# Bipolar Person Name Identification of Topic Documents Using Principal Component Analysis

**Chein Chin Chen**
Department of Information Management
National Taiwan University
paton@im.ntu.edu.tw

**Chen-Yuan Wu**
Department of Information Management
National Taiwan University
r97725035@ntu.edu.tw

## Abstract

In this paper, we propose an unsupervised approach for identifying bipolar person names in a set of topic documents. 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. Empirical evaluations demonstrate the efficacy of the proposed approach in identifying bipolar person names of topics.

## 1 Introduction

With the advent of Web2.0, many online collaborative tools, e.g., weblogs and discussion forums are being developed to allow Internet users to express their perspectives 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 a topic comprehensively. Since the essence of Web2.0 is knowledge sharing, collaborative tools are generally designed with few 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. Research subjects, such as topic threading and timeline mining (Nallapati et al., 2004; Feng and Allan, 2007; Chen and Chen, 2008), are thus being studied to help Internet users comprehend numerous topic documents efficiently.

A topic consists of a sequence of related events associated with a specific time, place, and person(s) (Nallapati et al., 2004). Topics that involve bipolar (or competitive) viewpoints are often attention-getting and attract a large number of topic documents. For such topics, identifying the polarity of the named entities, especially person names, in the topic documents would help readers learn the topic efficiently. For instance, for the 2008 American presidential election, Internet users can find numerous Web documents about the Democrat and Republican parties. Identifying important people in the competing parties would help readers form a balanced view of the campaign.

Existing works on topic content mining focus on extracting important themes in topics. In this paper, we propose an unsupervised approach that identifies bipolar person names in a set of topic documents automatically. We employ principal component analysis (PCA) (Smith, 2002) 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. In addition, we present two techniques, called off-topic block elimination and weighted correlation coefficient, to reduce the effect of data sparseness on person name bipolarization. The results of experiments based on two topic document sets written in English and Chinese respectively demonstrate that the proposed PCA-based approach is effective in identifying bipolar person names. Furthermore, the approach is language independent.



*Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 170–178,
Beijing, August 2010

## 2 Related Work

Our research is closely related to opinion mining, which involves identifying the polarity (or sentiment) of a word in order to extract positive or negative sentences from review documents (Ganapathibhotla and Liu, 2008). Hatzivassiloglou and McKeown (1997) validated 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%) 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 (2003) manually selected seven positive and seven negative words as a polarity lexicon and proposed using pointwise mutual information (PMI) to calculate the polarity of a word. 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 (2006) 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 datasets for the sentiments and trained eight sentiment classifiers to assign sentiment scores to a synset. Kanayama and Nasukawa (2006) posited that polar clauses with the same polarity tend to appear successively in contexts. The authors derived the coherent precision and coherent density of a word in a training corpus to predict the word's polarity. Ganapathibhotla and Liu (2008) 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 one-side association (OSA), which is a variant of PMI. OSA assigns a positive (negative) orientation to the 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 approaches in three respects. First, most works identify the polarity of adjectives and adverbs because the syntactic constructs generally express sentimental semantics. In contrast, our method identifies the polarity of person names. Second, to the best of our knowledge, all existing polarity identification methods require external information sources (e.g., WordNet, manually selected polarity words, or training corpora). However, our method identifies bipolar person names by simply analyzing person name usage patterns in topic documents without using external information. Finally, our method does not require any language constructs, such as conjunctions; hence, it can be applied to different languages.

## 3 Method

### 3.1 Data Preprocessing

Given a set of topic documents, we first decompose the documents into a set of non-overlapping *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. Let $U = \{u_1, u_2, \ldots, u_m\}$ be a set of *textual units* in $B$. In this study, a unit refers to a person name. Then, the document set can be represented as an *mxn unit-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 $u_j$ in $b_i$. In addition, a row in $A$, denoted as $\underline{u_i}$, represents a textual unit $i$; and it is an $n$-dimensional vector whose $j$'th entry, denoted as $u_{i,j}$, is the frequency of $u_i$ 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 dataset (Smith, 2002). In our research, it identifies the most important unit pattern in the topic blocks by first constructing an *mxm unit relation matrix* $R$, in which the $(i,j)$-entry (denoted as $r_{i,j}$) denotes the correlation coefficient of $u_i$ and $u_j$. The correlation is computed as follows:

$$r_{i,j} = corr(u_i, u_j) = \frac{\sum_{k=1}^{n}(u_{i,k} - \tilde{u_i})*(u_{j,k} - \tilde{u_j})}{\sqrt{\sum_{k=1}^{n}(u_{i,k} - \tilde{u_i})^2} * \sqrt{\sum_{k=1}^{n}(u_{j,k} - \tilde{u_j})^2}},$$

where $\tilde{u_i}=1/n\sum^{n}_{k=1}u_{i,k}$ and $\tilde{u_j}=1/n\sum^{n}_{k=1}u_{j,k}$ are the average frequencies of units $i$ and $j$ respectively.

The range of $r_{i,j}$ is within [-1,1] and the value represents the degree of correlation between $u_i$ and $u_j$ under the decomposed blocks. If $r_{i,j} = 0$, we say that $u_i$ and $u_j$ are uncorrelated; that is, occurrences of unit $u_i$ and unit $u_j$ in the blocks are independent of each other. If $r_{i,j} > 0$, we say that units $u_i$ and $u_j$ are positively correlated. That is, $u_i$ and $u_j$ tend to co-occur in the blocks; otherwise, both tend to be jointly-absent. If $r_{i,j} < 0$, we say that $u_i$ and $u_j$ are negatively correlated; that is, if one unit appears, the other tends not to appear in the same block simultaneously. Note that if $r_{i,j} \neq 0$, $|r_{i,j}|$ scales the strength of a positive or negative correlation. Moreover, since the correlation coefficient is commutative, $r_{i,j}$ will be identical to $r_{j,i}$ such that matrix $R$ will be symmetric.

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

max $\underline{v}^T R \underline{v}$,
s.t. $\underline{v}^T \underline{v} = 1$.

Without specifying any constraint on $\underline{v}$, the objective function becomes arbitrarily large with large entry values of $\underline{v}$. Constraint $\underline{v}^T \underline{v} = 1$ limits the search space within the set of length-normalized vectors. Chen and Chen (2008) show that the desired $\underline{v}$ for the above constrained optimization problem is the eigenvector of $R$ with the largest eigenvalue. Furthermore, as $R$ is a symmetric matrix, such an eigenvector always exists (Spence et al., 2000) and the optimization problem is solvable.

PCA is not the only method that identifies important textual patterns in terms of eigenvectors. For instance, Gong and Liu (2001), Chen and Chen (2008) utilize the eigenvectors of symmetric matrices to extract salient concepts and salient themes from documents respectively[1]. The difference between PCA and other eigenvector-based approaches lies in the way the unit relation matrix is constructed. PCA calculates $r_{i,j}$ by using the correlation coefficient, whereas the other approaches employ the inner product or cosine formula[2] (Manning et al., 2008) to derive the relationship between textual units. Specifically, the correlation coefficient is identical to the cosine formula if we normalize each unit with its mean:

$$corr(u_i, u_j) = \frac{\sum_{k=1}^{n} (u_{i,k} - \tilde{u_i}) * (u_{j,k} - \tilde{u_j})}{\sqrt{\sum_{k=1}^{n} (u_{i,k} - \tilde{u_i})^2} * \sqrt{\sum_{k=1}^{n} (u_{j,k} - \tilde{u_j})^2}}$$

$$= \frac{\sum_{k=1}^{n} u^*_{i,k} * u^*_{j,k}}{\sqrt{\sum_{k=1}^{n} u^{*\,2}_{i,k}} * \sqrt{\sum_{k=1}^{n} u^{*\,2}_{j,k}}}$$

$$= \text{cosine}(\underline{u}^*_i, \underline{u}^*_j),$$

where $\underline{u}^*_i = \underline{u}_i - \tilde{u_i}[1,1,\ldots,1]^T$; $\underline{u}^*_j = \underline{u}_j - \tilde{u_j}[1,1,\ldots,1]^T$; and are the mean-normalized vectors of $\underline{u}_i$ and $\underline{u}_j$, respectively. Conceptually, the mean normalization process is the only difference between PCA and other eigenvector-based approaches.

Since the eigenvectors of a symmetric matrix form an orthonormal basis of $R^m$, they may contain negative entries (Spence et al., 2000). Even though Kleinberg (1999) and Chen and Chen (2008) have shown experimentally that negative entries in an eigenvector are as important as positive entries for describing a certain unit pattern, the meaning of negative entries in their approaches is unexplainable. This is because textual units (e.g., terms, sentences, and documents) in information retrieval are usually characterized by frequency-based metrics, e.g., term frequency, document frequency, or TFIDF (Manning et al., 2008), which can never be negative. In PCA, however, the mean normalization process of the correlation coefficient gives bipolar meaning to positive and negative entries and that helps us partition textual units into bipolar groups in accordance with their signs in $\underline{v}$.

1 The right singular vectors of a matrix $A$ used by Gong and Liu (2001) are equivalent to the eigenvectors of a symmetric matrix $A^T A$ whose entries are the inner products of the corresponding columns of $A$.

2 The inner product is equivalent to the cosine formula when the calculated vectors are length normalized (Manning et al., 2008).

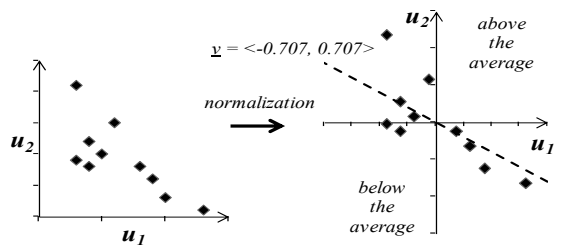


Figure 1. The effect of the mean normalization process.

The synthesized example in Figure 1 illustrates the effect of the normalization process. In this example, we are only interested in textual units $u_1$ and $u_2$; the corpus consists of ten blocks. Graphically, each block can be represented as a point in a 2-dimensional vector space. The mean normalization process moves the origin of the 2-dimensional vector space to the centroid of the blocks that makes negative unit values explainable. A negative unit of a block in this normalized vector space indicates that the number of occurrences of the unit in the block is less than the unit's average; by contrast, a positive unit means that the number of occurrences of the unit in a block is above the average. In the figure, the most important unit pattern $\underline{v}$ <-0.707, 0.707> calculated by PCA is represented by the dashed line. The signs of $\underline{v}$'s entries indicate that the occurrence of $u_1$ will be lower than the average if $u_2$ occurs frequently in a block. In addition, as the signs of entries in an eigenvector are invertible (Spence et al., 2000), the constituent of $\underline{v}$ also claims that if $u_1$ occurs frequently in a block, then the probability that we will observe $u_2$ in the same block will be lower than expected. The instances of bipolar word usage behavior presented in $\underline{v}$ are consistent with the distribution of the ten blocks. As mentioned in Section 2, Kanayama and Nasukawa (2006) validated that polar text units with the same polarity tend to appear together to make contexts coherent. Consequently, we believe that the signs in PCA's principal eigenvector are effective in partitioning textual units into bipolar groups.

### 3.3 Sparseness of Textual Units

A major problem with employing PCA to process textual data is the sparseness of textual units. To illustrate this problem, we collected 411 news documents about the 2009 NBA Finals from Google News and counted the frequency that each person name occurred in the documents. We also evaluate the documents in the experiment section to determine if the proposed approach is capable of bipolarizing the person names into the teams that played in the finals correctly. We rank the units according to their frequencies and list the frequencies in descending order in Figure 2. The figure shows that the frequency distribution follows Zipf's law (Manning et al., 2008); and for most units, the distribution in a block will be very sparse.

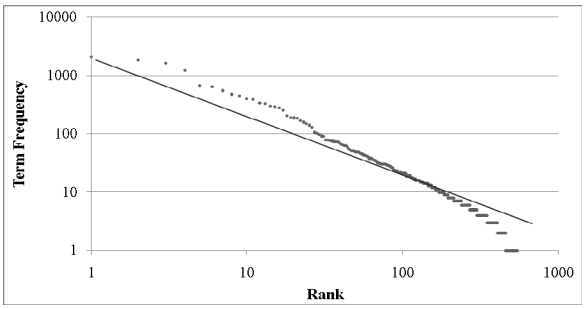


Figure 2. The rank-frequency distribution of person names on logarithmic scales (base 10).

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

#### Weighted Correlation Coefficient

The so-called data sparseness problem in scenario 2 affects many statistical information retrieval and language models (Manning et al., 2008). For units with the same polarity, data sparseness could lead to underestimation of their correlations because the probability that the units will occur together is reduced. Conversely, for uncorrelated units or units with opposite polarities,

data sparseness may lead to overestimation of their correlations because they are frequently jointly-absent in the decomposed blocks. While smoothing approaches, such as Laplace's law (also known as adding-one smoothing), have been developed to alleviate data sparseness in language models (Manning et al., 2008), they are not appropriate for PCA. This is because the correlation coefficient of PCA measures the divergence between units from their means, so adding one to each block unit will not change the divergence. To summarize, data sparseness could influence the correlation coefficient when units do not co-occur. Thus, for two units $u_i$ and $u_j$, we separate $B$ into co-occurring and non-co-occurring parts and apply the following weighted correlation coefficient:

$$corr_w(u_i,u_j) = \left( \begin{array}{l} (1-\alpha) \sum_{b \in co(i,j)} (u_{i,b} - \tilde{u_i}) * (u_{j,b} - \tilde{u_j}) + \\ \qquad \alpha \sum_{b \in B-co(i,j)} (u_{i,b} - \tilde{u_i}) * (u_{j,b} - \tilde{u_j}) \end{array} \right) / \sqrt{(1-\alpha) \sum_{b \in co(i,j)} (u_{i,b} - \tilde{u_i})^2 + \alpha \sum_{b \in B-co(i,j)} (u_{i,b} - \tilde{u_i})^2} * \sqrt{(1-\alpha) \sum_{b \in co(i,j)} (u_{j,b} - \tilde{u_j})^2 + \alpha \sum_{b \in B-co(i,j)} (u_{j,b} - \tilde{u_j})^2},$$

where $corr_w(u_i,u_j)$ represents the weighted correlation coefficient between units $i$ and $j$; and $co(i,j)$ denotes the set of blocks in which units $i$ and $j$ co-occur. The range of parameter $\alpha$ is within [0,1]. It weights the influence of non-co-occurring blocks when calculating the correlation coefficient. When $\alpha = 0.5$, the equation is equivalent to the standard correlation coefficient; and when $\alpha = 0$, the equation only considers the blocks in which units $i$ and $j$ co-occur. Conversely, when $\alpha = 1$, only non-co-occurring blocks are employed to calculate the units' correlation. In the experiment section, we will examine the effect of $\alpha$ on bipolar person name identification.

**Off-topic Block Elimination**

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

## 4 Performance Evaluations

In this section, we evaluate two topics with bipolar (or competitive) viewpoints to demonstrate the efficacy of the proposed approach.

### 4.1 The 2009 NBA Finals

For this experiment, we collected 411 news documents about the 2009 NBA Finals from Google News during the period of the finals (from 2009/06/04 to 2009/06/16). The matchup of the finals was Lakers versus Orlando Magic. In this experiment, a block is a topic document, as paragraph tags are not provided in the evaluated documents. First, we parsed the blocks by using Stanford Named Entity Recognizer[3] to extract all possible named entities. We observed that the parser sometimes extracted false entities (such as Lakers Kobe) because the words in the headlines were capitalized and that confused the parser. To reduce the effect of false extraction by the parser, we examined the extracted named entities manually. After eliminating false entities, the dataset comprised 546 unique named entities; 538 were person names and others represented organizations, such as basketball teams and basketball courts. To examine the effect of the weighted correlation coefficient, parameter $\alpha$ is set between 0 and 1, and increased in increments of 0.1; and the threshold $\beta$ used by off-topic block elimination is set at 0.3. The frequency distribution of the person names, shown in Figure 2, indicates that many of the person names rarely appeared in the examined blocks, so their distribution was too sparse for PCA. Hence, in the following subsections, we sum the frequencies of the 538 person names in the examined blocks. We select the first $k$ frequent person names, whose accumulated term frequencies reach 60% of the total frequencies, for evaluation. In other words, the evaluated person names account for 60% of the person name occurrences in the examined blocks.

For each parameter setting, we perform principal component analysis on the examined blocks and the selected entities, and partition the entities into two bipolar groups according to

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

their signs in the principal eigenvector. To evaluate the accuracy rate of bipolarization, we need to label the team of each bipolar group. Then, the accuracy rate is the proportion of the entities in the groups that actually belong to the labeled teams. Team labeling is performed by examining the person names in the larger bipolarization group. If the majority of the entities in the group belong to the Lakers (Magic), we label the group as Lakers (Magic) and the other group as Magic (Lakers). If the two bipolar groups are the same size, the group that contains the most Lakers (Magic) entities is labeled as Lakers (Magic), and the other group is labeled as Magic (Lakers). If both groups contain the same number of Lakers (Magic) entities, we randomly assign team labels because all assignments produce the same accuracy score. To the best of our knowledge, there is no similar work on person name bipolarization; therefore, for comparison, we use a baseline method that assigns the same polarity to all the person names.

| Magic | | Lakers | |
|---|---|---|---|
| Dwight Howard | 0.0884 | Derek Fisher | -0.0105 |
| Hedo Turkoglu | 0.1827 | Kobe Bryant | -0.2033 |
| Jameer Nelson | 0.3317 | Lamar Odom | -0.1372 |
| Jeff Van Gundy*+ | 0.3749 | LeBron James*^ | -0.0373 |
| Magic Johnson* | 0.3815 | Mark Jackson*^ | -0.2336 |
| Rafer Alston | 0.3496 | Pau Gasol | -0.1858 |
| Rashard Lewis | 0.1861 | Paul Gasol*+ | -0.1645 |
| Stan Van Gundy | 0.4035 | Phil Jackson | -0.2553 |

Table 1. The bipolarization results for NBA person names. ($\alpha = 0.8$ and $\beta = 0.3$)

Table 1 shows the bipolarization results for frequent person names in the dataset. The parameter $\alpha$ is set at 0.8 because of its superior performance. The left-hand column of the table lists the person names labeled as Magic and their entry values in the principal eigenvector; and the right-hand column lists the person names labeled as Lakers. It is interesting to note that the evaluated entities contain person names irrelevant to 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 finals. In addition, many documents misspell Pau Gasol as Paul Gasol. Even though the names refer to the same player, the named entity recognizer parses them as distinct entities. We propose two evaluation strategies, called *strict evaluation* and *non-strict evaluation*. The strict evaluation strategy treats the person names that do not refer to the players, coaches in the finals as false positives. Under the non-strict strategy, the person names that are closely related to Lakers or Magic players, such as a player's relatives or misspellings, are deemed true positives if they are bipolarized into the correct teams. In Table 1, 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. In addition, the symbol + indicates that the person name represents a relative of a member of the team he/she is bipolarized to; or the name is a misspelling, but it refers to a member of the bipolarized team. This kind of bipolarization is correct under the non-strict evaluation strategy. As shown in Table 1, the proposed method bipolarizes the important persons in the finals correctly without using any external information source. The accuracy rates of strict and non-strict evaluation are 68.8% and 81.3% respectively. The rates are far better than those of the baseline method, which are 37.5% and 43.8% respectively. If we ignore the neutral entities, which are always wrong no matter what bipolarization approach is employed, the strict and non-strict accuracies are 78.6% and 92.9% respectively. In the non-strict evaluation, we only mis-bipolarized Magic Johnson as Magic. The mistake also reflects a problem with person name resolution when the person names that appear in a document are ambiguous. In our dataset, the word 'Magic' sometimes refers to Magic Johnson and sometimes to Orlando Magic. Here, we do not consider a sophisticated person name 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 Cavaliers) 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-occur in the documents, LeBron James was wrongly classified as a member of Lakers.

Figures 3 and 4 illustrate the effects of the weighted correlation coefficient and off-topic block elimination on NBA person name bipolarization. As shown in the figures, eliminating off-topic blocks generally improves the system performance. It is noteworthy that, when off-topic blocks are eliminated, large $\alpha$ values produce good bipolarization performances. As mentioned in Section 3.3, a large $\alpha$ 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 specifically reveals opposing or jointly-absent relationships between entities. Therefore, the bipolarization performance improves as $\alpha$ increases. Conversely, when off-topic blocks are not eliminated, the set of non-co-occurring blocks will contain off-topic blocks. As both entities in a pair tend to be absent in off-topic blocks, a large $\alpha$ value will lead to overestimation of the correlation between bipolar entities. Consequently, the bipolarization accuracy decreases as $\alpha$ increases. It is also interesting to note that the bipolarization performance decreases as $\alpha$ decreases. We observed that some of the topic documents are recaps of the finals, which tend to mention Magic and Lakers players together. As a small $\alpha$ value makes co-occurrence blocks important, recap-style documents will overestimate the correlation between bipolar entities. Consequently, the bipolarization performance is inferior when $\alpha$ is small.

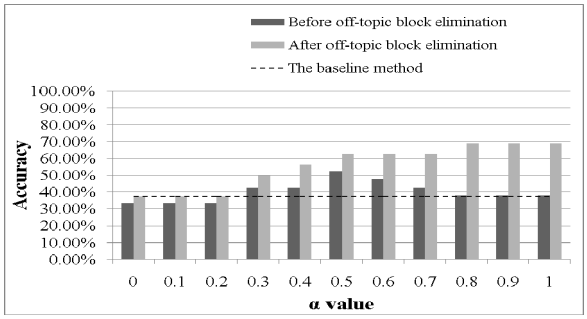


Figure 3. The effects of the weighted correlation coefficient and off-topic block elimination on NBA person name bipolarization. (Strict)

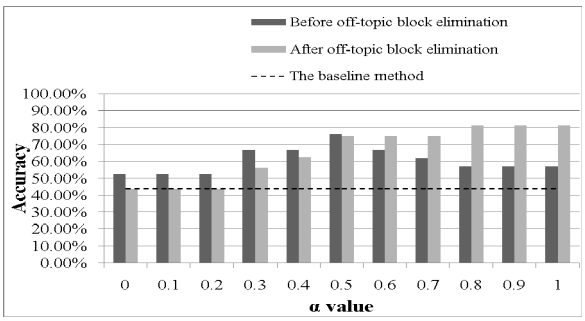


Figure 4. The effects of the weighted correlation coefficient and off-topic block elimination on NBA person name bipolarization. (Non-strict)

### 4.2 Taiwan's 2009 Legislative By-Elections

For this experiment, we evaluated Chinese news documents about Taiwan's 2009 legislative by-elections, in which two major parties, the Democratic Progressive Party (DPP) and the KouMin-Tang (KMT), campaigned for three legislative positions. Since the by-elections were regional, not many news documents were published during the campaign. In total, we collected 89 news documents that were published in The Liberty Times [4] during the election period (from 2009/12/27 to 2010/01/11). Then, we used a Chinese word processing system, called Chinese Knowledge and Information Processing (CKIP)[5], to extract possible Chinese person names in the documents. Once again, the names were examined manually to remove false extractions. The dataset comprised 175 unique person names. As many of the names only appeared once, we selected the first $k$ frequent person names whose accumulated frequency was at least 60% of the total term frequency count of the person names for evaluation. We calculated the accuracy of person name bipolarization by the same method as the NBA experiment in order to assess how well the bipolarized groups represented the KMT and the DPP. As none of the selected names were misspelled, we do not show the non-strict accuracy of bipolarization. The threshold $\beta$ is set at 0.3, and each block is a topic document.

Table 2 shows the bipolarization results for the frequent person names of the candidates of the respective parties, the party chair persons, and important party staff members. The accuracy rates of the bipolarization and the baseline me-

[4] http://www.libertytimes.com.tw/index.htm
[5] http://ckipsvr.iis.sinica.edu.tw/

thods are 70% and 50%, respectively. It is noteworthy that the chairs of the DPP and the KMT, who are Ing-wen Tsai and Ying-jeou Ma respectively, are correctly bipolarized. We observed that, during the campaign, the chairs repeatedly helped their respective party's candidates gain support from the public. As the names of the chairs and the candidates often co-occur in the documents, they can be bipolarized accurately. We also found that our approach bipolarized two candidates incorrectly if the competition between them was fierce. For instance, Kun-cheng Lai and Li-chen Kuang campaigned intensively for a single legislative position. As they often commented on each other during the campaign, they tend to co-occur in the topic documents. PCA therefore misclassifies them as positively correlated and incorrectly groups Kun-cheng Lai with the KMT party.

| KMT (國民黨) | | DPP (民進黨) | |
|---|---|---|---|
| Kun-cheng Lai (賴坤成)* | 0.39 | Wen-chin Yu (余文鈸)* | -0.56 |
| Li-chen Kuang (鄺麗貞) | 0.40 | Den-yih Wu (吳敦義)* | -0.03 |
| Li-ling Chen (陳麗玲) | 0.01 | Chao-tung Chien (簡肇棟) | -0.56 |
| Ying-jeou Ma (馬英九) | 0.05 | Ing-wen Tsai (蔡英文) | -0.17 |
| | | Tseng-chang Su (蘇貞昌) | -0.01 |
| | | Jung-chung Kuo (郭榮宗) | -0.01 |

Table 2. The bipolarization results for the election dataset. ($\alpha = 0.7$)

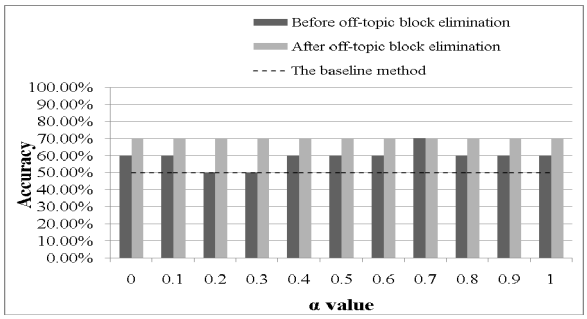


Figure 5. The effects of the weighted correlation coefficient and off-topic block elimination.

Figure 5 shows that off-topic block elimination is effective in person name bipolarization. However, the weighted correlation coefficient only improves the bipolarization performance slightly. We have investigated this problem and believe that the evaluated person names in the documents are frequent enough to prevent the data sparseness problem. While the weighted correlation coefficient does not improve the bipolarization performance significantly, the proposed PCA-based approach can still identify the bipolar parties of important persons accurately. Unlike the results in the last section, the accuracy rate in this experiment does not decrease as $\alpha$ decreases. This is because the topic documents generally report news about a single party. As the documents rarely recap the activities of parties, the co-occurrence blocks accurately reflect the bipolar relationship between the persons. Hence, a small $\alpha$ value can identify bipolar person names effectively.

The evaluations of the NBA and the election datasets demonstrate that the proposed PCA-based approach identifies bipolar person names in topic documents effectively. As the writing styles of topic documents in different domains vary, the weighted correlation coefficient may not always improve bipolarization performance. However, because we eliminate off-topic blocks, a large $\alpha$ value always produces superior bipolarization performances.

## 5 Conclusion

In this paper, we have proposed an unsupervised approach for identifying bipolar 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. The experiment results demonstrate that the proposed approach identifies bipolar person names of topics successfully without using any external knowledge; moreover, it is language independent. The results also show that off-topic block elimination along with a large $\alpha$ value for the weighted correlation coefficient generally produce accurate person name bipolarization. In the future, we will integrate text summarization techniques with the proposed bipolarization method to provide users with polarity-based topic summaries. We believe that summarizing important information about different polarities would help users gain a comprehensive knowledge of a topic.

## Acknowledge

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. This work was supported in part by NSC 97-2221-E-002-225-MY2.

## References


Chen, Chien Chin and Meng Chang Chen. 2008. TSCAN: a novel method for topic summarization and content anatomy. In *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 579-586.

Esuli, Andrea and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation*.

Feng, Ao and James Allan. 2007. Finding and Linking Incidents in News. In *Proceedings of the sixteenth ACM Conference on information and knowledge management*, pages 821-830.

Ganapathibhotla, Murthy and Bing Liu. 2008. Mining Opinions in Comparative Sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 241-248.

Gong, Yihong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19-25.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174-181.

Kanayama, Hiroshi and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355-363.

Kleinberg, Jon M.. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*, *5*, pages 604-632.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Nallapati, Ramesh, Ao Feng, Fuchun Peng and James Allan. 2004. Event Threading within News Topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 446-453.

Smith, Lindsay I.. 2002. *A Tutorial on Principal Components Analysis*. Cornell University.

Spence, Lawrence E., Arnold J. Insel and Stephen H. Friedberg. 2000. *Elementary Linear Algebra, A Matrix Approach*. Prentice Hall.

Turney, Peter D., and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, pages 315-346.