doi:10.3772/j.issn.1006-6748.2016.02.015

SCMR: a semantic-based coherence micro-cluster recognition algorithm for hybrid web data stream①

Wang Min(王 珉)②, Wang Yongbin, Li Ying
(School of Computer, Communication University of China, Beijing 100024, P. R. China)

Abstract

Data aggregation from various web sources is very significant for web data analysis domain. In addition, the recognition of coherence micro cluster is one of the most interesting issues in the field of data aggregation. Until now, many algorithms have been proposed to work on this issue. However, the deficiency of these solutions is that they cannot recognize the micro-cluster data stream accurately. A semantic-based coherent micro-cluster recognition algorithm for hybrid web data stream is proposed. Firstly, an objective function is proposed to recognize the coherence micro-cluster and then the coherence micro-cluster recognition algorithm for hybrid web data stream based on semantic is raised. Finally, the effectiveness and efficiency evaluation of the algorithm with extensive experiments is verified on real music data sets from Baidu inc. and Migu inc. The experimental results show that the proposed algorithm has better recall rate than the non-semantic micro cluster recognition algorithm and single source data flow micro cluster recognition algorithm.

Key words: hybrid web data stream, coherence micro-clustering, entity unified, object coherence, semantic computing

0 Introduction

Subject data aggregation is of great significance in the information processing domain. It requires continuous acquisition of the same type and subject data from different websites according to the specific objective. Nevertheless, the coherence search is still a difficult problem in the heterogeneous field towards the web information aggregation[1-3] due to such complex factors as multi-source, miscibility and heterogeneity.

In this study, a novel method is proposed for recognizing the micro-cluster coherence from multi-source heterogeneous data. Firstly, taking web sematic into consideration, a coherence micro-cluster recognition algorithm is proposed for hybrid web data stream, named as semantic based coherence micro-cluster recognition (SCMR). Therefore, SCMR algorithm makes full use of the soft subspace and data block partition technologies. In addition, practical applications show that our algorithm is effective for resolving the high dimensional data sparseness problem and is efficient for improving the coherence micro-cluster recognition.

1 Preliminaries

To formalize the proposed coherence micro-cluster recognition algorithm for multi-source heterogeneous web data, some formal definitions are put forward as follows.

Definition 1 Data stream model

The continuous obtained data from the same data source $X_t^1 \cdots X_t^k$ are called data stream[4]. Each data point X_t^i is a data object that includes D dimensions (Data Object, abbreviation DO), expressed as $X_t^i = DO_{ti}^1 \cdots DO_{ti}^d$.

Definition 2 Homogeneous data flow model

The data stream model of different data sources has data with the same dimensions, and the data specification is identical or similar. Assume stream data X, Y, and $\Leftrightarrow$ represents homogeneous relation, i.e. $\exists X, \exists Y$, $if(\text{Spec}(Xi) \Leftrightarrow \text{Spec}(Yi)) \cap (\text{Dim}(X) = \text{Dim}(Y)) \mapsto X \Leftrightarrow Y$. Spec( ) represents data attribute specification definition, Xi and Yi represent the i^{th} dimension of X, Y and Dim( ) represents the total number of data dimensions of the target data stream.

① Supported by the National High Technology Research and Development Programme of China (No. 2011AA120300, 2011AA120302) and the National Key Technology Support Program of China (No. 2013BAH66F02).

② To whom correspondence should be addressed. E-mail: wm_cuc@163.com

Received on Jan. 5, 2016

Definition 3 Heterogeneous data flow model

Such a model refers to the data stream model of different data sources that has data with different dimensions, or data specifications are not the same and do not match. Assume stream data X, Y, $\otimes$ represents heterogeneous relation, that is: $\exists X, \exists Y$, $if(\mathrm{Spec}(Xi) \otimes \mathrm{Spec}(Yi)) \cup (\mathrm{Dim}(X) \neq \mathrm{Dim}(Y)) \mapsto X \otimes Y$. Here Spec( ) represents data attribute specification definition, X_i and Y_i represent the i^{th} dimension of X and Y; Dim( ) represents the total number of data dimensions of the target data stream.

Definition 4 Compatible data flow model

Given heterogeneous data streams X, Y, namely $X \otimes Y$, profile $\mathrm{Prof}(X)$ of data stream X and the profile of Y have homogeneous relation $\mathrm{Prof}(X) \Leftrightarrow \mathrm{Prof}(Y)$, that is: $\exists X, \exists Y$, $if(\mathrm{Prof}(X) \Leftrightarrow \mathrm{Prof}(Y)) \cap (X \otimes Y) \mapsto X \oplus Y$. Here Prof( ) represents the data stream profile from the part of the dimension.

Definition 5 Time sliding window

When the data is continuously provided by the data source, the time sequence data stream is formed according to the order of arrival. Time data stream is the time sequence data stream, abbreviated as TDS. The storage pool of the cached data is called a sliding window, abbreviated as sW. The t time sequence of the sliding window is called $sW_t(TDS) = < X_t^1 \cdots X_t^k, RW, RD, RE >$, RW is the size of the sliding window, also called the upper limit of the size of the data object in the window, RD is the arrival time of window data, RE is the failure time of window data.

Definition 6 Single data stream micro-cluster

Assume that the current data stream is formed into micro-cluster set $mcS(sW)$ through the coherence micro-cluster recognition in time sliding window sW. Namely $mcS(sW) = < MCF_w^1 \cdots MCF_w^z, n, RD, RE >$. $MCF_w^1 \cdots MCF_w^z$ are called a series of micro clusters in a sliding window sW, $n(n \leqslant RW)$ is the actual size of the data objects in the window, arrival time and failure time of current stream data are (RD, RE).

Definition 7 Hybrid web data stream micro-cluster

Because of the existence of multiple parallel data streams, it is an assumption that there are β different sliding windows: $sW_1 \cdots sW_\beta$ at time t. In general, different single data stream sliding windows $sW_1 \cdots sW_\beta$ belong to heterogeneous data stream models, expressed as $sW_1 \otimes sW_2 \cdots \otimes sW_\beta$. Testing whether there is coherence micro-cluster between $sW_1 \cdots sW_\beta$ based on the following rules:

(1) If similar data stream profile exists among $sW_1 \cdots sW_\beta$, there may be coherence micro-cluster among $sW_1 \cdots sW_\beta$, i. e.,
$\forall sW_1 \cdots sW_\beta$, $sW_1 \otimes sW_2 \cdots \otimes sW_\beta \rightarrow \exists \mathrm{Prof}(sW_1 \cdots sW_\beta)$, $sW_1 \oplus sW_2 \cdots \oplus sW_\beta$;

(2) If a similar data stream profile does not exist among $sW_1 \cdots sW_\beta$, there is no coherence micro-cluster among $sW_1 \cdots sW_\beta$, i. e.,
$\forall sW_1 \cdots sW_\beta$, $sW_1 \otimes sW_2 \cdots \otimes sW_\beta \rightarrow ! \exists \mathrm{Prof}(sW_1 \cdots sW_\beta)$, $sW_1 \oplus sW_2 \cdots \oplus sW_\beta = false$.

If $sW_1 \oplus sW_2 \cdots \oplus sW_\beta$, after getting $\mathrm{Prof}(sW_1 \cdots sW_\beta)$, it is possible to obtain window micro-cluster set through the coherence micro-cluster recognition among similar data stream profiles $\mathrm{Prof}(sW_1 \cdots sW_\beta)$, namely $mcS(sW_1 \cdots sW_\beta)$. There will be WC $(w_1 \cdots w_\beta) = < MCF^1 \cdots MCF^\phi, m, RD, RE >$. A series of micro clusters $MCF^1 \cdots MCF^\phi$ are obtained from data streams in $w_1 \cdots w_\beta$ through the coherence micro-cluster recognition, $MCF^1 \cdots MCF^\phi$ are called a series of micro clusters between sliding windows $sW_1 \cdots sW_\beta$, $m(m \leqslant RW \times \beta)$ is the actual size of the data objects in windows $sW_1 \cdots sW_\beta$, arrival time and failure time of current stream data are (RD, RE).

2 Related work

Coherence micro-cluster recognition is the basic work of data duplication eliminating[4,5]. Related researches about data stream clustering are divided into two types: batch clustering algorithm and streaming clustering algorithm[7,8]. Reviewing the core essence of existing researches, the goal of clustering is data dividing rather than coherence search. Actually, coherence search on heterogeneous relational data sources mainly focus on conditions where data source is relatively clear and data structure is static. Therefore, the core objective of the two researches is different from the direct objective of this paper. Heterogeneous Web data stream batching does not realize the clustering of the data, but achieves the coherence micro-cluster recognition of data stream. The algorithm of clustering and the coherence micro-cluster recognition algorithm are different because of their different targets. Current studies have some limitations concerning methods and theories, and lack the suitable technologies for coherence micro-cluster recognition on heterogeneous data[9,10]. Concentrating on coherence micro-cluster recognition, this paper argues that three aspects need to be improved for the coherence search on heterogeneous data.

(1) Transformation from single data stream to hybrid Web data stream

Most existing coherence micro-clustering algorithms are used for single source, mostly deterministic

and continuous data[11,12]. However, since these algorithms only consider the distance between data objects while the heterogeneous data object, properties and other factors are ignored, they cannot be directly applied to the polymerization of heterogeneous Web data information. There are still stringent challenges in applying existing researches of heterogeneity elimination to multi-source heterogeneous Web data in the field.

(2) Transformation from cluster partition to coherence micro-cluster recognition

Existing clustering algorithms focus on data partition causing deviation from current research of coherence micro-cluster recognition[13,14]. The cluster partition will be carried out by data clustering algorithm and the latter will be carried out by data screening mechanism in order to filter out coherence micro-clusters. If the existing algorithms are used to deal with heterogeneous attribute data, non-continuous attributes have to be given up which will reduce original data information and cannot guarantee the accuracy of the clustering result.

(3) Transformation from data computing to semantic computing

Most of the existing coherence micro-cluster recognition algorithms can only process with the continuous data attribute and numeric attribute. They are limited in dealing with continuous data attribute and some of them are limited in dealing with data with tag attributes. The continuous attribute means the attribute with continuous values, such as length, weight; the tag attribute means the attribute valued for the limited state, such as color and occupation. Multi-source, heterogeneous and attribute mixed problems propose new requirements to the design and improvement of coherence micro-cluster recognition[14].

In summary, it is of great significance to design coherence micro-cluster recognition algorithm dealing with multi-source and hybrid Web data. Multi-source hybrid Web data check is an important means to ensure the quality of aggregated data. In multi-source heterogeneous Web data, finding the similarity between the heterogeneous data is the essence of coherence micro-cluster recognition problem on the basis of filtering.

3 Objective function formulation

Practice shows that there are high dimension and data sparse problems in the recognition process of multi-source heterogeneous Web data micro clusters. Analysis shows that the dimension number of the Web data objects description is often up to dozens or even more; numbers of attribute data object properties from different sources are varied and not completed and data sparse phenomenon is serious. One of the solutions of these problems is the soft subspace coherence micro-cluster recognition method, considering the influence of attributes of each micro cluster[11,12]. In this work, the core of soft subspace coherence micro-cluster recognition method dynamically selects the most important attribute, reducing redundant or secondary attributes and effectively improving the robustness, adaptability and efficiency.

The entropy that was first proposed in the Shannon information theory has been widely used in engineering technology, social economic and other fields. The basic idea of attribute entropy weight is to determine the attribute weight according to the range of the index variance. Based on the fuzzy entropy, the subspace coherence micro cluster recognition algorithm is more suitable for the processing of the current heterogeneous data. This algorithm can not only accurately deal with the different high dimensional data flow, but also effectively recognize the multiple data window using semantic strategy. To the multi-source heterogeneous data, the effect of micro cluster recognition based on SCMR is shown in Fig. 1.

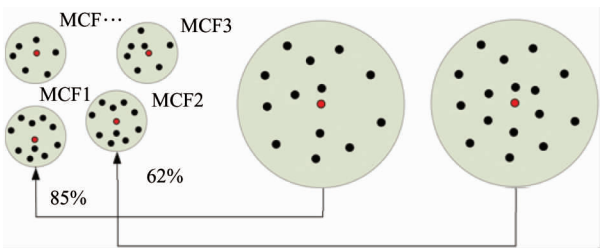

Fig. 1 Schematic diagram of the coherence micro-cluster recognition of multi-source heterogeneous data

Specifically, timing t of sequential data flow slide (TDS) window, named as sWt (TDS), which contains a data set $X = \{x_1, x_2, \cdots, x_N\} \subset R^D$ of N samples and D dimensions. Among them, x_{jk} means the value of the j sample and the k^{th} dimension. $MCF_w^1 \cdots MCF_w^c$ is a set of micro-cluster of the slide window. Using coherence micro-cluster recognition can get micro-cluster center $V = \{v_i, 1 \leqslant i \leqslant C\}$. v_{ik} means the value of the k^{th} dimension of the i^{th} micro-cluster center. The whole data set belongs to degree matrix $U = \{u_{ij} \mid 1 \leqslant i \leqslant C, 1 \leqslant j \leqslant N\}$. The u_{ij} means the j sample x_j belongs to the fuzzy membership degree of the i^{th} micro-cluster center. In order to better find micro-cluster's spatial structure characteristics, each micro cluster dimension is given a feature weighting coefficient w_{ik}, which means the

importance of the k^{th} dimensional feature to the i^{th} micro cluster center v_i.

(1) The objective function (FWCF) of the micro-clusters recognition in the continuous attribute space of the traditional fuzzy is weighted:

$$\begin{cases} FWCF(X) = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}\sum_{k=1}^{D} w_{ij}(x_{jk} - v_{ik})^2 \\ \text{s.t.}\ 0 \leqslant u_{ij} \leqslant 1, \sum_{i=1}^{C} u_{ij} = 1, 0 \leqslant w_{ik} = 1, \sum_{k=1}^{D} w_{ik} = 1 \end{cases} \tag{1}$$

In Eq. (1), the iterative method of fuzzy consensus degree u_{ij} and weight coefficient w_{ij} refers to the following Eqs(5) and (6). However, in Eq. (1), the function of FWCF does not completely consider the uncertainty effect caused by the uncertainty of micro cluster. In order to consider the uncertain factors, the entropy weighted index is introduced.

(2) The objective function (FWCF) of the recognition of the continuous attribute space of the coherence micro-cluster with entropy weight

The information entropy is introduced into the data stream of coherence micro-cluster recognition method, and the entropy means the uncertainty of the k^{th} attribute to the i^{th} micro-cluster. By introducing entropy weight index $\vartheta = -\sum_{i=1}^{C}\sum_{k=1}^{D} w_{ij} \times \ln w_{ij}$, the objective function can be shown as

$$\begin{cases} EWCF(X) = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}\left(\sum_{k=1}^{D} w_{ij}(x_{jk} - v_{ik})^2 - \sum_{k=1}^{D} w_{ij} \times \ln w_{ij}\right) \\ \text{s.t.}\ 0 \leqslant u_{ij} \leqslant 1, \sum_{i=1}^{C} u_{ij} = 1, 0 \leqslant w_{ik} = 1, \sum_{k=1}^{D} w_{ik} = 1 \end{cases} \tag{2}$$

The definition and metric calculation rules of algorithm EWCF fuzzy membership degree u_{ij}, the measure method of characteristic weighting factor w_{ij} are according to the calculation rule in Eqs(5) and (6). Because the web object attribute data type varies, data on the code, storage format, attributes and other aspects are different. Eq. (2) can only calculate continuous, numeric data. To extend support discrete data into definition 8, calculation rules of label data are defined as Eq. (3).

(3) The objective function (EWSCF) of the coherence micro cluster recognition of label attribute space

The objective function EWSCF is obtained by introducing semantic similarity measure strategy into the spatial clustering of the feature of non-continuous data:

$$\begin{cases} EWCF(X) = \sum_{j=1}^{N}\sum_{i=1}^{C} u_{ij}\left(\sum_{k=1}^{D} w_{ij}\text{SemanticSim}(x_j, v_i)^2 - \sum_{k=1}^{D} w_{ij} \times \ln w_{ij}\right) \\ \text{s.t.}\ 0 \leqslant u_{ij} \leqslant 1, \sum_{i=1}^{C} u_{ij} = 1, 0 \leqslant w_{ik} = 1, \sum_{k=1}^{D} w_{ik} = 1 \end{cases} \tag{3}$$

(4) The semantic density function of micro clusters

Assume the Web data stream in the sliding window of the unit time obtains $MCF^1 \cdots MCF^{\phi}$ micro-clusters by means of the recognition coherence micro-cluster. For micro-cluster MCF^i, assume its micro-cluster center is v_i, and x_j is the arbitrary ontology example of MCF^i. Then the semantic density of micro-cluster MCF^i is defined as:

$$\begin{cases} M_i = \sum_{x_j \in MCF^i} \text{SemanticSim}(x_j, v_i) / \| MCF^i \| \\ \text{s.t.}\ 1 \leqslant j \leqslant \| MCF^i \|, 1 \leqslant i \leqslant \phi \end{cases} \tag{4}$$

$\| MCF^i \|$ means the object scale of micro-cluster MCF^i. u_{ij} means the j^{th} sample belongs to the i^{th} micro-cluster's v_i fuzzy membership degree. The center of the set of micro-cluster $MCF^1 \cdots MCF^{\phi}$ is $V = \{v_i, 1 \leqslant i \leqslant \phi\}$. $m(m \leqslant RW \times \beta)$ is the total data object size of the window $w_1 \cdots w_{\beta}$. Ontology object instance in data stream x_j belonging to the i^{th} micro-cluster's fuzzy membership degree can be expressed as

$$\begin{cases} u_{ij} = \text{SemanticSim}(x_j, v_i) / \left(\sum_{v_\partial \in MCF^1 \cdots MCF^{\phi}} \text{SemanticSim}(x_j, v_\partial)\right) \\ \text{s.t.}\ 1 \leqslant j \leqslant m, 1 \leqslant i \leqslant \phi, 1 \leqslant \partial \leqslant \phi \end{cases} \tag{5}$$

The weighted coefficient of the k^{th} dimension importance characteristic to the i^{th} micro-cluster center can be expressed as

$$\begin{cases} w_{ik} = \sum_{x_j \in MCF^i \cap prop(k) \in x_j} \text{SemanticSim}(x_j, v_i) / \\ \qquad \sum_{x_j \in MCF^i} \text{SemanticSim}(x_j, v_i) \\ \text{s.t.}\ 1 \leqslant k \leqslant \| MCF^i \|, 1 \leqslant i \leqslant \phi, 1 \leqslant k \leqslant D \end{cases} \tag{6}$$

Further normalization of w_{ik} is performed. Now,

$$0 \leqslant w_{ik} = 1, \sum_{k=1}^{D} w_{ik} = 1$$

The recognition of the ith MCF^i micro-cluster center v_i uses existing micro-cluster MCF^i to recognize some entity λ, whose semantic similarity is greater than the other micro-cluster MCF^i critical value of the specified density radius. Measure strategy of micro-cluster v_i can be expressed as

$$\begin{cases} v_i = \lambda, \forall x_j \in MCF^i\ \text{SemanticSim}(x_j, \lambda) \geqslant r \\ \text{s.t.}\ 1 \leqslant k \leqslant \| MCF^i \|, 1 \leqslant i \leqslant \phi, r \text{ is the micro cluster radius} \end{cases} \tag{7}$$

4 Algorithm description

The coherence micro-cluster recognition process for Web-oriented multi-source heterogeneous data stream is complicated even though the algorithm SCMR still needs to meet the following requirements:

(1) **Semantic**: It is able to dynamically recognize and construct the structures of different sources, construct ontology conceptual model, evolve ontology conceptual model, and construct ontology instances dynamically.

(2) **Compatibility**: It will support a combination of heterogeneous ontology instances of the data model to recognize coherence micro-cluster recognition analysis to compute the semantic similarity between the models of heterogeneous ontology instances.

(3) **Capacity**: Because of infinite and continuous data, the limited memory and storage space for coherence micro-cluster recognition, it's impossible to store vast amounts of data, so the data need to be simplified or to be discarded selectively.

(4) **Efficiency**: The time to process each record should be as little as possible, which is able to keep up with the rate of data stream. On the condition of meeting the requirements for coherence micro-cluster recognition, it is better to scan data sets only once.

(5) **Stages**: It includes online coherence micro-cluster recognition and offline coherence micro-cluster recognition. The online coherence micro-cluster recognition algorithm does the micro-cluster recognition for data in sliding window and outputs the candidate micro-cluster sequence; The offline micro-cluster recognition algorithm does the micro-cluster recognition between the output of online algorithm and the data in database, and then outputs the coherence micro-cluster sequence.

Based on this, the process of coherence micro-cluster recognition for hybrid Web data stream is divided into three stages: semantic isomorphism, data block division, and coherence micro-cluster recognition. In the stage of semantic isomorphism, ontology concept model and instance for Web data stream need to be constructed. In addition, the mapping is created between the multi-source heterogeneous data stream ontology concept model and the standard ontology model automatically. In the stage of data block division, importance and uncertainties of entropy weight of data segmentation and division based on the properties need to be confirmed. In the stage of coherence micro-cluster recognition, a sliding window technique is used to achieve coherence micro-cluster recognition for hybrid Web data stream object. It is considered that the process of coherence micro-cluster recognition for composite satisfy the features of semantics, compatibility, capacity, efficiency and stage.

4.1 Semantic isomorphism

Web data stream is multi-source, hybrid, heterogeneous and massive, which is also presented as structured data, semi-structured data and non-structural data. That means, to construct Web entity dynamically, converting heterogeneous data to isomorphic data is required. In the polymerization process of multi-source heterogeneous data stream, three major changes need to be achieved, which are presented as follows:

(1) Converting a mixed mode to a unified model: using XML schema for structured data, semi-structured and non-structural data to make mode unity, which is benefit for the open and standardization of data structure, including but not limited to XML, TXT, CSV, Web Service, CGI, JSON, RDMS and other hybrid structure of the data stream.

(2) Converting non-semantic data to semantic data: constructing local-ontology dynamically. For different XML schema, firstly, determine whether there is a local-ontology. If it doesn't exist, it will be constructed as a temporary local-ontology according to structure information dynamically, and construct ontology instance according to structure information and temporary local-ontology; otherwise construct ontology instance directly in accordance with the existing local-ontology.

(3) Converting heterogeneous semantic data to isomorphic semantic data: First, according to field characteristics, construct a common standard ontology in the field; then achieve isomorphism which means local ontology semantic ontology map common standards ontology.

Fig. 2 describes the transformation process of the XML, TEXT, CSV, Web Service, CGI, JSON, RDMS and other hybrid structure of the data stream.

4.2 Data block division

According to compatible composite data stream window $sW_1 \oplus sW_2 \oplus sW_\beta$, common data stream profile $Prof(sW_1 Ls W_\beta)$ is first obtained, then data are processed using the above semantic isomorphism to obtain semantic space ontology instance data sets DataSet (E).

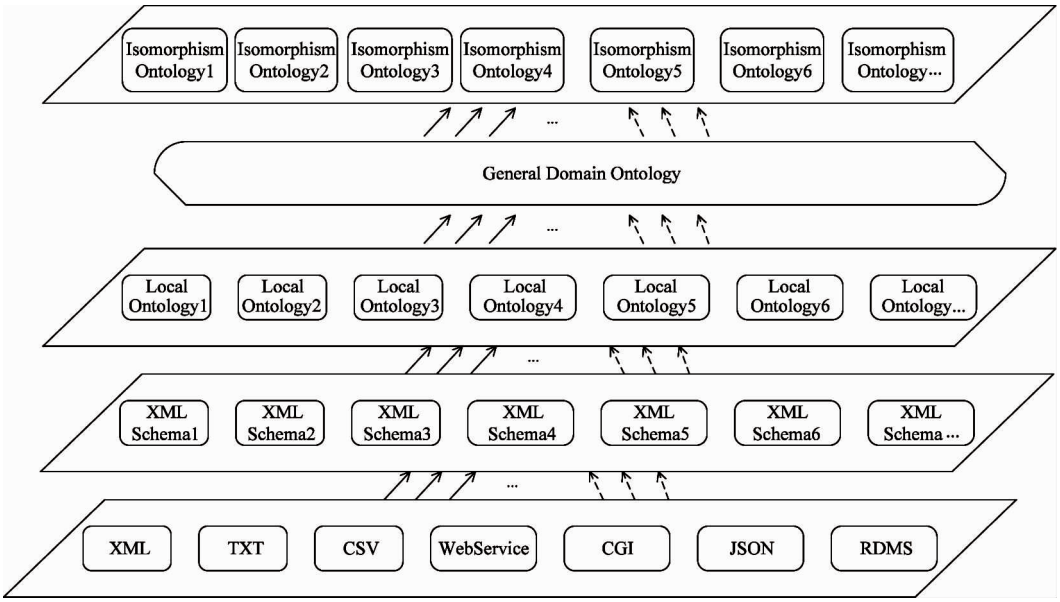

Fig. 2 The process of semantic isomorphism for multi-source heterogeneous data stream

The core principle of data block division is to consider the uncertainty of property features to the coherence micro-clusters recognition[10] and then the uncertainty is ordered by the property entropy weight index, make data set DataSet(E) for data block hierarchical division and do coherence micro-clusters recognition for the underlying data blocks again. The entropy weight index $\vartheta(prop_i) = -\sum_{j=1}^{C} w_{ij} \times \ln(w_{ij})$ of property features $prop_i$ is used to calculate the overall uncertainties of property features $prop_i$ in the history of coherence micro-clusters recognition. According to the results of coherence micro-clusters recognition, the actual impact to properties is analyzed and property entropy weight index $\vartheta(prop_i)$ is updated. In practice, the higher entropy weight index $\vartheta(prop_i)$ of property features $prop_i$ is more conducive to divide the data block, and the higher data discrimination. After finishing making data block division according to property features which has higher entropy weight index $\vartheta(prop_i)$, there is a possibility of the presence of micro-clusters among the internal data of the same data block, there is no possibility of micro clusters among different data blocks. After determining the set of data blocks, micro-cluster recognition in each data block iteratively is done. It is assumed that ontology instance data set DataSet(E) in accordance with property features which has higher entropy weight index $\vartheta(prop_i)$:

DataSet(E) = $\langle dataBlock_1, dataBlock_2, \cdots dataBlock_s \rangle$. s is the total size of the data-block sets.

4.3 Coherence micro-cluster recognition

This paper puts forward the coherence micro-cluster recognition algorithm for hybrid semantic Web data stream, *EWSCF* for short.

Algorithm SCMR

Input: Ontology instance dataset DataSet(E), the number of micro-cluster ϕ, the size of the sliding window S, and the max iterative times M of *EWSCF* algorithm.

Output: Ultimate micro-cluster center $V = \{v_i, 1 \leqslant i \leqslant \phi\}$ and micro-cluster character weighting coefficient matrix $W = \{w_{ik} \mid 1 \leqslant i \leqslant \varphi, 1 \leqslant k \leqslant D\}$. Initialize: For ϕ historical micro-cluster centers $v_i(0)$ initializes to null set, corresponding center weighted coefficient $p_i(0)$ is set as 0, and data sub-block index is set as $t = 1$.

Iterate:

(1) successively obtain a data block dataBlock$_m$($1 \leqslant m \leqslant S$) from dataset DataSet(E), and read the data sample therein;

When $t = 1$, initialize the attributive character weighting coefficient $w_{ik}(0)$, $1 \leqslant i \leqslant C, 1 \leqslant k \leqslant D$, and randomly select ϕ initial micro-cluster center $v_i(0)$, $1 \leqslant i \leqslant \phi$ from the data block dataBlock$_m$;

When $t > 1$, utilize the center and weighting coefficient, coming from a prior data subset coherence micro-cluster recognition, to initialize the micro-cluster center and micro-cluster character weighting coefficient, calculation rules see Eq. (5) and Eq. (6).

(2) For data sample and ϕ historical micro-cluster center $v_i(t = 1)$ in the data block dataBlock$_m$($1 \leqslant m \leqslant S$), weighting *EWSCF* algorithm for coherence micro-cluster recognition would be used. .

(a) Set iterator index $iteratorIndex = 1$.

(i) Use Eq. (5) to calculate the fuzzy member-

ship u_{ij} of the data sample and historical micro-cluster center to each micro-cluster center;

(ii) Use Eq. (6) to calculate the character weighting coefficient $w_{ik}(t)$ of each data cluster;

(iii) Use Eq. (7) to calculate ϕ new micro-cluster center $v_i(t)$.

(b) $iteratorIndex = iteratorIndex + 1$.

(c) Coherence step (a) and step (b), until the iterator index $iteratorIndex$ reaches maximum M or meets the stop condition of weighting $EWSCF$ algorithm.

(3) Proceed the on-line processing of micro-cluster

(a) Create micro-cluster: when t moment arrives, there's a new sample data x. According to the algorithm of Eq. (3), search for the micro-cluster which is similar to semantics of the sample data x in the existing ones. If there's no similar semantic micro-cluster, then regard the sample data as initial data, and create a micro-cluster MCF.

(b) Update micro-cluster: when t moment arrives, there's a new sample data x. According to the algorithm of Eq. (3), search for the micro-cluster which is similar to semantics of the sample data x in the existing ones. If there's more than one similar semantic micro-clusters, then according to the similarity sequence of x to the micro-cluster center $v_i(t)$, update micro-cluster MCF until its semantic similarity achieves maximum.

(c) Combine micro-cluster: when t moment arrives, according to the algorithm of Eq. (3), in the existing micro-cluster, there're more than two micro-clusters whose semantic similarity exceed a critical value, then combine the similar micro-clusters into the group micro-cluster MCF.

(d) Delete micro-cluster: as time goes on, heterogeneous model in the data stream is changing. When a micro-cluster has no new data put in for a long time, that micro-cluster MCF needs to be normalization processed, then it's called solidifying.

(4) For each micro-cluster, use Eq. (4) to calculate its semantic density.

(5) $t = t + 1$.

Until the data stream's terminal or the whole dataset is handled.

(6) According to the semantic density, sequence the micro-cluster and denoted by $MCF^1 \cdots MCF^{\phi}$.

Through above-mentioned iterative coherence micro-cluster recognition, screening for a micro-cluster sequence $MCF^1 \cdots MCF^{\phi}$ from hybrid Web data stream.

4.4 Algorithm complexity analysis

The coherence micro-cluster recognition algorithm computing complexity analysis is as follows:

D: Dimension of the data sample character space;

S: Size of the data entity scale arriving at sliding window at each moment;

ϕ: Number of the data cluster included in the whole data stream sample set;

s: Number of data block that the SCMR algorithm needs to visit.

For new arrived data sub-block, the needed computing complexity of grand design to calculate fuzzy membership u_{ij}, micro-cluster center v_i and character weighting coefficient w_{ik} of S new arriving data sample and ϕ weighted micro-cluster center, using the SCMR algorithm, is $O((S+\phi) \times \phi \times D)$. Because the maximum iterative times of the weighting $EWSCF$ algorithm is M, the computing complexity of single data sub-block to clustering partition is $O((S+\phi) \times \phi \times D \times M)$. It is supposed that coherence micro-cluster recognition algorithm for hybrid web data stream based on semantic needs to traverse s data sub-block, the ultimate computing complexity of $EWSCF$ algorithm is $O(s \times (S+\phi) \times \phi \times D \times M)$. And the computing complexity of $EWSCF$ algorithm doing semantic coherence micro-cluster recognition to the whole dataset or data stream is $O(s \times S \times \phi \times D \times M)$.

5 Experiments and results

This study, based on the Web music data aggregation and user unified access technology, has carried on the preliminary data accumulation. So far, Migu music 4.76 million data and Baidu music 8 million data are obtained. This paper adopts the way of Web search to access to volume Web music data continually. In order to ensure the quality of data after aggregation, SCMR algorithm mentioned in Section 5.3 is used to do the coherence micro-cluster recognition. To test the feasibility and innovation of the SCMR algorithm, many metrics, including semantic density within coherence micro-cluster, micro-cluster recall rate, algorithm performance etc. are verified and compared with the BloomFilter algorithm and the Streamlib algorithm. In the experiment, the data are obtained from different time intervals and different sliding windows. Several times comparisons are executed to ensure the credibility and effectiveness of the algorithm. In this experiment, RW is the width of the sliding window, also was the upper limit value of the data object scale in the

window. Suppose $RW = 50000$, there are 6 different Web music data stream and the dimension of the data stream structure is different, there existing semantic heterogeneity. RD is the arriving time of the window data, RE is the cut-off time of the window data, $T_1 \cdots T_{10} \in \{(RD, RW)\}$. And $T\Lambda = RW - RD$ is called a Web music data stream collection cycle, here set $T\Lambda = 10$h. In the 10 times of coherence micro-cluster recognition, its results show the diverse micro-cluster sequence $MCF^1 \cdots MCF^{\phi}$.

5.1 Comparison in recall and clustering purity

The clustering purity of coherence micro-cluster is defined as:

$$\begin{cases} Pur = \sum_{i=1..\phi} \left(\sum_{x_j \in MCF^i} \text{SemanticSim}(x_j, v_i) / \| MCF^i \| \right) / \phi \\ s.t.\ 1 \leqslant j \leqslant \| MCF^i \|,\ 1 \leqslant i \leqslant \phi \end{cases} \tag{8}$$

As shown in Fig. 3 and Fig. 4, by comparing the three algorithms ' capability of recognizing coherence micro-cluster the SCMR algorithm has the capability of processing heterogeneous semantic data with hybrid structure and could improve the recall capability by using entropy characteristics. Besides, comparing to BloomFilter algorithm and Streamlib algorithm, SCMR algorithm has a higher mean of coherence micro-clusters recall. After calculating the cluster purity of those three algorithms based on Eq. 8 separately, meanwhile, micro-cluster has the highest cluster purity with SCMR algorithm.

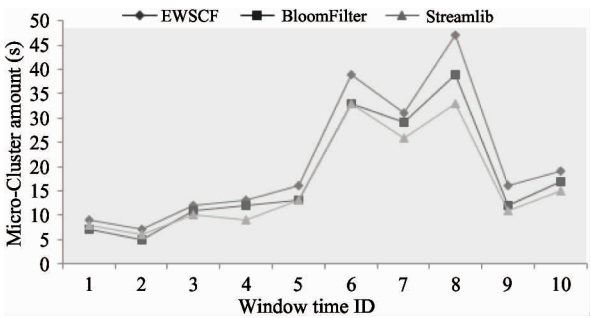

Fig. 3 Comparison of amount of micro-cluster during recognize

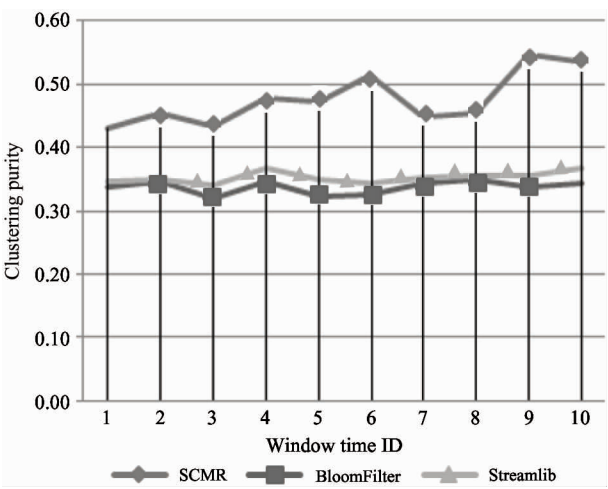

Fig. 4 Comparison of clustering purity of micro-cluster

5.2 Performance comparison of SCMR algorithm, bloom filter algorithm and streamlib algorithm

First, under 10 consecutive sliding window thresholds (each sliding window contains 50000 heterogeneous metadata), the average recognition time for coherence micro-cluster by three algorithms is compared, as shown in Fig. 5. From this figure, it is found that the time efficiency of SCMR algorithm and non-semantic streamlib algorithm are roughly the same, both of them are significantly higher than BloomFilter algorithm. This result demonstrates that SCMR is able to increase the capability of processing heterogeneous semantic data stream and the efficiency of recognizing micro-cluster by using the data block partitioning strategy based on entropy index simultaneously.

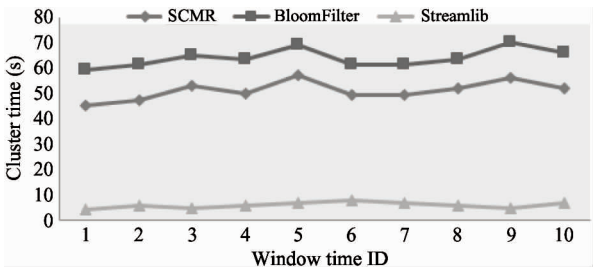

Fig. 5 Comparison of recognition time for micro cluster in sliding window

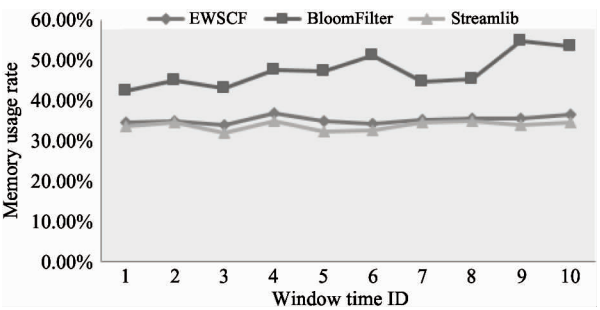

Fig. 6 Comparison of memory usage when recognizing micro cluster in sliding window

Second, the memory usage of different algorithms is compared. For each algorithm, it samples the memory resource consumption by processing 50000 heterogeneous metadata under 10 consecutive sliding window thresholds. As shown in Fig. 6, the memory usage of SCMR algorithm and non-semantic Streamlib algorithm are also roughly the same, which verifies the former

conclusion.

6 Conclusion

The coherence micro-cluster recognition for hybrid Web data stream based on semantic is still in an initial phase of the study. In order to solve the problem of coherence micro-cluster recognition with multi-source heterogeneous hybrid Web data stream, the process is divided into three phases, which are semantic isomorphic phase, data block division phase and online coherence micro-cluster recognition phase. Based on the mechanisms of fuzzy weight and entropy weight, a new optimization object function SCMR is proposed for coherence micro-cluster recognition with entropy characteristics, and by using SCMR, a coherence micro-cluster recognition algorithm is proposed for hybrid Web data stream based on semantic. The experimental results show that the proposed algorithm has better recall rate than the non-semantic micro cluster recognition algorithm and single source data flow micro cluster recognition algorithm.

Reference

[1] Wang P, Zhang L. Review and outlook of large scale data processing system for big data. *Chinese High Technology Letters*, 2015,25(08-09):793-801 (In Chinese)

[2] Yang Q K, Wang J, Ling W Q. Research on traffic heterogeneous data integration based on semantic web and cloud services. *Chinese High Technology Letters*, 2015, 25(7):694-702 (In Chinese)

[3] Hu K F, Xie J D, Zhao L. TP-mine: A partition cluster based incremental clustering algorithm for RFID trajectory data mining. *Chinese High Technology Letters*, 2014,24 (6):597-601 (In Chinese)

[4] Zhang J P, Chen F C, Li S M, et al. Data stream clustering algorithm based on density and affinity propagation. *Acta Automatica Sinica*, 2014, 40(2): 278-288 (In Chinese)

[5] Kong Q. Research on Entity Resolution Technologies in Web Data Integration: [Master dissertation]. Jinan: Computer School, Shangdong University, 2010

[6] Wang J R. Research on Mutual Promotion of Entity Resolution and Pattern Matching in Web Information Integration: [Ph. D dissertation]. Jinan: Computer School, Shangdong University, 2010

[7] Zhang C, Jin C Q, Zhou A Y. A uncertain data stream clustering algorithm. *Journal of Software*, 2010, 21(9): 2173-2182 (In Chinese)

[8] Xu Y, Huo H, Xi Ji J J, et al. A uncertain data stream subspace clustering algorithm. *Journal of Information Technology*, 2014,4:27-30 (In Chinese)

[9] Luo Q H, Peng Y, Peng X Y. A multidimensional uncertain data stream clustering algorithm. *Chinese Journal of Scientific Instrument*, 2013, 34(6): 1330-1338 (In Chinese)

[10] Yu X, Yin G S, Xu X D, et al. A data stream subspace clustering algorithm based on region partition. *Journal of Computer Research and Development*, 2014, 51(1): 88-95 (In Chinese)

[11] Zhu L, Lei J S, Bi Z Q, et al. A soft subspace clustering algorithm based on data stream. *Journal of Software*, 2013, 24(11): 2610-2627 (In Chinese)

[12] Chen L F, Guo G G, Jiang Q S. Self-adapting soft subspace clustering algorithm. *Journal of Software*, 2010, 21 (10): 2513-2523 (In Chinese)

[13] Iam-On N, Boongoen T, Garrett S, et al. A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(9):2396-2409

[14] Meesuksabai W, Kangkachit T, Waiyamaik M. HUE stream: evolution-based clustering technique for heterogeneous data streams with uncertainty. In: Proceeding of the 7th International Conference of Advanced Data Mining and Applications, Beijing, China, 2011. 27-40

Wang Min, born in 1979. She is now pursuing her Ph. D degree in Communication University of China and has received her M. S. degrees from Beijing Jiaotong University in 2004. Her research interests include the web data analysis algorithms for hybrid web data stream.