An Ant Colony-based Text Clustering System with Cognitive Situation Dimensions*

Yi Guo†
Department of Computer Science and Engineering,
East China University of Science and Technology, Shanghai, 200237, China
School of Information Science and Technology,
Shihezi University, Shihezi, Xinjiang, 832003, China
E-mail: yguo1110@ecust.edu.cn

Yan Li
School of Foreign Languages,
East China University of Science and Technology, Shanghai, 200237, China
E-mail: liyan@ecust.edu.cn,

Zhiqing Shao
Department of Computer Science and Engineering
East China University of Science and Technology, Shanghai, 200237, China
E-mail: zshao@ecust.edu.cn

Received 15 November 2013
Accepted 31 July 2014

Abstract
In order to build human cognition features into the procedure of clustering, this paper introduces a novel text clustering system, CogTCA (Cognitive Text Clustering with Ants), which (1) represents texts according to four cognitive situation dimensions in form of cognitive situation matrices and vectors rather than canonical sparse matrices of high dimensions, (2) proposes several new similarity measures among texts, and (3) implements a text clustering task as solving a combinatorial optimization problem with an encounter ant colony algorithm.

Keywords: Text Clustering; Cognitive Situation Dimensions; Situation Models; Ant Colony Optimization

* This work is financially supported by National Natural Science Foundation of China (Grant No. 61462073 and 61003126), and the Fundamental Research Funds for the Central Universities (Grant No.WH1114029) granted to Yi Guo.
† Corresponding author. Address: Department of Computer Science and Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai, 200237, China. Tel.: +86 21 64253682 / 15821071625. Email addresses: yguo71625@foxmail.com (Y. Guo), yguo1110@ecust.edu.cn (Y. Guo).
1. Introduction

Text clustering is an important processing task in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. Text clustering determines the intrinsic grouping information and partitions content relevant texts together according to similarity measures.

In the view of technology, text clustering is an unsupervised and automatic procedure of grouping text documents into clusters. Clustering algorithms group a collection of texts into natural clusters. Instances in the same cluster are similar to each other and share certain properties. The results of clustering are solely based on text representation, similarity measures and clustering algorithms.

Text clustering tasks with large vocabularies and dispersive datasets will extremely compromise the performance of clustering algorithms. Therefore, the common technique is feature selection from reduced feature dimensions. The absence of labels for guiding the clustering process is a prominent difficulty of feature selection in unsupervised text clustering. It is necessary to extract useful and typical features from high dimensional feature space with consideration of the clustering algorithm performance.

The classical clustering algorithms include the agglomerative method and the Ward's method. In current stage, the typical clustering algorithms applied (1) statistics and probability, e.g. word term frequency, word meaning frequency, and itemset frequency; (2) lexical and semantics analysis, and (3) even more complicated methods, e.g. ontologies, fuzzy clustering, self-organizing maps and knowledge-based vector space. Recent research works testify that the performance of a text clustering system can be improved with computation using conceptual, semantic and knowledge-based features.

On the one hand, the purpose of text clustering is to put similar text documents together efficiently to meet human interest in information searching and understanding. Therefore, it is essential to integrate human cognition features into the clustering procedure to present the cognitive process of text understanding or comprehension, which is one of the motivations of this research work.

An important function and property of the human cognitive system is the ability to extract important information out of textually described situations, which plays a vital role in human understanding. When humans read and comprehend a text or document, they try to build up a situation model or mental model, which describes the state of affairs in human minds. The theory of situation models, connecting many aspects of cognitive philosophy, linguistics and artificial intelligence, are the focal points in this research work.

On the other hand, a well-known challenge in text clustering is handling of text data in large volume, high dimensionality and complex semantics. If we attempt applying situation models to represent texts, the hierarchical clustering task is reformulated to a task of looking for a shortest “path” connecting all the texts in a given set (please refer to Section 3 for details).

As to the choice of computation algorithm, population-based optimization algorithms, such as genetic algorithm (GA) and ant colony optimization (ACO), have attracted a lot of attention among so many methods proposed for combinatorial optimization problems. These methods attempt to achieve better solutions by application of knowledge from previous iterations.

The genetic algorithms (GA) are stochastic global search methods, which provide a means in data mining, especially in searching poorly understood and irregular spaces. The ant colony optimization (ACO) has been proposed as a meta-heuristic approach for solving hard combinatorial optimization problems. A typical example that an ant system is apt to is the traveling salesman problem (TSP).

The basis mechanism is that ants exploit on their ground a substance called pheromone, while walking from food sources to the nest and vice versa. Ants can smell pheromone substance and, when choosing their way, they tend to choose, in probability, paths marked by strong pheromone concentrations. An important and interesting behavior of ant colonies appears to be their foraging behavior. In particular, ants are capable of finding the shortest paths between food sources and their nest without using visual cues. It has been proven in experiments that a
colony of ants can find shortest path employing this pheromone trail with following behaviors. The computational results of Aghdam and colleagues indicate that, (a) ACO algorithm achieves good enough solutions in a reasonable amount of computation time, and outperforms GA in most (9 out of 11) testing categories in the task of text feature selection.

This paper introduces a novel and effective system, CogTCA (Cognitive Text Clustering with Ants), a research effort for text clustering. Inspired by cognitive situation models, CogTCA represents texts according to four cognitive situation dimensions in form of cognitive situation matrices and vectors rather than canonical sparse matrices of high dimensions, proposes several new similarity measures among texts, and implements a text clustering task as solving a combinatorial optimization problem (to find a shortest "path" connecting all the texts in a given set) using the encounter ant colony system (E-ACS).

The rest of the paper is organized as follows. Section 2 introduces the cognitive situation models and inherent dimensions selected for text clustering. Section 3 solicits the method of converting text clustering into a traveling salesman problem. Section 4 introduces the structure of CogTCA and relevant processing details. Section 5 presents experimental results and performance evaluation. In final, Section 6 concludes this paper with concise remarks.

### 2. Cognitive Situation Models and Dimensions

The human cognitive system has the vital ability to extract important information from the textual information. After the information is extracted, how does human represent and use it for understanding? The answer is situation models, which represent the mental activities of human understanding and comprehension.

Many tasks based on language processing, including text clustering, are rationally annotated with the situation models. According to most researches in this area, a situation model consists of five different dimensions (Temporality, Spatiality, Protagonist, Causality and Intentionality), which refer to different information sources. When new information concerning any of the five dimensions is extracted, the situation model is updated according to the new information. The situation model is very useful for comprehension of texts or single sentences.

In addition, the combination and comprehension of several texts and sentences can be much better explained by the theory of situation models. For examples, situation models can integrate information across sentences; the explanation of similarities in comprehension performances across modalities can only be implemented with situation models; situation models have strong influence for effects of domain expertise on comprehension; situation models can explain the cognitive procedure of human multiple source learning.

Obviously, an important feature about situation models is the multidimensionality. To be specific, in a sentence, “Temporality” is the temporal information; “Spatiality” is the spatial information; “Protagonist” is the subject or a noun phrase (NP) that plays the role of subject, which might involve anaphoric inference with previous sentences; “Causality” is the causal connection between text elements of sentences when situation changes; “Intentionality” is the intentional connections among the goals of protagonists.

After examine above five dimensions, we conclude that (a) “Temporality”, “Spatiality” and “Protagonist” can be extracted at the syntactic level, (b) “Causality” and “Intentionality” can be perceived via analysis of predicates in sentences, as predicates take the main responsibility of information delivery in traditional English. Therefore, “Activity” is recruited as a complementary dimension to present predicates, the necessary “doing” information, in text clustering tasks. In the view of practice, the cognitive situation dimensions include Temporality, Spatiality, Protagonist, and Activity, which are implemented in CogTCA to represent texts in a concise manner.

### 3. Restate the computation procedure for text clustering task

Fig.1(a) presents two clusters, A1 and B1, in a two dimensional space for demonstration purpose. Fig.1(b) presents that, after A1 exchanges the closest point with B1, A1 becomes A2 and B1 turns into B2. Let’s define $A_1 = \{a_1, a_2, ..., a_m\}$, $B_1 = \{b_1, b_2, ..., b_n\}$, $m,N \in N$, $n \in N$. After exchange a pair of points, we get $A_2 = \{a_1, a_2, a_3, ..., a_m\}$, $B_2 = \{b_1, b_2, b_3, ..., b_n\}$.
$m \in \mathbb{N}$, $n \in \mathbb{N}$

$a_s$ and $b_s$ are the exchange points. This exchange operation illustrates that either A2 or B2 includes a heterogeneous point, which reduces the purity of either cluster before exchange. In order to measure the closeness and compactness of a cluster, we propose a parameter as follows to calculate the average distance between any pair of points in the cluster.

**Definition 1.** Define the average cluster internal distance (ACID) for cluster $A_1$ in Eq. (1).

$$\text{ACID}(A_1) = \frac{1}{C_m} \sum_{i<j} \|a_i - a_j\| \quad (1)$$

Here, $C_m$ is the number of 2-combinations in the $m$-element set ($A_1 = \{a_1, a_2, \ldots, a_m\}$). The total number of internal distances for any pair of points equals the combination of $m$ points taken 2 at a time without repetitions. $\|a_i - a_j\|$ is the Euclidean distance between $a_i$ and $a_j$ in a multi-dimension space. If $a_i$ and $a_j$ fall in a $z$ dimension space, $\|a_i - a_j\|$ is presented in Eq. (2).

$$\text{Euclid}(a_i, a_j)_z = \sqrt{\sum_{k=1}^{z} (a_{ijk} - a_{jik})^2} \quad (2)$$

The condition for sum up, $1 \leq j < k \leq n$, ensures that the Euclidean distance for a pair of points is used only once for calculating ACID. For example, as $\|a_j - a_i\|$ is equivalent to $\|a_j - a_i\|$, $\|a_j - a_i\|$ is redundant and excluded from Eq. (1). From definition 1, it is not difficult to prove (see Appendix A) that, ACID($A_1$) < ACID($A_2$) and ACID($B_1$) < ACID($B_2$), which means that, once a cluster encloses an impure element (impurity), its average cluster internal distance will increase in a notable manner.

**Definition 2.** Define the internal open route (IOR) as a sequence containing each point once and only once from a finite cluster $A_1 = \{a_1, a_2, \ldots, a_m\}$. Based on the knowledge of combinatorics, the number of IORs for cluster $A_1$ equals the number of such $m$-permutations of $m$ as denoted in Eq. (3a).

$$P_m^n = m! \quad (3a)$$

An example IOR of $A_1$, the sequence of “$a_1$-$a_2$-$\ldots$-$a_m$”, is composed with $(m-1)$ edges, “$a_1a_2$”, “$a_2a_3$”, “$a_m,a_1$”. The length of internal open route (LIOR) for current sequence is calculated with Eq. (3b).

$$\text{LIOR}(A_1) = \text{LIOR}("a_1-a_2-\ldots-a_m") = \text{Euclid}(a_1, a_2) + \text{Euclid}(a_2, a_3) + \ldots + \text{Euclid}(a_{m-1}, a_m) \quad (3b)$$

The shortest internal open route (SIOR) is denoted as the minimal value of these $m!$ LIORs.

$$\text{SIOR}(A_1) = \min_{1 \leq k \leq m!} \{\text{LIOR}(A_1)_k\} \quad (3c)$$

Based on definitions 1 and 2, the text clustering task is restated as (1) dividing all texts (points) of set $S$ in Fig.2(a) into several processed clusters ($S_1, S_2, \ldots, S_n$) in Fig.2(b) according to designated or intrinsic dimensions (“Temporality”, “Spatiality”, “Protagonist”, “Causality”, “Intentionality” and “Activity”), (2) searching the final optimal solution that satisfies following conditions as (3d).

When the average cluster internal distance (ACID) for cluster $S_i$ ($i=1,2,\ldots,n$) approaches its minimum, which indicates the most compact state of $S_i$, the length of internal open route (LIOR) for current cluster $S_i$ simultaneously turns into the shortest internal open route (SIOR) of $S_i$.

![Fig.1. Analysis of cluster internal distances](image-url)
4. Structure of CogTCA

4.1 Text representation

CogTCA is composed with seven modules (in Fig. 3). The phase of text representation includes modules “Sentence Parsing”, “Extraction of Cognitive Situation Dimensions”, “Construction of Cognitive Situation Vectors” and “Construction of Cognitive Situation Matrices”. In order to extract the four cognitive situation dimensions (“Protagonist”, “Temporality”, “Spatiality”, and “Activity” described in Section 3) in an accurate manner, the cognitive interactionist sentence parser is applied to parse each sentence and obtain a corresponding syntactic structure, which consists of a set of labeled links connecting pairs of words, and a constituent-tree containing conventional constituents (e.g., noun phrases, verb phrases, and prepositional phrases).

In the module of “Extraction of Cognitive Situation Dimensions”, WordNet is referred to facilitate semantic analysis. As a useful tool for computational linguistics and natural language processing, WordNet groups nouns, verbs, adjectives and adverbs into distinct sets of cognitive synonyms (synsets), which are interlinked with conceptual-semantic and lexical relations.

Temporality is the temporal information in each sentence. In most conditions, the temporal information contains a time-relevant noun phrase. Spatiality should satisfy two constraints: a noun phrase and location-relevance. Activity, normally identified as a verb phrase or predicate, is the dominant part of a sentence. Identification of activities relies on the extraction of verb phrases from constituent-trees. Protagonists also rely on the noun phrases, appearing mostly as subjects and partially as objects, and anaphoric inference is an unavoidable language phenomenon. The corresponding extraction algorithms are based on our previous work.

A notable capability of human cognitive systems is to extract the most dominant information from textual contents. Therefore, human beings can represent the context in multiple dimensions (including central concepts in theories of situated cognition) and monitor situational
information for understanding and comprehension. To simulate above cognitive features, each text is presented with a set of situation vectors, which are composed of four dimensions extracted from each sentence. A cognitive situation vector is defined in Eq. (4).

**Definition 3.** Define a cognitive situation vector (CSV) as
\[ SV_i = (P_i, A_i, T_i, S_i), \quad i \in \{1, 2, \ldots m\} \]  
\[ m \] is the total number of sentences in a text, \( i \) is the sequence of a sentence in a text.

For the \( i^{th} \) sentence, \( SV_i \) presents corresponding situation vector, \( P_i \) is the subject/object information, Protagonist, \( A_i \) is the Activity information, \( T_i \) is the Temporality information, \( S_i \) is the Spatiality information.

Based on definition 3, a text is represented with a cognitive situation matrix (CSM) in Eq. (5). A CSM \((M_p)\) is filled with a set of situation vectors \((SV_{p,i})\), each of which is composed with cognitive situation elements, e.g. words, phrases, and clauses.

**Definition 4.** Define a cognitive situation matrix (CSM) as
\[ M_p = (SV_{p,1}, SV_{p,2}, \ldots, SV_{p,i}, \ldots, SV_{p,m})^T, \quad 1 \leq i \leq m, \quad 1 \leq p \leq k \]  
\( m \) is the total number of sentences in a text, \( k \) is total number of texts in collection \( D \), \( p \) is a subscript to identify different texts in a collection, \( T \) presents the transpose operation.

**4.2 Similarity of cognitive situation matrices**

In order to compare any pair of texts, we need to define and calculate the correlation between their CSMs.

**Definition 5.** If the CSMs of any two texts are represented with \( M_p \) and \( M_q \) as Eq.s (6) and (7), the correlation between \( M_p \) and \( M_q \) is defined as \( Cor(M_p, M_q)^T \) and calculated with Eq. (8).

\[ M_p = (SV_{p,1}, SV_{p,2}, \ldots, SV_{p,i}, \ldots, SV_{p,m})^T \]  
\[ M_q = (SV_{q,1}, SV_{q,2}, \ldots, SV_{q,j}, \ldots, SV_{q,n})^T \]  
\[ 1 \leq i \leq m, \quad 1 \leq j \leq n, \quad 1 \leq p, q \leq k \]  
\[ Cor(M_p, M_q)^T \]  
\[ Cor(M_p, M_q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (SV_{p,i} \cdot SV_{q,j})}{\sqrt{\sum_{i=1}^{m} (SV_{p,i})^2 \cdot \sum_{j=1}^{n} (SV_{q,j})^2}} \]  
\[ \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} (SV_{p,i} \cdot SV_{q,j}) \]
In Eq. (8), the correlation between any pair of CSVs is expressed and calculated in Eq. (9), which heavily relies on the semantic relations upon these cognitive dimensions.

\[
Cor(SV_{p,i}, SV_{q,j}) = w_p \cdot \text{Sem}(P_{p,i}, P_{q,j}) + w_A \cdot \text{Sem}(A_{p,i}, A_{q,j}) + w_T \cdot \text{Sem}(T_{p,i}, T_{q,j}) + w_S \cdot \text{Sem}(S_{p,i}, S_{q,j})
\]  

(9)

\[w_p, w_A, w_T, w_S\] present the weights for dimension Protagonist, Activity, Temporality and Spatiality, and satisfy following conditions: \(0 < w_p, w_A, w_T, w_S < 1\), and \(w_p + w_A + w_T + w_S = 1\). The four weights are equally initialized with empiristic value 0.25, and also dynamically adjusted at the end of each main loop during the computing procedure (please refer to Algorithm-1 (3h)).

In Eq. (9), \(\text{Sem}(P_{p,i}, P_{q,j})\) present the semantic product between \(P_{p,i}\) and \(P_{q,j}\) which indicates whether \(P_{p,i}\) and \(P_{q,j}\) are semantically equal or not. The semantic product of \(P_{p,i}\) and \(P_{q,j}\) is described in Eq. (10). If and only if \(P_{p,i}\) and \(P_{q,j}\) are synonyms or phrases of same meanings, they are treated lexically or semantically equal.

\[
\text{Sem}(P_{p,i}, P_{q,j}) = 1, \text{ when } P_{p,i} \text{ and } P_{q,j} \text{ are lexically or semantically equal} \tag{10}
\]

\[\text{Sem}(P_{p,i}, P_{q,j}) = 0, \text{ otherwise}\]

Eq. (10) also applies to the dimensions of Activity \((A_{p,i}, A_{q,j})\), Temporality \((T_{p,i}, T_{q,j})\), and Spatiality \((S_{p,i}, S_{q,j})\).

In Eq. (8), if we use \(C_{ij}\) \((1 \leq i \leq m, 1 \leq j \leq n)\) to replace \(Cor(SV_{p,i}, SV_{q,j})\), Eq. (8) can be denoted as Eq. (8\(^a\)) in a simple form.

\[
Cor(M_p, M_q)^T = \begin{pmatrix}
SV_{p_1,1} & SV_{p_1,2} & \cdots & SV_{p_1,n} \\
SV_{p_2,1} & SV_{p_2,2} & \cdots & SV_{p_2,n} \\
\vdots & \vdots & \ddots & \vdots \\
SV_{p_m,1} & SV_{p_m,2} & \cdots & SV_{p_m,n}
\end{pmatrix} = \begin{pmatrix}
Cor(SV_{p_1,SV_{q_1}}) & Cor(SV_{p_1,SV_{q_2}}) & \cdots & Cor(SV_{p_1,SV_{q_n}}) \\
Cor(SV_{p_2,SV_{q_1}}) & Cor(SV_{p_2,SV_{q_2}}) & \cdots & Cor(SV_{p_2,SV_{q_n}}) \\
\vdots & \vdots & \ddots & \vdots \\
Cor(SV_{p_m,SV_{q_1}}) & Cor(SV_{p_m,SV_{q_2}}) & \cdots & Cor(SV_{p_m,SV_{q_n}})
\end{pmatrix} \tag{8}
\]

**Definition 6.** To measure the closeness between any pair of texts using their CSMs \((M_p, M_q)\), a new parameter **Similarity** is computed with Eq. (11) based on above correlation matrix in Eq. (8\(^a\)).

\[
\text{Similarity}(M_p, M_q) = \frac{1}{m} \left( \sum_{i=1}^{m} \left( \max_{1 \leq j \leq n} \{C_{ij}\} \right) \right) \tag{11}
\]

where \(\max_{1 \leq j \leq n} \{C_{ij}\}\) is the maximum value in the \(i^{th}\) row in \(Cor(M_p, M_q)^T\). In another word, when comparing the \(i^{th}\) sentence of \(M_p\) with each sentence of \(M_q\), only the \(i^{th}\) sentence of \(M_q\) and its most similar sentence of \(M_p\) are decisive elements for the similarity. The similarity is the average of maximum values for each row in matrix \(Cor(M_p, M_q)^T\). As \(C_{ij} \in [0,1]\), \(\max_{1 \leq j \leq n} \{C_{ij}\} \in [0,1]\).

Therefore, the value of **Similarity** falls in the range of \([0,1]\) as well.

**Definition 7.** Define another parameter **CogDist** to describe the cognitive semantic distance between \(M_p\) and \(M_q\) as follows.

\[
\text{CogDist}(M_p, M_q) = [\text{Similarity}(M_p, M_q)]^{-1} = \text{Similarity}^{-1}(M_p, M_q) \tag{12}
\]

The value of a **CogDist** is inversely proportional to the **Similarity**, which means that the more similar two texts are, the closer they are in the cognitive semantic distance. As **Similarity**(\(M_p, M_q\)) \(\in [0,1]\), **CogDist**(\(M_p, M_q\)) \(\in [1, \infty)\).

### 4.3 Construction of an ant colony system for clustering

Based on definitions 1, 2 and 7, the best solution for the clustering a text collection \(D\) into appropriate clusters \((D_1, D_2, ..., D_n)\) is interpreted as searching for the shortest internal open route (SIOR) of \(D\), which connects all points (texts). Meanwhile, each cluster \(D_i\) \((1 \leq i \leq n)\) reaches its minimum value ACID\(_{min}(D_i)\).

\[
\text{ACID}(D_i) = \frac{2}{m \cdot (m-1)} \sum_{i \neq j \leq n} \text{CogDist}(M_j, M_k) \rightarrow \text{minimum} \tag{13}
\]
m is the number of texts in \( D_i \), \( M_i \) and \( M_k \) are the CSMs for any pair of text in \( D_i \).

In this condition, SIOR(\( D_i \)) and SIOR(\( D \)) reach the state to produce the shortest internal open routes

\[
\text{SIOR}(D) = \left( \sum_{\delta \in \text{CSM}} \text{SIOR}(D_i) \right) + \text{BridgeSum} \tag{14}
\]

\( \text{BridgeSum} \) is the sum-up of cognitive distances in the shortest internal open route (SIOR) connecting clusters \( (D_1, D_2, \ldots, D_n) \) together.

So far, clustering texts is transformed into a combinatorial optimization problem, suitable for approaching with an ant colony optimization system\(^{24}\).

Traditional ant colony algorithm has advantages in computation convergence and limitations in computing speed and achieving diverse solutions\(^{25}\).

To avoid these limitations, based on the encounter phenomena of ants, this paper implements an improved ant colony system (E-ACS) constructed as follows. Its core optimization algorithm is presented in Algorithm-1 (pseudo code). In E-ACS, when an ant has visited more than half texts, it will try to exchange “information” with another ant. If the sum-up of visited texts exceeds the total text number, replica texts will be eliminated to update a new list of visited texts and the pheromone trail of the edges visited by the two ants.

Suppose that \( m \) artificial ants are assigned within \( n \) texts. An artificial ant \( k \) (at text \( i \)) chooses the text \( j \) to move to among those which do not belong to its working memory \( N_k \) by applying the following probabilistic formula:

\[
p^k_j(t) = \begin{cases} \frac{\tau^k_{ij}(t)\eta^\alpha_{ij}}{\sum_{i \in N_k} \tau^k_{ij}(t)\eta^\alpha_{ij}} & j \notin N_k \\ 0 & \text{otherwise} \end{cases}
\tag{15}
\]

\( p^k_j(t) \) is the probability with which ant \( k \) chooses to move from text \( i \) to text \( j \). \( \tau^k_{ij}(t) \) is the amount of pheromone trail on edge \((i, j)\) at time \( t \), \( \eta^\alpha_{ij} \) is a heuristic function, which was chosen to be the inverse of the cognitive semantic distance between texts \( i \) and \( j \). \( \alpha \) is a parameter which weighs the relative importance of pheromone trail and of closeness. \( \beta \) is a parameter which weighs the relative importance of the heuristic function. This formula favors those edges which are shorter and have a higher level of pheromone trail.

The pheromone trail is changed both locally and globally. Global updating is intended to reward edges belonging to shorter tours. As soon as all artificial ants have completed their tours in \( n \) texts, the best ant deposits pheromone on its visited edges. The other edges remain unchanged.

In the global trail updating formula as follows,

\[
\tau^k_{ij}(t + \Delta t) \leftarrow (1 - \rho)\tau^k_{ij}(t) + \rho \sum^m_{k=1} \Delta \tau^k_{ij}
\tag{16}
\]

the amount of pheromone \( \sum^m_{k=1} \Delta \tau^k_{ij} \) deposited on each visited edge \((i, j)\) by the best ant is inversely proportional to the length of the tour: the shorter the tour, the greater the amount of pheromone deposited on edges, which can be presented in Eq. (17), if \( T_k \) is defined as the length of the tour. \( \rho \) is the parameter for pheromone reinforcement, and \( 1 - \rho \) is the parameter for pheromone evaporation.

This manner of depositing pheromone is intended to emulate the property of differential pheromone trail accumulation, which in the case of real ants was due to the interplay between the length of the path and continuity of time.

\[
T_L \propto \left( \sum^m_{k=1} \Delta \tau^k_{ij} \right)^{-1},
\]

when \( \sum^m_{k=1} \Delta \tau^k_{ij} \rightarrow \text{maximum} \), \( T_L \rightarrow \text{minimum} \n\tag{17}\]

\[
\Delta \tau^k_{ij} = \begin{cases} Q \quad & \text{if ant } k \text{ visits the edge } (i, j) \text{ in this loop cycle} \\ 0 \quad & \text{otherwise} \end{cases}
\tag{18}
\]

where \( Q \) is a computation constant to mark the initial strength of pheromone when an ant starts a new loop, and \( L_k \) is the total distance that ant \( k \) visits in a loop. In this paper, the empiristic value of \( Q \) is 36.

\textbf{Algorithm-1: The encounter ant colony system (E-ACS)}

\textbf{Begin}

(1) Initialize and prepare computing parameters

\text{(texts, ant number, CogDist, } N_k, \alpha, \beta, \tau_0(0), \text{ } M, \rho, \text{ } w_p, W, W_2, W_3, \ldots)\text{;}

(2) \text{FOR } i = 1, 2, \ldots, M /* \text{main loop for iteration} */
FOR each ant /* Eq. (15)-(18) are used */
(3a) Building up its own route
(3b) IF (half texts have been visited)
(3c) THEN Loop Start:
(3d) IF ((the total number of texts visited by any two ants) > (the total text number))
THEN Label the two ants as "encounter ants";
Eliminate replica texts;
Build a new list of visited texts for the two ants;
Update the pheromone trail of the edges visited by the two ants;
ELSE continue loop
END IF
(3e) IF ((the number of encounter ants) > (the threshold ?))
THEN Build an encounter route;
Update the pheromone trail of the edges in the encounter route;
End and Exit
ELSE continue loop
END IF
(3f) Visit a new text
(3g) Loop End:
END IF
END FOR /* end of main loop for iteration */

4.4 Produce clustering results

The previous computing procedure builds up a shortest internal open route connecting all texts together. The length of any edge presents the cognitive semantic distance between texts on both ends. A top-down splitting operation is required to produce the final clustering result, which depends on the required number of clusters or the threshold of diameter of each cluster. Two different splitting algorithms, number-based (NBSA) and diameter-based (DBSA), are presented as follows. In NBSA, saying \( N = 20 \), which states that the target texts are required to be divided into 20 clusters. The more clusters required, the result is more fine-grained. In DBSA, the diameter threshold \( D_{\text{dthres}} \) has direct impact on the granularity of clustering results. The shorter the diameter threshold \( D_{\text{dthres}} \) is, the more fine-grained clusters we have. NBSA is more efficient and decisive (or arbitrary sometimes), while DBSA is more practical in control of clustering granularity as implemented in this paper.

**Algorithm-2**: The number-based splitting algorithm (NBSA)

**Begin**
(1) Set up the cluster number \( N \);
/* \( N \) is a pre-assigned number of required clusters for the splitting result */
(2) Search for the longest \((N-1)\) edges;
(3) Cut off these \((N-1)\) edges;
(4) Produce each segment as a cluster.
**End**

**Algorithm-3**: The diameter-based splitting algorithm (DBSA)

**Begin**
(1) Define the diameter as the longest cognitive semantic distance between any pair of texts in a cluster;
(2) Set up the diameter threshold \( D_{\text{dthres}} \);
/* \( D_{\text{dthres}} \) is a pre-assigned value to determine the granularity of the splitting result */
(3) Present the shortest internal open route as \( \{d_1, d_2, \ldots, d_n\} \), \( n \) is the total number of texts, \( d_i \) and \( d_e \) are the start and end points;
(4) Initialize the cluster counter \( k = 1 \);
(5) FOR \((i=1; i<n+1; i++)\)
Put \( d_i \) in queue \( Q_k \rightarrow \{d_1, d_2, \ldots, d_i\} \)
IF \( \max_{d_p,d_q \in Q_k} \text{CogDist}(d_p, d_q) > D_{\text{dthres}} \)
THEN \( Q_k \{d_i, \ldots, d_k\} ; \)
\( k = k + 1 \);
\( Q_k \{d_i\} \)
END IF
END FOR
(6) The final clusters \( \rightarrow \{Q_1, Q_2, \ldots, Q_k\} \)
**End**
5. Experiments and evaluation

5.1 Implementation and text corpus

CogTCA is developed with Perl/Java in Windows 7 Ultimate / Fedora 14 using a personal computer (Dell Precision T5500, Intel® Xeon® Quad Core (E5620) 2.4GHz, 12GB DDR3 RDIMM, 1333MHz). The following experiments are also implemented in the same hardware and software environment.

We use Reuters-21578 and RCV1-v2 in our experiments. Reuters-21578, currently the most widely used test collection for text processing research, includes 21578 tagged documents and 669 indexed categories spanning across 9 various domains (shown in table 2). RCV1-v2 contains 35 times as many newswire stories (804,414 for RCV1-v2) as the popular Reuters-21578 collection. These texts are organized in four hierarchical groups (top level topic categories): CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). RCV1-v2 provides benchmark data on all categories, which include: (1) 103 Topic categories—101 with one or more positive training examples on the training set and all 103 (including all the 101) have one or more positive test examples on the test set, (2) 354 Industry categories—313 with positive training examples, and 350 (including all of the 313) with positive test examples, and (3) 366 Region categories—228 with positive training examples and 296 (including all of the 228) with positive test examples. The standard clusters are constructed on the basis of the category label for each text.

5.2 Evaluation methods

Three metrics, Purity, F-Score and normalized mutual information (NMI)\(^{13}\), shown in Eq. (19)-(21), are used to measure the clustering performance. Purity is a simple measure, in the range of [0, 1], to compute the proportion of the documents which are correctly clustered. F-Score is a multiple evaluation method that combines recall and precision measures. A perfect clustering solution will be the one in which every cluster has a corresponding cluster containing the exactly same documents in the resulting hierarchical tree, in which case the F-Score will be one. In general, the higher the F-Score values, the better the clustering solution is. The NMI value is 1 when clustering results perfectly match the external category labels and close to 0 for a random partitioning. NMI is a better measure than entropy which is biased towards high K solutions\(^{36,37}\).

\[
\text{Purity} = \frac{1}{n} \sum_{i=1}^{k} \max_{j \in C} \left( \frac{n_i}{n} \right) \tag{19}
\]

\[
F-Score = \frac{1}{n} \sum_{i} \left[ \frac{n_i \log \left( \frac{n_i}{n_k} \right)}{n_k} \right] \sum_{i} \left[ \frac{n_i \log \left( \frac{n_i}{n_k} \right)}{n_k} \right] \tag{20}
\]

\[
NMI = \frac{1}{n} \sum_{i} \left[ \frac{n_i \log \left( \frac{n_i}{n_k} \right)}{n_k} \right] \sum_{i} \left[ \frac{n_i \log \left( \frac{n_i}{n_k} \right)}{n_k} \right] \tag{21}
\]

In these equations, relevant parameters are defined as follows. For an experimental set \(D\), texts are labeled in C classes, each being noted as \(L_j (1 \leq j \leq C)\). \(D\) is clustered into \(K\) clusters, each being noted as \(U_k (1 \leq k \leq K)\). \(n\) is the number of texts in \(D\); \(n_j\) is the number of texts in \(L_j\); \(n_k\) is the number of texts in \(U_k\); \(n_{jk}\) is the number of mutual texts for \(L_j\) and \(U_k\).

5.3 Experimental tracks and result analysis

Three reference clustering models, representing different technical strategies, are selected to evaluate CogTCA. These models are frequent term-based text clustering model (FT)\(^{10}\), a conceptional self-organizing map model (ConSOM, impact factor equals 0.6)\(^{18}\) and knowledge-based vector space model (KBVSM, based on Hirst and St-Onge’s ontology\(^{38}\))\(^{19}\).

To examine the performance of CogTCA with above three references our experiments are designed to implement in four tracks (shown in table 1):

1. Clustering with single domains (CSD) uses seven domains (shown in table 2) of Reuters-21578. In each domain, 120-150 texts are selected randomly to conduct the test for each domain 12 times. The average results is shown in Fig.4(a)-(c).

2. Clustering across domains (CAD) includes 10 scheduled experiments (identified with roman numbers: I,
II, III, ..., X) in each of which 300 texts are selected in random from the categories across 9 domains of Reuters-21578. The experimental results are displayed in Fig.4(d)-(f). The “Ave” columns indicate averages for each algorithm.

(3) Clustering with single groups (CSG) uses four hierarchical groups (CCAT, ECAT, GCAT and MCAT) of RCV1-v2. In each group, 20 thousand texts are selected randomly to conduct the test for each group 10 times. The average results is shown in Fig.5(a)-(c).

(4) Clustering across groups (CAG) includes 10 scheduled experiments (identified with roman numbers: I, II, III, ..., X) in each of which 25 thousand texts are selected in random from RCV1-v2. The experimental results are displayed in Fig.5(d)-(f). The “Ave” columns indicate averages for each algorithm.

The experimental result of track (1), Fig.4(a)-(c), presents that, for each algorithm in test, its performance varies on different domains, but is still not strongly domain-dependent or domain-sensitive. FT achieves lower scores than the other three (KBVSM, ConSOM and CogTCA), which could be due to FT heavily relies on term frequency, without involving other language factors, lexical or semantic. KBVSM and ConSOM are peer to peer on the measure of Purity in Fig.4(a), so are ConSOM and CogTCA on the measure of NMI in Fig.4(c). Overall, CogTCA scores higher than the other three reference systems on three standard measure parameters, Purity, F-Score and NMI.

### Table 1. List of experimental tracks

<table>
<thead>
<tr>
<th>Track ID and name</th>
<th>Target corpus</th>
<th>Number of texts for each test</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) CSD</td>
<td>Reuters-21578</td>
<td>120-150</td>
<td>All the test texts are randomly selected within a single domain, and very likely come from various categories</td>
</tr>
<tr>
<td>(2) CAD</td>
<td>Reuters-21578</td>
<td>300</td>
<td>All the test texts are randomly selected without constraints of domains, and definitely come from various categories.</td>
</tr>
<tr>
<td>(3) CSG</td>
<td>RCV1-v2</td>
<td>20,000</td>
<td>All the test texts are randomly selected within a single group (CCAT, ECAT, GCAT or MCAT) of RCV1-v2.</td>
</tr>
<tr>
<td>(4) CAG</td>
<td>RCV1-v2</td>
<td>25,000</td>
<td>All the test texts are randomly selected across the boundaries of groups (CCAT, ECAT, GCAT and MCAT) of RCV1-v2.</td>
</tr>
</tbody>
</table>

### Table 2. Statistic of Categories of Reuters-21578 used in experiments

<table>
<thead>
<tr>
<th>Codes for Categories (Domains)</th>
<th>Numbers of Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commodities</td>
<td>78</td>
</tr>
<tr>
<td>Countries</td>
<td>175</td>
</tr>
<tr>
<td>Currencies</td>
<td>27</td>
</tr>
<tr>
<td>Economic Indicators</td>
<td>16</td>
</tr>
<tr>
<td>Exchanges</td>
<td>39</td>
</tr>
<tr>
<td>Organizations</td>
<td>56</td>
</tr>
<tr>
<td>People</td>
<td>267</td>
</tr>
</tbody>
</table>
Yi Guo, Yan Li, Zhiqing Shao

Track (1) Clustering with single domains (CSD)

Track (2) Clustering across domains (CAD)

(a) Purity

(d) Purity

(b) F-Score

(e) F-Score

(c) NMI

(f) NMI

Fig. 4. Experimental results of CSD and CAD
An Ant Colony-based Text

Track (3) Clustering with single groups (CSG)

(a) Purity

(b) F-Score

(c) NMI

Track (4) Clustering across groups (CAG)

(d) Purity

(e) F-Score

(f) NMI

Fig. 5. Experimental results of CSG and CAG
The experimental result of track (2), Fig.4(d)-(f), states that, when processing mixed texts from various domains, FT still plays no better than any of other three algorithms (KBVSM, ConSOM and CogTCA). KBVSM and ConSOM still perform equally well on the measure of Purity in Fig.4(d), which might be brought on by application of knowledge bases in form of ontologies. ConSOM and CogTCA are even matched, while the latter one performs slightly better, on both measures of $F$-Score and $NMI$ in Fig.4(e)-(f). Note that ConSOM takes an impact factor of 0.6, which has been proven as one of the best two impact factors (the other is 0.7). When comparing (a, b, c) and (d, e, f) in Fig.4, we see that all the four algorithms achieve better evaluation score in track (2). This phenomenon is believed due to the fact that the content deviations of single-domain sets are more prominent than those of multiple-domain sets.

The experimental result of track (3), Fig.5(a)-(c), indicates that, in a much larger testing corpus, four algorithms show no salience of group (or topic) sensitivity. FT remains its previous status, and KBVSM and ConSOM entangle with each other still on measure of Purity. ConSOM acts slightly inferior to CogTCA, except its upswing at the group of MCAT on the measure of $F$-Score.

The experimental result of track (4), Fig.5(d)-(f), reveals that, when processing much more mixed texts from various groups (top level topics). ConSOM and CogTCA are notable competitive to each other on all three measures. FT and KBVSM remain their rankings as expected. When comparing (a, b, c) and (d, e, f) in Fig.5, once again, we see that all the four algorithms are of better performance when processing “mixed” texts. This phenomenon confirms that it is of more challenge to divide similar texts into subsections of finer granularity.

“Extraction of Cognitive Situation Dimensions” (ECSD), a prerequisite processing module shown as Fig.3, plays an indispensable role of extracting four cognitive situation dimensions (“Protagonist”, “Temporality”, “Spatiality”, and “Activity”) for subsequent modules in CogTCA. The time cost of this module has direct impact on the whole clustering system. Therefore, the elapsed time (ET) is recorded as a supplementary index for each experiment when the four tracks are implemented. The average elapsed time of ECSD, shown as Fig.6, is calculated on the basis of per hundred thousand ($10^5$) words according to domains (Reuters-21578) or groups (RCV1-v2).

![Fig.6 Average elapsed time of ECSD in four experimental tracks.](image)

**Fig.6 Average elapsed time of ECSD in four experimental tracks.** (a) The elapsed time is not sensitive to domains or groups, ranging from 979ms to 1128ms. (b) Texts of RCV1-v2 costs slightly more time in average than those of Reuters-21578.

### 6. Concluding remarks

This paper emphasizes the necessity of integrating human cognitive features into text processing procedures and advantages of artificial ant systems in solving hard combinatorial optimization problems, and presents the feasibility of converting a text clustering task into a shortest path searching problem.

Subsequently, this paper introduces a new text clustering
system, CogTCA, which represents a text in the form of cognitive situation matrix and searches for the shortest internal open routes with an optimized approach, the encounter ant colony system, within a set of texts. The final clustering result is delivered in a top-down splitting manner based on either number or diameter of expected clusters.

CogTCA condenses the cognitive situation models into four dimensions, which makes the computation with cognitive features applicable and practical. CogTCA processes texts at syntactical and semantic levels and complements research works heavily relying on statistics and probability.

CogTCA is examined with elaborately designed experimental tracks involving a broad range of sufficient texts of Reuters-21578 and RCV1-v2. The experimental results have testified the performance and effectiveness of CogTCA. The future research work will focus on the elevation of metrics in performance evaluation and the domain sensitivity of texts. Meanwhile, text processing always faces linguistic uncertainty in lexicology and semantics. The integration of solutions for linguistic uncertainty (Entropy, Fuzziness and Ambiguity) into E-ACS is another potential research interest in no time.
Appendix A. Analysis of cluster internal distances

For clusters \(A_1=\{a_1, a_2, \ldots, a_m\}\) and \(B_1=\{b_1, b_2, \ldots, b_n\}\), suppose all the nodes in \(A_1\) and \(B_1\) satisfy:

\[
\|a_j - a_i\| < \|b_j - a_i\|, \quad 1 \leq j \leq m, \quad 1 \leq i \leq n
\]

If exchange \(a_p\) with \(b_q\), \(A_1 \rightarrow A_2=\{a_1, a_2, \ldots, a_{p-1}, b_q, a_{p+1}, \ldots, a_m\}\)

\[
\sum_{i \leq j \leq m} \|x_j - a_i\| = \sum_{i \leq j \leq m} \|x_j - a_i\| + \sum_{p \leq i \leq m} \|x_j - a_i\| + \sum_{i \leq j \leq m} \|x_p - a_i\|
\]

\[
< \sum_{i \leq j \leq m} \|x_j - a_i\| + \sum_{p \leq i \leq m} \|x_j - a_i\| + \sum_{i \leq j \leq m} \|x_p - a_i\|
\]

If multiply a constant \(\frac{1}{C_m^2}\) to either side of above inequation, we get

\[
\frac{1}{C_m^2} \sum_{i \leq j \leq m} \|x_j - a_i\| < \frac{1}{C_m^2} \left\{ \sum_{i \leq j \leq m} \|x_j - a_i\| + \sum_{p \leq i \leq m} \|x_j - a_i\| + \sum_{i \leq j \leq m} \|x_p - a_i\| \right\}
\]

which also means that

\[
\text{ACID}(A_1) < \text{ACID}(A_2)
\]

Similarly,

\[
\text{ACID}(B_1) < \text{ACID}(B_2)
\]

QED.
Appendix B. The numerical results of experimental tracks (1)-(4)

**Track (1) Clustering with single domains (CSD)**

<table>
<thead>
<tr>
<th>Categories</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commodities</td>
<td>0.479</td>
<td>0.734</td>
<td>0.708</td>
<td>0.780</td>
</tr>
<tr>
<td>Countries</td>
<td>0.512</td>
<td>0.662</td>
<td>0.717</td>
<td>0.758</td>
</tr>
<tr>
<td>Currencies</td>
<td>0.519</td>
<td>0.671</td>
<td>0.644</td>
<td>0.759</td>
</tr>
<tr>
<td>Economic</td>
<td>0.467</td>
<td>0.743</td>
<td>0.638</td>
<td>0.815</td>
</tr>
<tr>
<td>Exchanges</td>
<td>0.509</td>
<td>0.684</td>
<td>0.665</td>
<td>0.765</td>
</tr>
<tr>
<td>Organizations</td>
<td>0.453</td>
<td>0.702</td>
<td>0.686</td>
<td>0.784</td>
</tr>
<tr>
<td>People</td>
<td>0.471</td>
<td>0.661</td>
<td>0.718</td>
<td>0.801</td>
</tr>
</tbody>
</table>

(a) Purity

<table>
<thead>
<tr>
<th>Categories</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commodities</td>
<td>0.358</td>
<td>0.456</td>
<td>0.688</td>
<td>0.743</td>
</tr>
<tr>
<td>Countries</td>
<td>0.420</td>
<td>0.483</td>
<td>0.655</td>
<td>0.747</td>
</tr>
<tr>
<td>Currencies</td>
<td>0.376</td>
<td>0.509</td>
<td>0.640</td>
<td>0.711</td>
</tr>
<tr>
<td>Economic</td>
<td>0.366</td>
<td>0.547</td>
<td>0.610</td>
<td>0.711</td>
</tr>
<tr>
<td>Exchanges</td>
<td>0.351</td>
<td>0.532</td>
<td>0.668</td>
<td>0.747</td>
</tr>
<tr>
<td>Organizations</td>
<td>0.434</td>
<td>0.504</td>
<td>0.670</td>
<td>0.700</td>
</tr>
<tr>
<td>People</td>
<td>0.431</td>
<td>0.493</td>
<td>0.668</td>
<td>0.745</td>
</tr>
</tbody>
</table>

(b) F-Score

<table>
<thead>
<tr>
<th>Categories</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commodities</td>
<td>0.315</td>
<td>0.440</td>
<td>0.650</td>
<td>0.709</td>
</tr>
<tr>
<td>Countries</td>
<td>0.326</td>
<td>0.490</td>
<td>0.637</td>
<td>0.715</td>
</tr>
<tr>
<td>Currencies</td>
<td>0.363</td>
<td>0.455</td>
<td>0.693</td>
<td>0.738</td>
</tr>
<tr>
<td>Economic</td>
<td>0.338</td>
<td>0.437</td>
<td>0.681</td>
<td>0.774</td>
</tr>
<tr>
<td>Exchanges</td>
<td>0.346</td>
<td>0.498</td>
<td>0.675</td>
<td>0.735</td>
</tr>
<tr>
<td>Organizations</td>
<td>0.309</td>
<td>0.458</td>
<td>0.725</td>
<td>0.731</td>
</tr>
<tr>
<td>People</td>
<td>0.331</td>
<td>0.450</td>
<td>0.770</td>
<td>0.746</td>
</tr>
</tbody>
</table>

(c) NMI

**Track (2) Clustering across domains (CAD)**

<table>
<thead>
<tr>
<th>Experiment IDs</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.502</td>
<td>0.672</td>
<td>0.700</td>
<td>0.806</td>
</tr>
<tr>
<td>II</td>
<td>0.331</td>
<td>0.760</td>
<td>0.671</td>
<td>0.829</td>
</tr>
<tr>
<td>III</td>
<td>0.559</td>
<td>0.669</td>
<td>0.687</td>
<td>0.827</td>
</tr>
<tr>
<td>IV</td>
<td>0.477</td>
<td>0.731</td>
<td>0.738</td>
<td>0.827</td>
</tr>
<tr>
<td>V</td>
<td>0.548</td>
<td>0.673</td>
<td>0.709</td>
<td>0.783</td>
</tr>
<tr>
<td>VI</td>
<td>0.459</td>
<td>0.674</td>
<td>0.740</td>
<td>0.826</td>
</tr>
<tr>
<td>VII</td>
<td>0.303</td>
<td>0.714</td>
<td>0.683</td>
<td>0.806</td>
</tr>
<tr>
<td>VIII</td>
<td>0.493</td>
<td>0.747</td>
<td>0.700</td>
<td>0.801</td>
</tr>
<tr>
<td>IX</td>
<td>0.313</td>
<td>0.699</td>
<td>0.689</td>
<td>0.799</td>
</tr>
<tr>
<td>X</td>
<td>0.454</td>
<td>0.674</td>
<td>0.505</td>
<td>0.764</td>
</tr>
<tr>
<td>Ave</td>
<td>0.502</td>
<td>0.698</td>
<td>0.703</td>
<td>0.827</td>
</tr>
</tbody>
</table>

(d) Purity

<table>
<thead>
<tr>
<th>Experiment IDs</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.400</td>
<td>0.541</td>
<td>0.736</td>
<td>0.757</td>
</tr>
<tr>
<td>II</td>
<td>0.445</td>
<td>0.537</td>
<td>0.739</td>
<td>0.719</td>
</tr>
<tr>
<td>III</td>
<td>0.394</td>
<td>0.497</td>
<td>0.653</td>
<td>0.738</td>
</tr>
<tr>
<td>IV</td>
<td>0.424</td>
<td>0.542</td>
<td>0.681</td>
<td>0.688</td>
</tr>
<tr>
<td>V</td>
<td>0.365</td>
<td>0.476</td>
<td>0.649</td>
<td>0.668</td>
</tr>
<tr>
<td>VI</td>
<td>0.399</td>
<td>0.497</td>
<td>0.709</td>
<td>0.741</td>
</tr>
<tr>
<td>VII</td>
<td>0.369</td>
<td>0.483</td>
<td>0.670</td>
<td>0.710</td>
</tr>
<tr>
<td>VIII</td>
<td>0.402</td>
<td>0.481</td>
<td>0.668</td>
<td>0.737</td>
</tr>
<tr>
<td>IX</td>
<td>0.394</td>
<td>0.497</td>
<td>0.709</td>
<td>0.747</td>
</tr>
<tr>
<td>X</td>
<td>0.443</td>
<td>0.502</td>
<td>0.671</td>
<td>0.742</td>
</tr>
<tr>
<td>Ave</td>
<td>0.404</td>
<td>0.506</td>
<td>0.669</td>
<td>0.719</td>
</tr>
</tbody>
</table>

(e) F-Score

<table>
<thead>
<tr>
<th>Experiment IDs</th>
<th>FT</th>
<th>KBVSM</th>
<th>ConSOM</th>
<th>CogTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.352</td>
<td>0.498</td>
<td>0.782</td>
<td>0.803</td>
</tr>
<tr>
<td>II</td>
<td>0.391</td>
<td>0.490</td>
<td>0.756</td>
<td>0.758</td>
</tr>
<tr>
<td>III</td>
<td>0.333</td>
<td>0.544</td>
<td>0.749</td>
<td>0.807</td>
</tr>
<tr>
<td>IV</td>
<td>0.268</td>
<td>0.543</td>
<td>0.720</td>
<td>0.705</td>
</tr>
<tr>
<td>V</td>
<td>0.388</td>
<td>0.473</td>
<td>0.701</td>
<td>0.763</td>
</tr>
<tr>
<td>VI</td>
<td>0.333</td>
<td>0.511</td>
<td>0.687</td>
<td>0.784</td>
</tr>
<tr>
<td>VII</td>
<td>0.293</td>
<td>0.436</td>
<td>0.728</td>
<td>0.754</td>
</tr>
<tr>
<td>VIII</td>
<td>0.351</td>
<td>0.540</td>
<td>0.740</td>
<td>0.770</td>
</tr>
<tr>
<td>IX</td>
<td>0.320</td>
<td>0.514</td>
<td>0.667</td>
<td>0.784</td>
</tr>
<tr>
<td>X</td>
<td>0.236</td>
<td>0.536</td>
<td>0.749</td>
<td>0.793</td>
</tr>
<tr>
<td>Ave</td>
<td>0.362</td>
<td>0.508</td>
<td>0.728</td>
<td>0.778</td>
</tr>
</tbody>
</table>

(f) NMI

Fig B.1 Experimental results of CSD and CAD
Fig. B.2 Experimental results of CSG and CAG
References


26. T. Stützle, M. Dorigo, The ant colony optimization
Yi Guo, Yan Li, Zhiqing Shao


39. X.Z. Wang, L.C. Dong, J.H. Yan, Maximum ambiguity based sample selection in fuzzy decision tree induction, *IEEE Transactions on Knowledge and Data Engineering*, IEEE