International Journal of Innovations in Engineering and Technology (IJIET)
http://dx.doi.org/10.21172/ijiet.133.13

Cluster Validity on Clustering Semantic Structural Documents

Hsu-Kuang Chang[1]
[1]Department of Information Engineering, I-Shou University, Kaohsiung, Taiwan

Abstract- Recently, the amount of XML data is in increasing as electronic document systems adopt XML as the standard format in document exchange. With great increase in online information, XML document clustering takes a critical role in efficient document organization, navigation, and retrieval of a large amount of documents. In order to analyze the information represented in XML documents efficiently, researches on XML document clustering are actively in progress. Clustering a XML document collection is an ambiguous task: A clustering, i. e. a set of XML document groups, depends on the chosen clustering algorithm as well as on the algorithm's parameter settings. To find the best among several clustering result, it is common practice to evaluate their internal structures with a cluster validity measure. A clustering is considered to be useful to a user if particular structural properties are well developed. Nevertheless, the presence of certain structural properties may not guarantee usefulness from an information retrieval standpoint, say, whether or not the found XML document groups resemble the classification of a human editor. The paper in hand investigates this point: Based on already classified XML document collections we generate clustering and compare the predicted quality to their real quality. Our analysis includes the classical cluster validity measures from Dunn and Davies-Bouldin as well as the new proposed measuring WGV (within-group-variance) and BGV (between-group-variance) combining the distance measure with the μ membership within the cluster.
Keywords – WGV, BGV, DB, MDB, WB, XML

I. INTRODUCTION

Clustering is a useful method to analyze large collections of XML documents. It has the potential to identify unknown classification schemes that highlight relations and differences between XML documents. Therefore an evaluation of the results is necessary to assess their quality. In clustering tasks the procedure of evaluating the results is known under the term cluster validity [1]. Most cluster validity measures assess certain structural properties of a clustering result. If the structural properties of the outcome are well developed, then the result is considered to be of interest to the user. Because the focus is on the structural properties of a data set, these measures are also called objective measures [2]. Research on subjective measures has not been as intensive as on objective measures, and there are not many texts that discuss subjective measures in the context of document clustering [3].
We perform various clustering experiments based on the ACM's SIGMOD Record [4] and Astronomical Dataset Markup Language DTD [5]. Some extensive and good overview of clustering algorithms can be found in the literature [6]. Many different cluster validity measures have been proposed [7], [8] such as the Dunn's separation measure [9], the Bezdek's partition coefficient [10], the Xie-Beni's separation measure [11], Davies-Bouldin's measure [12], etc. A comparative examination of thirty validity measures was presented in [13] and an overview of the various measures can be found in [14].
The remainder of this paper is organized as follows. Section II introduces the investigated cluster validity measures and our proposed cluster validity measure, and Section III presents the experimental analysis results. The Section IV provides some concluding remarks.

II. CLUSTER VALIDITY MEASURES INDEX AND PROPOSED METHOD

As we face on the cluster validity problem, we should focus on the two premises. (1) The longer distance the better result for the between two groups (clusters). That means the more separation between the XML documents within the ith group and the jth group which is the smaller similarity degree and the lower relationship between these two group XML documents. (2) The distance among XML documents within the same group is the shorter the better, that is the more similarity and the stronger relationship for all XML documents belong to the ith group. Moreover, the denser and more centralize for all XML documents within the same group represent the low separation and small variance within this group. So, we take these two measurements into the consideration for the cluster validity index and describe measures on the following next section. Consider a partition of the XML documents sets $X=\{xj;j=1,2,\ldots,N\}$, and the center of each cluster, vi (i=1,2,…,c) , where N is the XML document number and c is the cluster number. In the following, we use vi to denote both cluster i and its cluster center. Let uij (i=1,2,…,c;j=1,2,…,N) is the membership of XML document j in cluster i. The C by N matrix U = [uij] is called a membership matrix. The membership matrix U is allowed to have elements with values between 0 and 1.

2.1 The modified Davies Bouldin cluster separation measurement

We introduce the cluster separation measure of the Davies and Bouldin [12]. Basically, the indicator of Davies and Bouldin use the definition of within-group-variance/between-group-variance. We modify the DB (MDB) method and describe more detail as follows.

The within-group-variance in the ith group defined as

$$S_i = \left(\frac{1}{|X_i|} \sum_{x \in X_i} TED(x - V_i) \right) \tag{1}$$

where Vi is the central tree in the ith group, |Xi| is the total numbers of XML documents in the ith group, and Si is the variance within the ith group.

The between-group-distance between the ith and jth groups defined as

$$d_{i,j} = \left\{ \sum_{s=1}^{p} TED(V_{S,i} - V_{S,j}) \right\} = |V_i - V_j|_t \tag{2}$$

where di,j is the tree-edit-distance (TED) of the central documents V_i and V_j between the ith and jth groups.

The cluster separation measurement MDB defined as

$$MDB = \frac{Within - Group - Variance}{Between - Group - Variance} = \frac{1}{C} \sum_{i=1}^{C} R_i \quad \text{and} \quad R_i = \max_{j, j \neq i} \left\{ \frac{S_i + S_j}{d_{i,j}} \right\},$$

that is , $MDB = \frac{1}{C} \sum_{i=1}^{C} R_i$, $R_i = \max_{j, j \neq i} \left\{ \frac{S_i + S_j}{d_{i,j}} \right\}$ (3)

As numerator get smaller and denominator get larger of the MDB which is smaller DB value, we get the better resulting cluster, that means $\min_{C} \{MDB\}$ is desired.

2.2. Our Proposed Method

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 ith cluster (Ci)

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

, μij is the membership calculated from the Bayes theorem or the fuzzy C-mean and $S_i = \{ j \mid I_{i,j} = 1 \}$.

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

$$WGV(\mu, V; X) = \sum_{i=1}^{C} \sum_{j \in S_i} (C_i)^{-1} TED(x_j, V_i) \tag{4}$$

Ci is the denseness within the cluster which is defined as the membership μij with the ith cluster. μij is calculated from the Bayes theorem or the Fuzzy C-mean. μij represents the membership of the XML document xj 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. Ci is the denseness determined by the μij membership. Beside, from the distance of XML document and central document, TED (xi,Vj), defined the XML document diversity within the cluster. We defined the within-group-variance as combine both of these two criteria, the bigger Ci the more dense, and the smaller TED (xi,Vj) the better result for the within-group-variance, so we have min{WGV}.

Definition 3: Between-Group-Variance (BGV) among two clusters

$$BGV(\mu,V)=\frac{1}{\binom{C}{2}}\sum_{\lambda=1}^{C-1}\sum_{m=\lambda+1}^{C}\left(\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}\right)^{-1}TED(V_\lambda-V_m) \tag{5}$$

where $S_\lambda=\{j\,|\,I_{\lambda,j}=1\}$ and $S_m=\{j\,|\,I_{m,j}=1\}$, V_λ,V_m are the central document of the λ th and the mth cluster separately.

$$\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}$$

From the Definition 3, we know that 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 $TED(V_\lambda-V_m)$ is desired. So, BGV(μ,V) get larger get better. Finally, we combine BGV(μ,V) with WGV(μ,V;X) to define a WB(μ,V;X) membership cluster validity indicator so called WB(μ,V;X) as follows.

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

$$=\frac{\frac{1}{\binom{C}{2}}\sum_{\lambda=1}^{C-1}\sum_{m=\lambda+1}^{C}\left(\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}\right)^{-1}TED(V_\lambda,V_m)}{\sum_{i=1}^{C}\sum_{j\in S_i}(C_i)^{-1}TED(x_j,V_i)}$$

$$=\frac{\frac{1}{\binom{C}{2}}\sum_{\lambda=1}^{C-1}\sum_{m=\lambda+1}^{C}\frac{TED(V_\lambda,V_m)}{\frac{\sum_{j\in S_\lambda\cup S_m}\mu_{\lambda,j}*\mu_{m,j}}{|S_\lambda|+|S_m|}}}{\sum_{i=1}^{C}\sum_{j\in S_i}\frac{TED(x_j,V_i)}{C_i}} \tag{6}$$

The maximized $WB(\mu,V;X)$ is better that means $\max_{C}^{WB}$ is desired.

III. EXPERIMENTAL EVALUATION

The goal of our work is to find documents with structural similarity, that is, documents generated from a common DTD. The experiments were conducted as follows. The following five DTDs were downloaded from ACM's SIGMOD Record homepage [4]: OrdinaryIssuePage.dtd (O in short), ProceedingsPage1999.dtd (P-1999 in short), ProceedingsPage2002.dtd (P-2002 in short), IndexTerm1999.dtd (IT-1999 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 [5]:700 XML documents from adml.dtd (Astronomical Dataset Markup Language DTD). Also we download the nigara data[5]: 1500 XML documents from movie.dtd, department.dtd, club.dtd, and personnel.dtd. Based upon these sets of XML documents with similar characteristics, their cluster validity were computed, analyzed and

reported as follows. Table 1 first shows variant numbers of XML documents form originated 2 DTDs, 3 DTDs, 4 DTDs and 5 DTDs, also computes the value of Dunn, MDB and WB from variant clusters. As we know, maximized Dunn, minimized MDB and maximized WB are desired separately, the better clustering outcome. The cluster validity ratio result between the Dunn, MDB and WB are shown on the Table 1.

Table 1 Cluster Results of the Variant XMLs from Homogeneous DTDs

# of XML Docs	Proposed Clusters	Origin 2 DTDs P-1999 IT-1999	Origin 3 DTDs Ord-2005 P-1999 IT-1999	Origin 4 DTDs Ord-2005 P-1999 P-2002 IT-1999	Origin 5 DTDs Ord-2005 Ord-2002 P-1999 P-2002 IT-1999
		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	0.53 0.86 0.31	0.53 0.86 0.18	0.53 1.57 0.12	0.53 2.57 0.1
200	3	0.24 1.04 0.22	0.24 0.78 0.22	0.24 3.17 0.15	0.24 2.75 0.1
	4	0.17 25.2 0.09	0.17 3.99 0.09	0.17 3.00 0.23	0.17 4.76 0.09
	5	0.46 28.1 0.19	0.46 3.90 0.13	0.46 11.4 0.13	0.46 3.46 0.11
		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	3.14 0.4 0.4	1.19 1.22 0.09	1.12 1.30 0.07	1.12 1.30 0.06
300	3	2.15 1.45 0.17	3.17 0.89 0.14	1.21 1.29 0.09	1.19 1.22 0.07
	4	2.04 3.63 0.09	3.0 1.15 0.10	4.72 0.80 0.17	1.52 1.42 0.09
	5	1.98 4.4 0.05	2.65 3.90 0.08	3.55 1.17 0.11	2.72 0.86 0.10
		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	3.50 0.33 0.4	1.18 1.16 0.09	1.12 1.40 0.03	1.12 1.37 0.05
400	3	2.50 1.26 0.14	4.75 0.84 0.14	1.21 1.33 0.05	1.19 1.28 0.06
	4	2.18 2.78 0.03	3.87 1.26 0.10	4.70 0.84 0.09	1.52 1.41 0.07
	5	1.46 2.29 0.02	3.77 7.38 0.08	2.88 1.16 0.07	2.72 0.89 0.08

Proceeding1999.dtd (P-1999), Proceeding2002.dtd (P-2002), Ordinary2005.dtd (Ord-2005), Ordinary2002.dtd (Ord-2002), IndexTerm1999.dtd (IT-1999)

On the following Table 2, we show the Dunn index (Dunn), modified Davis Bouldin (MDB) and within-group-variance and between-group-variance (WB) values from different XML documents on the variant heterogeneous XML clustering result. As the same criteria, maximized Dunn, minimized MDB and maximized WB are desired separately, the better clustering outcome. The cluster validity ratio result between the Dunn, MDB and WB are shown on the Table 2. The O, IT, N, M, D indicate the shorted name OrdinaryIssuePage.dtd, IndexTermPage.dtd, Nasa.dtd, Move.dtd, Dept.dtd respectively.

Table 2 Cluster Results of the Variant XMLs from Heterogeneous DTDs

# of XML Docs	Proposed Clusters	Origin DTDs IT N	Origin DTDs O IT N	Origin DTDs O IT N M	Origin DTDs O IT N M D
		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	3.14 1.18 0.31	1.06 1.37 .027	1.00 1.33 .026	1.00 1.63 .018
700	3	2.16 1.58 0.22	3.84 1.07 .031	1.05 1.65 .027	1.00 1.53 .02
	4	0.4 11.3 0.09	2.57 1.37 .03	1.18 1.14 .029	1.05 1.64 .022
	5	0.5 12.1 0.19	1.1 9.18 .02	1.12 1.21 .028	1.20 1.27 .023
		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	3.7 1.1 .02	1.06 1.4 0.02	1.0 1.33 0.01	1.00 1.62 0.016
900	3	0.42 14.2 .007	3.87 1.07 0.21	1.05 1.65 0.02	1.00 1.52 0.018
	4	0.57 11.3 .013	2.89 1.38 0.01	1.18 1.13 0.021	1.07 1.39 0.019
	5	1.21 12.4 .008	3.01 9.17 0.01	0.94 1.21 0.02	2.50 1.26 0.02
1200		Dunn MDB WB	Dunn MDB WB	Dunn MDB WB	Dunn MDB WB
	2	3.85 1.17 .015	1.07 1.84 .016	1.00 1.08 0.016	1.00 1.62 0.012
	3	2.87 1.54 .014	3.69 1.07 .025	1.07 1.18 0.02	1.00 1.54 0.014
	4	1.65 11.25 .008	0.34 10.9 .013	1.67 1.04 0.022	1.07 1.39 0.014
	5	2.94 30.2 .005	0.48 9.18 .019	1.33 1.06 0.017	2.50 1.25 0.015

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

IV.CONCLUSION

We have presented a cluster validity method using Dunn index, modified DB (MDB) and WB tests, and have presented some very accurate results. The index of the MDB method is one kinds of the adaptive clustering index, but it is totally depending on the geometric principle (distance) to measure the two XML documents difference. We also 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 and have presented wonderful satisfied results.

V. REFERENCE

[1] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering Validity Checking Methods: Part II. ACM SIGMODRecord, 1(3):19–27, 2002. ISSN 0163-5808.
[2] A. Silberschatz and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. IEEE Trans. onKnowledge and Data Engineering, 8(6), 1996.
[3] Tuzhilin. Handbook of Data Mining and Knowledge Discovery, chapter Usefulness, Novelty, and Integration of Interestingness Measures. Oxford University Press, 2002.
[4] ACM SIGMOD Record home page [http://www.acm.org/sigmod/record/xml]
[5] http://www.cs.wisc.edu/niagara/data/
[6] A. K. Jain and R. C. Dubes, Algorithms for Clustering (Data. Englewood Cliffs, NJ: Prentice Hall, New Jersey, 1988).
[7] J. C. Dunn, A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters, Journal Cybern., 3(3) 1973, 32-57.
[8] Gath, and A. B. Geva, Unsupervised Optimal Fuzzy Clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence, 11, 1989, 773-781.
[9] C. Dunn, Well Separated Clusters and Optimal Fuzzy Partitions, Journal Cybern., 4, 1974,. 95-104..
[10] C. Bezdek, Numerical Taxonomy with Fuzzy Sets, J. Math. Biol., 1, 1974, 57-71.
[11] X. L. Xie and G. Beni, A Validity Measure for fuzzy Clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence, 13(8), 1991, 841-847.
[12] D. L. Davies and D. W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Analysis and Machine Intelligence, 1(4), 1979, 224-227.
[13] G. W. Milligan and M. C. Cooper, An Examination of Procedures for Determining the Number of Clusters in a Data Set, Psychimetrika, 50, 1985, 159-179.
[14] R. Dubes and A. K. Jain, Validity studies in clustering methodologies, Pattern Recognition, 11(1), 1979, 235-253.