Expert Systems with Applications 40 (2013) 7555–7561

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A novel noise filter based on interesting pattern mining for bag-of-features images

Zhiang Wu [a], Jie Cao [a,b,*], Haicheng Tao [b], Yi Zhuang [c]

[a] Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China
[b] College of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[c] College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, China

ARTICLE INFO

Keywords:
Image data
Bag-of-Features
Noise filtering
Interesting pattern mining
Shortest cosine interesting pattern

ABSTRACT

Improving the quality of image data through noise filtering has gained more attention for a long time. To date, many studies have been devoted to filter the noise inside the image, while few of them focus on filtering the instance-level noise among normal images. In this paper, aiming at providing a noise filter for bag-of-features images, (1) we first propose to utilize the cosine interesting pattern to construct the noise filter; (2) then we prove that to filter noise only requires to mine the shortest cosine interesting patterns, which dramatically simplifies the mining process; (3) we present an in-breadth pruning technique to further speed up the mining process. Experimental results on two real-life image datasets demonstrate effectiveness and efficiency of our noise filtering method.

1. Introduction

Filtering noise and thus improving the quality of image data have been a focus of attention in a number of diverse fields, including medical science (Leipsic et al., 2010), remote sensing (Al-amri, Kalyankar, & Khamitkar, 2010), biology, and image processing (Hindle, Shao, Lin, Lu, & Zhang, 2011; Sevgen & Arik, 2011; Wang, Li, & Song, 2013). Many studies have been devoted to filtering the undesired information (i.e., noise) inside an image (Al-amri et al., 2010; Buades, Coll, & Morel, 2008; Hajiaboli, 2011; Lin, 2011). Besides the noise contained in images, there exist irrelevant, vague, and/or incomplete images, which constitute noise images at instance-level. However, few studies have been done on removing such instance-level noise images.

The instance-level noise has posed challenges to conduct further analysis on the image data. In this paper, we consider the clustering analysis for the image data containing significant noise. The robustness against noise is a long-standing but yet very difficult problem of clustering (Li, Liu, Chen, & Tang, 2007): (1) how to obtain correct clustering from noisy data; or further, (2) how to remove the noise and obtain correct clustering from noisy data. This paper attempts to solve the second problem in image clustering. That is, the proposed method performs denoising and clustering simultaneously in order to provide noise-free clusters.

In this paper, we limit our scope to images represented by bag-of-features model, which extracts features of an image, learns "visual vocabulary", quantizes features using visual vocabulary, and finally represents images by frequencies of "visual words" (Gu, Zhao, & Zhu, 2011; Li & Perona, 2005). It is natural to convert such bag-of-features data to market basket transactions data, in which "visual word" corresponds to "item". We propose to utilize a pattern mining method for noise filtering, since the interesting pattern (e.g., cosine interesting pattern) can sketch the essential feature of an object which is unchanged even as the images of the object vary, which will be shown in Section 5. Based on the set of interesting patterns, the noise image is then defined as the images including none of the patterns, i.e., the images does not contain any essential feature of the object. To enhance the efficiency of pattern mining, we prove only mining the shortest cosine interesting patterns are sufficient for the application of noise filtering, by virtue of Conditional Anti-Monotone Property of cosine similarity (Wu, Zhu, Liu, & Xia, 2012). A novel method is presented to mine the shortest cosine interesting patterns. It combines an "in-breadth" traversing method with a pruning technique in terms of the upper bound of cosine similarity during the exhaustive search of all candidates.

The rest of this paper is organized as follows. The related work is reviewed in the next section. Section 3 presents the definition of noise filtering based on cosine interesting patterns and proves the shortest cosine interesting patterns (SCIPs) are sufficient to noise filtering. In Section 4, we elaborate details of the SCIP mining algorithm. In Section 5, we sketch the regions represented by SCIPs in order to show the significance of SCIPs on images. Experimental

* Corresponding author at: Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China.
E-mail address: Jie.Cao@njue.edu.cn (J. Cao).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.058

results will be given in Section 6. Finally, Section 7 concludes this paper.

2. Related work

Image enhancement through noise filtering or reduction is a fundamental problem in image processing. Most of the existing methods (Buades et al., 2008; Hajiaboli, 2011; Lin, 2011; Van De Ville et al., 2003) take noise filtering as an image restoration problem in which it attempts to recover an underlying perfect image from a degraded copy. Such studies do not conflict with our work. They try to filter or reduce the noise inside an image and recover its relatively clear copy, while our work tries to filter noise images among a set of normal images. To the best of our knowledge, no work has been done on filtering the instance-level noise images.

Following the pioneering work by Agrawal, Imielinski, and Swami (1993), there has been a vast amount of research on developing a theory for the association analysis problem. For brevity and emphasis, we only concentrate on interesting pattern mining. A lot of objective measures of interestingness have been proposed or borrowed including symmetric measures such as correlation, cosine, Jaccard, etc. and asymmetric measures such as mutual information, J-measure, Gini index, Laplace, etc. (Geng & Hamilton, 2006; Tan, Steinbach, & Kumar, 2005). Among them, cosine similarity extracted particular attention. Indeed, cosine similarity has been widely used as a popular proximity measure in text mining (Zhao & Karypis, 2004), information retrieval (Bayardo, Ma, & Srikant, 2007), and bio-informatics (Mistry & Pavlidis, 2008) to avoid the "curse of dimensionality". Wu et, al. defined the Conditional Anti-Monotone Property for cosine similarity (Wu et al., 2012), which is employed to prove that only mining SCIPs is sufficient to noise filtering. In summary, in spite of plenty of theoretical research on interesting pattern mining problem, it is necessary to bring forth more and more application domains for interesting pattern mining, such as noise filtering in this paper.

3. The cosine interesting pattern for noise filtering

Let $\mathbb{D} = \{v_1, \ldots, v_d\}$ be the set of all visual words in the visual vocabulary and $\mathbb{I} = \{i_1, \ldots, i_n\}$ be the set of all images. Every image i_p can be regraded as a transaction containing a subset of visual words chosen from $\mathbb{D}$. The image dataset $\mathbb{I}$ is similar to market basket transactions, and therefore the pattern mining methodology can be applied. We first formally define the set of noise images as follows:

Definition 1 (*Noise images*). Let $\mathbb{I}^N$ be a set of noise images, $\mathbb{I}^N \subseteq \mathbb{I}$, and $P = \{C_1, \ldots, C_p\}$ denote the set of cosine interesting patterns mined from $\mathbb{I}$. We have:

$$\mathbb{I}^N = \{i_s | \forall C_j \in P, C_j \not\subseteq i_s, i_s \in \mathbb{I}\}. \quad (1)$$

Definition 1 implies that the noise images are a set of images containing none of cosine interesting patterns. So the core task of noise filtering is to discovery cosine interesting patterns. In what follows, we firstly give cosine interestingness measure and its Conditional Anti-Monotone Property (CAMP), based on which we prove only mining the shortest CIPs is enough for noise filtering. Finally, we derive one of the upper bounds of cosine and its anti-monotone property (AMP) to pave the way for in-breadth pruning.

3.1. Cosine interestingness measure and its CAMP

By using the 2-way contingency table for two variables v_p and $v_{p'}$ (Tan et al., 2005), cosine similarity of the 2-itemsets $X = \{v_p, v_{p'}\}$ can be defined as

$$\cos(X) = \frac{supp(\{v_p, v_{p'}\})}{\sqrt{supp(\{v_p\})supp(\{v_{p'}\})}}, \quad (2)$$

where $supp(\cdot)$ is the support of an itemset. It is obvious that Eq. (2) can be naturally extended to the multi-itemsets case. For any set of multi-itemsets $X = \{v_1, \ldots, v_K\}$ ($K \geqslant 2$), the cosine similarity of X is defined as:

$$\cos(X) = \frac{supp(X)}{\sqrt[K]{\prod_{p=1}^{K} supp(\{v_p\})}}. \quad (3)$$

As is known, the anti-monotone property (AMP) is a famous property for applying the Apriori principle (Agrawal et al., 1993). The cosine similarity does not hold the AMP. However, here the Conditional Anti-Monotone Property (CAMP) is presented and the cosine similarity is proved to satisfy the CAMP.

Definition 2 (*Conditional Anti-Monotone Property*). Let $J = 2^{\mathbb{D}}$ be the power set of $\mathbb{D}$. A measure M holds the Conditional Anti-Monotone Property, if $\forall X, Y \subseteq J$, given that (1) $X \subseteq Y$, and (2) if $Y \backslash X \neq \emptyset$, $\forall v_p \in X$ and $v_{p'} \in Y \setminus X, supp(\{v_p\}) \leqslant supp(\{v_{p'}\})$, we have $M(X) \geqslant M(Y)$.

According to this definition, CAMP differs from AMP only in adding one more condition on the items in the difference set $Y \backslash X$. That is, the items in the difference set must have higher support values than the items in X. Based on this definition, we have:

Theorem 1. *Cosine similarity holds the Conditional Anti-Monotone Property.*

Proof. Without loss of generality, we assume that $X = \{v_1, \ldots, v_K\}$ is a K-itemset ($K \geqslant 1$), $Y = X \bigcup \{v_{K+1}, \ldots, v_{K+L}\}$ is a $(K+L)$-itemset ($L \geqslant 0$), with $supp(\{v_{K+l}\}) \geqslant supp(\{v_p\})$, $\forall 1 \leqslant l \leqslant L$, $1 \leqslant p \leqslant K$. Now it remains to show $\cos(X) \geqslant \cos(Y)$.

Note that If $L = 0$, we have $X = Y$, thus $\cos(X) = \cos(Y)$. Now assume $L \neq 0$, i.e., $Y \backslash X \neq \phi$. It is easy to note that $supp(X) \geqslant supp(Y)$.

Moreover, according to the definition of geometric mean, we have

$$\sqrt[K]{\prod_{p=1}^{K} supp(\{v_p\})} \leqslant \sqrt[K+L]{\prod_{p=1}^{K+L} supp(\{v_p\})}, \quad (4)$$

for $supp(\{v_{K+l}\}) \geqslant supp(\{v_p\})$, $\forall 1 \leqslant l \leqslant L$, $1 \leqslant p \leqslant K$. Accordingly, we finally have

$$\cos(X) = \frac{supp(X)}{\sqrt[K]{\prod_{p=1}^{K} supp(\{v_p\})}} \geqslant \frac{supp(Y)}{\sqrt[K+L]{\prod_{p=1}^{K+L} supp(\{v_p\})}} = \cos(Y), \quad (5)$$

which completes the proof. □

Theorem 1 implies that if an itemset is uninteresting in terms of cosine all of its supersets will be uninteresting. This important property of cosine makes it to work as *support* to prune uninteresting patterns before examining them, such as the Apriori-like algorithm named CosMiner proposed in Wu et al. (2012). However, in this paper, the CAMP of cosine is used to illustrate we only need to mine the shortest cosine interesting patterns directing at the application of noise filtering, which will be shown in the next subsection.

3.2. The shortest cosine interesting patterns

Let *min_supp*be the minimum support threshold and *min_cos* be the minimum cosine threshold. A set of items X is called a cosine interesting pattern, if $supp(X) \geqslant min\_supp$ and $\cos(X) \geqslant min\_cos$.

Based upon Eq. (2), the shortest CIPs definitely contain *two* items, since the cosine of *one* item is identically equal to 1. We have the following theorem:

Theorem 2. *For any CIP containing more than 2 items, at least one of its shortest sub-itemsets is also the cosine interesting pattern.*

Proof. Without loss of generality, let a CIP $X = \{v_1, v_2, \ldots, v_K\}$ is a K-itemset ($K \geqslant 3$), and items in X are sorted in support-ASCENDING order, i.e., $supp(v_1) \leqslant supp(v_2) \leqslant \cdots \leqslant supp(v_K)$. So, the subset $X' = \{v_1, v_2\}$ of X is one of the shortest patterns. Due to the CAMP of cosine, we have $\cos(X') \geqslant \cos(X)$, since X' contains two items with the lowest support count. X' is a CIP, which completes the proof. □

Recalling the process of noise filtering, that is, an image is regraded as the noise once it does not contain any CIP. Therefore, if an image contains a CIP, it naturally contains all of its subsets, and thus we have the following corollary:

Corollary 1. *Only mining the shortest cosine interesting patterns is enough for noise filtering.*

Corollary 1 is true due to the CAMP of cosine and the problem definition of the noise filtering. As the increase of total number of items d, the number of patterns will increase exponentially, which exerts high computational cost for mining algorithms. Corollary 1 inspires us to examine the 2-itemsets only, and thus helps to improve the efficiency of mining algorithms.

3.3. Cosine upper bound and its AMP

To discover the shortest cosine interesting patterns, we have to examine all 2-itemsets candidate patterns. Due to the high-dimensionality of images, examining C_d^2 2-itemsets is still not a trivial task. In this section, we take a further look to cosine interestingness measure to pruning some uninteresting 2-itemsets. Specially, we derive one of its upper bounds and study its anti-monotone property. We assume $X = \{v_p, v_{p'}\}$ is a 2-itemset and $supp(\{v_p\}) \leqslant supp(\{v_{p'}\})$. We have:

$$\cos(X) = \frac{supp(\{v_p, v_{p'}\})}{\sqrt{supp(\{v_p\})supp(\{v_{p'}\})}} \leqslant \frac{supp(\{v_p\})}{\sqrt{supp(\{v_p\})supp(\{v_{p'}\})}} = \sqrt{\frac{supp(\{v_p\})}{supp(\{v_{p'}\})}}. \quad (6)$$

Accordingly we have the following definition:

Definition 3 (*The cosine upper bound of 2-itemsets*). For a 2-itemset $X = \{v_p, v_{p'}\}$, the upper bound of the cosine similarity is defined as

$$u_c(X) = \sqrt{\frac{\min(supp(\{v_p\}), supp(\{v_{p'}\}))}{\max(supp(\{v_p\}), supp(\{v_{p'}\}))}}. \quad (7)$$

We then have the following theorem:

Theorem 3. *The upper bound of cosine as shown in Definition 3 holds the anti-monotone property.*

Proof. Without loss of generality, given a sequence of 2-itemsets $\mathcal{X} : X_1 = \left\{v_p, v_{p'}^1\right\}, X_2 = \left\{v_p, v_{p'}^2\right\}, \ldots, X_L = \left\{v_p, v_{p'}^L\right\}$ with $supp(\{v_p\}) \leqslant supp\left(\left\{v_{p'}^1\right\}\right) \leqslant supp\left(\left\{v_{p'}^2\right\}\right) \leqslant \cdots \leqslant supp\left(\left\{v_{p'}^L\right\}\right)$. Since $u_c(X_l) = \sqrt{\frac{supp(\{v_p\})}{supp\left(\left\{v_{p'}^l\right\}\right)}}$ and $supp\left(\left\{v_{p'}^l\right\}\right)$ is monotone in l, we have $\forall 1 \leqslant k \leqslant l \leqslant L$, $u_c(X_k) \geqslant u_c(X_l)$, which completes the proof. □

Theorem 3 paves the way for pruning uninteresting sibling itemsets, which is called *in-breadth pruning*. To prune uninteresting supersets using the AMP of support or the CAMP of cosine works as an in-depth style. Since our task is to discovery SCIPs, if we explore the 2-itemsets in support-ascending order, we might prune subsequent 2-itemsets with the same suffix in advance. This is why the cosine upper bound can be used to further improve the efficiency of SCIP mining.

4. The shortest cosine interesting pattern mining

In this section, we propose a novel algorithm for mining the shortest cosine interesting patterns. To understand it, we first illustrate the *in-breadth* pruning technique based on the AMP of cosine upper bound. Then we present the algorithmic details.

4.1. In-breadth pruning technique

Let us begin from a lattice view of itemsets, which is often used to illustrate the in-breadth traversing methods such as Apriori-like algorithms. Given the visual vocabulary $\mathbb{D}$ with d visual words in total, the search space of pattern mining consists of 2^d different subsets. However, for SCIP mining, we only have to examine C_d^2 candidate 2-itemsets. For instance, Fig. 1 shows the search space of five items *A, B, C, D, E* in a Hasse lattice.

It is not necessary to design Apriori-like or FP-growth-like algorithms to discover SCIPs, since both of them will lead to quite a lot of extra overhead. The straightforward way is to adopt the brute-force method, which examines all the candidate 2-itemsets one by one and checks whether their support and cosine satisfy the thresholds. So, the number of candidates is C_d^2 initially. Now, we aim to make use of Theorem 3 to further reduce the search space.

To this end, we should specify an appropriate examination sequence for all the candidate 2-itemsets. The items should be ranked in the increasing order of support so that they are added to the 2-itemsets in strict accordance with this order. Fig. 2 illustrates the sequence by a small visual vocabulary containing only five words with $supp(\{A\}) \leqslant \cdots \leqslant supp(\{E\})$. Then we firstly examine the item with lowest support and generate candidate 2-itemsets by adding items with higher support. For example, examining A can generate four candidate 2-itemsets $\{A,B\}$, $\{A,C\}$, $\{A,D\}$, and $\{A,E\}$. In this examination sequence, we can utilize the AMP of cosine upper bound shown in Definition 3 for in-breadth pruning. For instance, in Fig. 2, if we find $u_c(\{A,C\}) < min\_cos$, remaining two candidates $\{A,D\}$ and $\{A,E\}$ can be pruned safely, since cos

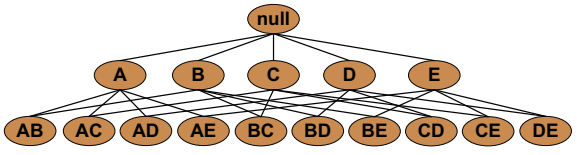

Fig. 1. A 2-itemsets lattice of five items.

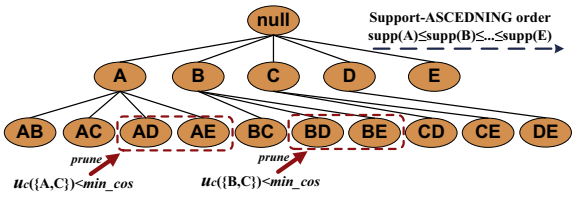

Fig. 2. Examination sequence and in-breadth pruning.

$(\{A,D\}) \leqslant u_c(\{A,D\}) \leqslant u_c(\{A,C\}) < min\_cos$ and cos $(\{A,E\})$ is the same.

4.2. Algorithm details

Algorithm 1. The SCIP mining algorithm

Input: Image dataset $\mathbb{I}$;
min_supp: the minimum support threshold;
min_cos: the minimum cosine threshold;
Output: P_2: the set of SCIPs , and $\emptyset$ is initially;
1 Scan $\mathbb{I}$ and compute the support of all items;
2 Remove infrequent items based on *min_supp*, and obtain the $\mathbb{D}$'s subset $\mathbb{D}^s = \{v_1^s, \cdots, v_{d'}^s\}, d' \leq d, supp(\{v_1^s\}) \leq \cdots \leq supp(\{v_{d'}^s\})$;
3 Create $d-1$ lists denoted as $*L[v_i^s], 1 \leq i \leq d-1$;
4 **for** *each line in* $\mathbb{I}$ **do**
5 For all pairs in this line, update L, i.e., $\forall p_{ij} = \{v_i^s, v_j^s\}, i \leq j, L[v_i^s][v_j^s] \leftarrow L[v_i^s][v_j^s] + 1$;
6 **end**
7 **for** *each* v_i^s *in* $\mathbb{D}^s$ **do**
8 **for** *each* v_j^s *in* $*L[v_i^s]$ **do**
9 **if** $supp(\{v_i^s, v_j^s\}) < min\_supp$ **then** continue;
10 **else if** $u_c\{(v_i^s, v_j^s\} < min\_cos$ **then** break;
11 **else if** $\cos(v_i, v_j) \geq min\_cos$ **then** $P_2 \leftarrow P_2 \bigcup \{v_i, v_j\}$;
12 **end**
13 **end**

Generally, the proposed algorithm for mining the shortest CIPs employs a brute-force strategy to enumerate all candidate 2-itemsets, and then pioneers the use of cosine upper bound based in-breadth pruning. The pseudocode for the shortest CIP mining algorithm is shown in Algorithm 1. Let $P_2 \subseteq P$ denote the set of CIPs with 2 items, i.e., the shortest CIPs. Some notable details are as follows:

- The algorithm initially makes a single pass over the image dataset $\mathcal{I}$ to determine the support of each item. Upon completion of the scan, we obtain a subset of all frequent 1-itemsets $\mathbb{D}^s$ that is sorted in support-ASCENDING order (see Lines 1–2).
- To store all candidate 2-itemsets, the algorithm creates $d-1$ link lists for every node except the dth node. Each element of link lists $L[v_i^s]\left[v_j^s\right]$ stores the support count of a 2-itemset of which the linked item v_j^s is more frequent than the head item v_i^s (see Line 3).
- To count the support of the candidate 2-itemsets, the algorithm needs to make an additional pass over the dataset (see Lines 4–6). Since the support count of every item has been obtained, the cosine and its upper bound can be computed after this step.
- Next, the algorithm traverses the link lists once to mine the shortest CIPs (see Lines 7–13). If the support and cosine of each candidate exceed the thresholds, the candidate is added to P_2. However, if the cosine upper bound of a candidate is lower than *min_cos*, the algorithm will omit the examination of remaining linked items, and turn to the next head item (see Line 10).

Let's now examine the efficiency of the algorithm. Specially, the space and time requirements are $O\left(\frac{d(d+1)}{2}\right)$ and $O\left(n \cdot C_{\bar{d}}^2\right)$, respectively, where $\bar{d}$ is the average number of non-empty features of each instance. The in-breadth pruning technique indeed can further speed up the mining process, which will be shown in Section 6.4.

5. The significance of CIPs on images

Based on the definition of the CIP, we actually use two measures, i.e., "the number of co-occurrences" and "the ratio of co-occurrences", to describe words in the CIP. That is, if the words in the CIP frequently appear as features of images and they always appear together, the pattern consisting of the words is considered to be interesting. Therefore, we argue that only the basic ingredients of the object can always appear together in its images with different shapes.

In this section, we try to sketch the regions represented by CIPs in order to show the significance of CIPs on images. To this end, we take `Oxford_5K` dataset as an example, and the detailed information about the dataset will be described in Section 6.1. Each visual word of `Oxford_5K` is an interest point extracted by the Harris detector (Mikolajczyk & Schmid, 2004), and thus the geometry feature of visual word is described by the feature centroid and three parameters of the ellipse. The shortest CIPs are then mined by the proposed algorithm with *min_supp* = 0.0988% and *min_cos* = 0.6, and four shortest CIPs of different landmarks are selected, as shown in Table 1.

By using the feature-displaying tool (http://www.robots.ox.ac.uk/vgg/research/affine/detectors.html), we marked the elliptic regions of words in SCIPs, as shown in Fig. 3. Note that some visual words corresponds to two or more ellipses in Fig. 3 due to the affine invariance (Mikolajczyk & Schmid, 2004). Regions marked in Fig. 3 can also be regarded as "interesting" points or regions in the media (Xie, Huang, Shen, Zhou, & Pang, 2012). As can be seen, the SCIP can successfully capture the key regions of the landmarks, which well reveal the feasibility of using SCIP for noise filtering. That is, no matter how the shapes of landmark in images change, the key regions of this landmarks will be maintained, and thus if an image does not contain any key region of the landmark, it can be confidently deemed as the noise.

6. Experimental analysis

In this section, we demonstrate the effectiveness of the proposed noise filtering method on two real-life image datasets.

6.1. Datasets

The evaluation is done on two real image datasets `Oxford_5K` (http://www.robots.ox.ac.uk/vgg/data/oxbuildings/index.html) and `LFW` (http://vis-www.cs.umass.edu/lfw/). Some characteristics of the used datasets are tabulated in Table 2. `Oxford_5K` comprises 11 different Oxford landmarks that actually are a particular part of buildings. Since `Oxford_5K` was retrieved from Flickr using 17 queries among which there are some general queries such as "Oxford" and "New Oxford", there exist a large number of noise images that are irrelevant to Oxford landmarks. A set of labels measuring the degree of clarity, i.e., "Good", "OK", "Junk", and "Absent" were manually assigned in Philbin, Chum, Isard, Sivic, and Zisserman

Table 1
Four marked shortest CIPs.

Landmarks	SCIP	*supp* (%)	cos
`All Souls`	{654840,595886}	0.0988	0.674
`Radcliffe Cam`	{143705,129703}	0.0988	0.913
`Hertford`	{864808,532455}	0.1186	0.600
`Bodleian`	{847144,827670}	0.0988	0.630

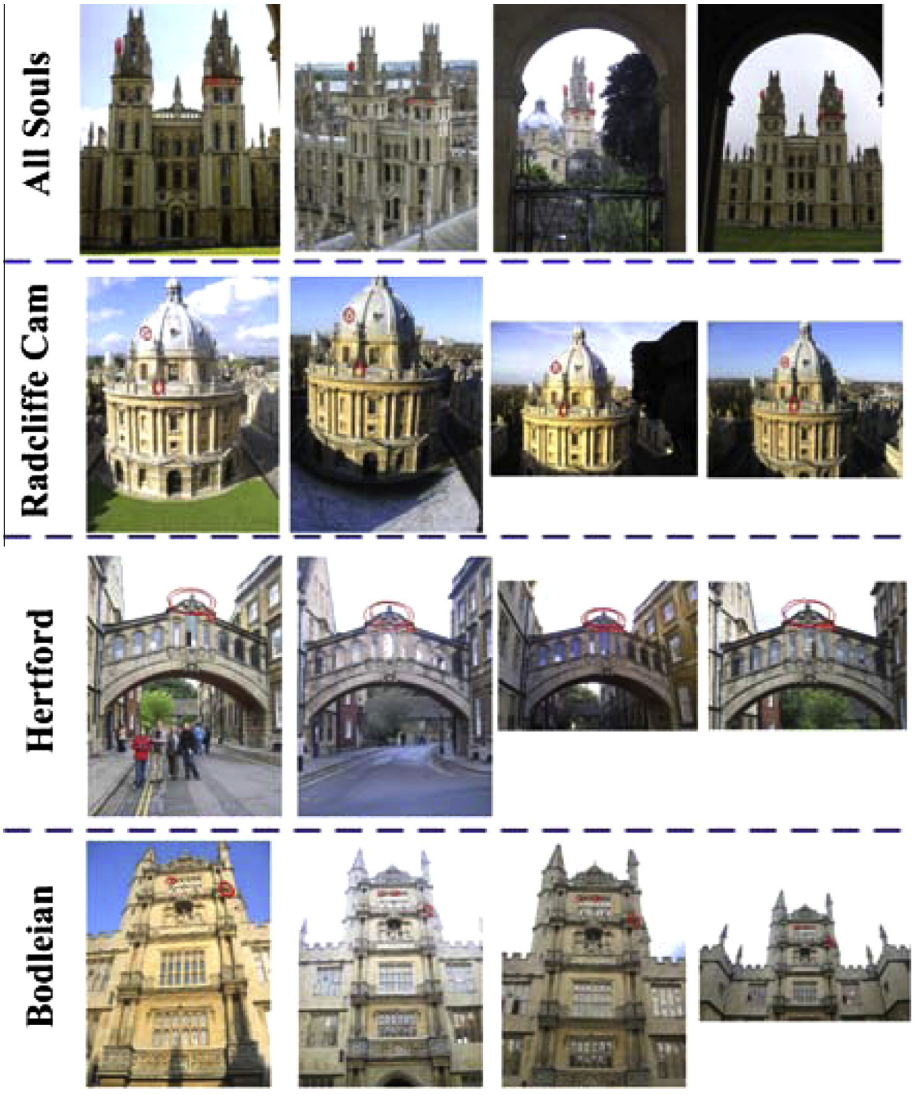

Fig. 3. The shortest CIPs marked on the images.

Table 2
Some characteristics of image datasets.

Name	#Instances	#Features	#Class	Density (%)	Noise (%)
Oxford_5K	5060	658346	11	0.0228	88.2
LFW	13233	3456	1680	11.886	30.1

(2007), and therefore if we take the images with more than 25% of the object clearly visible as normal instances, i.e., labeled 'Good" or "OK", there are only 568 normal instances in Oxford_5K, or equivalently, roughly 88.2% instances are noise.

The Labeled Faces in the Wild (LFW) dataset (http://vis-www.cs.umass.edu/lfw/) contains 13,233 faces detected in images downloaded from *Yahoo! News* in 2002–2003 (Huang, Ramesh, Berg, & Learned-Miller, 2007; Guillaumin, Verbeek, & Schmid, 2009). Each face image is labeled by the name of the person, and thus there are 5749 people appear in the images, and 1680 people having two or more images. From the clustering perspective, it is hard to group the people with only one face image into any cluster. Therefore, the 4069 people having only one image are considered to be the noise which accounts for 30.1%.

6.2. Performance on noise filtering

Here we illustrate the effect of noise filtering on two datasets. Table 3 shows four experimental cases on Oxford_5K, and in each case parameter setting of the SCIP mining algorithm, the number of SCIPs, the number of remaining images after noise filtering, and the number of clear images (labeled "Good" or "OK") among remaining images are given (http://www.robots.ox.ac.uk/vgg/data/oxbuildings/gt_files_170407.tgz/). We set *min_cos* around 0.48 to guarantee the mined patterns are interesting, and decrease *min_supp* to retain more images. As can be seen, as the increase of remaining images, the clear images also increase steadily. In

Table 3
The number of reserved clear images on Oxford_5K.

	Parameter Setting		#SCIPs	#Images	#Clear_Images
	min_supp (%)	*min_cos*			
Oxf_0	–	–	–	5060	568
Oxf_1	0.10	0.48	839	559	166
Oxf_2	0.09	0.48	3263	1001	368
Oxf_3	0.08	0.48	3375	1061	418
Oxf_4	0.08	0.45	3553	1352	**568**

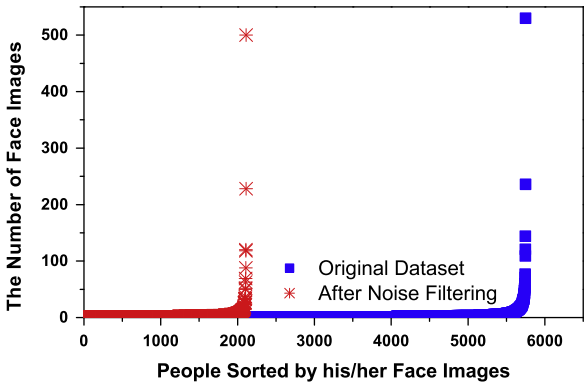

Fig. 4. The distribution of face images on LFW.

Table 4
The comparison on clustering performance.

Datasets	Parameter Setting		#SCIPs	#Images	NMI
	min_supp (%)	*min_cos*			
Oxford_5K (*K* = 11)	–	–	–	5060	0.157
	0.10	0.48	839	559	0.331
	0.09	0.48	3263	1001	0.483
	0.08	0.45	3553	1352	**0.534**
LFW (*K* = 1680)	–	–	–	13233	0.184
	40	0.60	21	3653	**0.703**
	35	0.55	321	6235	0.692
	20	0.50	907	11389	0.678

case Oxf_4, we can get all the clear images among 1352 remaining images. That is to say, we have successfully filtered 3708 noise images from 4492 noise images, i.e., about 82.5% noise images have been filtered, but retaining all clear images.

Fig. 4 depicts the distribution of face images on LFW. The red points are obtained by setting *min_supp* = 0.35% and *min_cos* = 0.55. As can be seen, we have totally filtered 7186 images including 2832 people with only one face image. That is to say, roughly 70% noise images, i.e., people with only one face image, have been successfully removed.

6.3. Improvement on clustering performance

In this section, we investigate the clustering performance on the image datasets before/after noise filtering, attempting to address the application value of the proposed noise filter. As the clustering algorithms (or tools) are not our emphasis, we select CLUTO (http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview) with default setting (i.e. *clmethod = direct, crfun = i2, sim = cosine, ntrials = 10*, and *colmodel = none*) as our clustering tool. This setting in CLUTO is indeed an implementation of Kmeans. Normalized Mutual Information (*NMI*) is used to evaluate the clustering performance (Cao, Wu, Wu, & Xiong, 2013), and it can be computed as: $NMI = I(U,V)/\sqrt{H(U)H(V)}$, where the random variables U and V denote the cluster and class sizes, respectively, $I(U,V)$ is the mutual information between U and V, and $H(U)$ and $H(V)$ are the Shannon entropies of U and V, respectively. *NMI* is in the range of [0,1], and a larger *NMI* indicates a better clustering performance.

Table 4 shows the clustering performance on two datasets. The clustering quality in original datasets is extremely poor due to the bad influence exerted by the vast noise (see the underlined data in Table 4). After filtering the noise using SCIPs, the clustering quality has been improved significantly. Based on the clustering results, if we label the cluster by the name of the most frequent landmark or person, we can get the number of images that were misclassified. Fig. 5 shows four sampling clusters from two datasets in the case of the highest *NMI* (see the bold data in Table 4). Note that Tommy Franks (12/19) indicates there are 12 right images among 19 images, and some misclassified images are marked in the rectangle. As can be seen, most of the images are correctly classified in every cluster. Meanwhile, the quality of some easy cases such as All Souls comes near to perfection.

6.4. Inside the SCIP mining

In this subsection, we take a further step to explore the effectiveness of the in-breadth pruning inside the SCIP mining algorithm. Since the in-breadth pruning is irrelevant to *min_supp*, we set *min_supp* = 0 to mine interesting patterns from the whole visual vocabulary. We then compare the runtime of SCIP mining with and without in-breadth pruning. Note that the runtime includes the I/O time for traversing the dataset and the execution time for examining the candidate 2-itemsets.

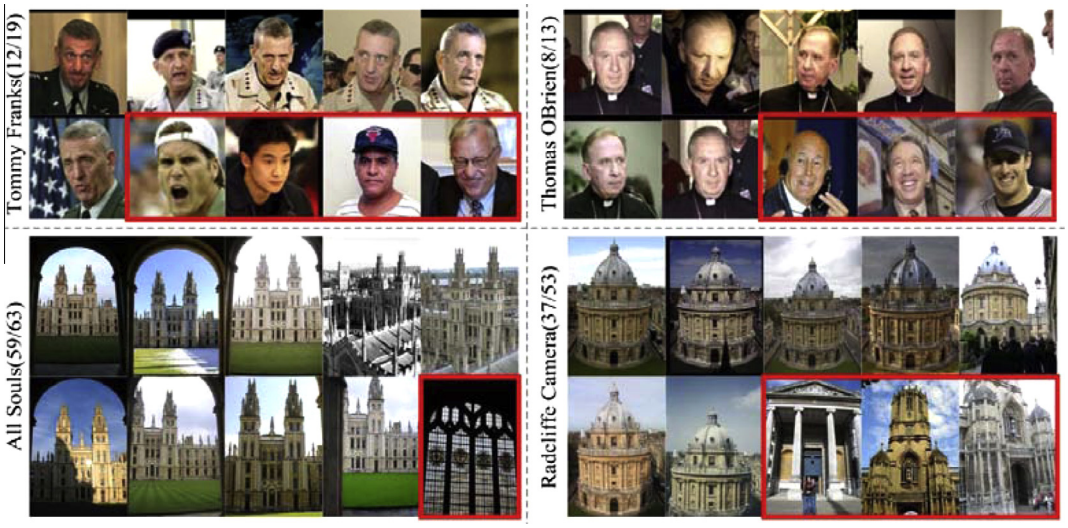

Fig. 5. Sample clusters by CLUTO on Oxford_5K and LFW.

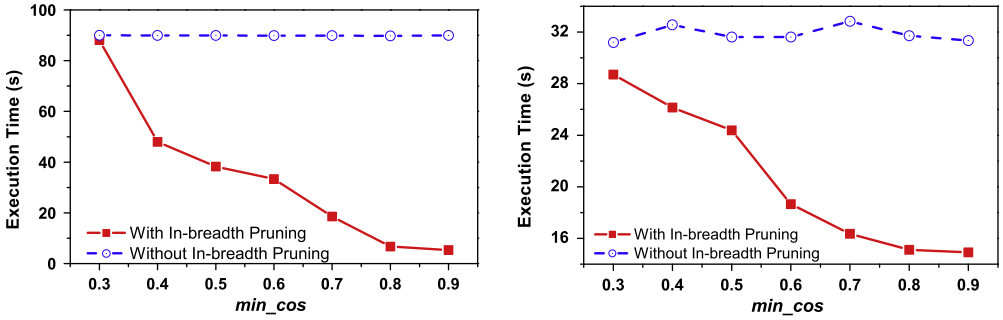

Fig. 6. Runtime comparison between SCIP mining algorithm with and without in-breadth pruning.

Fig. 6 shows the results of runtime with *min_cos* values varying from 0.3 to 0.9. As can be seen, the case without in-breadth pruning is less sensitive to the change of *min_cos*, since it always examine all of the candidates. However, the in-breadth pruning technique reduces the runtime remarkably as the increase of *min_cos*, since there will be more and more 2-itemsets of which the cosine upper bounds are lower than *min_cos*. Meanwhile, we notice that the runtime on `Oxford_5K` is higher than that on `LFW`. This is because the number of features of `Oxford_5K` is much higher than `LFW`, which leads to a higher cost for traversing the dataset twice.

7. Conclusion

This paper studied the problem of filtering noise images, rather than noise inside images, from a large number of normal images. In particular, we reveal the cosine interesting pattern can capture essential feature of the object in images, and thus present a novel noise filtering method based on the cosine interesting pattern mining. We prove that to filter noise only requires to mine the shortest cosine interesting patterns, which dramatically simplifies the mining process. Furthermore, we exploit the anti-monotone property of cosine upper bound for in-breadth pruning, which makes the SCIP mining algorithm more efficient. Experimental results on two real-life image datasets have demonstrated our noise filter can remove most of irrelevant, vague, and/or incomplete images and thus improve the performance of further analysis (e.g., cluster analysis).

There are a wealth of research directions that we are currently considering, such as expanding other interestingness measures, improving other data mining techniques, and applying the noise filter to real-life case study, and more.

Acknowledgment

This research was partially supported by National Natural Science Foundation of China (Nos. 71072172, 61103229, 61003074), National Key Technologies R& D Program of China (Nos. 2013BAH16F01, 2013BAH16F04), Jiangsu Provincial Colleges and Universities Outstanding S& T Innovation Team Fund (No. 2011013), Key Project of Natural Science Research in Jiangsu Provincial Colleges and Universities (No. 12KJA520001), the Natural Science Foundation of Jiangsu Province of China (No. BK2012863), The Program of Natural Science Foundation of Zhejiang Province (No. LY13F020008), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, Washington, DC, USA.

Al-amri, S. S., Kalyankar, N. V., & Khamitkar, S. D. (2010). A comparative study of removal noise from remote sensing image. Compuing Research Repository. abs/1002.1148. arXiv:1002.1148.

Bayardo, R. J., Ma, Y., & Srikant, R. (2007). Scaling up all pairs similarity search. In *Proceedings of the 16th international conference on world wide web*, Banff, Alberta, Canada.

Buades, A., Coll, B., & Morel, J. M. (2008). Nonlocal image and movie denoising. *International Journal of Computer Vision, 76*(2), 123–139.

Cao, J., Wu, Z., Wu, J., & Xiong, H. (2013). SAIL: Summation based incremental learning for information-theoretic text clustering. *IEEE Transactions on Cybernetics, 43*, 570–584.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys, 38*(3), 9.

Guillaumin, M., Verbeek, J., & Schmid, C. (2009). Is that you? metric learning approaches for face identification. In *IEEE 12th international conference on computer vision* (pp. 498–505).

Gu, G., Zhao, Y., & Zhu, Z. (2011). Integrated image representation based natural scene classification. *Expert Systems with Applications, 38*(9), 11273–11279.

Hajiaboli, M. (2011). An anisotropic fourth-order diffusion filter for image noise removal. *International Journal of Computer Vision, 92*(2), 177–191.

Hindle, A., Shao, J., Lin, D., Lu, J., & Zhang, R. (2011). Clustering web video search results based on integration of multiple features. *World Wide Web, 14*(1), 53–73.

Huang, G., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report. University of Massachusetts, Amherst.

Leipsic, J., LaBounty, T. M., Heilbron, B., Min, J. K., Mancini, G. B. J., Lin, F. Y., et al. (2010). Adaptive statistical iterative reconstruction: Assessment of image noise and image quality in coronary ct angiography. *American Journal of Roentgenology, 195*(3), 649–654.

Li, F.F., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *IEEE conference on computer vision and pattern recognition* (pp. 524–531).

Li, Z., Liu, J., Chen, S., & Tang, X. (2007). Noise robust spectral clustering. In *IEEE 11th international conference on computer vision* (pp. 1–8).

Lin, T. C. (2011). Decision-based fuzzy image restoration for noise reduction based on evidence theory. *Expert Systems with Applications, 38*(7), 8303–8310.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal Computer Vision, 60*(1), 63–86.

Mistry, M., & Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics, 9*(1), 327.

Philbin, J., Chum, O., Isard, M., Sivic, J., & Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Sevgen, S., & Arik, S. (2011). On-chip template training system and image processing applications using iterative annealing on ace16 k chip. *Expert Systems with Applications, 38*(10), 12900–12905.

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison-Wesley Publication.

Van De Ville, D., Nachtegael, M., Van der Weken, D., Kerre, E. E., Philips, W., & Lemahieu, I. (2003). Noise reduction by fuzzy image filtering. *IEEE Transactions on Fuzzy Systems, 11*(4), 429–436.

Wang, C., Li, Y., & Song, X. (2013). Video-to-video face authentication system robust to pose variations. *Expert Systems with Applications, 40*, 722–735.

Wu, J., Zhu, S., Liu, H., & Xia, G. (2012). Cosine interesting pattern discovery. *Information Sciences, 184*(1), 176–195.

Xie, Q., Huang, Z., Shen, H. T., Zhou, X., & Pang, C. (2012). Quick identification of near-duplicate video sequences with cut signature. *World Wide Web, 15*(3), 355–382.

Zhao, Y., & Karypis, G. (2004). Criterion functions for document clustering: Experiments and analysis. *Machine Learning, 55*, 311–331.