = \log \frac{N}{DF_j} \tag{3}$$

Table 2 shows PEWF, PEIDF, and PESSW on an example trees in Figure 1.

Table 2. An example of PTWF, PTIDF and PESSW

Path / Element	PEWF			PEIDF			PESSW		
	doc_1	doc_2	doc_3	doc_1	doc_2	doc_3	doc_1	doc_2	doc_3
/B/I/T/D	1.0	0.0	0.0	1.1	0.0	0.0	1.1	0.0	0.0
/B/I/A/F	1.0	1.0	0.0	0.41	0.41	0.0	0.41	0.41	0.0
/B/I/A/L	1.0	1.0	0.0	0.41	0.41	0.0	0.41	0.41	0.0
/M/I/A/L	0.0	0.0	1.0	0.0	0.0	1.1	0.0	0.0	1.1
/B/I/T/	2.0	1.0	0.0	0.41	0.41	0.0	0.81	0.41	0.0
/B/I/A	1.0	1.0	0.0	0.41	0.41	0.0	0.41	0.41	0.0
/B/I/	2.0	1.0	0.0	0.41	0.41	0.0	0.81	0.41	0.0
/B	1.0	1.0	0.0	0.41	0.41	0.0	0.41	0.41	0.0
/M/I/T	0.0	0.0	1.0	0.0	0.0	1.1	0.0	0.0	1.1
/M	0.0	0.0	1.0	0.0	0.0	1.1	0.0	0.0	1.1
/M/I	0.0	0.0	2.0	0.0	0.0	1.1	0.0	0.0	2.2
/I/T/D	0.5	0.0	0.0	1.10	0.0	0.0	0.5	0.0	0.0
/I/A/F	0.5	0.5	0.0	0.41	0.41	0.0	0.21	0.21	0.0
/I/A/L	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
/I/T	1.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
/I/A	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
/I	1.0	0.5	1.0	0.0	0.0	0.0	0.0	0.0	0.0
/T/D	0.33	0.0	0.0	1.1	0.0	0.0	0.33	0.0	0.0
/A/F	0.33	0.33	0.0	0.41	0.41	0.0	0.14	0.14	0.0
/A/L	0.33	0.33	0.33	0.0	0.0	0.0	0.0	0.0	0.0
/T	0.67	0.33	0.33	0.0	0.0	0.0	0.0	0.0	0.0
/A	0.33	0.33	0.33	0.0	0.0	0.0	0.0	0.0	0.0
D	0.25	0.0	0.0	1.1	0.0	0.0	0.27	0.0	0.0
F	0.25	0.25	0.0	0.41	0.41	0.0	0.1	0.1	0.0
L	0.25	0.25	0.25	0.0	0.0	0.0	0.0	0.0	0.0

Let d_x and d_y be two vectors that represent a XML document doc_x and doc_y. Cosine similarity is defined as the angle between two vectors and quantified by equation (4) and (5).

$$\cos\theta = \frac{d_x \cdot d_y}{|d_x| \cdot |d_y|} \text{ , that is} \tag{4}$$

$$sim(doc_x, doc_y) = \frac{d_x \cdot d_y}{|d_x| \times |d_y|} = \frac{\sum_{k=1}^{t} w_{kx} \times w_{ky}}{\sqrt{\sum_{k=1}^{t} w_{kx}^2} \times \sqrt{\sum_{k=1}^{t} w_{ky}^2}} \tag{5}$$

where $d_x = (w_{1x}, w_{2x}, ..., w_{tx})$ and $d_y = (w_{1y}, w_{2y}, ..., w_{ty})$, t is the total number of path elements in d_x, d_y respectively[16].

3.1 Distance Dimension Reduction (DDR) via the QR Factorization

Let us deal with n XML documents whose dimension is m s.t. $m \succ n$. We compute the QR factorization of the XML document matrix $D \in \Re^{m \times n}$:

$$D = QR = Q\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1 R_1,$$

where $Q \in \Re^{m\times m}$ is an orthogonal matrix and $R_1 \in \Re^{n\times n}$ is an upper triangular matrix. Then, $Q_1 \in \Re^{m\times n}$ can be considered as a dimensionality transformation matrix when m > n and the lower dimensional representation $\hat{x} \in \Re^{n\times 1}$ of a vector $x \in \Re^{m\times 1}$ can be computed as $\hat{x} = Q_1^T x$. Thus, the lower dimensional representation d_i of each XML document $\hat{d}_i = Q_1^T d_i = r_i$, where d_i is the i-th column of D and r_i is the i-th column of R_1.

Theorem 1. QR-Factorization

Let $\left(\vec{d}_1, \vec{d}_2, ..., \vec{d}_n\right)$ be a basis of D a subspace of $\Re^m$. For j=2,3,…,n, where we write $\vec{d}_j = \vec{d}_j^{\,II} + \vec{d}_j^{\,\perp}$, with respect to span $\left(\vec{d}_1, \vec{d}_2, ..., \vec{d}_{j-1}\right)$.

Then, $\vec{q}_1 = \dfrac{1}{\left\|\vec{d}_1\right\|}\vec{d}_1$

$\vec{d}_2^{\,\perp} = \vec{d}_2 - \vec{d}_2^{\,II} = \vec{d}_2 - (\vec{q}_1 \cdot \vec{d}_2)\vec{q}_1$, and $\vec{q}_2 = \dfrac{1}{\left\|\vec{d}_2^{\,\perp}\right\|}\vec{d}_2^{\,\perp}$,

$\vec{d}_3^{\,\perp} = \vec{d}_3 - \vec{d}_3^{\,II} = \vec{d}_3 - (\vec{q}_1 \cdot \vec{d}_3)\vec{q}_1 - (\vec{q}_2 \cdot \vec{d}_3)\vec{q}_2$, and $\vec{q}_3 = \dfrac{1}{\left\|\vec{d}_3^{\,\perp}\right\|}\vec{d}_3^{\,\perp}$,

……, , and ……

$\vec{d}_j^{\,\perp} = \vec{d}_j - \vec{d}_j^{\,II} = \vec{d}_j - (\vec{q}_1 \cdot \vec{d}_j)\vec{q}_1 - ... - (\vec{q}_{j-1} \cdot \vec{d}_j)\vec{q}_{j-1}$, and $\vec{q}_j = \dfrac{1}{\left\|\vec{d}_j^{\,\perp}\right\|}\vec{d}_j^{\,\perp}$

……, , and ……

$\vec{d}_m^{\,\perp} = \vec{d}_m - \vec{d}_m^{\,II} = \vec{d}_m - (\vec{q}_1 \cdot \vec{d}_m)\vec{q}_1 - ... - (\vec{q}_{m-1} \cdot \vec{d}_m)\vec{q}_{m-1}$, and $\vec{q}_m = \dfrac{1}{\left\|\vec{d}_m^{\,\perp}\right\|}\vec{d}_m^{\,\perp}$

The Gram-Schmidt process represents a change of basis from the old basis $D = (\vec{d}_1, \vec{d}_2, ..., \vec{d}_n)$ to a new, orthonormal basis $Q = (\vec{q}_1, \vec{q}_2, ..., \vec{q}_n)$ of vector *V*. If R is the change basis matrix form D to *Q*, then

D=QR

$$(\vec{d}_1, \vec{d}_2, ..., \vec{d}_n) = (\vec{q}_1, \vec{q}_2, ..., \vec{q}_n) \begin{bmatrix} r_{11} & r_{12} & ... & ... & r_{1n} \\ 0 & r_{22} & ... & ... & r_{2n} \\ 0 & 0 & r_{33} & ... & r_{3n} \\ ... & ... & ... & .. & .. \\ 0 & 0 & 0 & 0 & r_{nn} \end{bmatrix}$$

This equation is called the QR-factorization of the matrix *D*. Using Gram-Schmidt algorithm, we know that

$$\vec{d}_j = \vec{d}_j^{\,\mathrm{II}} + \vec{d}_j^{\,\perp}$$

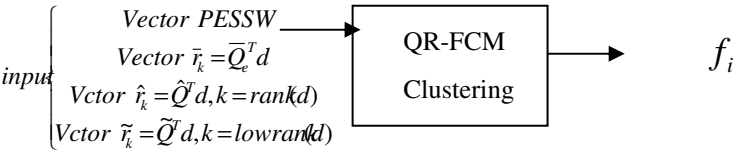
$$= (\vec{q}_1 \cdot \vec{d}_1)\vec{q}_1 + ... + (\vec{q}_i \cdot \vec{d}_j)\vec{q}_i + ... + (\vec{q}_{j-1} \cdot \vec{d}_j)\vec{q}_{j-1} + \left\|\vec{d}_j^{\,\perp}\right\|\vec{q}_j$$

$$= r_{11}\vec{q}_1 + ... + r_{ij} + ... + r_{j-1,j}\vec{q}_{j-1} + r_{jj}\vec{q}_j$$

It follows that $r_{ij} = \vec{q}_i \vec{d}_j$ if i < j, $r_{jj} = \left\|\vec{d}_j^{\,\perp}\right\| \geq 0$; and r_{ij}=0 if i > j. The last equation implies that R is an upper triangular matrix. The first diagonal entry is $r_{11} = \left\|\vec{d}_1\right\|$, since $\vec{d}_1 = \left\|\vec{d}_1\right\|\vec{q}_1$. $(\because \vec{q}_1 = \frac{1}{\left\|\vec{d}_1\right\|}\vec{d}_1)$

3.2 Our Proposed QR-FCM Algorithm

From the QR factorization as the previous section described, we have the originated document vector PESSW, document vector $\bar{r}$ (economic QR factoring), document vector $\hat{r}$ (rank of PESSW), and document vector $\tilde{r}$ (low rank of QR) based on the XML documents, which taken as the QR-FCM input data and then go through the clustering.



F(I)=[f_1,f_2,…..,f_c] , where $f_i = \sum_{j=1}^{N} \mu_{ij} P_j = \frac{1}{N} \sum_{j=1}^{N} \mu_{ij}$.

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} \mu_{ij}^m \left\| d_i - c_j \right\|^2 , 1 \le \mathrm{m} \le \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of d_i in the cluster j, d_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| d_i - c_j \right\|}{\left\| d_i - c_k \right\|} \right)^{\frac{2}{m-1}}}, \quad c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m . d_i}{\sum_{i=1}^{N} \mu_{ij}^m}$$

This iteration will stop when $\max_{ij} \{ \left| \mu_{ij}^{(k+1)} - \mu_{ij}^{(k)} \right| \} < \xi$, where ξ is a termination criterion between 0 and 1, whereas k are the iteration steps. This procedure converges to a local minimum or a saddle point of J_m. The QR-FCM algorithm is composed of the following steps:

1. *Input the number of clusters c, the weighting exponent m, and error tolerance ε.*
2. *Initialize the cluster centers* $c = \{c_i\}$ *, for* $1 \le i \le C$.
3. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
4. *At k-step: calculate the centers vectors* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m . d_i}{\sum_{i=1}^{N} \mu_{ij}^m}$$

5. Update $U^{(k)}$, $U^{(k+1)}$

$$\mu_{ij} = \frac{\left(\frac{1}{\left\| d_j, \upsilon_i \right\|_2} \right)^{\frac{1}{m-1}}}{\sum_{i=1}^{C} \left(\frac{1}{\left\| d_j, c_i \right\|_2} \right)^{\frac{1}{m-1}}} \quad , \quad \mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| d_i - c_j \right\|}{\left\| d_i - c_k \right\|} \right)^{\frac{2}{m-1}}}$$

6. If || $U^{(k+1)}$ - $U^{(k)}$||< ξ then STOP; otherwise return to step 4.

4 Experiment Result

4.1 Within-Group-Variance and Between-Group-Variance Using Membership μ

The purpose of any clustering technique [17], [18], [19], [20], [21] is to evolve a K x n partition matrix (U)=D of a data set D $(D = \{d_1, d_2, ..., d_n\})$ in $\Re^N$, representing its partitioning into a number, say K, of clusters (C_1; C_2; . . . ; C_K). The partition matrix (U) = D may be represented as U=[μ_{kj}], k=1,..,K, and j= 1,..., n, where u_{kj} is the membership of pattern d_j to clusters C_k. In crisp partitioning of the data, the following condition holds: u_{kj}=1 if $d_j \in C_k$; otherwise, u_{kj}=0. Right now, we use the membership *u* as separation measuring degree to figure out the variance of the within-group-variance and between-group-variance among the XML documents in the different cluster. First, we define the useful definitions as follows.

Definition 1. The closeness (denseness) of the XML documents in the i^{th} cluster (C_i)

$$C_i = \frac{\sum_{j \in S_i} \mu_{ij}}{|S_i|} \quad \textbf{where} \quad I_{i,j} = \begin{cases} 1 & if \quad \mu_{i,j} = \max_{1 \le k \le C} \mu_{k,j} \\ 0 & if \quad \mu_{i,j} \neq \max_{1 \le k \le C} \mu_{k,j} \end{cases},$$

, μ_{ij} is the membership calculated from the QR-FCM (QR fuzzy C-mean) and $S_i = \{j \mid I_{i,j} = 1\}$.

Definition 2. Within-Group-Variance (WGV) of the cluster

$$WGV(\mu, V; D) = \sum_{i=1}^{C} \sum_{j \in S_i} (C_i)^{-1} \left\| d_j, \upsilon_i \right\|_2 \tag{6}$$

C_i is the denseness within the cluster which is defined as the membership μ_{ij} with the ith cluster. μ_{ij} is calculated from QR-FCM (QR fuzzy C-mean). μ_{ij} represents the membership of the XML document d_j within the cluster i. If the all membership within the ith cluster is large, then each XML document is close to the central document. That means each XML document is close to the central document within the cluster, each close to the central document, high denseness, low variance. C_i is the denseness determined by the μ_{ij} membership. Beside, from the distance of XML document and central document, $\left\| d_j, \upsilon_i \right\|_2$, defined the XML document diversity within the cluster. We defined the within-group-variance as combine both of these two criteria, the bigger C_i the more dense, and the smaller $\left\| d_j, \upsilon_i \right\|_2$ the better result for the within-group-variance, so we have $_{\min}\{\text{WGV}\}$.

Definition 3. The contribution of the separation XML document d_j among λcluster and *m* cluster defined as $\mu_{\lambda,j} * \mu_{m,j}$.

Example 1

$$m == \begin{matrix} C_1 \\ C_2 \end{matrix} \begin{bmatrix} d_1 & d_2 & d_3 \\ \mu_{11} & \mu_{12} & \mu_{31} \\ \mu_{12} & \mu_{22} & \mu_{32} \end{bmatrix} = \begin{matrix} C_1 \\ C_2 \end{matrix} \begin{bmatrix} d_1 & d_2 & d_3 \\ 0.8 & 0.1 & 0.4 \\ 0.2 & 0.9 & 0.6 \end{bmatrix},$$

$$\mu_{11} * \mu_{21} = 0.16, \mu_{12} * \mu_{22} = 0.09, \mu_{13} * \mu_{23} = 0.2$$

From the membership matrix *m*, the smaller the $\mu_{\lambda,j} * \mu_{m,j}$, the bigger contribution of the separation for the λ cluster and *m* cluster.

Example 2

$$m = \begin{matrix} \\ C_1 \\ C_2 \\ C_3 \end{matrix} \begin{bmatrix} d_1 & d_2 \\ \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \\ \mu_{31} & \mu_{32} \end{bmatrix} = \begin{matrix} \\ C_1 \\ C_2 \\ C_3 \end{matrix} \begin{bmatrix} d_1 & d_2 \\ 0.8 & 0.5 \\ 0.1 & 0.3 \\ 0.1 & 0.2 \end{bmatrix} \quad \mu_{11}*\mu_{21}=0.08 \ \mu_{12}*\mu_{22}=0.15$$

Shows the separation contribution of the XML document d_1 with cluster 1 and cluster 2 (C_1 and C_2) and the separation contribution of the XML document d_2 with cluster 1 and cluster 2 (C_1 and C_2).

Definition 4. Between-Group-Variance (BGV) among two clusters is defined as the following

$$BGV(\mu,\upsilon) = \frac{1}{\binom{C}{2}} \sum_{\lambda=1}^{C-1} \sum_{m=\lambda+1}^{C} \left(\frac{\sum_{j \in S_\lambda \cup S_m} \mu_{\lambda,j} * \mu_{m,j}}{|S_\lambda| + |S_m|} \right)^{-1} \left\| \upsilon_\lambda, \upsilon_m \right\|_2 \tag{7}$$

where $S_\lambda = \{ j \mid I_{\lambda,j} = 1 \}$ **and** $S_m = \{ j \mid I_{m,j} = 1 \}$, $\upsilon_\lambda, \upsilon_m$ are the vector of the central document on the λ^{th} and the *m-th* cluster separately.

From the Example 2 on the Definition 3, we know that $\frac{\sum_{j \in S_\lambda \cup S_m} \mu_{\lambda,j} * \mu_{m,j}}{|S_\lambda| + |S_m|}$ represents the separation contribution for all XML documents in the cluster λ and m. Combine the representing separation contribution using membership with between-group distance to define a between-group-variance (BGV), the smaller value of $\frac{\sum_{j \in S_\lambda \cup S_m} \mu_{\lambda,j} * \mu_{m,j}}{|S_\lambda| + |S_m|}$, the more separation and the more separation contribution. The maximized $\left\| \upsilon_\lambda, \upsilon_m \right\|_2$ is desired. So, $BGV(\mu,\upsilon)$ get larger get better. Finally, we

combine $BGV(\mu,\upsilon)$ with $WGV(\mu,\upsilon;D)$ to define a $WB(\mu,\upsilon;D)$ membership cluster validity indicator so called $WB(\mu,\upsilon;D)$ as follows.

$$WB(\mu,\upsilon;D)=\frac{(B)etweeb-(G)roup-(V)ariance(\mu,\upsilon)}{(W)ithin-(G)roup-(V)ariance(\mu,\upsilon;D)}=\frac{BGV(\mu,\upsilon)}{WGV(\mu,\upsilon;D)}$$

$$=\frac{\frac{1}{\binom{C}{2}}\sum_{\lambda=1}^{C-1}\sum_{m=\lambda+1}^{C}\left(\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}\right)^{-1}\left\|\upsilon_\lambda,\upsilon_m\right\|_2}{\sum_{i=1}^{C}\sum_{j\in S_i}(C_i)^{-1}\left\|d_j,\upsilon_i\right\|_2}$$

$$=\frac{\frac{1}{\binom{C}{2}}\sum_{\lambda=1}^{C-1}\sum_{m=\lambda+1}^{C}\frac{\left\|\upsilon_\lambda,\upsilon_m\right\|_2}{\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}}}{\sum_{i=1}^{C}\sum_{j\in S_i}\frac{\left\|d_j,\upsilon_i\right\|_2}{C_i}} \tag{8}$$

The maximized $WB(\mu,\upsilon;D)$ is better that means $\max_C WB$ is desired.

Definition 5. Precision, Recall, and *F*-Measure

Let *X* represent the set of XML documents and let $C = \{C_1, \ldots, C_k\}$ be a clustering of *X*. Moreover, let $C^\Delta=\{C_1^\Delta,\ldots C_\ell^\Delta\}$ designate the human reference classification. Then the recall of cluster C_i with respect to class C_i^Δ, R_i, is defined as $|C_i\cap C_i^\Delta|/|C_i^\Delta|$. The precision of cluster C_i with respect to class C_i^Δ, PR_i, is defined as $|C_i\cap C_i^\Delta|/|C_i|$. The *F*-Measure combines the precision and recall measures from information retrieval [22]. The *F*-Measure combines both values as follows:

$$F-Measure=\frac{2}{\frac{1}{PR_i}+\frac{1}{R_i}} \tag{9}$$

and uses the formula to evaluate the QR-FCM clustering result on the following section.

4.2 Working on Real Data Sets

We have developed a prototype and performed extended evaluation of our framework for validity clustering XML documents. We tested the performance as well as the

quality of the clustering results using real data. The prototype testbed is a java-based software that can (a) generate synthetic XML documents or use existing ones, (b) extract feature (structural summaries) from XML documents, (c) norm distance calculate pair-wise structural distances between these summaries, (d) perform the QR-Fuzzy C-mean (QR-FCM). The goal of our work is to find documents with structural similarity, that is, documents generated from a common DTD. For any choice of distance metric, we can evaluate how closely the reported lower rank dimension document matrix corresponding to the actual originated XML. The experiments were conducted as follows. The following five DTDs were downloaded from ACM's SIGMOD Record homepage [23]: OrdinaryIssuePage.dtd(O in short), ProceedingsPage1999.dtd(P-1999 in short), ProceedingsPage2002.dtd(P-2002 in short), IndexTerm1999.dtd (IT in short), Ordinary2002.dtd (Ord-2002 in short) , and Ordinary2005.dtd (Ord-2005 in short). For another real data set we used the documents on ADC/NASA [24]:70 XML documents from adml.dtd (Astronomical Dataset Markup Language DTD). Also we download the nigara data [24]: 150 XML documents from movie.dtd, department.dtd, club.dtd, and personnel.dtd. Based upon these sets of XML documents with similar characteristics, their accuracy of low rank dimension information retrieval were computed, analyzed and reported as follows. Table 1 first shows the results of QR fuzzy C-mean (QR-FCM) on the variant numbers of PESSW XML documents (50, 70, 90, and 120) from originated 3 DTDs, 4 DTDs and 5 DTDs which we called the heterogeneous XML documents, then Table 2 shows the results of QR fuzzy C-mean (QR-FCM) on the variant numbers of economic QR factorization $R = \overline{Q}^T d$ XML documents (50, 70, 90, and 120) from originated 3 DTDs, 4 DTDs and 5 DTDs , and Table 3 shows the results of QR fuzzy C-mean (QR-FCM) on the variant numbers of $\hat{r}_k = Q_k^T d$ (k=rank of PESSW) XML documents (50, 70, 90, and 120) from originated 3 DTDs, 4 DTDs and 5 DTDs and finally Table 4 shows the results of QR fuzzy C-mean (QR-FCM) on the variant numbers of $\tilde{r}_k = Q_k^T d$ (k=low rank of k) XML documents (50, 70, 90, and 120) from originated 3 DTDs, 4 DTDs and 5 DTDs. We also compute the value of F-measure on the clustering quality measure.

Table 3. QR Fuzzy C-mean (QR-FCM) on PESSW XML from variant DTDs

PESSW	3 Origin DTDs				4 Origin DTDs				5 Origin DTDs			
	O	IT	N		O	IT	N	M	O	IT	N	M	D
# XML	50	70	90	120	50	70	90	120	50	70	90	120
Space	125k	397k	484k	619k	144k	427k	564k	713k	142k	497k	518k	689k
WB	0.14	0.31	.21	.025	.09	.029	.021	.022	.08	.023	.02	.015
PR_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
R_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F-Measure	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(O)rdinaryIssuePage.dtd, (I)ndex(T)ermPage.dtd,(N)asa,(M)ove.dtd,(D)ept.dtd

Table 4. QR Fuzzy C-mean (QR-FCM) on $\bar{r}_k = \bar{Q}_k^T d$ XML from variant DTDs

$\bar{r}_k = \bar{Q}_k^T d$ k=economic rank	3 Origin DTDs				4 Origin DTDs				5 Origin DTDs				
	O	IT	N		O	IT	N	M	O	IT	N	M	D
# XML	50	70	90	120	50	70	90	120	50	70	90	120
Space	23k	45k	75k	93k	23k	44k	76k	166k	23k	133k	75k	133k
WB	0.14	.031	.21	.025	.09	.029	.02	.022	.08	.023	.02	.015
PR_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
R_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F-Measure	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(O)rdinaryIssuePage.dtd, (I)ndex(T)ermPage.dtd,(N)asa,(M)ove.dtd,(D)ept.dtd

Table 5. QR Fuzzy C-mean (QR-FCM) on $\hat{r}_k = \hat{Q}_k^T d$ XML from variant DTDs

$\hat{r}_k = \hat{Q}_k^T d$ k= rank(D)	3 Origin DTDs				4 Origin DTDs				5 Origin DTDs				
	O	IT	N		O	IT	N	M	O	IT	N	M	D
# XML	50	70	90	120	50	70	90	120	50	70	90	120
Space	4k	18k	23k	35k	6k	19k	28k	48k	7k	40k	28k	40k
WB	0.14	.031	.21	.025	.09	.029	.02	.022	.08	.023	.02	.015
PR_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
R_q	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F-Measure	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(O)rdinaryIssuePage.dtd, (I)ndex(T)ermPage.dtd,(N)asa,(M)ove.dtd,(D)ept.dtd

Table 6. QR Fuzzy C-mean (QR-FCM) on $\tilde{r}_k = \tilde{Q}_k^T d$ XML from variant DTDs

$\tilde{r}_k = \tilde{Q}_k^T d$ k= low rank	3 Origin DTDs O IT N				4 Origin DTDs O IT N M				5 Origin DTDs O IT N M D			
# XML	50	70	90	120	50	70	90	120	50	70	90	120
Space	2k	9k	11k	17k	3k	10k	14k	24k	3k	20k	14k	20k
WB	0.14	.031	.21	.025	.09	.029	.02	.022	.08	.023	.02	.015
P_R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
R_R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
F-Measure	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(O)rdinaryIssuePage.dtd, (I)ndex(T)ermPage.dtd,(N)asa,(M)ove.dtd,(D)ept.dtd

5 Conclusion and Future Work

The original XML documents $D_N=[d_1,d_2,..,d_N]$ are modeled on the vector space model according to the path element on each document, that is D_N=PESSW, then do QR factorization on the PESSW we derived the PESSW=QR ($D = \overline{Q}_k \overline{R}_k$), or $\overline{Q}_k{}^T D = \overline{R}_k$, and then take the rank on the $\hat{r}_k = \hat{Q}_k^T d$ with the rank of PESSW, and finally $\tilde{r}_k = \tilde{Q}_k^T d$ with the low-rank of QR on PESSW. We passed the 4 resulting vectors (PESSW, $\overline{R}_k$, $\hat{R}_k$, and $\tilde{R}_k$) into the QR-FCM clustering algorithm to get the clustering result. From the clustering results on the section experiment, we found the same clustering result from the variant PESSW, $\overline{R}_k$, $\hat{R}_k$ and $\tilde{R}_k$ vectors. We conclude that use the low-rank vector $\tilde{R}_k$ instead of PESSW original document, not only saving the space on the input vector but also spending less time to cluster on the documents. Based on the clustering, the next issues will be arisen such as how to index the clustered XML documents for speeding up query response, for example like R-tree and B^+-tree, and how to manage the dynamic-updating XML document for adding XML document, adding a path element, and modifying a path element weight.

References

[1] Pei, J., Han, J., Asi, B.M., Pinto, H.: PrefixSpan: Mining Sequenctial Pattern efficiently by Prefix-Projected Pattern Growth. In: Int. Conf. Data Engineering, ICDE (2001)

[2] Hwang, J.H., Ryu, K.H.: XML A New XML clustering for Structural Retrieval. In: International Conference on Conceptual Modeling (2004)

[3] Hwang, J.H., Ryu, K.h.: Clustering and retrieval of XML documents by structure. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3481, pp. 925–935. Springer, Heidelberg (2005)

[4] Lian, W., Wai-lok, D.: An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. IEEE Computer Society Technical Committee on Data Engineering (2004)
[5] Massay, W.F.: Principal components regression in exploratorystatistical research. J. Amer Statist. Assoc. 60, 234–246 (1965)
[6] Torgerson, W.S.: Theory & Methods of Scaling. Wiley, New York (1958)
[7] Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
[8] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
[9] Saul, L.K., Roweis, S.T.: Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research 4, 119–155 (2003)
[10] Donoho, D.L., Grimes, C.E.: Hessian eigenmaps: locally embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA 100, 5591–5596 (2003)
[11] Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)
[12] Zhang, Z., Zha, H.: Principal manifolds and nonlinear dimension reduction via tangent space alignment. SIAM Journal of Scientific Computing 26(1), 313–338 (2004)
[13] Kim, H., Park, H., Zha, H.: Distance preserving dimension reduction for manifold learning. In: Proceedings of the 2007 SIAM International Conference on Data Mining, SDM 2007 (2007)
[14] Dalamagas, T., Cheng, T., Winkel, K.J., Sellis, T.: A Methodology for Clustering XML Documents by Structure. Information Systems 31(3), 187–228 (2006)
[15] Gao, J., Zhang, J.: Clustered SVD strategies in latent semantic indexing. Inf. Process. Manag. 41(5), 1051–1063 (2005)
[16] Berry, M.W., Shakhina, A.P.: Computing sparse reduced-rank approximation to sparse matrices. ACM Trans. Math. Software 31(2), 252–269 (2005)
[17] Tou, J.T., Gonzalez, R.C.: Pattern Recognition Principles. Addison-Wesley, Reading (1974)
[18] Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
[19] Frigui, H., Krishnapuram, R.: A Robust Competitive Clustering Algorithm with Application in Computer Vision. IEEE Trans. Pattern Analysis and Machine Intelligence 21(1), 450–465 (1999)
[20] Everitt, B.S.: Cluster Analysis, 3rd edn. Halsted Press (1993)
[21] Maulik, U., Bandyopadhyay, S.: Genetic Algorithm Based Clustering Technique. Pattern Recognition 33, 1455–1465 (2000)
[22] Ye, Y.Q.: Comparing matrix methods in text-based information retrieval. — Tech. Rep., School of Mathematical Sciences, Peking University (2000)
[23] ACM SIGMOD Record home page, `http://www.acm.org/sigmod/record/xml`
[24] `http://www.cs.wisc.edu/niagara/data/`