

Fuzzy Clustering Approach for Star-Structured Multi-Type Relational Data

Jian-Ping Mei and Lihui Chen
Division of Information Engineering, School of EEE
Nanyang Technological University
50 Nanyang Avenue, Singapore
Email: meij0002@e.ntu.edu.sg, elhchen@ntu.edu.sg

Abstract—Recently, mining interrelated data among multiple types of objects attracts a lot of attention due to its importance in many real-world applications. Despite of extensive study on fuzzy clustering of vector space data and homogeneous relational data, very limited exploration has been made on fuzzy clustering of relational data involving several object types. In this paper, we propose FC-SMR, a fuzzy approach for clustering star-structured multi-type relational data, where the central type is related to multiple attribute types. In FC-SMR, objects of the central type are clustered based on the rankings of objects of different attribute types. We formulate the clustering problem as a constrained maximization problem and give an efficient algorithm for finding local solutions of the defined objective function. Experimental studies conducted on real-world document data show the effectiveness of the new approach.

Index Terms—Fuzzy clustering; co-clustering; multi-type; heterogeneous relational data; document categorization

I. INTRODUCTION

Compared with hard clustering, fuzzy clustering which allows overlaps among clusters is generally agreed to be able to provide a more accurate and natural description of the underlying structure of real-world data. The same as k-means, most existing studies on fuzzy clustering such as fuzzy c-means (FCM) [1] deal with vector space data, of which each object is represented as a vector in some feature space, e.g. a document is represented as a vector where each feature or dimension of the vector is a distinctive word. Other than the vector-feature represented data, the pairwise relational data also has been studied for a long time [2], [3], [4]. Most typically, the relationships are similarities or dissimilarities between each pair of objects from a given data set, e.g. similarity between two documents. These document-document relation is referred as homogeneous as it exist between objects of the same type, i.e., document. Although the vector space data and homogeneous relational data are two basic data representation models based on which extensive studies have been carried out, new challenges come out when the technique of clustering is applied to the data mining domain for handling various real-world data sets, which is more natural to be represented as multi-type relational data, where inter-relations exist among objects of several types.

In some recent studies [5], [6], fuzzy co-clustering approaches have been proposed for document categorization which simultaneously cluster documents and words. Different from the two basic data representations, in co-clustering,

document and *word* are treated as two object types, and the *document-word* co-occurrence data is a bi-type heterogeneous relational data. While traditional clustering generates clusters of one type of objects, co-clustering simultaneously clusters objects of each of the two types. It has been shown in the literature that co-clustering approaches are always more effective and informative than those traditional one-side clustering approaches [7].

However, co-clustering is not applicable on data involving more than two object types. For document data sets, other than *document* and *word*, which form the *document-word* relation, document may also be related to other object types. For example, in hierarchical taxonomy mining, each document belongs to a category, which forms a *document-category* relation. The relations among objects of different types may form different structures. A star-structure is a special case where the central type is related to several attribute types as illustrated in Fig. 1. For document clustering, *document* is the central type and other object types, such as *word* and *category* are attribute types, and the *document-word* relation corresponds to a bi-type star-structure, while the *document-word* together with the *document-category* relations form a tri-type star-structure. It is possible to transform multi-type relational data to one of the basic data representation forms with respect to the object type that is interested, and then use an existing approach to get the clusters of the objects of this type. For example, a vector representation for the central-type objects may be obtained by concatenating all attribute-type objects as its features. However, this leads to a very sparse data in a high dimensional space which is a very difficult problem for most clustering approaches to solve.

To fully make use of those central-attributes relations, multi-type relational clustering approaches have been proposed in [8], [9], [10], [11], [12], [13], [14]. Most of these approaches are formulated as high-level co-clustering approaches, where [9], [11], [15], [12] formulate clustering as matrix factorization problems, [14], [13] are model-based approaches solved by the EM algorithm, [8] solves the problem with semi-definite programming and [10] is an iterative spectral clustering which needs eigenvalue decomposition at each iteration. In this paper, we develop a fuzzy clustering approach for star-structured multi-type relational data mining (FC-SMR). The proposed approach is a natural extension of the fuzzy co-clustering

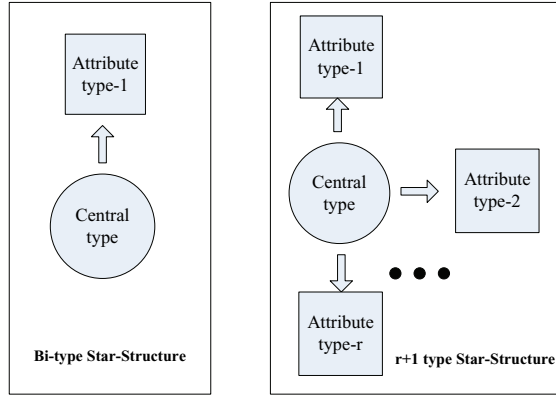


Fig. 1. Multi-type relational data with a star-structure

approach (FCoDok) in [5]. In FCoDok, a co-cluster is a document cluster associated with a word cluster, while in FC-SMR, each document cluster is associated with r clusters with each of them corresponding to an attribute type. This multi-type relational clustering problem has some similarity to the multi-view clustering [16] and collaborative clustering [17] in the sense that all these three problems make use of more than one data descriptions or views, i.e. relational matrices between the central type and different attribute types, views by different feature sets or data source distributed in multiple places. However, in [16] and [17], clustering is performed locally on each data view during which the partition information on other views are collaborated to some degree, while in the problem of this study, only a global partition of the documents is produced as an optimal solution for taking care of all relations. Moreover, the clusters of objects of each of the attribute types are also obtained.

The rest of the paper is arranged as follows: we first review some related work in Section 2. The detailed formulation of the proposed approach is then presented in Section 3. Experimental results on benchmark document data sets are reported and discussed in Section 4. Finally, we make the conclusion of this paper in Section 5.

II. FUZZY CO-CLUSTERING FOR DOCUMENTS

Several fuzzy co-clustering approaches have been proposed in the literature, and here we only review the most related one referred as FCoDok proposed in [5]. This approach is an improved version of [18] for document categorization. Given the document-word co-occurrence matrix $r_{ij} \in \mathbf{R}_{n \times m}$ of n documents and m words, the objective of FCoDok is to maximize the following criterion:

$$J_{FCoDok} = \sum_{c=1}^k \sum_{i=1}^n \sum_{j=1}^m u_{ic} r_{ij} v_{jc} - T_u \sum_{c=1}^k \sum_{i=1}^n u_{ic}^2 - T_v \sum_{c=1}^k \sum_{j=1}^m v_{jc}^2 \quad (1)$$

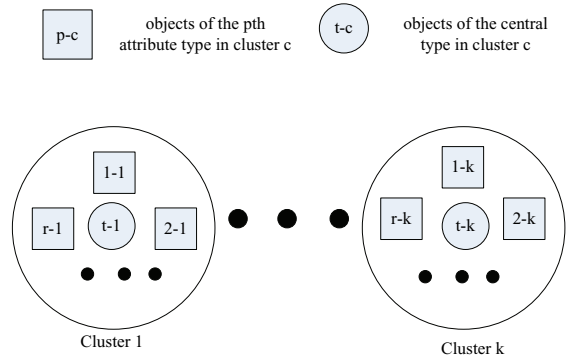


Fig. 2. The clustering problem for multi-type star-structured relational data.

subject to

$$\forall c, i, j, \quad u_{ic} \geq 0, v_{jc} \geq 0 \quad (2)$$

$$\forall i, \sum_{c=1}^k u_{ic} = 1; \quad \forall c, \sum_{j=1}^m v_{jc} = 1 \quad (3)$$

where k is the number of clusters, u_{ic} is the membership of document i with respect to cluster c , and v_{jc} is the membership of word j with respect to cluster c . From (3), it can be observed that u_{ic} is the standard fuzzy membership as for each document, the sum of its u values in all clusters is 1; differently, for each word j , the sum of its v values in all clusters is not required to be 1, instead, the sum of v values over all words in each cluster is 1. To differentiate these two kinds of measures, from now on, we refer u the membership while v the ranking. The first term of (1) called degree of aggregation measures the quality of the clustering and the second and third terms of (1) are quadratic regularization of u and v , respectively. This quadratic regularization has also been used as an alternative way to generate the fuzziness of memberships in fuzzy c-means [19].

The updating equations of u_{ic} and v_{jc} are derived by the method of Lagrange Multiplier, which is shown as below [5]:

$$u_{ic} = \frac{1}{2T_u} g_{ic} + \frac{1}{k} \left(1 - \frac{1}{2T_u} \sum_{f=1}^k g_{if}\right) \quad (4)$$

$$v_{jc} = \frac{1}{2T_v} h_{jc} + \frac{1}{m} \left(1 - \frac{1}{2T_v} \sum_{p=1}^m h_{pc}\right) \quad (5)$$

where $g_{ic} = \sum_{j=1}^m v_{jc} r_{ij}$ and $h_{jc} = \sum_{i=1}^n u_{ic} r_{ij}$. Since the above updating formulas do not guarantee nonnegativity of u_{ic} and v_{jc} , during each iteration, negative values of u_{ic} and v_{jc} are set to 0, and re-normalization is performed accordingly to satisfy the summation constraints in (3). Solutions that guarantee nonnegativity are also derived in [5] with the KKT (Karush-Kuhn-Tucker) conditions.

III. FUZZY CLUSTERING FOR STAR-STRUCTURED MULTI-TYPE RELATIONAL DATA (FC-SMR)

The FCoDok handles the *document-word* co-occurrence data, which can be regarded as bi-type star-structured relational

data, where *word* is the only attribute type. Now, we present the proposed approach FC-SMR for star-structured relational data with $r > 1$ attribute types.

A. Problem Formulation

Given a document data involving $r > 1$ attribute types, we have r matrices $\{\mathbf{R}_p\}_{p=1}^r$, where $r_{ij}^p \in \mathbf{R}_p$ is the relationship between the i th object of the central type, i.e., document x_i , and the j th object of the p th attribute type denoted as y_j^p . We assume that each relation matrix is weighted by β_p . As illustrated in Fig. 2, the problem is to produce k document clusters based on r relation matrices $\{\mathbf{R}_p\}_{p=1}^r$, and each document cluster is associated with a cluster of objects of each of the attribute types. We use u_{ic} to represent the fuzzy membership of x_i in cluster c , and v_{jc}^p the ranking of y_j^p in cluster c . The proposed approach is to maximize the following objective:

$$J_{FC-SMR} = \sum_{p=1}^r \sum_{c=1}^k \sum_{i=1}^n \sum_{j=1}^{m_p} \beta_p u_{ic} r_{ij}^p v_{jc}^p \quad (6)$$

$$- \frac{\rho}{2} \sum_{c=1}^k \sum_{i=1}^n (u_{ic})^2 - \frac{\eta_p}{2} \sum_{p=1}^r \sum_{c=1}^k \sum_{j=1}^{m_p} (v_{jc}^p)^2 \quad (7)$$

subject to

$$\forall i, c, \quad u_{ic} \geq 0; \quad \forall p, j, c, \quad v_{jc}^p \geq 0; \quad (8)$$

$$\forall i, \sum_{c=1}^k u_{ic} = 1, \quad \forall p, j, \sum_{c=1}^k v_{jc}^p = 1. \quad (9)$$

where n and m_p are the number of documents and the number of objects of the p th attribute type, respectively. The document clusters obtained as the solution of the above problem may not be the optimal with respect to the relation between *document* and each individual attribute type, but it is an overall optimum considering the relations between *document* and all attribute types. We use the classic method of Lagrange multiplier to derive the local solutions of the constrained optimization problem formulated above. The Lagrangian is formed as

$$L = J_{FC-SMR} + \sum_{i=1}^n \gamma_i \left(\sum_{c=1}^k u_{ic} - 1 \right) + \sum_{p=1}^r \sum_{c=1}^k \lambda_c^p \sum_{j=1}^{m_p} (v_{jc}^p - 1) \\ + \sum_{c=1}^k \sum_{i=1}^n \alpha_{ic} u_{ic} + \sum_{p=1}^r \sum_{c=1}^k \sum_{j=1}^{m_p} \theta_{jc}^p v_{jc}^p \quad (10)$$

where all γ_i and λ_c^p , α_{ic} , θ_{jc}^p are Lagrange multipliers. According to the KKT (Karush-Kuhn-Tucker) conditions, after some algebraic manipulations, u_{ic} can be derived as

$$u_{ic} = \begin{cases} \frac{1}{\rho} g_{ic} + \frac{1}{|\mathcal{K}_i^+|} \left(1 - \frac{1}{\rho} \sum_{f \in \mathcal{K}_i^+} g_{if} \right), & \text{for } c \in \mathcal{K}_i^+ \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

with

$$g_{ic} = \sum_{p=1}^r \sum_{j=1}^{m_p} \beta_p v_{jc}^p r_{ij}^p \quad (12)$$

where

$$\mathcal{K}_i^+ = \{f : u_{if} > 0\}$$

is the set that consists of cluster ids of which the memberships of x_i is positive, and

$$v_{jc}^p = \begin{cases} \frac{1}{\eta_p} h_{jc}^p + \frac{1}{|\mathcal{N}_c^+|} \left(1 - \frac{1}{\eta_p} \sum_{l \in \mathcal{N}_c^+} h_{lc}^p \right), & \text{for } j \in \mathcal{N}_c^+ \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

with

$$h_{jc}^p = \beta_p \sum_{i=1}^n r_{ij}^p u_{ic} \quad (14)$$

where

$$\mathcal{N}_c^+ = \{l : v_{lc}^p > 0\}$$

is the set of object ids of which the rankings in cluster c is positive. The derivation is similar as in FCoDok and we can use Procedure-K and Procedure-N in [4] to decide $\mathcal{K}_i^+$ and $\mathcal{N}_c^+$.

The first term of (11) decides the membership distribution of each x_i in k clusters while the second term is a normalization to ensure the summation constraint to be satisfied. From (12), it can be observed that given the rankings of attribute objects in each cluster, the membership of x_i in a cluster c is larger than in cluster f if more objects of r attribute types related to x_i have large ranking values in cluster c than in cluster f . Similarly, the first term in (13) decides the distribution of rankings among objects of the p th attribute type in cluster c , and the second term is there to ensure the satisfaction of the summation constraints on the ranking values. From (14), we may know that the rankings of objects of the same attribute type are only decided by the memberships of the documents. Given the memberships of documents in a cluster c , the ranking of attribute object y_i^p in this cluster is larger than that of another attribute object y_j^p if more documents related to y_i^p have larger memberships in this cluster.

B. Algorithm

For implementation, we can update $u_{ic} \in \mathbf{U}_{n \times k}$ and $v_{jc}^p \in \mathbf{V}_p(m_p \times k)$ in matrix form as below and use the same simplified way as in [5] to deal with negative memberships and rankings.

Calculate $\mathbf{U}$:

$$\mathbf{U} = \frac{1}{\rho} \mathbf{G} + \frac{1}{k} \left(-\frac{1}{\rho} \mathbf{G} \right) \quad (15)$$

$$\mathbf{G} = \sum_{p=1}^r \beta_p \mathbf{R}_p \mathbf{V}_p \quad (16)$$

For each row i of $\mathbf{U}$, find

$$\mathcal{K}_i^+ = \{f : u_{if} > 0\} \quad (17)$$

modify

$$u_{ic} = \begin{cases} 0, & \text{for } c \notin \mathcal{K}_i^+ \\ \frac{u_{ic}}{\sum_{f \in \mathcal{K}_i^+} u_{if}}, & \text{for } c \in \mathcal{K}_i^+ \end{cases} \quad (18)$$

for each p , calculate $\mathbf{V}_p$:

$$\mathbf{V}_p = \frac{1}{\eta_p} \mathbf{H}_p + \frac{1}{m_p} \left(-\frac{1}{\eta_p} \mathbf{H}_p \right) \quad (19)$$

$$\mathbf{H}_p = \beta_p \mathbf{R}_p^T \mathbf{U} \quad (20)$$

For each column c of $\mathbf{V}_p$, find

$$\mathcal{N}_c^+ = \{l : v_{lc}^\mu > 0\} \quad (21)$$

modify

$$v_{jc}^p = \begin{cases} 0, & \text{for } j \notin \mathcal{N}_c^+ \\ \frac{v_{jc}^p}{\sum_{l \in \mathcal{N}_c^+} v_{lc}^p}, & \text{for } j \in \mathcal{N}_c^+ \end{cases} \quad (22)$$

where is a matrix with a proper size, of which every entry is 1. The detailed algorithm is given as below:

Algorithm 1 FC-SMR

Input: $n \times m_p$ Relation matrix $\{\mathbf{R}_p\}_{p=\{1,2,\dots,r\}}$, the number of clusters k , parameter ρ , $\{\eta_p\}_{p=\{1,2,\dots,r\}}$, and positive weights $\{\beta_p\}_{p=\{1,2,\dots,r\}}$.

Output: Fuzzy memberships of the central objects $\mathbf{U}$ and rankings of the attribute objects $\mathbf{V}_p$ for each $p = \{1, 2, \dots, r\}$.

Method:

- 1: Initialize $\mathbf{U}^{(0)}$ with nonnegative values.
 - 2: **repeat**
 - 3: **for** each $p = \{1, 2, \dots, r\}$ **do**
 - 4: Update $\mathbf{V}_p^{(l)}$ as (19) with $\mathbf{H}_p^{(l)}$ calculated from (20).
 - 5: Modify $\mathbf{V}_p^{(l)}$ according to (21) and (22).
 - 6: **end for**.
 - 7: Update $\mathbf{U}^{(l+1)}$ as (15) with $\mathbf{G}^{(l+1)}$ calculated from (16).
 - 8: Modify $\mathbf{U}^{(l+1)}$ according to (17) and (18).
 - 9: **until** convergence
-

IV. EXPERIMENTAL RESULTS

This section presents experimental studies of the proposed FC-SMR approach for document clustering. We use the benchmark data set 20newsgroups [20].

A. Data

The original data of 20newsgroups downloaded at <http://people.csail.mit.edu/jrennie/20Newsgroups/> contains 18828 non-duplicated documents, which are categorized into 20 topics. Following [8] and [10], we generate three subsets consisting of different subtopics as listed in Table I. For example, data set TM1 consists of clusters C_1 and C_2 , where C_1 is the *rec.sport* cluster which contains two subtopics or categories on *baseball* and *hockey*, and C_2 is the *talk.politics* cluster that has three subtopics on *guns*, *mid-east* and *misc*. Each of the three data sets is formed by randomly selecting 100 documents from each of the chosen subtopics. For each data set, we obtain two relation matrices *document-word* and *document-category* which are denoted as $\mathbf{R}_1$ and $\mathbf{R}_2$, respectively. The *rainbow* toolkit [21] is used to

TABLE I
STRUCTURE OF NEWSGROUP DATASETS

TM1	C_1 : {rec.sport.baseball, rec.sport.hockey}
	C_2 : {talk.politics.guns, talk.politics.mid-east, talk.politics.misc}
TM2	C_1 : {comp.graphics, comp.os.ms-windows.misc}
	C_2 : {rec.autos, rec.motorcycles}
	C_3 : {sci.crypt, sci.electronics}
TM3	C_1 : {comp.sys.ibm.pc.hardware, comp.sys.mac.hardware}
	C_2 : {rec.autos, rec.motorcycles}
	C_3 : {sci.med, sci.space}
	C_4 : {talk.politics.guns, talk.politics.mid-east}

get the *document-word* occurrence data. Stop-words have been removed and 2000 words with the largest information gain are kept. Non-text documents are skipped by *rainbow*. After preprocessing, documents consisting of less than two words are removed and finally we have 497, 598 and 794 documents for TM1, TM2 and TM3, respectively. The *tf-idf* weighting [22] is used for weighting the words in each document. Each subtopic is a category and $\mathbf{R}_2$ is a binary matrix with an entry is 1 if the document is from that subtopic otherwise is 0.

B. Setting Up

We tested the proposed approach FC-SMR on the TM1, TM2 and TM3 three data sets. For comparison, we also implemented four existing approaches: Hyperspherical Fuzzy c-Means (HFCM) [23], fuzzy co-clustering (FCoDok) [5], Non-negative matrix factorization approach referred as (NMF-H) [12] and spectral relational clustering (SRC) [10]. HFCM is a modified fuzzy c-means clustering with cosine distance for documents; NMF-H and SRC are two existing multi-type relational clustering approaches. The results of HFCM and FCoDok are obtained based on $\mathbf{R}_1$, the *document-word* bi-type relational data. Three multi-type relational clustering approaches, FC-SMR, NMF-H and SRC are based on $\mathbf{R}_1$ and $\mathbf{R}_2$, i.e. the *word-document-category* tri-type relational data.

In our experiment, the number of clusters k is set to be equal to the real number of classes for each data set. We set $T_u = 0.001$, $T_v = 1$ for FCoDok and $\rho = 0.01$, $\eta_1 = \eta_2 = 1$ for FC-SMR. The fuzzifier $m = 1.02$ is used for HFCM. We let $\beta_1 + \beta_2 = 1$ in FC-SMR. For all fuzzy approaches, truncated clusters are obtained by assigning each object to the cluster with the largest membership.

C. FC-SMR vs. HFCM and FCoDok

We first compare the performance of the proposed FC-SMR with existing fuzzy approaches HFCM and FCoDok. Table II to Table IV give the best clustering results that HFCM, FCoDok and FC-SMR can achieve among 30 trials with random initializations on the three data sets, respectively.

It can be seen from these contingency tables that with the additional *document-category* relation, FC-SMR is able to achieve 100% accuracy on all the three data sets. The results of FCoDok and HFCM on TM1 and TM2 are close, but FCoDok performs better than HFCM on TM3. Although FCoDok and HFCM make use of the same *document-word* matrix, HFCM treats each document as a vector and use the cosine similarity to measure the closeness of two documents, while FCoDok

treats document and word as two different object types and documents are clustered based the ranking of words. There is no explicit similarity measure defined in FCoDok and FC-SMR.

Fig. 3 plots $\mathbf{u}_1 = \{u_{1,1}, u_{2,1}, \dots, u_{500,1}\}$, the membership values of documents in C_1 of TM1 produced by HFCM and FC-SMR, respectively. According to the summation constraints, the memberships in the other cluster follow $\mathbf{u}_2 = 1 - \mathbf{u}_1$. For TM1, the ground truth partition is that the first 198 documents are in C_1 : *rec.sport* and the rest are in C_2 : *talk.politics*. It can be seen from Fig. 3a that in HFCM, for the 21th document, $u_{21,1} < u_{21,2}$, although this document should be in C_1 . At the same time, six documents (labeled by circles) which should be in C_2 are assigned larger memberships in C_1 . These results indicate that the representation of these documents in terms of word frequency is not sufficient enough for correctly assigning those documents into the right clusters. The *document-category* relation help to clarify the uncertainty or correct the misleading information in the *document-word* relation. This can be seen from Fig. 3b that, when the *document-category* relation are further incorporated, those previously mis-clustered documents are now assigned reasonable memberships in two clusters which are consistent with the ground truth. Although it can be told from the document-category matrix that which documents are in the same subtopic, this matrix alone delivers no information on the hierarchy structure of the subtopics, i.e. which subtopics should be combined into a high-level cluster. Therefore, clustering only based on the *document-category* relation does not produce any meaningful result as it merely combines those subgroups into the specified number of clusters in a random way.

D. Comparison of NMI Values

Now we compare FC-SMR with four other approaches in terms of *Normalized Mutual Information* (NMI) [24]. The higher the NMI value, the more closer of the algorithm generated results to the ground truth. The NMI value equals to 1 only when the partition produced by an algorithm is identical to the ground truth of the partition. Each approach is run 30 trials on each data set. Random initializations are used for HFCM, and other approaches are initialized with document fuzzy memberships produced by HFCM. Table V shows the mean and standard deviation (numbers in brackets) of NMI values over 30 trials. For FC-SMR, we give results with different values of β_1 and β_2 .

It can be seen that FC-SMR achieves significant better results than other approaches on all three data sets. Generally, three multi-type relational clustering approaches making use of both $\mathbf{R}_1$ and $\mathbf{R}_2$ perform better than HFCM and FCoDok where $\mathbf{R}_2$ is not used. FCoDok and HFCM give close results on TM2 but HFCM outperforms FCoDok on TM1, while FCoDok performs better than HFCM on TM3. Among three multi-type relational approaches, FC-SMR is the best and NMF-H is better than SRC. Like FC-SMR, NMF-H also produces the ideal result on TM1, but its performances on the other two data sets are not as good as those of FC-SMR.

TABLE V
COMPARISON OF NMI VALUES

	TM1	TM2	TM3
HFCM	0.8862 (0.0149)	0.7183 (0.0076)	0.7862 (0.0395)
FCoDok	0.8506 (0.0138)	0.7127 (0.0024)	0.8129 (0.0408)
NMF-H	1.0000 (0.0000)	0.8008 (0.1757)	0.8925 (0.1032)
SRC	0.9595 (0.1541)	0.7207 (0.1679)	0.8053 (0.1331)
FC-SMR ($\beta_2 = 1 - \beta_1$)			
($\beta_1 = 0.1$)	1.0000 (0.0000)	1.0000 (0.0000)	0.9866 (0.0525)
($\beta_1 = [0.2, 0.8]$)	1.0000 (0.0000)	1.0000 (0.0000)	0.9917 (0.0455)
($\beta_1 = 0.9$)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

It also shows that the results of FC-SMR on three data sets do not change much with respect to β_1 in $[0.1, 0.9]$.

E. Ranking of Words and Categories

Together with the documents clusters, we also obtain the word rankings and category rankings. The ranking values of each category in each cluster is shown in the three matrices given in (23) to (25) for TM1, TM2 and TM3, respectively, where each row indicates a category and each column indicates a cluster or a topic. Top 10 words of each cluster are also listed in Table VI. As attribute types, *category* and *word* can be used for description or characterization of the document cluster. We can see that those key categories and key words which have large ranking values provide meaningful information to tell what topic each document cluster possibly related to. If we take C_1 of TM1 for example, from its associated key categories named *baseball* and *hockey*, and key words, such as *game*, *hockey*, *team*, *players*, *baseball*, we may guess that documents of this cluster are about sports, possibly on hockey and baseball. Such kind of information can be used as a short summary of the cluster which provides a quick sense of the whole document cluster without looking through each of the long documents.

$$\begin{matrix} & \textit{rec.sport} & \textit{talk.politics} \\ \textit{baseball} & 0.4327 & 0 \\ \textit{hockey} & 0.5673 & 0 \\ \textit{guns} & 0 & 0.3067 \\ \textit{mideast} & 0 & 0.3506 \\ \textit{misc} & 0 & 0.3428 \end{matrix} \quad (23)$$

$$\begin{matrix} & \textit{comp} & \textit{rec} & \textit{sci} \\ \textit{graphics} & 0.4716 & 0 & 0 \\ \textit{os.ms-windows.misc} & 0.5284 & 0 & 0 \\ \textit{autos} & 0 & 0.4579 & 0 \\ \textit{motorcycles} & 0 & 0.5421 & 0 \\ \textit{crypt} & 0 & 0 & 0.7018 \\ \textit{electronics} & 0 & 0 & 0.2982 \end{matrix} \quad (24)$$

TABLE II
CONTINGENCY TABLE OF CLUSTERING RESULTS ON TM1

	HFCM		FCoDok		FC-SMR	
	Clust 1	Clust 2	Clust 1	Clust 2	Clust 1	Clust 2
Class 1	197	1	196	2	198	0
Class 2	6	293	4	295	0	299

TABLE III
CONTINGENCY TABLE OF CLUSTERING RESULTS ON TM2

	HFCM			FCoDok			FC-SMR		
	Clust 1	Clust 2	Clust 3	Clust 1	Clust 2	Clust 3	Clust 1	Clust 2	Clust 3
Class 1	193	3	3	194	4	1	199	0	0
Class 2	2	188	10	2	190	8	0	200	0
Class 3	20	9	170	21	11	167	0	0	199

TABLE IV
CONTINGENCY TABLE OF CLUSTERING RESULTS ON TM3

	HFCM				FCoDok				FC-SMR			
	Clust 1	Clust 2	Clust 3	Clust 4	Clust 1	Clust 2	Clust 3	Clust 4	Clust 1	Clust 2	Clust 3	Clust 4
Class 1	189	4	5	1	194	1	4	0	199	0	0	0
Class 2	3	188	5	3	4	190	4	1	0	199	0	0
Class 3	4	9	180	6	6	6	182	5	0	0	199	0
Class 4	0	6	2	189	1	6	4	186	0	0	0	197

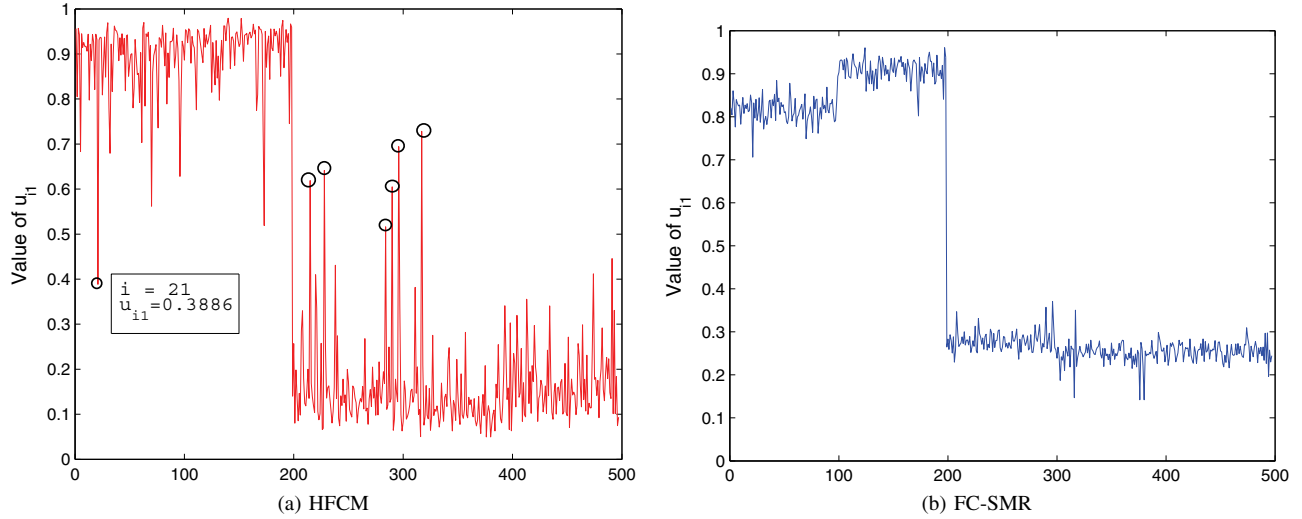


Fig. 3. Fuzzy memberships of documents with respect to C_1 of TM1 by HFCM and FC-SMR. The (horizontal) x-axis denotes the document id ($i = 1, \dots, 500$) and the (vertical) y-axis shows the value of the membership u_{i1} . In (a), small circles label objects that are clustered incorrectly.

TABLE VI
TOP 10 WORDS IN EACH CLUSTER OF THREE DATA SETS GENERATED BY FC-SMR

TM1		TM2			TM3			
C_1 : rec.sport	C_2 : talk.politics	C_1 : comp	C_2 : rec	C_3 : sci	C_1 : comp.sys	C_2 : rec	C_3 : sci	C_4 : talk.politics
game	jews	windows	car	key	mac	bike	space	israel
hockey	israel	file	bike	clipper	drive	car	nasa	israeli
team	people	graphics	dod	chip	card	dod	moon	jews
games	government	files	bmw	keys	scsi	bikes	launch	arab
players	gun	tiff	dog	nsa	apple	list	pat	arabs
ca	state	version	cars	encryption	disk	ride	shuttle	turkish
baseball	children	ftp	insurance	netcom	mb	honda	henry	jewish
season	cramer	site	bikes	phone	software	bmw	orbit	mr
win	fbi	program	audi	des	drives	oil	alaska	armenians
espn	cosmo	dos	motorcycle	government	driver	miles	mission	muslim

$$\begin{matrix}
& \text{comp.sys} & \text{rec} & \text{sci} & \text{talk.politics} \\
\begin{matrix} \text{ibm.pc.hardware} \\ \text{mac.hardware} \\ \text{autos} \\ \text{motorcycles} \\ \text{med} \\ \text{space} \\ \text{guns} \\ \text{mideast} \end{matrix} & \begin{pmatrix} 0.5265 & 0 & 0 & 0 \\ 0.4735 & 0 & 0 & 0 \\ 0 & 0.3749 & 0 & 0 \\ 0 & 0.6251 & 0 & 0 \\ 0 & 0 & 0.4056 & 0 \\ 0 & 0 & 0.5944 & 0 \\ 0 & 0 & 0 & 0.3530 \\ 0 & 0 & 0 & 0.6470 \end{pmatrix}
\end{matrix} \quad (25)$$

V. CONCLUSION

In this paper, we work on a fuzzy clustering approach for star-structured multi-type relational data analysis, where relations exist between the central type and r attribute types. The main idea of the new approach is to estimate the fuzzy memberships of the central-type objects based on the rankings of all related attribute-type objects, and the memberships of central-type objects are in turn used for updating the rankings of all attribute-type objects. Our experimental study on benchmark document data sets demonstrates the benefit of using information from multiple relation matrices for clustering as well as the feasibility and effectiveness of the proposed approach.

REFERENCES

- [1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.
- [2] R. J. Hathaway, J. W. Devenport, and J. C. Bezdek, "Relational duals of the c-means clustering algorithms," *Pattern Recogn.*, vol. 22, pp. 205–212, 1988.
- [3] R. Krishnapuram, A. Joshi, O. Nasraoui, and L. Yi, "Low-complexity fuzzy relational clustering algorithms for web mining," *IEEE Trans. Fuzzy Syst.*, vol. 9, pp. 595–607, 2001.
- [4] J.-P. Mei and L. Chen, "Fuzzy clustering with weighted medoids for relational data," *Pattern Recognition*, vol. 43, pp. 1964–1974, 2010.
- [5] K. Kummamuru, A. Dhawale, and R. Krishnapuram, "Fuzzy co-clustering of documents and keywords," in *12th IEEE Int. Conf. Fuzzy Systems*, 2003.
- [6] W.-C. Tjhi and L. Chen, "Dual fuzzy-possibilistic coclustering for categorization of documents," *IEEE Trans. Fuzzy Syst.*, vol. 17, pp. 532–543, 2009.
- [7] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proc. KDD '03*, 2003, pp. 89–98.
- [8] B. Gao, T.-Y. Liu, X. Zheng, Q.-S. Cheng, and W.-Y. Ma, "Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering," in *Proc. KDD'05*.
- [9] B. Long, X. Wu, Z. Zhang, and P. S. Yu, "Unsupervised learning on k-partite graphs," in *KDD'06*.
- [10] B. Long, Z. Zhang, X. Wu, and P. S. Yu, "Spectral clustering for multi-type relational data," in *Proc. 23th Int. Conf. Machine Learning*, 2006, pp. 585–592.
- [11] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. KDD'08*.
- [12] Y. Chen, L. Wang, and M. Dong, "Non-negative matrix factorization for semisupervised heterogeneous data coclustering," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1459–1474, 2010.
- [13] B. Long, Z. Zhang, and P. S. Yu, "A probabilistic framework for relational clustering," in *KDD'07*.
- [14] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *Proc. KDD'09*.
- [15] F. Wangy, T. Li, and C. Zhang, "Semi-supervised clustering via matrix factorization," in *SIAM International Conference on Data Mining, SDM'08*, 2008.
- [16] G. Cleuziou, M. Exbrayat, L. Martin, and J.-H. Sublemontier, "Cofkm: a centralized method for multiple-view clustering," in *Proc. 9th IEEE Int. Conf. Data Mining*, 2009.
- [17] W. Pedrycz, "Collaborative fuzzy clustering," *Pattern Recogn. Lett.*, no. 14, pp. 1675–1686, 2002.
- [18] C.-H. Oh, K. Honda, and H. Ichihashi, "Fuzzy clustering for categorical multivariate data," in *Proc. Joint 9th IFSA World Cong. and 20th NAFIPS Int. Conf.*, Vancouver, 2001, pp. 2154–2159.
- [19] S. Miyamoto and K. Umayahara, "Fuzzy clustering by quadratic regularization," in *IEEE Int. Conf. Fuzzy Systems*, 1998.
- [20] K. Lang, "NewsWeeder: learning to filter netnews," in *Proc. 12th Int. Conf. Machine Learning*, 1995, pp. 331–339.
- [21] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," 1996. [Online]. Available: <http://www.cs.cmu.edu/mccallum/bow/>
- [22] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [23] M. E. S. Mendes and L. Sacks, "Evaluating fuzzy clustering for relevance-based information access," in *12th IEEE Int. Conf. Fuzzy Systems*, 2003, pp. 648–653.
- [24] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal on Machine Learning Research*, vol. 3, pp. 583–617, 2002.