2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Extension of Similarity Measures in VSM: from Orthogonal Coordinate System to Affine Coordinate System

Junyu Xuan, Jie Lu, Guangquan Zhang and Xiangfeng Luo

***Abstract*— Similarity measures are the foundations of many research areas, e.g. information retrieval, recommender system and machine learning algorithms. Promoted by these application scenarios, a number of similarity measures have been proposed and proposing. In these state-of-the-art measures, vector-based representation is widely accepted based on Vector Space Model (VSM) in which an object is represented as a vector composed of its features. Then, the similarity between two objects is evaluated by the operations on two corresponding vectors, like cosine, extended jaccard, extended dice and so on. However, there is an assumption that the features are independent of each others. This assumption is apparently unrealistic, and normally, there are relations between features, i.e. the co-occurrence relations between keywords in text mining area. In this paper, a space geometry-based method is proposed to extend the VSM from the orthogonal coordinate system (OVSM) to affine coordinate system (AVSM) and OVSM is proved to be a special case of AVSM. Unit coordinate vectors of AVSM are inferred by the relations between features which are considered as angles between these unit coordinate vectors. At last, five different similarity measures are extended from OVSM to AVSM using unit coordinate vectors of AVSM. Within the numerous application fields of similarity measures, the task of text clustering is selected to be the evaluation criterion. Documents are represented as vectors in OVSM and AVSM, respectively. The clustering results show that AVSM outweighs the OVSM.**

I. Introduction

SIMILARITY measures are the foundations of many research areas, like information retrieval, recommender system [1], some machine learning algorithms, i. e. case based reasoning [2], and so on. For examples, in the information retrieval area, the returned document list should be ranked by the similarity with the query given by a user. The similarity measure here is used to make sure the returned documents satisfying user's requirement; In the recommender system area, the items of a user will be recommended to users who are similar with him/her. The similarity measure here is used to make sure that recommendation is taken between two users with same interest; For some machine learning algorithms, like k-nearest neighbors algorithm, the similarity measure is the fundamental operation for these state-of-art classification or clustering algorithms.

Due to the broad application scenarios, plenty of similarity measures are proposed [3], including L_p Minkowski family, L_1 family, Intersection family, Inner Product family, Fidelity family, Squared L_2 family and Shannon's entropy family. In these methods, vector representation methods, based on Vector Space Model (VSM), have been accepted by many researchers and adopted by many works [4], [5], including cosine, extended jaccard, extended dice and person coefficient.

Classical VSM is under an Orthogonal Coordinate System (OVSM) which is composed by a number of orthogonal unit coordinate vectors/features of objects. One problem of OVSM is its assumption that the features are independent of each others. However, in most of cases, there will be some kind of relations between features and then this assumption is broken. For example, two keywords may be synonyms in a document or two users may be friends. Apparently, these relations will also impact on the similarity between two objects. It seems that there are only two options: one is to ignore these relations for continuing to use the classical similarity measures; the other is to abandon the widely accepted classical similarity measures.

The final goal of this paper is to extend the classical similarity measures by considering the relations between features of objects. After this extension, the similarity measures under up-mentioned situations will be more accurate and their original physical meanings will be kept. For example, cosine similarity is the still the cosine value of angle between two vectors.

In this paper, five classical similarity measures are extended to incorporate the relations between features which are assumed to be in hand. At first, Affine Coordinate System [6], in which unit coordinate vectors do not have to be orthogonal with each others, is introduced by us to replace the orthogonal coordinate system of VSM. To our knowledge, it is the first time that Affine Coordinate System is introduced for similarity measures. Then, the relations between features are considered as the angles between the unit coordinate vectors of VSM under affine coordinate system (AVSM). Through these angles, the unit vectors are inferred by seeing them as the normal vectors in OVSM. At last, five different similarity measures are extended by the unit coordinate vectors of AVSM. The merit of the extension from OVSM to AVSM is that it keeps the definitions and physical meanings of classical similarity measures and incorporates the relations between features as well. Text mining is selected as the background throughout this paper to keep consistent with classical VSM paper [7].

The rest of paper is organized as follows. Some related work are summarized in Section II. In Section III we intro-

Junyu Xuan, Jie Lu and Guangquan Zhang are with the Centre for Quantum Computation and Intelligent Systems (QCIS), School of Software, Faculty of Engineering and Information Technology, University of Technology, Sydney (UTS), Australia (email: Junyu.Xuan@student.uts.edu.au, {Jie.Lu, Guangquan.Zhang}@uts.edu.au). Xiangfeng Luo is with the School of Computer Engineering and Science, Shanghai University, China (e-mail: luoxf@shu.edu.cn).

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

duce traditional VSM under orthogonal coordinate system (OVSM). Keyword network of a corpus is constructed as a kind of relations between keywords in Section IV. Our extended VSM under affine coordinate system (AVSM) is proposed in Section V and extended five different similarity measures based on AVSM are given in Section VI, respectively. In Section VII, we compare different similarity measures on task of document clustering, including ones under OVSM and ones under AVSM. At last, Section VIII concludes.

II. Related Work

After the emergence of VSM in 1975 [7], the non-orthogonal problem of it has been attracting attentions of researchers. Wong [8] proposes GVSM to get the term correlations in a corpus by considering the non-orthogonal problem of VSM. In GVSM, all the combinations of all dimensions are seen as new dimensions which are all orthogonal with each other. LSI [9] also concentrate this problem but from the term-document matrix view. Single Value Composition is adopted by LSI to find the core matrix of term-document matrix. In this core matrix, each dimension is orthogonal with others, which are considered as new dimensions of documents. Some similar works [10]–[12] try to reduce the computational complexity of LSI by seeing it as an optimization problem. These methods only use feature-object data and have no idea about the relations between features of data. So, the utilized latent relation between features is only co-occurrence relations. Other relations cannot be directly incorporated into these methods, like the synonyms relation between keywords in documents or friend relation between users in social networks. Therefore, some works [13], [14] try to utilize outer information, like user-supplied information used to improve the performance of information retrieval [15]. Topic-based Vector Space Model (TVSM) [16] [17] is proposed to incorporate the similarity relations, which are from WordNet, between keywords and cosine similarity measure is revised. However, there is no theoretical explanation behind the equation, like how to get the basic unit coordinate vectors.

III. OVSM - VSM under orthogonal coordinate system

In the Vector Space Model under orthogonal coordinate system (OVSM), a document d_i in a corpus is represented by a vector of keywords,

$$d_i =< w_{k_0}, w_{k_1}, \cdots, w_{k_{n-1}} > \quad (1)$$

where n is the number of keywords in this corpus and w_{k_i} is the weight of keyword k_i in this document. Each keyword is considered as a dimension and then all keywords together form a space S_o,

$$\begin{aligned} \boldsymbol{e_0} &: (1, 0, \cdots, 0) \\ \boldsymbol{e_1} &: (0, 1, \cdots, 0) \\ &\vdots \\ \boldsymbol{e_{n-1}} &: (0, 0, \cdots, 1) \end{aligned} \quad (2)$$

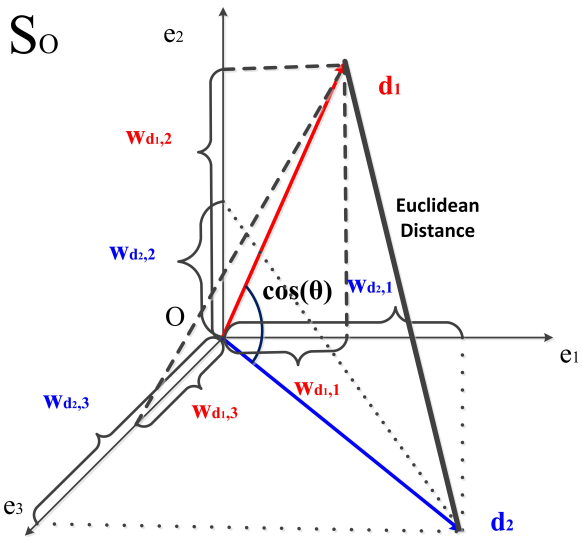


Fig. 1. An example of VSM under orthogonal coordinate system (OVSM) with three dimensions. d_1 and d_2 are two vector-represented documents by OVSM and θ is the angle between these two documents. e_i is the unit coordinator vector. and $w_{d_i,j}$ is the weight of document d_i on dimension j. Two similarity measures are shown in figure, one is cosine and the other is Euclidean distance.

where $\boldsymbol{e_i}$ is a unit coordinate vector of dimension i of this space. It can be seen from Equation 2 that each dimension in space S_o is orthogonal with others, which can be formally represented as,

$$\boldsymbol{e_i} \times \boldsymbol{e_j} = 0, \quad i \neq j \quad (3)$$

This assumption is necessary because the orthogonal coordinate system can be used to describe this space only when they are independent. This part is just what we want to release in this paper. By these unit coordinate vectors, the document d_i can also be represented as,

$$\boldsymbol{d_i} = w_{k_0} \cdot \boldsymbol{e_0} + w_{k_1} \cdot \boldsymbol{e_1} + \cdots + w_{k_{n-1}} \cdot \boldsymbol{e_{n-1}} \quad (4)$$

Here, an example of space S_o with three dimensions is given in Fig. 1. It can be seen that e_i is orthogonal with each others and the coordinates of vector/document d_i are just the weights on different dimensions/keywords. Physical meanings of two basic similarity measures, $Cosine$ and $Euclidean$, are shown in this figure.

IV. Keyword Network of a corpus

In VSM, documents in a corpus D are represented as vectors of keywords which are assumed to be independent of each others. In fact, there are many relations between keywords which can be mined to enhance the expressing of document semantics [18]. Here, the co-occurrence relation is selected as a example of keyword relations. The value of co-occurrence relation of two keywords k_i and k_j in a corpus D is,

$$f_{k_i,k_j} = \frac{D(k_i, k_j)}{|D|}$$

where $D(k_i, k_j)$ is the number of documents which contains keyword k_i and k_j simultaneously and $|D|$ is the number of

documents in corpus D. From the definition, we can see that the value scope of f_{k_i,k_j} is $[0,1]$ and $f_{k_i,k_j} = f_{k_j,k_i}$. For example, assume that there are two documents in a corpus:

d_1 : *'Semantic Web is a collaborative movement.'*

d_2 : *'The Semantic Web can drive the evolution of the current Web.'*

In this corpus, '$semantic$' and 'web' exist in all documents (d_1 and d_2), so $f_{semantic,web} = 1$. In the OVSM, these two keywords will generate two independent dimensions. In this example, they can be merged to one word '$semanticweb$' and single dimension should be used to describe them in OVSM. The merit of this merge is to magnify the difference between two documents and then improve the performance of information retrieval.

By combining keywords and their relations, a corpus can be represented by another way: Keyword Network M_D, in which nodes denote keywords and links denote co-occurrence relations between keywords, f_{k_i,k_j}. For co-occurrence relation is non-directional, M_D is a symmetrical matrix. Keyword Network is a flat representation of a corpus comparing to the VSM. How to incorporate this network into the VSM is what we are trying to do in the next section.

V. AVSM - VSM UNDER AFFINE COORDINATE SYSTEM

The keyword network introduced in Sectioin 4 has broken the assumption of OVSM that the keywords are independent of each others. The Equation 3 is no longer true. In this paper, the relations between keywords are considered as the angles between the unit coordinate vectors to form an affine coordinate system, then VSM under affine coordinate system (AVSM) is introduced to extend OVSM to non-orthogonal vector space. We think AVSM is the most natural extension of OVSM because it keeps the features of vectors by space geometry method compared with other methods, like Single Value Decomposition of LSI [9].

In order to represent documents and compute similarity between them in AVSM, the unit coordinate vectors of AVSM are first inferred by the relative geometry positions of them constrained by angles between keywords.

In affine coordinate system, the keywords do not have to be orthogonal with each others. So, the affine coordinate system is more appropriate to represent keywords with relations in VSM. We call it AVSM in this paper. The unit coordinate vectors of ASVM $\{\boldsymbol{a_i}\}$ also represent keywords. However, not like OVSM, there are angles A_D between the unit coordinate vectors of AVSM. An element of A_D is $\gamma_{i,j}$ that is the angle between unit coordinate vectors $\boldsymbol{a_i}$ and $\boldsymbol{a_j}$.

These angles express the associated degrees of unit coordinate vectors. They have the same meaning with the keyword relations in Keyword Network introduced in Section 4. So, the keyword network is used to compute the angles between all unit coordinate vectors as follows,

$$\gamma_{i,j} = \arccos(f_{k_i,k_j}) \tag{5}$$

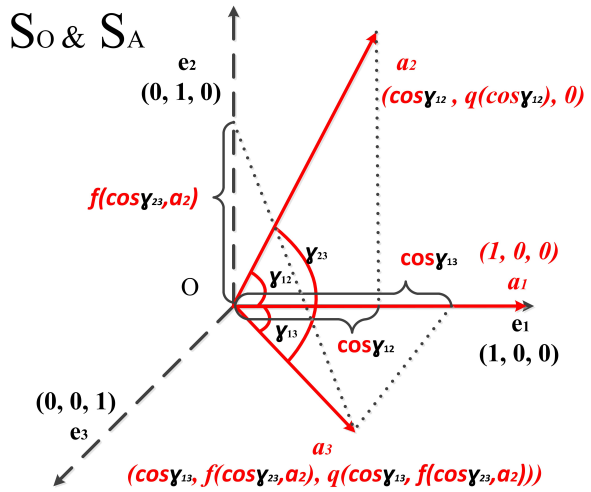


Fig. 2. An example of getting unit coordinate vectors of AVSM in three dimensions. There are two coordinate systems in this figure: orthogonal one (black and dashed lines) and affine one (red and real lines). Two sets of coordinates: orthogonal one $\{\boldsymbol{e_i}\}$ (non-italic) and affine one $\{\boldsymbol{a_i}\}$ (italic). $\boldsymbol{a_1}$ and $\boldsymbol{e_1}$ are completely coincide with each other. γ_{ij} is the angle between two affine unit coordinate vectors $\boldsymbol{a_i}$ and $\boldsymbol{a_j}$. The function f is used to keep the angles between $\{\boldsymbol{a_i}\}$ and q is used to make sure $|\boldsymbol{a_i}| = 1$.

Algorithm 1 Computation of coordinate vectors of AVSM

Input: $\{\boldsymbol{e_i}\}$ of OVSM and angle matrix A_D of $\{\boldsymbol{a_i}\}$ of AVSM
Output: $\{\boldsymbol{a_i}\}$ of AVSM

```
set a_0 = e_0 = (1, 0, ..., 0)
for i = 1; i ≤ n − 1 do
    float sum = 0
    for j = 0; j ≤ i − 1 do
        a_i[j] = cos(A_D[i][j]) − Σ_{t=0}^{j−1} a_j[t] a_i[t]
        a_i[j] = a_i[j] / a_j[j]
        sum = sum + (a_i[j])^2
    end for
    set a_i[i] = sqrt(1 − sum)
    for j = i + 1; j ≤ n − 1 do
        a_i[j] = 0
    end for
end for
return {a_i}
```

With these angles in hand, the unit coordinate vectors can be computed. The detailed procedure is described in Algorithm 1.

Here, an example of Algorithm 1 is given to explain this procedure. As shown in Fig. 2, there are two coordinate systems: OVSM with three dimensions $\{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ and AVSM with three dimensions $\{\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}\}$. Notice that $\{\boldsymbol{a_i}, \boldsymbol{e_i}\}$ are all unit vectors, so their norms all equal one, $|\boldsymbol{a_i}| = 1$ and $|\boldsymbol{e_i}| = 1$.

Firstly, the $\boldsymbol{a_1}$ is set to overlap with $\boldsymbol{e_1}$,

$$\boldsymbol{a_1} = \boldsymbol{e_1} = (1, 0, 0)$$

This is one part of work to fix the position of AVSM. Then, plane $< \boldsymbol{a_1}, \boldsymbol{a_2} >$ is set to be coplanar with plane

$< \boldsymbol{e_1}, \boldsymbol{e_2} >$, so $\boldsymbol{a_2}[2] = 0$. According to the norm of $\boldsymbol{a_2}$ and the angle $\gamma_{1,2}$ between $\boldsymbol{a_2}$ and $\boldsymbol{a_1}$, we can get the coordinate of $\boldsymbol{a_2}$ in OVSM on $\boldsymbol{e_1}$,

$$\boldsymbol{a_2}[0] = \cos(\gamma_{1,2})$$

In order to keep its norm equals one,

$$\boldsymbol{a_2}[1] = q(\cos(\gamma_{1,2})) = \sqrt{1 - \cos^2(\gamma_{1,2})} = \sin(\gamma_{1,2})$$

Finally, we get

$$\boldsymbol{a_2} = (\cos(\gamma_{1,2}), \sin(\gamma_{1,2}), 0)$$

Same as $\boldsymbol{a_2}$, we can get the coordinate of $\boldsymbol{a_3}$ in OVSM on $\boldsymbol{e_1}$ according to the norm of $\boldsymbol{a_3}$ and the angle $\gamma_{1,3}$ between $\boldsymbol{a_3}$ and $\boldsymbol{a_1}$,

$$\boldsymbol{a_3}[0] = \cos(\gamma_{1,3})$$

As Fig. 2 shows, we can infer the coordinate of $\boldsymbol{a_3}$ in OVSM on $\boldsymbol{e_2}$. By the angle of $\boldsymbol{a_2}$ and $\boldsymbol{e_2}$, the coordinate of $\boldsymbol{a_3}$ in OVSM on $\boldsymbol{e_2}$ is got,

$$\boldsymbol{a_3}[1] = \frac{\cos(\gamma_{2,3}) - \cos(\gamma_{1,2})\cos(\gamma_{1,3})}{\sin(\gamma_{1,2})}$$

The coordinate of $\boldsymbol{a_3}$ in OVSM on $\boldsymbol{e_3}$ is computed by considering its norm.

$$\begin{aligned}\boldsymbol{a_3}[2] &= q\left(\cos(\gamma_{1,3}), \frac{\cos(\gamma_{2,3}) - \cos(\gamma_{1,2})\cos(\gamma_{1,3})}{\sin(\gamma_{1,2})}\right) \\ &= \sqrt{1 - \left(\cos^2(\gamma_{1,3}) + \left(\frac{\cos(\gamma_{2,3}) - \cos(\gamma_{1,2})\cos(\gamma_{1,3})}{\sin(\gamma_{1,2})}\right)^2\right)}\end{aligned} \tag{6}$$

To sum up, the space S_a is $\boldsymbol{a_1}, \boldsymbol{a_2}, \boldsymbol{a_3}$. Notice that if the angle $\gamma_{1,2}$ between $\boldsymbol{a_1}$ and $\boldsymbol{a_2}$ is 0°, $\boldsymbol{a_2} = \boldsymbol{a_1}$ = (1, 0, 0). $\gamma_{1,2} = 0^\circ$ means the co-occurrence frequency of keyword k_1 and k_2 is 1. This is same with our example in Section 4. These two keywords/unit coordinate vectors are merged into one keyword/unit coordinate vector.

Before the similarity computation, Measure Coefficient $g_{i,j} = \boldsymbol{a_i}\boldsymbol{a_j}$ of the affine coordinate system is given,

$$(g_{i,j}) = \begin{pmatrix} a_0 a_0 & a_0 a_1 & \cdots & a_0 a_{n-1} \\ a_1 a_0 & a_1 a_1 & \cdots & a_1 a_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1} a_0 & a_{n-1} a_1 & \cdots & a_{n-1} a_{n-1} \end{pmatrix} \tag{7}$$

Apparently, $g_{i,j}$ equals to $g_{j,i}$ and $g_{j,i} \in [0, 1]$. Since $\boldsymbol{a_i}$ and $\boldsymbol{a_j}$ are two unit coordinate vectors, their product is equal to cosine value of them, $g_{i,j} = \boldsymbol{a_i}\boldsymbol{a_j} = \cos\gamma_{i,j}$. So, we get,

$$(g_{i,j}) = M_D$$

From Algorithm 1, it can be seen that the unit coordinate vectors are sensitive to the sequence of keywords. As discussed former, all the keywords with fixed angles together like an 'object' in the OVSM, which could have different positions. The different sequence of keywords in Algorithm 1 will give different positions to this 'object'. However, the 'shape' of this 'object', angles between these keywords, will keep unchanged wherever the position they will be.

Actually, there is a condition to get unit coordinate vectors of AVSM. It means that the relations between keywords/features should satisfy a condition before $\{\boldsymbol{a_i}\}$ can be computed by Algorithm 1. In this algorithm, there is a equation, $\boldsymbol{a_i}[i] = \sqrt{1 - sum}$, which is used for keeping the norm of $\boldsymbol{a_i}$. In fact, the sum may be bigger than 1. The condition is just to make sure $sum \leq 1$. Again, take three dimension as an example. Let's have a look about this 'condition'. Considering Equation 6, the condition can be written as,

$$\cos^2(\gamma_{1,3}) + \left(\frac{\cos(\gamma_{2,3}) - \cos(\gamma_{1,2})\cos(\gamma_{1,3})}{\sin(\gamma_{1,2})}\right)^2 \leq 1$$

After the derivation of this formula, we can get,

$$|\gamma_{1,3} - \gamma_{1,2}| \leq \gamma_{2,3} \leq (\gamma_{1,3} + \gamma_{1,2})$$

This condition means that if k_1 has a relation $\gamma_{1,2}$ with k_2 and another relation $\gamma_{1,3}$ with k_3, there should be a relation between k_2 and k_3 and the weight of this relation is within interval $[|\gamma_{1,3} - \gamma_{1,2}|, \gamma_{1,3} + \gamma_{1,2}]$. It is interesting that this condition also has the function to restrict the relative positions of three vectors, k_1, k_2 and k_3, in the space from the geometry view, like 'Triangle Inequality'. 'Triangle Inequality' is the inequality equation of lengths of three arcs, which defines the condition of forming a triangle. Our condition is an 'Angle Triangle Inequality', which defines when some vectors can form a space. If this condition is broken, they cannot even be placed in three-dimension space simultaneously. When considered in n-dimension, this condition has the same meaning. For now, we know that this condition is just 'Angle Triangle Inequality' in n-dimensions. Even though this condition may not be satisfied and the unit vectors cannot be computed for a specified kind of relation, we can still use the Equation 7 to compute the similarity measures introduced next for arbitrary relations.

VI. Similarity Measures under AVSM

After getting unit coordinate vectors of AVSM, we have a new Space under AVSM $S_a : \{O, a_0, a_1, \cdots, a_{n-1}\}$ in which two documents in a corpus can be represented as,

$$\begin{aligned}\boldsymbol{d_i} &= w_{k_0}^{d_i} \cdot \boldsymbol{a_0} + w_{k_1}^{d_i} \cdot \boldsymbol{a_1} + \cdots + w_{k_{n-1}}^{d_i} \cdot \boldsymbol{a_{n-1}} \\ \boldsymbol{d_j} &= w_{k_0}^{d_j} \cdot \boldsymbol{a_0} + w_{k_1}^{d_j} \cdot \boldsymbol{a_1} + \cdots + w_{k_{n-1}}^{d_j} \cdot \boldsymbol{a_{n-1}}\end{aligned}$$

where $w_{k_n}^{d_i}$ is the weight of d_i on $\boldsymbol{a_n}$ and $\{\boldsymbol{a_i}\}$ are unit coordinate vectors of AVSM. In following subsections we will extend five different similarity computation methods for these two documents from OVSM to AVSM.

A. Euclidean Similarity under AVSM

Based on the Euclidean distance based similarity between documents in OVSM [19], Euclidean similarity under AVSM

is,

$$
\begin{aligned}
E(\boldsymbol{d_i}, \boldsymbol{d_j}) &= e^{-\|\boldsymbol{d_i}-\boldsymbol{d_j}\|^2} \\
&= e^{-\sum_m^n \sum_l^n \left(\left(w_{k_m}^{d_i} - w_{k_m}^{d_j}\right)\cdot \boldsymbol{a_m} \left(w_{k_l}^{d_i} - w_{k_l}^{d_j}\right)\cdot \boldsymbol{a_l}\right)^2} \\
&= e^{-\sum_m^n \sum_l^n \left(\left(w_{k_m}^{d_i} - w_{k_m}^{d_j}\right)\cdot \left(w_{k_l}^{d_i} - w_{k_l}^{d_j}\right)\cdot g_{m,l}\right)^2}
\end{aligned} \tag{8}
$$

The Euclidean distance is actually the shortest distance between two points and is also the norm of the vector connecting these two points. Through this way, the euclidean distance in OVSM is extended by defining the norm of difference vector of two vectors/documents in AVSM as the euclidean distance in AVSM. The $exp(-x)$ is just a way to transfer distance to similarity.

B. Cosine Similarity under AVSM

The cosine similarity between documents is defined by the cosine value of the angle between two documents. Under AVSM, its meaning is unchanged,

$$
\begin{aligned}
C(\boldsymbol{d_i}, \boldsymbol{d_j}) = \cos(\theta_{d_i,d_j}) &= \frac{\mathbf{d_i} \cdot \mathbf{d_j}}{\|\mathbf{d_i}\| \cdot \|\mathbf{d_j}\|} \\
&= \frac{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot \boldsymbol{a_m} \cdot w_{k_l}^{d_j} \cdot \boldsymbol{a_l}\right)}{\|\mathbf{d_i}\| \cdot \|\mathbf{d_j}\|} \\
&= \frac{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot w_{k_l}^{d_j} \cdot g_{m,l}\right)}{\|\mathbf{d_i}\| \cdot \|\mathbf{d_j}\|}
\end{aligned} \tag{9}
$$

where

$$
\begin{aligned}
\|\mathbf{d_i}\| &= \sqrt{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot \boldsymbol{a_m} \cdot w_{k_l}^{d_i} \cdot \boldsymbol{a_l}\right)} \\
&= \sqrt{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot w_{k_l}^{d_i} \cdot g_{m,l}\right)}
\end{aligned} \tag{10}
$$

C. Pearson Correlation under AVSM

Pearson Correlation has been widely and successfully used in recommender system, information retrieval and so on. Under AVSM, it can be represented as,

$$
\begin{aligned}
&P(\boldsymbol{d_i}, \boldsymbol{d_j}) \\
&= \frac{1}{2}\left(\frac{\left(\boldsymbol{d_i} - \bar{d}_i\right)\cdot\left(\boldsymbol{d_i} - \bar{d}_i\right)}{\|\boldsymbol{d_i} - \bar{d}_i\| \cdot \|\boldsymbol{d_j} - \bar{d}_j\|} + 1\right) \\
&= \frac{1}{2}\left(\frac{\sum_m^n \sum_l^n \left(\left(w_{k_m}^{d_i} - \bar{w^{d_i}}\right)\boldsymbol{a_m}\left(w_{k_l}^{d_j} - \bar{w^{d_j}}\right)\boldsymbol{a_l}\right)}{\|\boldsymbol{d_i} - \bar{d}_i\|\|\boldsymbol{d_j} - \bar{d}_j\|} + 1\right) \\
&= \frac{1}{2}\left(\frac{\sum_m^n \sum_l^n \left(\left(w_{k_m}^{d_i} - \bar{w^{d_i}}\right)\left(w_{k_l}^{d_j} - \bar{w^{d_j}}\right)g_{m,l}\right)}{\|\boldsymbol{d_i} - \bar{d}_i\|\|\boldsymbol{d_j} - \bar{d}_j\|} + 1\right)
\end{aligned} \tag{11}
$$

where

$$
\begin{aligned}
&\|\boldsymbol{d_i} - \bar{d}_i\| \\
&= \sqrt{\sum_m^n \sum_l^n \left(\left(w_{k_m}^{d_i} - \bar{w^{d_i}}\right)\cdot\left(w_{k_l}^{d_i} - \bar{w^{d_i}}\right)\cdot g_{m,l}\right)}
\end{aligned} \tag{12}
$$

where $\bar{d}_i = \bar{w^{d_i}}$ is the average value of weights on $\{\boldsymbol{a_i}\}$ of document d_i. The form of equation is used to keep $P(d_i, d_j) \in [0, 1]$. Comparing with cosine similarity measure, Pearson Correlation removes the effect of average value of two vectors/documents and concentrates the trends of two vectors/documents.

D. Jaccard Similarity under AVSM

The Jarccard similarity is originally used to measure the similarity of two sets. In order to adopt it for vectors, extended Jaccard similarity is proposed [19], [20]. Here, we further extend this extended Jaccard similarity to AVSM,

$$
\begin{aligned}
J(\boldsymbol{d_i}, \boldsymbol{d_j}) &= \frac{\boldsymbol{d_i} \cdot \boldsymbol{d_j}}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2 - \boldsymbol{d_i} \cdot \boldsymbol{d_j}} \\
&= \frac{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot \boldsymbol{a_m} \cdot w_{k_l}^{d_j} \cdot \boldsymbol{a_l}\right)}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2 - \sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot \boldsymbol{a_m} \cdot w_{k_l}^{d_j} \cdot \boldsymbol{a_l}\right)} \\
&= \frac{\sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot w_{k_l}^{d_j} \cdot g_{m,l}\right)}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2 - \sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot w_{k_l}^{d_j} \cdot g_{m,l}\right)}
\end{aligned} \tag{13}
$$

where $\|\boldsymbol{d_i}\|$ is same with Equation 10.

E. Dice Similarity under AVSM

Similar to Jarccard, original Dice Coefficient is for sets. Here, the extended form of Dice similarity of two documents under AVSM is given,

$$
\begin{aligned}
D(\boldsymbol{d_i}, \boldsymbol{d_j}) &= \frac{2 \cdot \boldsymbol{d_i} \cdot \boldsymbol{d_j}}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2} \\
&= \frac{2 \cdot \sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot \boldsymbol{a_m} \cdot w_{k_l}^{d_j} \cdot \boldsymbol{a_l}\right)}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2} \\
&= \frac{2 \cdot \sum_m^n \sum_l^n \left(w_{k_m}^{d_i} \cdot w_{k_l}^{d_j} \cdot g_{m,l}\right)}{\|\boldsymbol{d_i}\|^2 + \|\boldsymbol{d_j}\|^2}
\end{aligned} \tag{14}
$$

where $\|\boldsymbol{d_i}\|$ is same with Equation 10.

To sum up, five similarity measures have their own forms and preoccupations. For example, cosine method assumes that if two different documents $\boldsymbol{d_i}$ and $\boldsymbol{d_j}$ have same angles with another document $\boldsymbol{d_h}$, they will have the same similarities with this document, $C(\boldsymbol{d_i}, \boldsymbol{d_h}) = C(\boldsymbol{d_j}, \boldsymbol{d_h})$ by Equation 9. Cosine method ignores the norms of two vectors. However, Euclidean similarity measure is sensitive to the norms of two vectors from Equation 8. Person Similarity ignores the average of vectors. In all the similarity measures under AVSM, the differences from their original definitions in OVSM are that their definitions and equations all have $g_{m,l}$. According to the definition, $g_{m,l}$ is the product of two unit coordinate vectors, a_m and a_l, and just reflects the angle between these two unit vectors. This suggests that the similarity measures shown in this section all have considered the relations between keywords. And It can be seen that OVSM is just a special case of AVSM at the situation $\boldsymbol{a_i} \times \boldsymbol{a_j} = 0, i \neq j$. At this situation, the similarity

measures all degenerate to their original definitions under OVSM.

From all the proposed extended similarity measures, we can see that the basic computation, computing product or norm of vectors, between the two vectors on the same dimension/unit coordinate vector/keyword is extended to the computation between values on each pair of dimensions/unit coordinate vectors/keywords. For example, for two vectors d_i and d_j, the basic computation between them is between $w_k^{d_i}$ and $w_k^{d_j}$ in OVSM. If there are n keywords in all, there will be n times computations and then the time complexity is $O(n)$. But, in AVSM, the basic computation between these two vectors is between $w_{k_m}^{d_i}$ and $w_{k_l}^{d_j}$. If there are also n keywords, there will be $n \times n$ times computations and the time complexity is $O(n^2)$. Therefore, a conclusion is drawn that the extension of these classical similarity measures is at the expense of complexity. Is that worthwhile to do so? we think it is application dependent. Here, we just show a feasible way to do that.

VII. Experiments

In order to compare the five similarity measures under both OVSM and AVSM, a common task, document clustering, is adopted here to compare their efficiency on measuring the similarity between documents. Except the mentioned five measures, Latent Semantic Indexing (LSI) [9], which also try to resolve the problem of non-orthogonal of keywords in VSM, is also implemented to join the comparison and its factor number is set as the number of topics in corresponding datasets.

A. Datasets

There are two datasets. One is documents from Reuters-21578[1], in which documents have been labeled with topics. 8 topics are selected, including *interest*, *coffee*, *crude*, *trade*, *ship*, *money-supply*, *money-fx*, and *sugar*. After filtering the stop words by a standard stopword list, each document in this dataset is represented by top 90% of keywords (only considering noun and verb) descendingly ranked by tf-idf in that document for removing the waist words. Some statistics are shown in Table I after removing documents which have less than 10 keywords (only considering noun and verb).

Another dataset is from DBLP dataset[2]. Paper abstracts from three different conferences, *ICCV*, *SIGCOMM* and *ICSE*, are extracted and the topics of these documents are set as their source conference names. Three conferences are selected to represent three different research areas, *Computer Vision*, *Computer Network* and *Software Engineering*. Different from Dataset I, all keywords (including stop words) of documents are preserved. Some statistics are shown in Table I.

[1]http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2]http://dblp.uni-trier.de/xml/

TABLE I
Statistics of datasets

dataset I		
Topic name	document number	all keyword number
interest	211	7786
coffee	114	
crude	355	
trade	333	
ship	156	
money-supply	98	
money-fx	260	
sugar	135	
dataset II		
Topic name	document number	all keyword number
ICCV	458	6982
SIGCOMM	372	
ICSE	458	

B. Evaluation Metrics

Since the document clustering is selected as the comparative method, three evaluation metrics for clustering results are introduced here, including Jaccard Coefficient (JC), Folkes&Mallows (FM) and F1 measure (F1).

Given a clustering result,

- a is the number of two points which are in same cluster of both benchmark and clustering result;
- b is the number of two points which are in same cluster of benchmark but in different cluster of clustering result;
- c is the number of two points which are not in same cluster of both benchmark but in same cluster of clustering result;
- d is the number of two points which are not in same cluster of both benchmark and clustering result.

and three metrics are,

- Jaccard Coefficient $JC = \frac{a}{a+b+c}$
- Folkes & Mallows $FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{1/2}$
- F1 measure $F1 = \frac{2a^2}{2a^2+ac+ab}$

The bigger three metrics are, the better this clustering result is.

C. Results and Discussions

In Table II, the clustering results of nine similarity measures are listed on three metrics introduced former, including four measures under OVSM, four corresponding extended measures under AVSM and LSI. It can be seen from this table that the best measure under OVSM is Dice and the best measure in AVSM is Pearson Correlation, respectively and these two methods all outweigh LSI. The more detailed comparisons are shown in Figures 4(a) and 3. From Figure 4(a), we can see that the methods under AVSM are better than ones under OVSM except Euclidean. The pairwise comparisons of methods under AVSM and OVSM are shown in Figure 3. From this figure, we can get that Cosine, Dice, Jaccard and Pearson under AVSM methods are all better than the ones under OVSM and Euclidean under AVSM is

TABLE II

CLUSTERING RESULTS OF DATASET I AND DATASET II

	Dataset I			Dataset II		
Methods	**JC**	**FM**	**F1**	**JC**	**FM**	**F1**
LSI	0.337	0.528	0.504	0.307	0.514	0.470
OVSM-C	0.477	0.646	0.646	0.446	0.617	0.617
OVSM-D	**0.485**	**0.653**	**0.653**	0.430	0.603	0.602
OVSM-J	0.466	0.643	0.636	0.429	0.601	0.600
OVSM-E	0.150	0.377	0.261	0.316	0.531	0.480
OVSM-P	0.378	0.551	0.549	**0.449**	**0.620**	**0.620**
AVSM-C	0.509	0.683	0.674	0.450	0.621	0.621
AVSM-D	0.501	0.676	0.668	**0.555**	**0.743**	**0.714**
AVSM-J	0.470	0.648	0.636	0.555	0.733	0.712
AVSM-E	0.147	0.362	0.257	0.309	0.518	0.473
AVSM-P	**0.526**	**0.690**	**0.689**	0.463	0.634	0.633

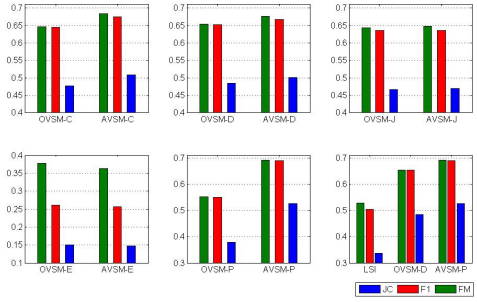
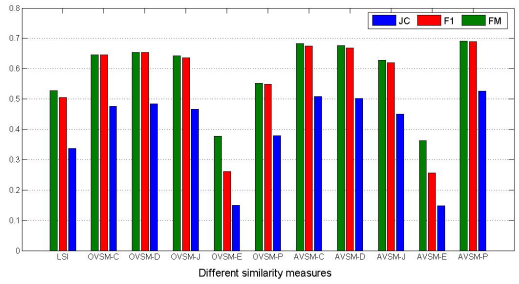
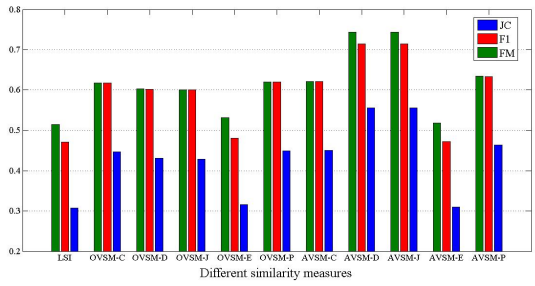


Fig. 3. Comparisons between similarity measures between OVSM and AVSM on Dataset I. The first five subfigures are pairwise comparisons between five measures in OVSM and AVSM. The last one is the comparision between two best measures in OVSM (OVSM-D) and AVSM (AVSM-P) with LSI.



(a) Dataset I



(b) Dataset II

Fig. 4. Clustering results of different similarity measures on Dataset I and II.

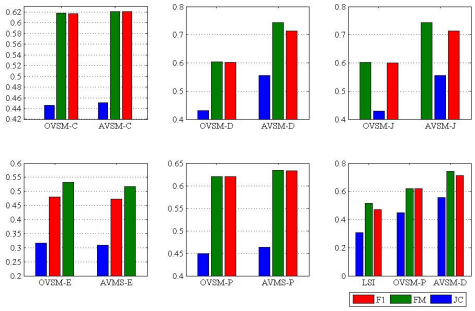
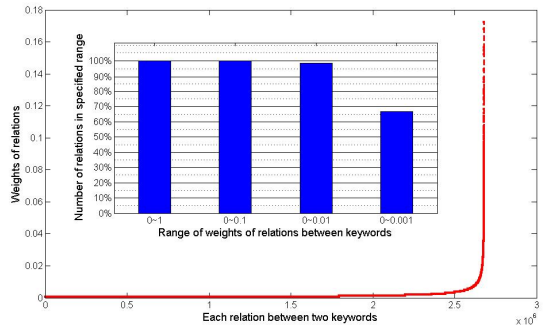


Fig. 5. Comparisons between similarity measures between OVSM and AVSM on Dataset II. The first five subfigures are pairwise comparisons between five measures in OVSM and AVSM. The last one is the comparision between two best measures in OVSM (OVSM-P) and AVSM (AVSM-D) with LSI.

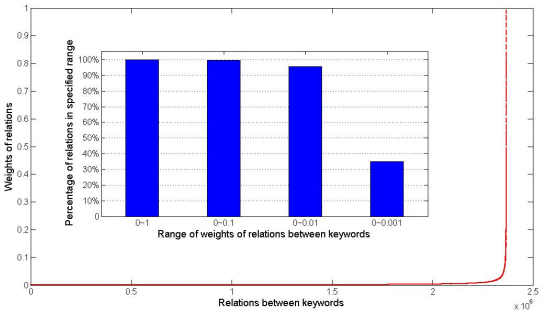
worse than it under OVSM. Notice that the performances of Jaccard under AVSM and OVSM are almost same although the value under AVSM is a little better than the one under OVSM. The last subfigure in Figure 3 shows the comparison between LSI, Dice under OVSM (best one under OVSM) and Pearson Coefficient under AVSM (best one under AVSM). The comparison suggests that these two methods are all better than LSI. The reason may be that LSI reduce the number of dimensions and this dimensionality reduction may influence the clustering results. From this result, the similarity measures in AVSM may be more appropriate for dealing with non-orthogonal problem of VSM than Single Value Decomposition of LSI with the angles (relations between features) in hand.

There is another problem: although the clustering results in Figure 3 show the efficiency of similarity measures in AVSM, the improvements of all methods are not of significance. Does that mean the influence of AVSM is not important? The answer is no. The reason is that the impact of AVSM comes from the angles between features (co-occurrence relations between keywords here). The weights of co-occurrence relations in Dataset I are drawn in Figure 6(a). Although the theoretical maximum value of this relation is 1, We notice that the maximum weight of co-occurrence relations here is only 0.17 and most of weights of relations are very small. In turn, the angles between keywords are also very small. This is normal, because there is little number of pairs of keywords existing in all documents after removing the stop words (i.e. 'is', 'the', 'a' and so on). However, to reiterate, co-occurrence relation is only one kind of relations between keywords, which is selected as the example to show the idea of AVSM because of its merit of easy implementation and understanding. The value of relations may be big if other kinds of relations are used and then the influence of AVSM will be enhanced in turn.

The clustering results of Dataset II is shown in Table II. Different from Dataset I, the best ones in dataset II are OVSM-P and AVSM-D. We think the relative performances of these different similarity measures depend on the data. There is no conclusion that OVSM-P must be better than



(a) Dataset I



(b) Dataset II

Fig. 6. Weight distribution of co-occurrence relations between keywords. The outer curve shows the weight distribution of relations and the inner subfigure shows the percentages of relations in each specified ranges.

OVSM-C or OVSM-D. They all have their own application scenarios. The same as ones in AVSM.

The average improvement of different similarity measures (except OVSM-E and AVSM-E) on dataset II, 0.80253, is bigger than the one on Dataset I, 0.586813. The reason may be that the stop words are kept in Dataset II. As shown in Fig. 6(b), there are less relations which are in the range of [0, 0.001] than in Dataset I. On the contrary, the number of relations which has weights in [0.001, 1] is bigger than it in Dataset I. So, the average values of relations in Dataset II is bigger than it in Dataset I. To emphasize again, the larger values of relations are, the bigger these relations' influence under AVSM.

VIII. Conclusions and Further Work

In this paper, we have extended the classical Vector Space Model from under orthogonal coordinates system to affine coordinates system. AVSM has released the assumption of OVSM that the features of an object are independent of each others and OVSM has been proved to be a special case of AVSM. The new unit coordinate vectors of AVSM has been computed as the normal vectors in OVSM by considering the relations between features of objects as angles. By these unit coordinate vectors, five different similarity measures have been extended to AVSM. Documents have been selected as the example throughout whole paper. In the experiments, text clustering has been selected to compare different methods under OVSM and AVSM and experiment results have proved our idea. In the future, we will consider to do more experiments on different datasets, like image, and find some datasets with more strong relations between features.

References

[1] Z. Zhang, H. Lin, K. Liu, D. Wu, G. Zhang, and J. Lu, "A hybrid fuzzy-based personalized recommender system for telecom products/services," *Information Sciences*, 2013.

[2] K. Amailef and J. Lu, "Ontology-supported case-based reasoning approach for intelligent m-government emergency response services," *Decision Support Systems*, 2013.

[3] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.

[4] C. Lee and T. Kawahara, "Hybrid vector space model for flexible voice search," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[5] K. Kesorn and S. Poslad, "An enhanced bag-of-visual word vector space model to represent visual content in athletics images," *Multimedia, IEEE Transactions on*, vol. 14, no. 1, pp. 211–222, 2012.

[6] H. S. M. Coxeter, *Introduction to geometry*. Wiley New York, 1969, vol. 6, no. 6.8.

[7] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[8] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. Wong, "On modeling of information retrieval concepts in vector spaces," *ACM Transactions on Database Systems (TODS)*, vol. 12, no. 2, pp. 299–321, 1987.

[9] S. T. Dumais, G. Furnas, T. Landauer, S. Deerwester, S. Deerwester *et al.*, "Latent semantic indexing," in *Proceedings of the Text Retrieval Conference*, 1995.

[10] N. Liu, B. Zhang, J. Yan, Q. Yang, S. Yan, Z. Chen, F. Bai, and W.-Y. Ma, "Learning similarity measures in non-orthogonal space," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 2004, pp. 334–341.

[11] M. Schultz and T. Joachims, "Learning a distance metric from relative comparisons," *Advances in neural information processing systems*, vol. 16, p. 41, 2004.

[12] D. Cai, X. He, and J. Han, "Tensor space model for document analysis," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 625–626.

[13] P. D. Turney, "Domain and function: A dual-space model of semantic relations and compositions," *Journal of Artificial Intelligence Research (JAIR)*, vol. 44, pp. 533–585, 2012.

[14] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," 2012.

[15] X. Tai, M. Sasaki, Y. Tanaka, and K. Kita, "Improvement of vector space information retrieval model based on supervised learning," in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*. ACM, 2000, pp. 69–74.

[16] J. Becker and D. Kuropka, "Topic-based vector space model," in *Proceedings of the 6th International Conference on Business Information Systems*, 2003, pp. 7–12.

[17] A. Polyvyanyy and D. Kuropka, *A Quantitative Evaluation of the Enhanced Topic Based Vector Space Model*. Univ.-Verlag, 2007.

[18] X. Luo, N. Fang, B. Hu, K. Yan, and H. Xiao, "Semantic representation of scientific documents for the e-science knowledge grid," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 7, pp. 839–862, 2008.

[19] A. Strehl, J. Ghosh, and R. Mooney, "Impact of similarity measures on web-page clustering," in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58–64.

[20] A. Strehl and J. Ghosh, "Value-based customer grouping from large retail data sets," in *AeroSense 2000*. International Society for Optics and Photonics, 2000, pp. 33–42.