An Unsupervised Approach for Person Name Bipolarization Using Principal Component Analysis

Chien Chin Chen, Zhong-Yong Chen, and Chen-Yuan Wu

Abstract—A topic is usually associated with a specific time, place, and person(s). Generally, topics that involve bipolar or competing viewpoints are attention getting and are thus reported in a large number of documents. Identifying the association between important persons mentioned in numerous topic documents would help readers comprehend topics more easily. In this paper, we propose an unsupervised approach for identifying bipolar person names in a set of topic documents. Specifically, we employ principal component analysis (PCA) to discover bipolar word usage patterns of person names in the documents, and show that the signs of the entries in the principal eigenvector of PCA partition the person names into bipolar groups spontaneously. To reduce the effect of data sparseness, we introduce two techniques, called the weighted correlation coefficient and off-topic block elimination. We also present a timeline system that shows the intensity and activeness development of the identified bipolar person groups. Empirical evaluations demonstrate the efficacy of the proposed approach in identifying bipolar person names in topic documents, while the generated timelines provide comprehensive storylines of topics.

Index Terms—Topic mining, sentiment analysis, bipolar timeline

1 Introduction

With the advent of Web 2.0, many online collaborative tools, such as web logs and discussion forums, are being developed to allow Internet users to express their opinions on a wide variety of topics via web documents. One benefit is that the web has become an invaluable knowledge base for Internet users to learn about topics. Since the essence of Web 2.0 is knowledge sharing, collaborative tools are designed with the minimum of constraints so that users will be motivated to contribute their knowledge. As a result, the number of topic documents on the Internet is growing exponentially. To help Internet users comprehend numerous topic documents quickly and easily, topic mining techniques, such as timeline mining [1], are essential.

Existing topic mining approaches focus on extracting important themes in documents of interest. Basically, a topic consists of a sequence of related events associated with a specific time, place, and person(s) [2]. Topics that involve bipolar (or competing) viewpoints are often attention-getting and generate a large number of documents. However, if people are not familiar with the topics, they may have to expend a great deal of time figuring out the association between important persons mentioned in the documents in order to fully comprehend the topics. Identifying the polarity of the named entities in topic documents, especially person names, would help readers comprehend the topic quickly and easily. For instance, for American presidential elections, Internet users can find numerous web documents about the Democratic and Republican parties. Identifying the names of important people in the competing parties would help readers form a balanced view of the campaign.

In this paper, we define a topic person name bipolarization research method. Given a topic that involves bipolar viewpoints, the method clusters important persons mentioned in the topic documents into sentiment-coherent groups. For instance, if the method is applied to a set of documents about an American presidential election, it processes the person names mentioned in the documents and identifies important members of the Democratic and Republican parties automatically. Although our research is closely related to sentiment analysis [3], which focuses on discovering bipolar text units mentioned in a set of documents, it differs in a number of respects. First, most sentiment analysis approaches identify the polarity of adjectives, adverbs, and verbs. Comparatively few works consider the polarity of named entities. To the best of our knowledge, this is the first work that considers the polarity of person names. Second, sentiment analysis methods normally classify text units in terms of positive orientation or negative orientation, but the polarity of persons may not have positive or negative meanings. Specifically, persons with different polarities hold opposite opinions about a certain topic (or issue), while persons in the same polarity group reach a consensus or have the same goal. Finally, sentiment analysis usually requires external knowledge sources or human-composed sentiment lexicons, such as WordNet [4] and General Inquirer,[1] to determine the orientation of a text unit. However, the polarity

• *The authors are with the Department of Information Management, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei City 106, Taiwan, R.O.C.*
E-mail: paton@im.ntu.edu.tw, {d98725003, r97725035}@ntu.edu.tw.

Manuscript received 22 Oct. 2010; revised 18 July 2011; accepted 1 Aug. 2011; published online 5 Aug. 2011.
Recommended for acceptance by X. Zhu.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2010-10-0561.
Digital Object Identifier no. 10.1109/TKDE.2011.177.

1. http://www.wjh.harvard.edu/~inquirer/.

of a person name is dynamic and context-dependent, so no external knowledge source is available for person name bipolarization research. For instance, politicians may agree (or disagree) about a particular topic, but that does not mean they are permanent friends (enemies). The property of context-dependence makes the person name bipolarization task a particularly challenging research issue.

To resolve the problem, we propose an unsupervised approach that identifies bipolar groups of person names in a set of topic documents automatically. Specifically, we use principal component analysis (PCA) [5] to discover bipolar word usage patterns of important person names in a set of topic documents, and show that the signs of the entries in the principal eigenvector of PCA partition the person names in bipolar groups spontaneously. We also present two techniques, called off-topic block elimination and weighted correlation coefficient, to reduce the effect of data sparseness on person name bipolarization. Finally, the occurrences of the identified bipolar person names are organized chronologically to form an activeness timeline of the topic of interest. As the approach simply analyzes word usage patterns of person names in topic documents, it can be applied to different topic domains and languages. The results of experiments based on 12 topic document sets written in English and Chinese demonstrate that the proposed PCA-based approach is effective in identifying bipolar groups of person names. Moreover, the generated activeness timelines describe the storylines of topics comprehensively.

The remainder of this paper is organized as follows: Section 2 contains a review of related works on sentiment analysis, person name clustering, and topic timeline mining. We describe the proposed person name bipolarization approach in Section 3, and evaluate its performance in Section 4. Then, in Section 5, we present our conclusions.

2 Related Work

Our survey of the literature on person name bipolarization revealed that there are surprisingly few related works. This is probably because the research subject is relatively new. Essentially, the technique clusters person names in topic documents into bipolar groups. In this section, we consider two closely related research subjects, namely, person name clustering and sentiment analysis. We also discuss topic timeline mining and explain how it differs from person name bipolarization.

2.1 Person Name Clustering

Person name clustering has attracted a considerable amount of attention in recent years because using person names to search for information is one of the most popular types of searches on the Internet [6]. However, when a person name is input to a search engine, the returned webpages may contain information about more than one person, so it may be difficult for the user to find the desired information. The goal of person name clustering (a.k.a. person name disambiguation) is to facilitate searching with person names (called person name searches hereafter) by partitioning the returned webpages into clusters, each of which represents a specific person. The Web People Search (WePS) evaluation workshops [6] provide various data sets to promote the development of efficient and effective person name clustering methods. Most person name clustering methods are based on the assumption that each returned page refers to a single person [6]. In general, personal attributes, such as email addresses are extracted from webpages to cluster contextually similar pages into clusters. Wan et al. [7] developed the WebHawk system to facilitate person name searches on the web. The system uses information extraction techniques to obtain names, job titles, organizations, and e-mail addresses from a webpage. The attributes are then combined with lexical features (e.g., bigrams in the documents) to disambiguate person names. Kalashnikov et al. [8] utilized search engines to measure the social similarity of webpages, and demonstrated that webpages referring to the same person often mention similar named entities, such as organizations. Specifically, the representative named entities mentioned in a pair of pages are submitted to a search engine, and the number of returned pages indicates the degree of social similarity of the pages. Song et al. [9] employed latent semantics analysis techniques to disambiguate person names and observed that namesakes usually have different interests. By comparing the distributions of interests, modeled by probabilistic latent variables in the webpages, namesakes can be disambiguated. To avoid merging persons with similar interests, the string differences between person names are considered.

The proposed approach differs from person name clustering because it generates clusters that possess bipolar orientations. The bipolar property makes person name bipolarization a unique and challenging research subject.

2.2 Sentiment Analysis

Our research is closely related to sentiment analysis, which attempts to identify the polarity (or sentiment) of a word in order to extract positive or negative sentences from documents [3]. Hatzivassiloglou and McKeown [10] showed that language conjunctions, such as *and*, *or*, and *but*, are effective indicators for judging the polarity of conjoined adjectives. The authors observed that most conjoined adjectives (77.84 percent) have the same orientation, while conjunctions that use *but* generally connect adjectives of different orientations. They proposed a log-linear regression model that learns the distributions of conjunction indicators from a training corpus to predict the polarity of conjoined adjectives. Turney and Littman [11] manually selected seven positive and seven negative words as a polarity lexicon and used pointwise mutual information (PMI) to calculate a word's polarity. A word has a positive orientation if it tends to co-occur with positive words; otherwise, it has a negative orientation. More recently, Esuli and Sebastiani [12] developed a lexical resource, called SentiWordNet, which calculates the degrees of objective, positive, and negative sentiments of a synset in WordNet. The authors employed a bootstrap strategy to collect training data sets for the sentiments and trained eight sentiment classifiers to assign sentiment scores to a synset. Meanwhile, Kanayama and Nasukawa [13] posited that polar clauses with the same polarity tend to appear successively in contexts. Their approach derives the coherent precision and coherent density of a word in a training

corpus to predict the word's polarity. Ganapathibhotla and Liu [14] investigated comparative sentences in product reviews. To identify the polarity of a comparative word (e.g., longer) with a product feature (e.g., battery life), the authors collected phrases that describe the pros and cons of products from Epinions.com and proposed using one-side association (OSA), which is a variant of PMI. OSA assigns a positive (resp. negative) orientation to a comparative-feature combination if the synonyms of the comparative word and feature tend to co-occur in the pros (resp. cons) phrases.

Our research differs from existing sentiment analysis approaches in a number of respects. First, most works on sentiment analysis identify the polarity of adjectives, adverbs, and verbs because the syntactic constructs generally express sentimental semantics. In contrast, our approach identifies the polarity of person names. Second, to the best of our knowledge, all existing polarity identification methods require external information sources, such as WordNet, manually selected polarity words, or training corpora. However, our approach identifies bipolar person names by simply analyzing person name usage patterns in topic documents without using external information. Finally, the proposed approach does not require any language constructs, such as conjunctions; hence, it can be applied to different languages.

2.3 Topic Timeline Mining

Topic timeline mining involves constructing a timeline to describe the development of a topic reported by a number of topic documents. As a topic is associated with seminal themes [15], mining methods need to identify the core themes in topic documents. Then, the activeness of the themes is measured chronologically to depict a topic's development. Kleinberg [1] proposed a mining technique that constructs a tree-like topic timeline from a series of topic documents. If the documents contain bursty information, hidden Markov models are used to model the activeness status of the topic and split it into active themes, modeled as tree nodes and branches. Nallapati et al. [2] and Feng and Allan [16] considered topic timeline mining as a document clustering problem. To identify the active themes of a topic, the topic documents are first grouped into significant clusters; then the clusters are connected chronologically to form a topic timeline. Mei and Zhai [15] also employed hidden Markov models to construct topic timelines. They modeled a theme as a language model and developed an EM algorithm to extract important themes from topic documents. The extracted themes are regarded as states of hidden Markov models and used to search for the best state sequence in a set of topic documents. The state sequence reveals the variation in the themes' strengths and depicts the activeness trend of the themes over the topic's lifespan. Chen and Chen [17] proposed an eigenvector-based approach to identify important themes in topic documents. They showed that the amplitude of an entry in an eigenvector determines the degree of correlation between a topic block (e.g., a set of topic sentences) and a theme, and described the activeness trend of a theme in terms of amplitude variations.

Existing works on topic timeline mining focus on extracting important themes from topic documents. To the best of our knowledge, no timeline mining approach considers the concept of polarity activeness. In an attempt to bridge this research gap, we identify bipolar person groups in topic documents and produce a timeline of the groups.

3 Method

In this section, we present our data model and the PCA-based approach for bipolar person name identification.

3.1 Data Model

Given a set of documents related to a bipolar or competing topic, we first decompose the documents into a set of nonoverlapping blocks $B = \{b_1, b_2, \ldots, b_n\}$. A block can be a paragraph or a document, depending on the granularity of PCA sampling. As there are no constraints on web document writing, counter bipolarization examples may exist in B, which would affect the bipolarization performances. To initiate the research of topic person name bipolarization, we assume that B does not contain a counter example. Moreover, in our evaluation we used web news documents to avoid counter examples, since news documents are written by well-trained journalists. Let $P = \{p_1, p_2, \ldots, p_m\}$ be a set of person names in B. Then, the document set can be represented as an $m \times n$ person-block association matrix A. A column in A, denoted as $\underline{b}_i$, represents a decomposed block i. It is an m-dimensional vector whose j'th entry, denoted as $b_{i,j}$, is the frequency of p_j in b_i. Meanwhile, a row in A, denoted as $\underline{p}_i$, represents a person i. It is an n-dimensional vector whose j'th entry, denoted as $p_{i,j}$, is the frequency of p_i in b_j.

3.2 PCA-Based Person Name Bipolarization

Principal component analysis is a well-known statistical method that is used primarily to identify the most important feature pattern in a high-dimensional data set [5]. In our research, we use PCA to identify the most important person pattern in the topic blocks by first constructing an $m \times m$ person relation matrix C, in which the (i,j)-entry (denoted as $c_{i,j}$) denotes the correlation coefficient of p_i and p_j. The correlation is computed as follows:

$$\begin{aligned} c_{i,j} &= corr(\underline{p}_i, \underline{p}_j) \\ &= \frac{\sum_{k=1}^{n} \left(p_{i,k} - p_i^{\sim}\right) * \left(p_{j,k} - p_j^{\sim}\right)}{\sqrt{\sum_{k=1}^{n} \left(p_{i,k} - p_i^{\sim}\right)^2} * \sqrt{\sum_{k=1}^{n} \left(p_{j,k} - p_j^{\sim}\right)^2}}, \end{aligned} \quad (1)$$

where $p_i^{\sim} = 1/n \sum_{k=1}^{n} p_{i,k}$ and $p_j^{\sim} = 1/n \sum_{k=1}^{n} p_{j,k}$ are the average frequencies of persons i and j, respectively. The range of $c_{i,j}$ is within [-1,1] and the value represents the degree of correlation between p_i and p_j under the decomposed blocks. If $c_{i,j} = 0$, we say that p_i and p_j are uncorrelated; that is, occurrences of person i and person j in the blocks are independent of each other. However, if $c_{i,j} > 0$, we say that persons i and j are positively correlated. That is, p_i and p_j tend to co-occur in the blocks; otherwise, both tend to be absent simultaneously. Conversely, if $c_{i,j} < 0$, we say that p_i and p_j are negatively correlated; that is, if one person appears in a block, the other tends not to appear in the block at the same time. Note that if $c_{i,j} \neq 0, |c_{i,j}|$ scales the strength of a positive or negative correlation. Moreover,

since the correlation coefficient is commutative, $c_{i,j}$ will be identical to $c_{j,i}$ such that the matrix C will be symmetric [18].

A person pattern is represented as a vector $\underline{v}$ of dimension m in which the i'th entry v_i indicates the weight of person i in the pattern. Since the matrix C depicts the correlation of the persons in the topic blocks, given a constitution of $\underline{v}$, $\underline{v}^T C \underline{v}$ computes the variance of the pattern to characterize the persons. A pattern is deemed important if it characterizes the variance of the persons specifically. PCA can then identify the most important person pattern by using the following object function:

$$\max \ \underline{v}^T C \underline{v}, \tag{2}$$

$$\text{s.t. } \underline{v}^T \underline{v} = 1. \tag{3}$$

Without specifying any constraint on $\underline{v}$, the object function becomes arbitrarily large with large values of $\underline{v}$. The constraint $\underline{v}^T \underline{v} = 1$ limits the search space to within the set of length-normalized unit vectors. Lagrange multiplier techniques [19] solve the above constrained optimization problem by constructing the following Lagrangian function Z:

$$Z(\underline{v}, \lambda) = \underline{v}^T C \underline{v} + \lambda\big(1 - \underline{v}^T \underline{v}\big). \tag{4}$$

Then, the stationary points of the function can be derived as follows:

$$\partial Z / \partial \nu = 2C\underline{v} - 2\lambda \underline{v} = 0. \tag{5}$$

Equation (5) implies that $C\underline{v} = \lambda \underline{v}$. In other words, $\underline{v}$ is a unit eigenvector of C and λ is the corresponding eigenvalue. The following theorem of symmetric matrices [18] shows that C always contains unit eigenvectors, so the constrained optimization problem is solvable.

Theorem 1. *Any mxm matrix C is symmetric if and only if there is an orthonormal basis V for* R^{m} *(i.e., the m-dimensional vector space) and a diagonal matrix D, such that* $C = VDV^{-1}$. $V = \{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_m\}$ *consists of the unit eigenvectors of C; and the diagonal entries of D are eigenvalues that correspond to the respective columns of V.*

PCA is not the only method that identifies important feature patterns in terms of eigenvectors. For instance, Gong and Liu [20], and Chen and Chen [17] utilized the eigenvectors of symmetric feature relation matrices to extract salient concepts and salient themes from documents, respectively.[2] A major difference between our PAC-based approach and other eigenvector-based pattern mining approaches lies in the way the relation matrix is constructed. We calculate $c_{i,j}$ by using the correlation coefficient, whereas other approaches employ the inner product or cosine similarity [21] to derive the relationships between text features. Specifically, by converting each $\underline{p}_i$ into a mean-normalized unit vector, the correlation coefficient is identical to the inner product and the cosine similarity:[3]

$$\begin{aligned} corr(\underline{p}_i, \underline{p}_j) &= \frac{\sum_{k=1}^{n} \left(p_{i,k} - p_i^{\sim}\right) * \left(p_{j,k} - p_j^{\sim}\right)}{\sqrt{\sum_{k=1}^{n} \left(p_{i,k} - p_i^{\sim}\right)^2} * \sqrt{\sum_{k=1}^{n} \left(p_{j,k} - p_j^{\sim}\right)^2}} \\ &= \frac{\sum_{k=1}^{n} p_{i,k}^* * p_{j,k}^*}{\sqrt{\sum_{k=1}^{n} p_{i,k}^{*2}} * \sqrt{\sum_{k=1}^{n} p_{j,k}^{*2}}} \\ &= \text{cosine}\big(\underline{p}_i^*, \underline{p}_j^*\big) \\ &= \underline{\underline{p}}_i \cdot \underline{\underline{p}}_j, \end{aligned} \tag{6}$$

where $\underline{p}_i^* = \underline{p}_i - p_i^{\sim}[1, 1, \ldots, 1]^T$ and $\underline{p}_j^* = \underline{p}_j - p_j^{\sim}[1, 1, \ldots, 1]^T$ are the mean-normalized vectors of $\underline{p}_i$ and $\underline{p}_j$, respectively; and $\underline{p}_i = \underline{p}_i^* / |\underline{p}_i^*|$ and $\underline{p}_j = \underline{p}_j^* / |\underline{p}_j^*|$ denote the unit vectors of $\underline{p}_i^*$ and $\underline{p}_j^*$, respectively. Specifically, the mean normalization process differentiates our approach from other eigenvector-based approaches. Before discussing the effect of the mean normalization process on person name bipolarization, we consider another difference between our approach and other eigenvector-based approaches. As mentioned in Section 2, the person name bipolarization task is a clustering problem. Existing eigenvector-based approaches perform clustering by treating each eigenvector as a pattern and use more than one eigenvector to identify feature clusters. However, in our approach, we use a single eigenvector (i.e., the principal eigenvector) of C to cluster persons into bipolar groups. As the correlation coefficient is the inner product of the mean-normalized unit vectors, the matrix C can be computed as follows:

$$C = \underline{\underline{A}}\,\underline{\underline{A}}^T, \tag{7}$$

where $\underline{\underline{A}}$ is the mean-normalized unit matrix of A and a row i in $\underline{\underline{A}}$ is $\underline{p}_i$. The theorem of singular value decomposition [18] indicates that the eigenvalues of any matrix multiplied by its transpose (i.e., HH^T; H is a matrix) must be greater than or equal to zero. Consequently, the diagonal entries of D in our matrix C are nonnegative. In addition, as V is an orthonormal basis of R^m, its inverse is identical to its transpose, i.e., $V^{-1} = V^T$ [18]. Therefore, the matrix C can be represented as follows:

$$\begin{aligned} C &= VDV^{-1} = VDV^T \\ &= [\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_m][d_{1,1}\underline{e}_1, d_{2,2}\underline{e}_2, \ldots, d_{m,m}\underline{e}_m]V^T \\ &= [d_{1,1}\underline{v}_1, d_{2,2}\underline{v}_2, \ldots, d_{m,m}\underline{v}_m][\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_m]^T \\ &= d_{1,1}\underline{v}_1\underline{v}_1^T + d_{2,2}\underline{v}_2\underline{v}_2^T + \cdots + d_{m,m}\underline{v}_m\underline{v}_m^T, \end{aligned} \tag{8}$$

where $\{d_{1,1}, d_{2,2}, \ldots, d_{m,m}\}$ are the diagonal entries of D and the set of $\underline{e}_i$'s are the standard vectors of R^m. Specifically, the matrix C is the weighted sum of m matrices spanned by its eigenvectors; and the scale of the nonnegative eigenvalues determines the strength of an eigenvector in characterizing the variance of the topic persons. Hence, we select the eigenvector with the largest eigenvalue (i.e., the principal eigenvector) for person name bipolarization.

According to Theorem 1, the eigenvectors of a symmetric matrix form an orthonormal basis of R^m; therefore, they contain negative entries [18]. Even though Kleinberg [23] and Chen and Chen [17] showed experimentally that the negative entries in an eigenvector are as important as the positive entries for describing a certain feature pattern, the meaning of negative entries in their approaches is unexplainable. This is

2. The right singular vectors of a matrix H used by Gong and Liu [20] are equivalent to the eigenvectors of the symmetric matrix $H^T H$ whose entries are the inner products of the corresponding columns of H [18].

3. The inner product is equivalent to the cosine similarity when the lengths of the calculated vectors are normalized; that is, the vectors are unit vectors [22].

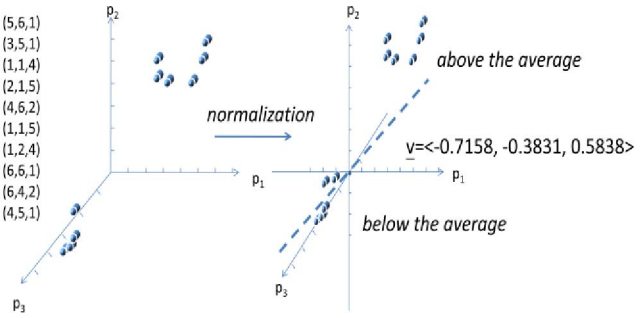

Fig. 1. The effect of the mean normalization process.

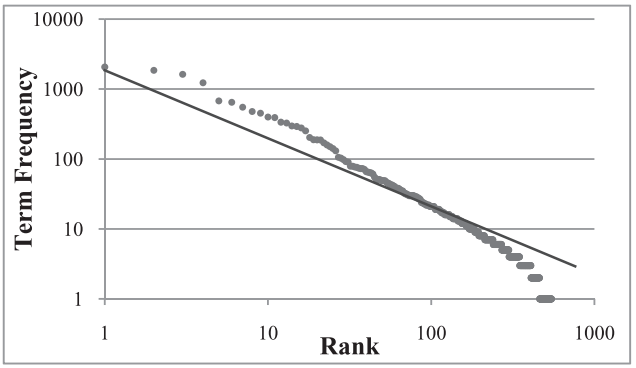

Fig. 2. The rank-frequency distribution of person names on logarithmic scales (base 10).

because text features (e.g., terms, sentences, or documents) in information retrieval or text mining are usually characterized by frequency-based metrics, such as the term frequency, document frequency, or TFIDF [21], which can never be negative. In PCA, however, the mean normalization process of the correlation coefficient assigns a bipolar meaning to positive and negative entries and that helps us partition person names into bipolar groups in accordance with their signs in $\underline{v}$.

The synthesized example in Fig. 1 illustrates the effect of the mean normalization process. In this example, there are three person names p_1, p_2, and p_3; and the corpus consists of 10 blocks. Graphically, each block can be represented as a point in a 3D vector space. The mean normalization process moves the origin of the 3D vector space to the centroid of the blocks, which makes the negative entry values explainable. A negative entry of a block vector in the mean-normalized vector space indicates that the number of occurrences of a person in the block is less than the person's average; conversely, a positive entry means that the number of occurrences of a person in a block is above the average count. In the figure, the principal eigenvector $\underline{v} = <-0.7158, -0.3831, 0.5838>$ calculated by PCA is represented by the dashed line. The signs of $\underline{v}$'s entries indicate that if p_3 occurs frequently in a block, then the probability of observing p_1 and p_2 in the same block will be lower than expected. In addition, as the signs of entries in an eigenvector are invertible [18], the constitution of $\underline{v}$ also claims that the occurrence of p_2 will be higher than the average if p_1 occurs frequently in a block; however, p_3 tends not to occur in the same block simultaneously. The instances of bipolar word usage behavior identified in $\underline{v}$ are consistent with the distribution of the 10 blocks. As mentioned in Section 2, Kanayama and Nasukawa validated that polar text units with the same polarity tend to appear together to make the contexts coherent. Consequently, we believe that the signs in PCA's principal eigenvector are effective in partitioning person names into bipolar groups.

3.3 Sparseness of Text Features

When PCA is used to process textual data, the sparseness of text features is a major problem. To demonstrate the problem, we collected 411 news documents related to the 2009 NBA Finals from Google News[4] and counted how often each person name occurred in the documents. We also evaluate the NBA topic in the experiment section to determine if the proposed approach is capable of correctly bipolarizing the person names into the teams that played in the finals. In Fig. 2, we rank the person names in descending order according to their frequency. The figure shows that the frequency distribution follows Zipf's law [22]; and most person names rarely appear in the documents and blocks.

We observe that a person name will not appear in a block in the following three scenarios: 1) The polarity of the block is the opposite of the polarity of the person name. For instance, if the person name represents a player in one team and the block contains information about the other team, the block's author would not mention the person in the block to ensure that the block's content is coherent. 2) Even if the polarity of a block is identical to that of the person name, the length of the block may not be sufficient to contain the person name. 3) The block is off-topic, so the person name will not appear in the block. In the last two scenarios, the absence of person names will impact the estimation of the correlation coefficient. To alleviate the problem, we propose two techniques, the weighted correlation coefficient and off-topic block elimination, which we describe in the following sections.

3.3.1 Weighted Correlation Coefficient

The data sparseness problem described in scenario 2 affects many statistical text mining and language models [22]. For person names with the same polarity, data sparseness could lead to underestimation of their correlations because the probability that the names will occur together is reduced. Conversely, for uncorrelated persons or persons with opposite polarities, data sparseness may lead to overestimation of their correlations because they are frequently absent simultaneously from the decomposed blocks. While smoothing approaches, such as Laplace's law (also known as added-one smoothing), have been developed to alleviate data sparseness in language models [22], they are not appropriate for PCA. This is because the correlation coefficient of PCA measures the divergence of person names from their means, so adding one to each person vector entry will not change the divergence. To summarize, data sparseness may influence the correlation coefficient when person names do not co-occur. Thus, for two person names, p_i and p_j, we separate B into co-occurring and non-co-occurring parts and apply the following weighted correlation coefficient:

4. http://news.google.com/.

$$corr_w(\underline{p}_i, \underline{p}_j) = \begin{pmatrix} (1-\alpha)\sum_{b\in co(i,j)}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim}) \\ +\alpha\sum_{b\in B-co(i,j)}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim}) \end{pmatrix} \Big/ \begin{pmatrix} \sqrt{(1-\alpha)\sum_{b\in co(i,j)}(p_{i,b}-p_i^{\sim})^2+\alpha\sum_{b\in B-co(i,j)}(p_{i,b}-p_i^{\sim})^2}* \\ \sqrt{(1-\alpha)\sum_{b\in co(i,j)}(p_{j,b}-p_j^{\sim})^2+\alpha\sum_{b\in B-co(i,j)}(p_{j,b}-p_j^{\sim})^2} \end{pmatrix}, \quad (9)$$

where $corr_w(\underline{p}_i, \underline{p}_j)$ represents the weighted correlation coefficient between person names i and j; and $co(i,j)$ denotes the set of blocks in which person names i and j co-occur. The range of parameter α is within [0,1]. It weights the influence of non-co-occurring blocks when calculating the correlation coefficient. When $\alpha = 0$, the equation only considers the blocks in which p_i and p_j co-occur. Conversely, when $\alpha = 1$, only non-co-occurring blocks are used to calculate the persons' correlation. It is noteworthy that when $\alpha = 0.5$, the equation is equivalent to the standard correlation coefficient, as shown by the following equation:

$$\begin{aligned} &corr_w(\underline{p}_i, \underline{p}_j) \\ &= \frac{\begin{pmatrix} 0.5\sum_{b\in co(i,j)}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim}) \\ +0.5\sum_{b\in B-co(i,j)}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim}) \end{pmatrix}}{\begin{pmatrix} \sqrt{0.5\sum_{b\in co(i,j)}(p_{i,b}-p_i^{\sim})^2+0.5\sum_{b\in B-co(i,j)}(p_{i,b}-p_i^{\sim})^2}* \\ \sqrt{0.5\sum_{b\in co(i,j)}(p_{j,b}-p_j^{\sim})^2+0.5\sum_{b\in B-co(i,j)}(p_{j,b}-p_j^{\sim})^2} \end{pmatrix}} \\ &= \frac{0.5*\sum_{b\in B}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim})}{\sqrt{0.5*\sum_{b\in B}(p_{i,b}-p_i^{\sim})^2}*\sqrt{0.5*\sum_{b\in B}(p_{j,b}-p_j^{\sim})^2}} \\ &= \frac{0.5*\sum_{b\in B}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim})}{\sqrt{0.5}*\sqrt{\sum_{b\in B}(p_{i,b}-p_i^{\sim})^2}*\sqrt{0.5}*\sqrt{\sum_{b\in B}(p_{j,b}-p_j^{\sim})^2}} \\ &= \frac{\sum_{b\in B}(p_{i,b}-p_i^{\sim})*(p_{j,b}-p_j^{\sim})}{\sqrt{\sum_{b\in \mathrm{B}}(p_{i,b}-p_i^{\sim})^2}*\sqrt{\sum_{b\in \mathrm{B}}(p_{j,b}-p_j^{\sim})^2}} \\ &= corr(\underline{p}_i, \underline{p}_j). \end{aligned} \quad (10)$$

We examine the effect of α on person name bipolarization in the experiment section.

3.3.2 Off-Topic Block Elimination

Including off-topic blocks in PCA will lead to overestimation of the correlation between person names. This is because person names are usually absent simultaneously from off-topic blocks that make uncorrelated or even negatively correlated persons positively correlated. To eliminate the effect of off-topic blocks on person name bipolarization, we construct a centroid of all the decomposed blocks by averaging $\underline{b}_i$'s. Then, the blocks whose cosine similarity to the centroid is lower than a predefined threshold β are excluded from the calculation of the correlation coefficient.

3.4 Activeness Timeline of Bipolar Groups

In Section 3.2, we explained how the signs of the entries in the principal eigenvector of PCA form bipolar groups of person names in a set of topic documents. To help users comprehend the storylines of a topic, it would be useful to analyze the activeness trend of each bipolar group. Mei and Zhai [15] observed that the activeness of an event in a time interval is positively correlated with the number of words related to the event. Thus, we measure the activeness of a bipolar group g at time t, denoted as $activeness_{g,t}$, by the following equation:

$$activeness_{g,t} = \frac{1}{|B_t|}\sum_{p_i\in P_g}\sum_{b_k\in B_t} p_{i,k}, \quad (11)$$

where $B_t \subseteq B$ and represents the set of blocks published at time t; $|B_t|$ indicates the number of blocks in B_t; $P_g \subseteq P$ and is the set of person names bipolarized to group g; and $p_{i,k}$ is the frequency of person name p_i in block b_k. Basically, $activeness_{g,t}$ is the number of occurrences of person names bipolarized to g at time t, normalized by the number of topic blocks at t. A large $activeness_{g,t}$ score means that group g is mentioned frequently at time t so the group is active. By contrast, a small $activeness_{g,t}$ score indicates that the group does not attract many reports and is therefore inactive. In the experiment section, we demonstrate that the trend of $activeness_{g,t}$ accurately reflects the development of a bipolar group.

4 Performance Evaluations

4.1 Data Corpus and Evaluation Metric

In the text mining field, evaluations are normally based on official corpora. However, to the best of our knowledge, there are no official corpora for the person name bipolarization task because the research subject is relatively new. We therefore compiled our own data corpus for the performance evaluations. The derived corpus is comprised of 12 topics (i.e., Topics $A \sim L$) with bipolar (or competing) viewpoints. Table 1 shows the statistics of the 12 topics. To demonstrate that the proposed approach can be applied to different languages and diverse topic domains, eight of the topics are in English, and four are in Chinese. English Topics $A \sim D$ related to four sports tournaments. We collected 411 news documents about the 2009 NBA Finals and 87 news documents on the 2010 NBA Finals from Google News. The matchups in the 2009 and 2010 Finals were Lakers versus Orlando Magic and Lakers versus Celtics, respectively. The opening game of the 2010 MLB season was between the Washington Nationals and the Philadelphia Phillies, and President Barack Obama threw the opening pitch. For the evaluations, we collected 33 news documents related to the opening game from Google News. We also collected 166 news documents related to the 2010 World Cup Final. The matchup in the final was the Netherlands versus Spain. Topics $E \sim H$ are in English and are related to four business issues: "Smartphone manufacturers deny Apple reception claims (Topic E)," "Google-Verizon deny tiered-web deal report (Topic F)," "Prudential's AIG deal (Topic G)," and "Google ends four years of censoring the web for China (Topic H)." The four topics comprise, respectively, 123, 74, 154, and 48 news documents, all downloaded from Google News. There are two bipolar groups in Topic E; one is Apple Computer and the other is a group of smartphone manufacturers that Apple CEO Steve Jobs criticized. In

TABLE 1
The Statistics of the Evaluation Corpus

ID	Topic Title	Dates	# of topic documents	# of extracted person names	# of evaluated person names for K cumulative frequency		
					K=50%	K=60%	K=70%
A	The 2009 NBA Finals	2009/6/4-2009/6/16	411	531	13	16	27
B	The 2010 NBA Finals	2010/6/4-2010/6/19	87	99	9	12	16
C	The 2010 MLB opening	2010/4/1-2010/4/5	33	61	6	10	14
D	The 2010 World Cup Final	2010/7/10-2010/7/12	166	163	13	16	21
E	Apple vs. Smartphone	2010/7/18-2010/7/22	123	74	3	6	6
F	Google-Verizon	2010/8/4-2010/8/6	74	53	4	5	6
G	Pru-AIG	2010/6/1-2010/6/3	154	102	3	5	7
H	Google vs. China	2010/1/13-2010/1/15	48	103	8	12	12
I	Taiwan's 2008 presidential election	2008/3/2-2008/3/23	37	86	4	7	13
J	Taiwan's 2009 county commissioner elections	2009/11/2-2009/12/5	50	97	3	4	5
K	Taiwan's 2009 legislative by-elections	2009/12/27-2010/1/11	89	172	7	7	14
L	Taiwan's 2010 legislative by-elections	2010/2/3-2010/2/28	46	60	3	4	5

Topic *F*, the Federal Communications Committee (FCC) strongly opposed the cooperation of Google and Verizon because it would have violated the principle of network neutrality. Thus, the bipolar groups in the topic are FCC and the union of Google and Verizon. In Topic *G*, Prudential wanted to buy AIG's Asian Unit, but a large number of Prudential's shareholders opposed the deal. The bipolar groups in this topic are the Prudential shareholders on the one hand and the executives of Prudential and AIG on the other. Topic *H* relates to Google's decision to quit the China market because of web censorship issues. The bipolar groups in this case are Google on the side and China government officials on the other. Topics $I \sim L$ are in Chinese and are related to four political elections in Taiwan. Topic *I* relates to Taiwan's 2008 presidential election; Topic *J* relates to the county commissioner elections in 2009; and the last two relate to Taiwan's legislative by-elections. The four topics contained 37, 50, 89, and 46 Chinese news reports, respectively. The reports were published by the Liberty Times[5] during the respective election periods. In the election covered by Topic *I*, two major political parties, namely, the Democratic Progressive Party (DPP) and the KouMinTang (KMT), competed for the position of President; and in the elections covered by the third and fourth topics, the parties competed for positions in the Legislative Yuan. It is noteworthy that, in political topics, people generally change their polarities for the sake of expediency. For instance, in Topic *J*, a group of KMT members were expelled from the party, so they campaigned against the KMT for one of the county commissioner positions. Subsequently, some of the expelled people were reconciled with the KMT in the 2010 elections (Topic *L*) and helped the party campaign for legislative positions. As mentioned in Section 1, the person name bipolarization task is difficult because the polarity of a person is dynamic and context dependent. In the following experiments, we show that our unsupervised approach is capable of identifying the dynamics of such polarity.

As paragraph tags are not provided in the evaluated documents, in this study, a block presents a topic document. When evaluating a topic, we first parsed its blocks by using a named entity recognizer to extract all possible person names. For Chinese documents, we used the Chinese Knowledge and Information Processing (CKIP) tool[6]; and for English documents, we used the Stanford Named Entity Recognizer.[7] Given an input text, the Stanford Named Entity Recognizer extracts all possible named entities from the text. The recognizer also tags an extracted entity as a person name, a location name, or an organization name. We used the extracted person names for evaluation. Since there is no perfect named entity recognition approach, we identified false person name entities. Most of the false entities were person name typos. To evaluate the true bipolarization performance, we removed the false entities comprised of the name of a person and the name of an organization (or a location) because they were ambiguous. For instance, the extracted entity Lakers Kobe may refer to the player Kobe Bryant or the team Lakers. We did not remove any typo entities because they refer to specific (unambiguous) persons and retaining them for the evaluations helps us test the robustness of our approach. As mentioned in Section 3, the frequency of extracted person names followed Zipf's law. Since many of the person names rarely appeared in the blocks, their distribution was too sparse for PCA. Hence, for each evaluated topic, we computed the frequency of each

5. http://www.libertytimes.com.tw/index.htm.

6. http://ckipsvr.iis.sinica.edu.tw/.

7. http://nlp.stanford.edu/software/CRF-NER.shtml.

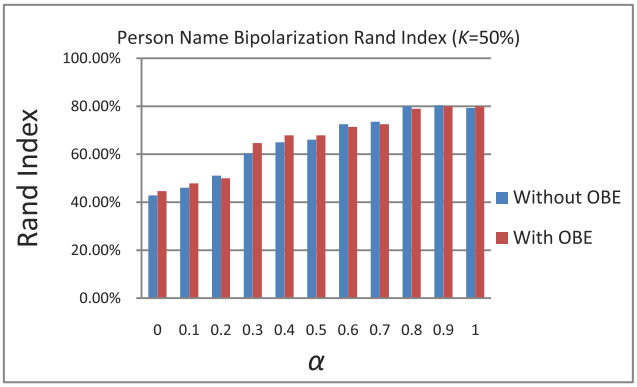

Fig. 3. Person name bipolarization rand indices when $K = 50\%$.

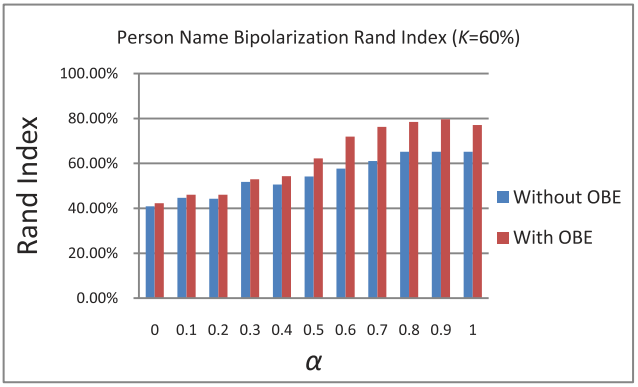

Fig. 4. Person name bipolarization rand indices when $K = 60\%$.

extracted person name in the examined blocks and ranked all the names in descending order according to their frequency. Then, in the evaluation step, we accumulated the frequencies of the most frequent person names and selected the names whose accumulated frequency reached K percent of the total frequency of all the extracted person names. In other words, the evaluated person names accounted for K percent of the person name occurrences in the examined blocks. In the following experiments, we assess the system performance under $K = 50, 60, \text{and } 70\%$. The numbers of extracted and evaluated person names in the topics are shown in Table 1. All the evaluated person names represent important topic persons. We adopted a two-phase annotation process to annotate the polarity of the evaluated person names. In the first phase, two experts were asked to read all the topic documents and then annotate the person polarity independently. In the second phase, discussions were held with the experts to resolve inconsistent annotations and establish a ground truth for the evaluations. As most of the topic persons had a clear stance, the interagreement between the experts was high. Specifically, the agreement rate was 95.2 percent, which was good enough to conduct reliable evaluations.

For each experimental setting (i.e., K and α), we performed principal component analysis on the examined blocks and the evaluated person names. We partitioned the names into two bipolar groups according to their signs in the principal eigenvector and utilized the rand index [22], a conventional evaluation metric frequently used to compare clustering algorithms, to evaluate the bipolarization performance. Specifically, the rand index is based on name pairs. After a set of person names are partitioned into two clusters, the index measures the percentage of clustering decisions that are correct (e.g., placing a name pair with the same polarity in the same cluster). For global performance comparisons, we adopted the microaverage scheme to average the bipolarization rand indices of the evaluated topics.

4.2 Effect of System Components

To examine the effect of the weighted correlation coefficient, the parameter α is set between 0 and 1, and increased in increments of 0.1. For each setting of α, we also examine the rand index with and without off-topic block elimination to determine the influence of noisy blocks on person name bipolarization. When off-topic block elimination is used, the threshold β is set at 0.3. Figs. 3, 4, and 5 show the rand indices of the proposed approach under various experimental settings. OBE is the acronym for off-topic block elimination.

As shown in the figures, eliminating off-topic blocks improves the system performance. In addition, the improvement under $K = 60 \text{ and } 70\%$ is more significant than that under $K = 50\%$. This is because a large K would include infrequent person names in the bipolarization process. Since the correlations between infrequent person names are easily affected by noisy blocks, eliminating off-topic blocks is an effective way to identify the relationships between persons whose names appear infrequently. Hence, the system performance is improved significantly. A large K also implies that the person name bipolarization task is difficult because the infrequent person names included in K would not be sufficient for PCA bipolarization. Consequently, the bipolarization rand index decreases as K increases, as shown in the above figures. It is noteworthy that, when off-topic blocks are eliminated, large α values produce good bipolarization results. As mentioned in Section 3.3, a large α implies that non-co-occurring blocks are important for calculating the correlation between a pair of person names. When off-topic blocks are eliminated, the set of non-co-occurring blocks reveal either opposing relationships between entities or the absence of any relationships. Therefore, the bipolarization performance improves as α increases.

Figs. 6, 7, and 8 show the average bipolarization rand indices when $K = 50,\ 60 \text{ and } 70\%$ for the sports, political, and business topics in our data set. As shown in Fig. 6, the bipolarization rand index for sports topics decreases as α

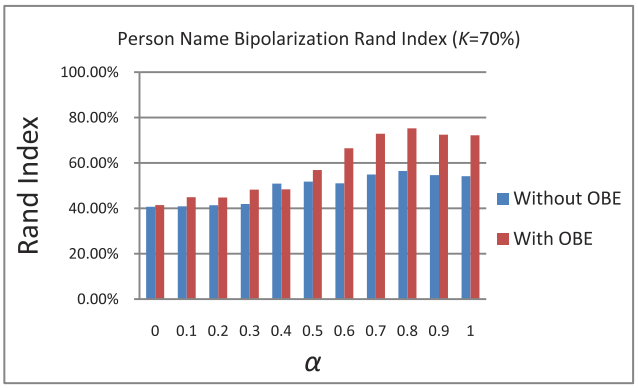

Fig. 5. Person name bipolarization rand indices when $K = 70\%$.

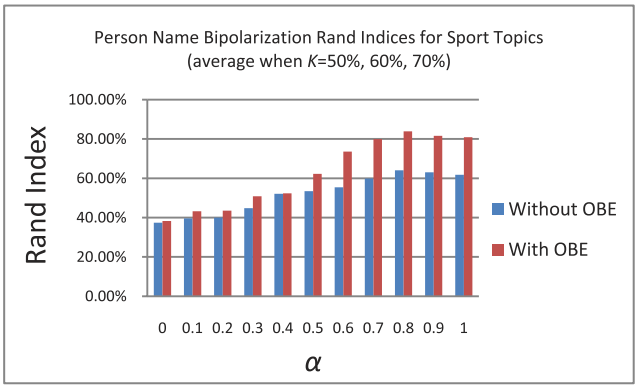

Fig. 6. Person name bipolarization rand indices for sports topics.

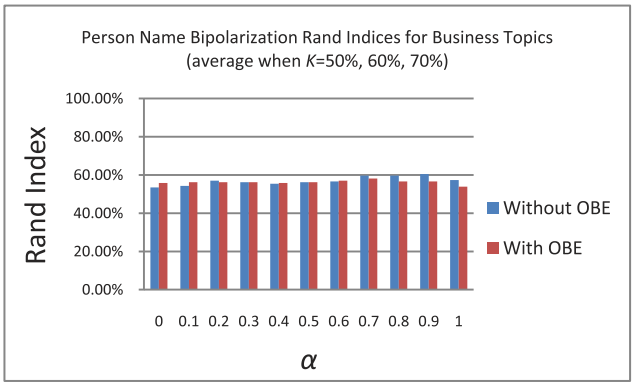

Fig. 8. Person name bipolarization rand indices for business topics.

decreases. We observe that some of the sports documents are recaps of the final game or the opening game, so they tend to mention players in the match together. As a small α value makes co-occurrence blocks important, recap-style documents overestimate the correlation between bipolar person names. Consequently, the bipolarization performance is inferior when α is small. Similarly, the bipolarization rand index for political topics also decreases as α decreases. This is because politicians of different parties often comment on each other during campaigns. Newspapers like to report such events to attract readers, so persons belonging to different parties frequently co-occur in the topic blocks. Therefore, the bipolarization performance under a small α is also inferior. By contrast, the rand index of business topics does not decline as α decreases. In fact, we observed that the improvement in the bipolarization performance of the business topics derived by the weighted correlation coefficient was insignificant. As shown in Table 1, the number of evaluated persons in the business topics under $K = 70\%$ is almost the same as that under $K = 50\%$. Hence, many of the evaluated person names are frequent enough to prevent the data sparseness problem. Although the weighted correlation coefficient does not improve the bipolarization performance of the business topics significantly, the proposed PCA-based approach can still identify the bipolar groups of important persons accurately.

As the number of political topic persons is not large, an instance of misbipolarization (e.g., placing two people with the same polarity in different clusters) will have a significant impact on the rand index. Hence, the rand indices for political topics are only about 0.6. However, in Section 4.4, we show that our method outperforms well-known clustering algorithms and achieves the best bipolarization performance for political topics. Thus, it is effective for such topics. Compared with sports topics, the rand indices for business topics are low. To analyze the performance differences, we calculate the correlation coefficient of the topic persons who have the same or different polarities in terms of their occurrences in topic blocks (i.e., (1)). We observe that most of the topics have a high and positive correlation coefficient between the persons with the same polarity. The phenomenon corresponds well with the observation in [13] that polar text units with the same polarity tend to appear together to make the content coherent. Table 2 shows the average correlation coefficient between persons with different polarities for the business, political, and sports topics. Surprisingly, the correlation coefficients for the business topics are not as negative as we expected. Under $K = 50\%$, the business topics even have a positive correlation coefficient. While business topic documents generally report events about a single polarity (which yields a high correlation coefficient between persons with the same polarity, i.e., 0.264), journalists often leave comments with different polarities to the end of the documents to produce a balanced report. For instance, for Topic H, many documents about the perspectives of China government officials also concluded with the opinions of David Drummond, who is the chief legal official of Google. For Topic F, the documents about the stance of the FCC are generally mixed with the comments of David Fish, who is the spokesperson for Verizon. The writing style blurs the correlation coefficient between persons with different polarities, and affects the performance of our approach for the business topics. The evaluated political and sports topics also contain documents that report the opinions of people with different polarities. For instance, the recap style documents of the NBA Finals report on players in the matches together. However, all the political and sports topics have a definite winning polarity, but the business topics do not have a clear winner. For instance, the Lakers won the championship title in Topics A and B. When a polarity won, a large number of the topic documents contained reports about the wining polarity. Because the documents mentioned the members of the winning polarity extensively, the negative correlation coefficient between

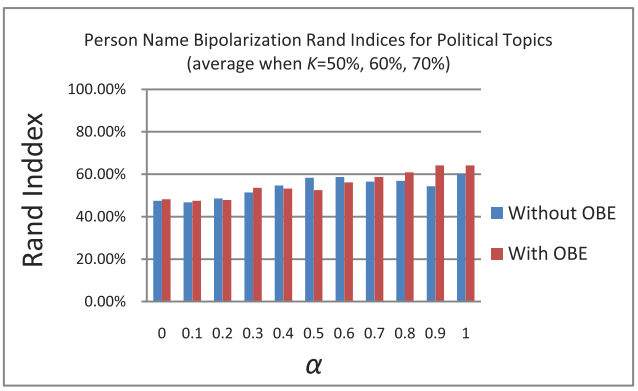

Fig. 7. Person name bipolarization rand indices for political topics.

TABLE 2
The Average Correlation Coefficients between Persons with Different Polarities

	$K = 50\%$	$K = 60\%$	$K = 70\%$
Business topics	0.082	-0.031	-0.006
Political topics	-0.099	-0.124	-0.092
Sports topics	-0.085	-0.083	-0.057

persons with different polarities increased. Hence, our approach achieved a superior performance on the political and sports topics.

The evaluations demonstrate that the proposed PCA-based approach can identify bipolar person names in topic documents effectively. In addition, eliminating off-topic blocks produces superior bipolarization results. The weighted correlation coefficient also improves the bipolarization performance. However, as the writing styles of topic documents in different domains vary, there is no universal α value of the weighted correlation coefficient for various topic domains.

4.3 Examples of Person Name Bipolarization

In this section, we consider two sports topics, namely, the 2009 NBA Finals and the 2010 MLB opening game, to demonstrate the outcomes of person name bipolarization. We select these topics because they are global news stories, so readers can understand them without background knowledge of a specific culture. In addition, we present the bipolarization results of two Chinese political topics to show that the proposed approach can identify polarity dynamics.

Table 3 shows the bipolarization results for the evaluated person names in the 2009 NBA finals data set. The left-hand columns of the table list the person names labeled as Magic and their entry values in the principal eigenvector; and the right-hand columns list the person names labeled as Lakers and their entry values. It is noteworthy that the evaluated entities contain person names that are not associated with the players in the NBA finals. For instance, the frequency of Magic Johnson, an ex-Lakers player, is high because he constantly spoke in support of the Lakers during the lead-up to the final games. In addition, many documents misspell players' names (e.g., Pau Gasol as Paul Gasol and Mickael Pietrus as Michael Pietrus). Even though the names refer to the same player, the named entity recognizer parses them as distinct entities. In Table 3, a person name annotated with the symbol * indicates that the entity is bipolarized incorrectly. For instance, Magic Johnson is not a member of Magic. The symbol indicates that the person name is neutral (or irrelevant) to the teams in the finals; and the symbol + indicates that the person name is misspelled, but it refers to a member of the bipolarized team. When evaluating the bipolarization performance, we treat the person names that refer to the players or coaches of a team as true positives if they are placed in the same cluster. Person names that are closely related to Lakers or Magic players, such as a player's relatives or misspellings, are also deemed true positives if they are bipolarized into the correct teams. The results in the table show that the proposed approach bipolarizes the important persons in the final game correctly without using any external information source. The rand index is 81.77 percent; however, if we ignore the neutral entities, which are always wrong irrespective of the bipolarization approach employed, the rand index is 93.73 percent. In this case, we only misbipolarized Magic Johnson as Magic. The mistake also reflects a problem with named entity resolution when the person names that appear in a document are ambiguous. In the topic documents, the word "Magic" sometimes refers to Magic Johnson and sometimes to Orlando Magic. Here, we do not consider a sophisticated named entity resolution scheme; instead, we simply assign the frequency of a person name to all its specific entities (e.g., Magic to Magic Johnson, and Kobe to Kobe Bryant) so that specific person names are frequent enough for PCA. As a result, Magic Johnson tends to co-occur with the members of Magic and is incorrectly bipolarized to the Magic team. Another interesting phenomenon is that LeBron James (a player with Miami Heat) is incorrectly bipolarized to Lakers. This is because Kobe Bryant (a player with Lakers) and LeBron James were rivals for the most valuable player (MVP) award in the 2009 NBA season. The documents that mentioned Kobe Bryant during the finals often compared him with LeBron James to attract the attention of readers. As the names often co-occurred in the documents, LeBron James was wrongly classified as a member of Lakers.

TABLE 3
The Bipolarization Results for the 2009 NBA Finals

Magic		Lakers	
Anthony Johnson	0.0308	Andrew Bynum	-0.0684
Courtney Lee	0.0663	Derek Fisher	-0.0244
Dwight Howard	0.1579	Jordan Farmar	-0.1084
Hedo Turkoglu	0.1752	Kobe Bryant	-0.2432
Jameer Nelson	0.221	Lamar Odom	-0.1241
Jeff Van Gundy	0.3268	LeBron James*^	-0.0408
Magic Johnson*	0.4444	Mark Jackson*^	-0.2714
Michael Pietrus+	0.003	Michael Jordan*^	-0.1396
Mickael Pietrus	0.0231	Pau Gasol	-0.148
Mikael Pietrus+	0.0055	Paul Gasol+	-0.1236
Rafer Alston	0.2351	Phil Jackson	-0.2989
Rashad Lewis+	0.1498	Shaquille O'Neal*^	-0.0972
Rashard Lewis	0.213	Trevor Ariza	-0.0899
Stan Van Gundy	0.3632		

$K = 70\%$; $a = 1$; with OBE

TABLE 4
The Bipolarization Results for the 2010 MLB Opening Game

Washington Nationals		Philadelphia Phillies	
Adam Dunn	-0.3591	Brad Lidge	0.0229
Barack Obama	-0.2066	Charlie Manuel	0.2177
Ivan Rodriguez	-0.4083	Chase Utley	0.3319
John Lannan	-0.0653	Jimmy Rollins	0.2915
Nyjer Morgan	-0.3131	Placido Polanco	0.3653
Ryan Zimmerman	-0.3221	Roy Halladay	0.0028
William Howard Taft*^	-0.0074	Ryan Howard	0.281

$K = 70\%$; $a = 0.8$; with OBE

Table 4 shows the bipolarization results for the frequent person names in the MLB data set. It is interesting that President Barack Obama is one of the evaluated persons. The reason is that he was invited to throw the opening

TABLE 5
The Bipolarization Results for Taiwan's 2009 County Commissioner Elections

KMT		Persons expelled from KMT	
Ying-jeou Ma (馬英九)	0.493	Pi-chin Chang (張碧琴)	-0.5215
Ching-chun Chiu (邱鏡淳)	0.3736	Yung-chin Cheng (鄭永金)	-0.5877
Shao-chi Peng (彭紹瑾)*^	0.0012		

$K = 70\%$; $a = 0.8$; with OBE

TABLE 6
The Bipolarization Results for Taiwan's 2010 Legislative Elections

KMT		DPP	
Ying-jeou Ma (馬英九)	0.3777	Shao-chi Peng(彭紹瑾)	-0.4969
Ching-chun Chiu (邱鏡淳)	0.5565	Ing-wen Tsai (蔡英文)	-0.5167
Yung-tang Cheng (鄭永堂)	0.1837		

$K = 70\%$; $a = 0.8$; with OBE

pitch of the 2010 MLB season. Many topic documents reported the event, so Barack Obama is classified as a frequent person name. Although the president is a White Sox fan, he supported the Nationals in this case because the team is based in Washington, DC. We therefore treated him as a member of the Nationals. In addition to President Obama, the evaluated person names included important players of the Nationals and the Phillies. The bipolarization rand index is 93.41 percent. President Obama is successfully bipolarized as a member of the Nationals because Ryan Zimmerman, a Nationals player, was selected to catch the first pitch. Thus, their names co-occurred frequently in the topic documents and were bipolarized together. The successful bipolarization of President Obama also demonstrates that the proposed bipolarization approach is context-oriented. That is, the bipolarization result depends on the given topic documents and the corresponding context. In this experiment, we only misbipolarized William Howard Taft as a Nationals player. The convention of inviting the president to throw the opening pitch was initiated by William Howard Taft. Since many of the topic documents reported the story, William Howard Taft frequently co-occurred with Barack Obama and Ryan Zimmerman in the documents; hence, they were bipolarized together. If we ignore this neutral person, we find that the important persons of the opening game are bipolarized perfectly.

Next, we consider the bipolarization results for Taiwan's 2009 county commissioner elections and 2010 legislative elections. In the county commissioner elections, the bipolar groups are the KMT and a group of persons expelled from the KMT. As shown in Table 5, the proposed approach identifies the bipolar groups correctly. Ying-jeou Ma and Ching-chun Chiu were the KMT's chair person and election candidate, respectively, while Pi-chin Chang and Yung-chin Cheng represented the expelled persons running for election. It is noteworthy that the KMT reconciled with Yung-chin Cheng prior to the legislative elections and nominated his brother, Yung-tang Cheng, to run for election. As shown in Table 6, we identified this polarity dynamic successfully and bipolarized Yung-tang Cheng and other KMT members together without using any external knowledge source. Additionally, important persons in the KMT and DPP, i.e., the candidates of the two parties, the party chair persons, and important party staff members, were bipolarized correctly, as shown in Table 6.

The bipolarization examples demonstrate that the proposed approach can identify bipolar groups of persons in topic documents accurately. Moreover, as the approach analyzes word usage patterns of important person names in a set of topic documents, it is context-oriented; hence, it does not require external knowledge sources.

4.4 Comparison with Other Methods

As mentioned in Section 2, the person name bipolarization task is a clustering problem that groups items into concept-coherent clusters. Here, we compare the proposed approach with three well-known text clustering algorithms, namely, the PLSI algorithm [24], the K-means algorithm [22] and the HAC algorithm [21]. Under K-means and HAC, a person name is represented by a high-dimensional term frequency vector (i.e., a row of the person block association matrix), where a vector entry indicates the frequency of the person name in a block. We use the traditional cosine similarity metric to cluster similar person names. To ensure that the comparisons are fair, each clustering algorithm partitions the evaluated person names into two clusters. In [24], the author treats each latent variable z of PLSI as a concept and groups the terms (or documents) of a text corpus into clusters according to $P(z \mid w)$ (or $P(z \mid d)$). In our experiment, a term w is a person name and there are two latent variables. As the clustering performance of PLSI and K-means depends on cluster initialization, we randomly initialize both algorithms 20 times and select the best, worst, and average results for comparison. We also iterate the algorithms until the clustering results become stable. Since the results are local optima, they are suitable for comparison. For HAC, we consider four well-known intercluster similarity strategies, namely, single-link, complete-link, average-link, and centroid-link strategies [21]. In addition, a naive method, which considers all the person names as a single polarity, serves as a baseline to evaluate the efficiency of the clustering-based bipolarization approaches. As mentioned in Section 4.2, different topic domains have different writing styles. Hence, there is no universal α value of the weighted correlation coefficient for various topic domains. Based on the experiment results in Section 4.2, we adopt a fixed α for each topic domain; and set it at 0.8, 0.7, and 0.9 for the sports, business, and political topics respectively. In addition, we employed the off-topic block elimination technique during the evaluations.

Table 7 compares the bipolarization results. For business topics, some of the top K-means and PLSI results are slightly better than our fixed setting results. In all other cases, our approach yields the best rand indices and outperforms the compared algorithms by a significant margin. The results demonstrate that the proposed approach can identify bipolar persons in different domains efficiently. As shown in the table, some of the best K-means results are superior; however the algorithm's average rand indices are low. Similarly, although the best results of PLSI are superior to those of the

TABLE 7
The Bipolarization Results of the Compared Methods

	Method	Sports Topics	Business Topics	Political Topics	All Topics
K=50%	Our approach	85.89%	67.50%	60.61%	80.36%
	PLSI (best)	68.60%	70.00%	51.52%	66.79%**
	PLSI (average)	51.57%	58.38%	45.15%	51.79%**
	PLSI (worst)	43.48%	50.00%	39.39%	43.93%**
	K-means (best)	78.26%	67.50%	57.58%	74.29%*
	K-means (average)	62.27%	51.25%	51.82%	59.46%**
	K-means (worst)	50.72%	47.50%	42.42%	49.29%**
	HAC (single)	50.72%	57.50%	42.42%	50.71%**
	HAC(complete)	72.46%	57.50%	48.48%	67.50%**
	HAC (average)	61.35%	75.00%	48.48%	61.79%**
	HAC (centroid)	62.32%	57.50%	48.48%	60.00%**
	Baseline	39.61%	25.00%	39.39%	37.50%**
K=60%	Our approach	88.89%	52.48%	79.63%	80.63%
	PLSI (best)	69.80%	60.40%	61.11%	67.00%**
	PLSI (average)	50.83%	50.30%	49.17%	50.54%**
	PLSI (worst)	44.73%	45.54%	42.59%	44.66%**
	K-means (best)	72.36%	63.37%	55.56%	68.77%**
	K-means (average)	57.83%	43.86%	50.46%	54.26%**
	K-means (worst)	48.43%	30.69%	42.59%	44.27%**
	HAC (single)	43.87%	33.66%	50.00%	42.49%**
	HAC(complete)	70.37%	43.56%	50.00%	62.85%**
	HAC (average)	63.25%	54.46%	53.70%	60.47%**
	HAC (centroid)	60.40%	43.56%	53.70%	56.32%**
	Baseline	39.32%	19.80%	38.89%	35.38%**
K=70%	Our approach	83.68%	59.83%	64.02%	77.64%
	PLSI (best)	67.49%	64.96%	56.08%	65.21%**
	PLSI (average)	50.69%	51.28%	48.52%	50.35%**
	PLSI (worst)	46.89%	40.17%	45.50%	45.92%**
	K-means (best)	71.50%	58.97%	57.14%	67.63%**
	K-means (average)	55.69%	45.64%	52.78%	54.09%**
	K-means (worst)	44.30%	33.33%	43.39%	42.95%**
	HAC (single)	44.95%	34.19%	44.97%	43.78%**
	HAC(complete)	52.85%	41.03%	52.38%	51.48%**
	HAC (average)	67.23%	50.43%	48.68%	62.15%**
	HAC (centroid)	49.74%	41.03%	50.26%	48.89%**
	Baseline	37.95%	24.79%	41.27%	37.11%**

The results marked with * and ** show, respectively, the proposed approach's improvements over the compared methods with 95% and 99% confidence levels based on the Z-statistic for two proportions [26].

HAC algorithm, its average results are inferior. The inferior average results of PLSI and K-means indicate that the algorithms are sensitive to their cluster initializations. However, initializing the algorithms appropriately for various topics is difficult because the bipolar relationships between important persons are context-dependent. The situation is even worse under PLSI because its initialization process must determine the values of $P(z)$, $P(d|z)$, and $P(w|z)$ (here, d represents a topic block, z is a latent variable; and w represents a person name) and there are infinite ways to initialize the distributions [25]. Due to the lack of an effective initialization process, the average results of PLSI and K-means are inferior.

We observe that K-means produces inferior bipolarization results when popular persons are selected as the initial cluster centroids. The phenomenon highlights a problem with using the cosine similarity score for person name bipolarization. Here, a person name is considered popular if it appears in several topic blocks, so the corresponding term frequency vector contains a large number of nonzero entries. As the cosine similarity calculates the normalized inner product of two frequency vectors, it tends to produce a high similarity score when the calculated vectors contain several nonzero entries. While popular person names tend to have high cosine similarity scores, they may be negatively correlated in the topic blocks. Table 8 shows some examples from our evaluation corpus. In the first example, Kobe Bryant and Dwight Howard are franchise players of Lakers and Magic, respectively, and they are popular person names in the 2009 NBA finals data set. We observe that their cosine similarity score (0.25) is much higher than the average score (0.17) in the data set because their term frequency vectors contain a lot of nonzero entries. Similarly, in the second and third examples, the cosine similarity scores of Ying-jeou Ma, Chan-ting Hsien, and Shao-chi Peng are high because they are popular persons in Taiwan's election data sets. Selecting one of them as the initial cluster centroid for the K-means algorithm would group cosine-similar but bipolar persons incorrectly in the same cluster, and thus impact the bipolarization performance. The performances of the single-link strategy of HAC also reflect the shortcoming of the cosine similarity metric. As the strategy calculates the similarity of two clusters by examining the most similar item pairs in the clusters, a high similarity score between popular but bipolar person names would merge bipolar person groups into a single cluster. Therefore, the performance of the single-link strategy is inferior. The other HAC clustering strategies consider all the pairs of similarities between clusters to compensate for the shortcoming of the cosine similarity, and thus produce better bipolarization results. The PLSI algorithm also groups popular person names together. This is because the object function of PLSI (i.e., $\sum_d \sum_w n(d,w)^* \log(\sum_z P(z)P(w|z)P(d|z))$, where $n(d, w)$ denotes the frequency of w in d) tends to compute a high $P(z|w)$ to person names that co-occur frequently in topic blocks. Consequently, recap-style documents of sports games and balanced reports of business and political topics group popular but bipolar person names together and thus affect PLSI's performance. The proposed approach determines the relationships between person names by using the correlation coefficient. Unlike the cosine similarity, the correlation coefficient indicates how the occurrences of two person names vary jointly in a set of topic blocks. As shown in Table 8, the metric correctly identifies bipolar relationships between popular persons and thus outperforms the compared methods.

TABLE 8
Examples of Negatively Correlated Person Names with High Cosine Similarity Scores

Bipolar person names	Cosine similarity	Correlation coefficient
<Kobe Bryant, Dwight Howard>	0.25	-0.13
<Chan-ting Hsieh, Ying-jeou Ma>	0.34	-0.08
<Shao-chi Peng, Ying-jeou Ma>	0.28	-0.34

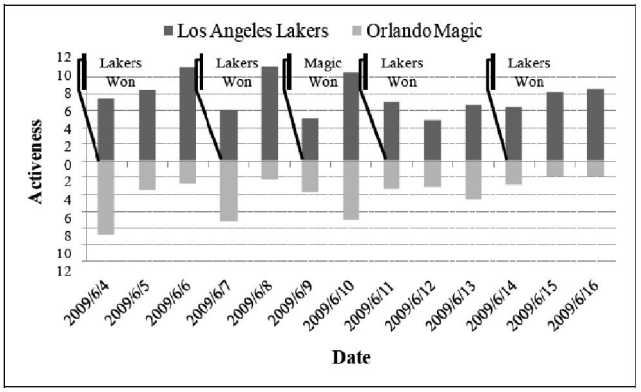

Fig. 9. The activeness timeline of the 2009 NBA finals ($\alpha = 0.8, K = 70\%$, with OBE).

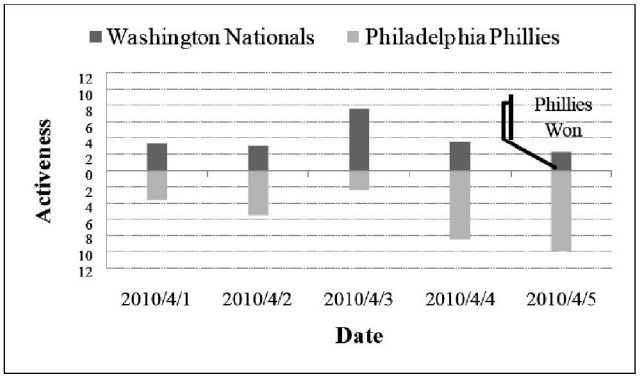

Fig. 10. The activeness timeline of the 2010 MLB opening game ($\alpha = 0.8$, $K = 70\%$, with OBE).

4.5 Activeness Timeline Evaluations

Timeline evaluation is difficult because there are no official benchmarks and metrics for the task. Thus, case studies (e.g., [1], [15], and [17]) are often used to demonstrate the benefit of timeline mining. In this section, we take the 2009 NBA Finals and the 2010 MLB opening game as case studies for activeness timeline evaluations. Fig. 9 shows the activeness timeline of the NBA data set. The upper and lower bars indicate the activeness of the Lakers and Magic, respectively, during the finals. The figure also shows the date and result of each game.

We observe that the activeness of the identified bipolar groups corresponds closely with the development of the finals. On 6/4/2009, the first day of the 2009 NBA finals, a large number of news documents from various news agencies reported and analyzed the game. As the documents frequently mentioned the players in the teams, both bipolar groups had a high activeness value. Interestingly, when a team won a game, its activeness score was high the next day. For instance, the activeness values of Lakers on 6/5/2009 and 6/8/2009 were high after Lakers won game1 and game2; and Magic had a high activeness score on 6/10/2009 after it won game3. As the games were often played at night, many of the documents related to the games were published the next day. The documents tend to highlight the winning team, especially the performance of the team's players, so the identified bipolar players and their activeness values successfully describe the development of the finals. The activeness trend shows that Lakers gradually dominated Magic, which corresponds with the outcome of the finals because Lakers won the championship title.

Fig. 10 shows the activeness timeline of the 2010 MLB opening game, which was held on 4/5/2010. Once again, the activeness of the identified bipolar groups, i.e., the Phillies and the Nationals, describes the development of the opening game. As shown in the figure, the activeness score of the Phillies is higher than that of the Nationals. This is because the Phillies is a famous team with a long history; so it is often the subject of news reports. On 4/3/2010, the Nationals had an activeness burst, which corresponded with the announcement that Nationals player Ryan Zimmerman had been selected to catch the opening pitch thrown by President Obama. Numerous news documents published that day reported the announcement. The Phillies won the opening game and many of the news documents on 4/5/2010 highlighted the performance of the Phillies' players; thus, the team's activeness value was high on that day.

The case studies demonstrate that the proposed timeline system can describe developments of bipolar topics successfully. As a result, the activeness trend helps readers understand the intensity of each bipolar group.

5 Conclusion

Topics involving bipolar viewpoints are usually reported by a large number of documents. Thus, identifying bipolar person names in the topic documents should help readers comprehend the topics in a more balanced manner. In this paper, we propose an unsupervised approach that identifies the polarity of person names in topic documents. We show that the signs of the entries in the principal eigenvector of PCA can partition person names into bipolar groups spontaneously. In addition, we introduce two techniques, namely the weighted correlation coefficient and off-topic block elimination, to address the data sparseness problem. Our experiment results demonstrate that the proposed approach can identify bipolar person names in topic documents correctly without using any external knowledge source. Moreover, the approach is context-oriented, and it can be applied to different languages and diverse topic domains. The results of the present study suggest areas for future research. For example, we observed that some of the evaluated person names possessed neutral orientations. Developing an effective method to identify neutral persons in topics would be worthwhile. Finally, since a topic may have more than two polarities, modeling the multipolarity identification problem would also be an interesting research subject.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This research was supported in part by NSC 97-2221-E-002-225-MY2 and NSC 99-2221-E-002-182 from the National Science Council, Republic of China.

References

[1] J.M. Kleinberg, "Bursty and Hierarchical Structure in Streams," *Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 91-101, 2002.

[2] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event Threading within News Topics," *Proc. 13th ACM Int'l Conf. Information and Knowledge Management,* pp. 446-453, 2004.
[3] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval,* vol. 2, nos. 1/2, pp. 1-135, 2008.
[4] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller, "Introduction to WordNet: An On-line Lexical Database," *Int'l J. Lexicography,* vol. 3, no. 4, pp. 235-244, 1990.
[5] L.I. Smith, *A Tutorial on Principal Components Analysis.* Cornell Univ., 2002.
[6] J. Artiles, J. Gonzalo, and S. Sekine, "The Semeval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task," *Proc. Int'l Workshop Semantic Evaluations,* pp. 64-69, 2007.
[7] X. Wan, J. Gao, M. Li, and B. Ding, "Person Resolution in Person Search Results: WebHawk," *Proc. 14th ACM Int'l Conf. Information and Knowledge Management,* pp. 163-170, 2005.
[8] D.V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra, "Towards Breaking the Quality Curse: A Web-Querying Approach to Web People Search," *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 27-34, 2008.
[9] Y. Song, J. Huang, I.G. Councill, J. Li, and C.L. Giles, "Efficient Topic-Based Unsupervised Name Disambiguation," *Proc. ACM/IEEE CS Seventh Joint Conf. Digital Libraries,* pp. 342-351, 2007.
[10] V. Hatzivassiloglou and K.R. McKeown, "Predicting the Semantic Orientation of Adjectives," *Proc. Eighth Conf. European Chapter of the Assoc. for Computational Linguistics,* pp. 174-181, 1997.
[11] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Trans. Information Systems,* vol. 21, pp. 315-346, 2003.
[12] A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," *Proc. Fifth Conf. Language Resources and Evaluation,* pp. 417-422, 2006.
[13] H. Kanayama and T. Nasukawa, "Fully Automatic Lexicon Expansion for Domain-Oriented Sentiment Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 355-363, 2006.
[14] M. Ganapathibhotla and B. Liu, "Mining Opinions in Comparative Sentences," *Proc. 22nd Int'l Conf. Computational Linguistics,* pp. 241-248, 2008.
[15] Q. Mei and C.X. Zhai, "Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining," *Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining,* pp. 198-207, 2005.
[16] A. Feng and J. Allan, "Finding and Linking Incidents in News," *Proc. 16th ACM Conf. Information and Knowledge Management,* pp. 821-830, 2007.
[17] C.C. Chen and M.C. Chen, "TSCAN: A Novel Method for Topic Summarization and Content Anatomy," *Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 579-586, 2008.
[18] L.E. Spence, A.J. Insel, and S.H. Friedberg, *Elementary Linear Algebra, A Matrix Approach.* Prentice Hall, 2000.
[19] W.L. Winston, *Operations Research.* Thomson, 2004.
[20] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," *Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 19-25, 2001.
[21] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, 1989.
[22] C.D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval.* Cambridge Univ. Press, 2008.
[23] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *J. ACM,* vol. 46, no. 5, pp. 604-632, 1999.
[24] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. 22nd Ann. Int'l SIGIR Conf. Research and Development in Information Retrieval,* pp. 50-57, 1999.
[25] A. Farahat and F. Chen, "Improving Probabilistic Latent Semantic Analysis with Principal Component Analysis," *Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics (EACL),* pp. 105-112, 2006.
[26] G. Keller and B. Warrack, *Statistics for Management and Economics.* Duxbury, 1999.

Chien Chin Chen received the PhD degree in electrical engineering from National Taiwan University, Taiwan, in 2007. He is currently an assistant professor in the Department of Information Management at National Taiwan University. His papers have appeared in *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *ACM Transactions on Information Systems (TOIS)*, SIGIR, SIGKDD, COLING, etc. His current research interests include text mining, information retrieval, and knowledge discovery.

Zhong-Yong Chen received the MS degree in information management from the National Kaohsiung University of Applied Sciences, Taiwan, in 2009. He is currently working toward the PhD degree in information management at the National Taiwan University, Taiwan. His research interests include information retrieval and text mining.

Chen-Yuan Wu received the MS degree in information management from National Taiwan University, Taiwan, in 2010. His current research interests include topic person name mining and information retrieval.