# Mining Outlier Participants: Insights Using Directional Distributions in Latent Models

Didi Surian[1] and Sanjay Chawla[2]

[1,2]University of Sydney and [1,2]NICTA
dsur5833@uni.sydney.edu.au, sanjay.chawla@sydney.edu.au

**Abstract.** In this paper we will propose a new probabilistic topic model to score the expertise of participants on the projects that they contribute to based on their previous experience. Based on each participant's score, we rank participants and define those who have the lowest scores as *outlier participants*. Since the focus of our study is on outliers, we name the model as **M**ining **O**utlier **P**articipants from **P**rojects (MOPP) model. MOPP is a topic model that is based on directional distributions which are particularly suitable for outlier detection in high-dimensional spaces. Extensive experiments on both synthetic and real data sets have shown that MOPP gives better results on both topic modeling and outlier detection tasks.

## 1 Introduction

We present a new topic model to capture the interaction between participants and the projects that they participate in. We are particularly interested in outlier projects, i.e., those which include participants who are unlikely to join in based on their past track record.

**Example:** Consider the following example. Three authors $A1, A2$ and $A3$ come together to write a research paper. The authors and the paper profiles are captured by a "category" vector as shown in Table 1. A category can be a "word" or "topic" and is dependent upon the model we use.

**Table 1.** Example: "category" vectors for authors and paper profiles. The dot products determine the "outlierness" of the authors to the paper

| Entity | Category 1 | Category 2 | Category 3 | Dot.P |
|---|---|---|---|---|
| Paper | 0.1 | 0.1 | 0.8 | |
| A1 | 0.1 | 0.2 | 0.7 | 0.59 |
| A2 | 0.2 | 0.2 | 0.6 | 0.52 |
| A3 | 0.8 | 0.1 | 0.1 | 0.17 |

Then we can compute the dot products $< Paper, A1 >$, $< Paper, A2 >$ and $< Paper, A3 >$ as shown in the last column. Based on the dot product, $A3$ is the most outlier participant in the paper.

*The challenge we address in this paper is to develop a new topic model to accurately form categories which can be used to discover outlier behavior as illustrated in Table 1.*

A natural approach is to use Latent Dirichlet Allocation (LDA) to model the track record of the participants and also the project descriptions. The advantage of using LDA is that we carry out the analysis in the "topic" space, which is known to be more robust compared to a word-level analysis. However, LDA is not particularly suitable for outlier detection as we illustrate in the following example.

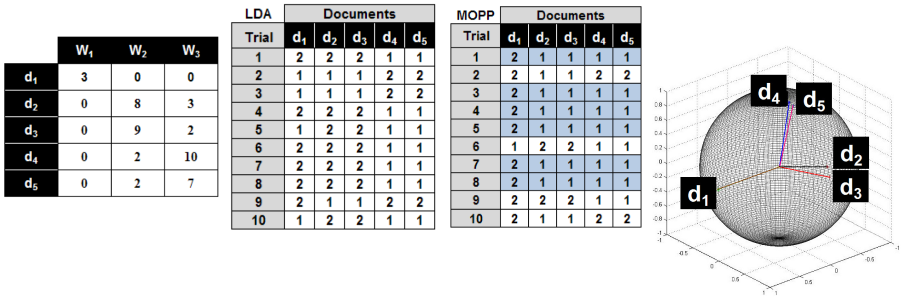
Assume that we want to cluster five documents into two clusters ($C1$, $C2$) and there are three unique words in the vocabulary (i.e. the dimension is three). Let $d$ = [$n_1$,$n_2$,$n_3$] represents number of occurrences of word $w_1$, $w_2$, and $w_3$ in document $d$. Assume we have: $d_1$ = [3,0,0], $d_2$ = [0,8,3], $d_3$ = [0,9,2], $d_4$ = [0,2,10], and $d_5$ = [0,2,7] (illustrated in Fig. 1(a)). It is likely that $d_1$ does not share any similarity with the other documents because $w_1$ in $d_1$ does not appear in the other documents. On the other hand, $d_2$, $d_3$, $d_4$ and $d_5$ have some common words, i.e., $w_2$ and $w_3$. Intuitively, $d_1$ should be clustered separately ($C1$), while $d_2$, $d_3$, $d_4$ and $d_5$ should be clustered together (in $C2$). However, because LDA is mainly affected by the word counts, $d_2$ will be clustered together with $d_3$ (in $C1$), and $d_4$ with $d_5$ (in $C2$); while $d_1$ will be clustered either to $C1$ *or* $C2$. Figure 1(b) shows the results of running ten consecutive trials with LDA[1]. As we can observe that **none** of the ten consecutive trials follows our first intuition (i.e. $d_1$ in a separate cluster). On the other hand, our proposed model gives a better solution which is shown in Fig. 1(c). Notice from Fig. 1(d) that if we represent the documents as unit vectors on a sphere, document $d_1$ is well separated from $d_2$ and $d_3$ or $d_3$ and $d_4$.

As the above example illustrates, the weakness of LDA is that it is not fully sensitive to the directionality of data and is essentially governed by word counts. Our proposed approach extends LDA by integrating directional distribution and treating the observations in a vector space. Specifically, we represent very high dimensional (and often sparse) observations as unit vectors, where direction plays a pivotal role in distinguishing one entity from another.

We highlight the importance of our proposed model from two perspectives. First, due to the integration of directional distribution, the resulting clusters are more robust against outliers and it could potentially give a better clustering solution. Secondly, because of the robustness, the outliers are well separated from the rest of the data, which could be used as a base for outlier detection. We present more details about the directional distribution that we use, the von Mises-Fisher (vMF) distribution, in Sect. 2. We summarize our contributions as follows:

1. We introduce a novel problem for discovering outlier participants in projects based on their previous working history.

[1] As LDA treats a document as a finite mixture over a set of topics, we assume a topic with the largest proportion as the topic of a document.



|  | W$_1$ | W$_2$ | W$_3$ |
|---|---|---|---|
| d$_1$ | 3 | 0 | 0 |
| d$_2$ | 0 | 8 | 3 |
| d$_3$ | 0 | 9 | 2 |
| d$_4$ | 0 | 2 | 10 |
| d$_5$ | 0 | 2 | 7 |

| LDA | Documents |  |  |  |  |
|---|---|---|---|---|---|
| Trial | d$_1$ | d$_2$ | d$_3$ | d$_4$ | d$_5$ |
| 1 | 2 | 2 | 2 | 1 | 1 |
| 2 | 1 | 1 | 1 | 2 | 2 |
| 3 | 1 | 1 | 1 | 2 | 2 |
| 4 | 2 | 2 | 2 | 1 | 1 |
| 5 | 1 | 2 | 2 | 1 | 1 |
| 6 | 2 | 2 | 2 | 1 | 1 |
| 7 | 2 | 2 | 2 | 1 | 1 |
| 8 | 2 | 2 | 2 | 1 | 1 |
| 9 | 2 | 1 | 1 | 2 | 2 |
| 10 | 1 | 2 | 2 | 1 | 1 |

| MOPP | Documents |  |  |  |  |
|---|---|---|---|---|---|
| Trial | d$_1$ | d$_2$ | d$_3$ | d$_4$ | d$_5$ |
| 1 | 2 | 1 | 1 | 1 | 1 |
| 2 | 2 | 1 | 1 | 2 | 2 |
| 3 | 2 | 1 | 1 | 1 | 1 |
| 4 | 2 | 1 | 1 | 1 | 1 |
| 5 | 2 | 1 | 1 | 1 | 1 |
| 6 | 1 | 2 | 2 | 1 | 1 |
| 7 | 2 | 1 | 1 | 1 | 1 |
| 8 | 2 | 1 | 1 | 1 | 1 |
| 9 | 2 | 2 | 2 | 1 | 1 |
| 10 | 2 | 1 | 1 | 2 | 2 |

(a) (b) (c) (d)

**Fig. 1.** Example: (a) the word counts for five documents; Topic assignments for the five documents: (b) LDA (c) MOPP; (d) Distribution of the documents on the unit sphere. Notice from (c) that MOPP separates $d_1$ from the other documents 6 out of 10 times.

2. We model the proposed problem using a topic-based hierarchical generative model based on the von Mises-Fisher (vMF) directional distribution. The model is named MOPP: **M**ining **O**utlier **P**articipants from **P**rojects.
3. We have implemented MOPP and compared it with a variation of Latent Dirichlet Allocation (LDA) on several synthetic and real data sets. We show that MOPP improves both the ability to detect outliers and form high quality clusters compared to LDA.

## 2 Related Work

The outlier or anomaly detection problem has been extensively researched in the data mining, machine learning and statistical communities. The survey by Chandola et. al. [1], provides an overview of contemporary data mining methods used for outlier detection. Our proposed model is mainly inspired by the concept of hierarchical structure in topic model used in Latent Dirichlet Allocation (LDA) [2]. In this section we present LDA and some work in directional distributions.

### 2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model proposed by Blei et al. [2]. LDA describes the generative process and captures the latent structure of topics in a text corpus. LDA is now widely used for the clustering and topic modeling tasks. The graphical representation of LDA is shown in Fig. 2.

The plate M represents documents and the plate Nm represents words in document $m$. W$_{m,n}$ represents the observed word $n$ in document $m$. Z$_{m,n}$ represents topic assignment of word $n$ in document $m$. $\theta_m$ represents topic mixture in

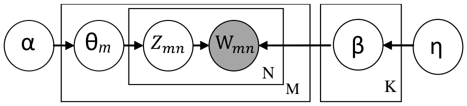


**Fig. 2.** LDA: Graphical Representation

document $m$. $\beta$ represents the underlying latent topics of those documents. Both $\alpha$ and $\eta$ represent the hyperparameters for the model. The generative process of LDA is summarized as follows:

| | |
|---|---|
| Topic mixture: | $\theta_m \| \alpha \sim \text{Dirichlet}(\alpha)$, $m \in \text{M}$ |
| Topic: | $\beta_k \| \eta \sim \text{Dirichlet}(\eta)$, $k \in \text{K}$ |
| Topic assignment: | $\text{Z}_{m,n} \| \theta_m \sim \text{Multinomial}(\theta_m)$, $n \in \text{Nm}$ |
| Word: | $\text{W}_{m,n} \| \beta_{\text{Z}_{m,n}} \sim \text{Multinomial}(\beta_{\text{Z}_{m,n}})$ |

### 2.2 Directional Distribution

Mardia [3,4] and Fisher [5] discussed von Mises-Fisher (vMF) distribution as a natural distribution and the simplest parametric distribution for directional statistics. The vMF distribution has support on $\mathbb{S}^{d-1}$ or unit ($d$-1)-sphere embedded in $\mathbb{R}^d$, and has properties analogous to the multi-variate Gaussian distribution. More details about vMF distribution can be found in [3]. The probability density function of vMF distribution is described as follows:

$$f(\mathbf{x}|\mu, \kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)} e^{\kappa \mu^T \mathbf{x}} \tag{1}$$

where $\mu$ is called *mean direction*, $\|\mu\| = 1$; $\kappa$ is called *concentration parameter* and characterizes how strongly the unit random vectors are concentrated[2] about the mean direction ($\mu$), and $\kappa \geq 0$. $I_r(\cdot)$ represents the modified Bessel function of the first kind of order $r$.

A body of work has shown the effectiveness of directional distribution for modeling text data. Zhong et al. [6] shown that vMF distribution gives superior results in high dimensions comparing to Euclidean distance-based measures. Banerjee et al. [7,8] proposed the use of EM algorithm for a mixture of von Mises-Fisher distributions (movMF). Banerjee et al. [9] also observed the connection between vMF distributions in a generative model and the spkmeans algorithm [10] which is superior for clustering high-dimensional text data. Reisinger et al. [11] proposed a model named SAM that decomposes spherically distributed data into weighted combinations of component vMF distributions. However, both movMF and SAM lack a hierarchical structure and cannot be scaled-up for domains involving multiple levels of structure.

The effectiveness of vMF distribution has also been studied in many outlier detection studies. Ide et al. [12] proposed an eigenspace-based outliers detection in computer systems, especially in the application layer of Web-based systems.

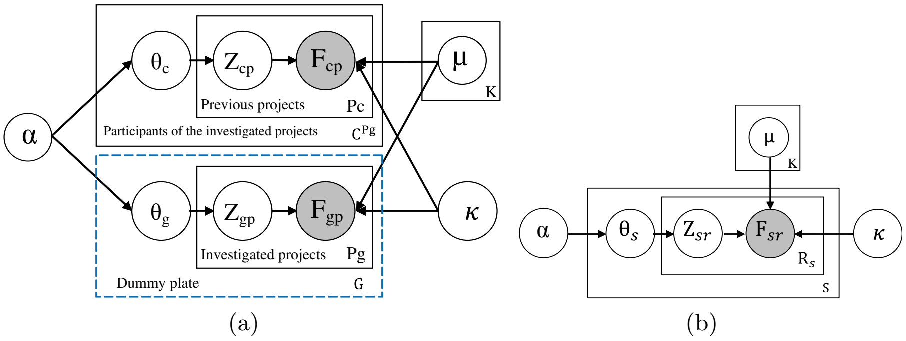
[2] Specifically, if $\kappa = 0$, the distribution is uniform and, if $\kappa \to \infty$, the distribution tends to concentrate on one density.

Fujimaki et al. [13] proposed the use of vMF distribution for spacecraft outliers detection. Both of these two papers use a single vMF distribution and compute the angular difference of two vectors to determine outliers. Kriegel et al. [14] proposed an approach to detect outliers based on angular deviation. Our proposed model uses a mixture of $K$-topics as latent factors that underlie the generative process of observations to detect outlier participants from projects.

## 3 Mining Outlier Participants

### 3.1 The Proposed Model

Figure 3(a) shows the graphical model of our proposed model. We use the following assumption: rectangular with solid line represents replication of plates, rectangular with dashed-line represents a single plate (named as a dummy plate), a shaded circle represents an observable variable, an unshaded circle represents latent/unobservable variable, and a directed arrow among circles represents a dependency among them.



(a) (b)

**Fig. 3.** Graphical representation: (a) MOPP, (b) The simplified MOPP

Recall that each project has a profile vector and each participant has a set of profile vectors. We refer the project as the *investigated project*. Each investigated project is represented by an $L_2$-normalized TF-IDF unit vector and each participant of the investigated project at time $t$ has a number of previous projects before $t$. The subscripts on the participants and the investigated projects plates in Fig. 3(a) are used to differentiate between the two plates. We keep a record about the respective subscripts and merge them to simplify MOPP. We call the new model as simplified MOPP and use it in the learning and inference process. The record will be used later after the learning and inference process to recover information about which plates the learnt latent values originally belong to. We show the simplified MOPP in Fig. 3(b). Notice that in the simplified MOPP, we "stack" the dummy plate G (with $\mathrm{P}_g$) and the participants of the investigated projects plates $\mathrm{C}^{P_g}$ (with $\mathrm{P}_c$). We then refer the stacked G and $\mathrm{C}^{P_g}$ as $S$. $\mathrm{P}_g$ and $\mathrm{P}_c$ on the respective G and $\mathrm{C}^{P_g}$ plates are referred as $\mathrm{R}_s$. We also rename the subscripts $c$ and $g$ as $s$, and the subscripts $p$ as $r$. Table 2 summarizes the main symbols we use in this work. We present the generative process of the proposed model in Algorithm 1.

**Table 2.** Main symbols and their definitions used in this work

| Symbol | Definition | Symbol | Definition |
|---|---|---|---|
| $\alpha$ | Hyperparameter | $\theta_g$ | Topic proportions for dummy plate |
| K | Numb. of topics | $\mathrm{P}_g$ | Investigated projects |
| $\mu$ | Mean direction of vMF | $\lvert\mathrm{P}_g\rvert$ | Numb. of $\mathrm{P}_g$ |
| $\kappa$ | Concentration parameter of vMF | $\mathrm{Z}_{gp}$ | Topic assignment of project $p$∈$\mathrm{P}_g$ |
| V | Dimension of unit vector F | $\mathrm{F}_{gp}$ | $L_2$-normalized TF-IDF unit vector of project $p$∈$\mathrm{P}_g$ |
| $\mathrm{C}^{\mathtt{P}_g}$ | Participants (of the investigated projects at snapshot $t$) plate | S | C + a dummy plate, C=∀$\mathrm{C}^{\mathtt{P}_g}$ |
| $\lvert\mathrm{C}^{\mathtt{P}_g}\rvert$ | Numb. of $\mathrm{C}^{\mathtt{P}_g}$ | $\lvert$S$\rvert$ | $\lvert$C$\rvert$ + 1 |
| $\theta_c$ | Topic proportions for participant $c$ | $\mathrm{R}_s$ | Projects on plate $s$, $\mathrm{R}_s$∈ {∀$\mathrm{P}_g$,∀$\mathrm{P}_c$} |
| $\mathrm{P}_c$ | Projects of participant $c$ before $t$ | $\lvert\mathrm{R}_s\rvert$ | Numb. of $\mathrm{R}_s$ |
| $\lvert\mathrm{P}_c\rvert$ | Numb. of $\mathrm{P}_c$ | $\theta_s$ | Topic proportions for $s$, $s$∈S |
| $\mathrm{Z}_{cp}$ | Topic assignment of project $p$∈$\mathrm{P}_c$ | $\mathrm{Z}_{sr}$ | Topic assignment of project $r$∈$\mathrm{R}_s$ |
| $\mathrm{F}_{cp}$ | $L_2$-normalized TF-IDF unit vector of project $p$∈$\mathrm{P}_c$ | $\mathrm{F}_{sr}$ | $L_2$-normalized TF-IDF unit vector of project $r$∈$\mathrm{R}_s$ |
| G | Dummy plate | | |

**Algorithm 1** Generative process of the proposed model

```
for $s$=1 to |S|, $s \in$ S do
    Choose topic proportions $\theta_s \sim \mathcal{D}ir(\alpha)$
        for $r$=1 to $|\mathrm{R}_s|$, $r \in \mathrm{R}_s$ do
            Choose a topic $\mu_k$, {1..K} $\ni \mathrm{Z}_{sr} \sim$ Multinomial($\theta_s$)
            Compute $L_2$-normalized TF-IDF unit vector $||\mathrm{F}_{sr}||_2 \sim$ vMF($\mu_k$, $\kappa$)
```

### 3.2 Learning and Inference Process

The joint distribution for a plate $s$ is given as follows:

$$p(\theta, \mathtt{Z}, \mathtt{F}|\alpha, \kappa, \mu) = p(\theta|\alpha) \prod_{r=1}^{|\mathtt{R}_s|} p(\mathtt{Z}_r|\theta) p(\mathtt{F}_r|\mathtt{Z}_r, \mu, \kappa) \tag{2}$$

We introduce variational parameters $\gamma$ and $\phi$ in the following variational distribution $q$ for the inference step in (3).

$$q(\theta, \mathtt{Z}|\gamma, \phi) = q(\theta|\gamma) \prod_{r=1}^{|\mathtt{R}_s|} q(\mathtt{Z}_r|\phi_r) \tag{3}$$

Application of Jensen's inequality for a plate $s$ [2] results in:

$$\log p(\mathtt{F}|\alpha, \mu, \kappa) = L(\gamma, \phi; \alpha, \mu, \kappa) + KL(q(\theta, \mathtt{Z}|\gamma, \phi)||p(\theta, \mathtt{Z}|\mathtt{F}, \alpha, \mu, \kappa)) \tag{4}$$

where KL represents the Kullback-Leibler divergence notation. By using the factorization of $p$ and $q$, we then expand the lower bound $L(\gamma,\phi;\alpha,\mu,\kappa)$ in (5):

$$\begin{aligned} L(\gamma, \phi; \alpha, \mu, \kappa) = {} & E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathtt{Z}|\theta)] + E_q[\log p(\mathtt{F}|\mathtt{Z}, \mu, \kappa)] \\ & - E_q[\log q(\theta)] - E_q[\log q(\mathtt{Z})] \end{aligned} \tag{5}$$

The derivation from (4) to (5) is similar with the derivation from (13) to (14) in ([2] p.1019) except for the third terms on the right hand side of (5) (highlighted), which now includes the vMF distribution. Due to the space constraint, interested readers should refer to [2] for the expansion details of the first, second, fourth, and fifth terms on the right hand side of (5).

The variational parameter $\gamma_k$ is calculated by maximizing (5) w.r.t. variational parameter $\gamma_k$. Using the same approach, the variational parameter $\phi$ is computed by maximizing (5) w.r.t. variational parameter $\phi_{rk}$ and introducing Lagrange multiplier, $\sum_{k=\mathit{1}}^{\mathtt{K}} \phi_{rk} = \mathbf{1}$. Both of these two steps result in (6).

$$\begin{aligned} \gamma_k^* &= \alpha_k + \sum_{r=1}^{|\mathtt{R}_s|} \phi_{rk} \\ \phi_{rk}^* &\propto exp((\Psi(\gamma_k) - \Psi(\sum_{k=1}^{\mathtt{K}} \gamma_k)) + \log p(\mathtt{F}_r|\mu_k, \kappa)) \end{aligned} \tag{6}$$

where $\Psi$ is the digamma function (the first derivative of the log Gamma function). To compute $\mu$, we need to calculate:

$$\mu^* = \arg\max_{\mu_k} \sum_{s=1}^{|\mathtt{S}|} \sum_{r=1}^{|\mathtt{R}_s|} \sum_{k=1}^{\mathtt{K}} \phi_{srk} \log p(\mathtt{F}_{sr}|\mu_k, \kappa) \tag{7}$$

Equation (7) is the same as fitting von Mises-Fisher distributions in a mixture of von Mises-Fisher distributions [7], where $\phi_{srk}$ is the mixture proportions. The complete process for variational EM algorithm in the learning and inference process includes the following iterative steps:

**E-step:** Compute the optimized values of $\gamma$ and $\phi$ for each plate $s$ using (6).
**M-step:** Maximize the lower bound w.r.t. to the model parameters $\alpha$ and $\mu$ described in the standard LDA model [2] and (7) respectively.

### 3.3 Scoring the Expertise to Project's Topic

The learning and inference process (Sect. 3.2) assigns a topic to each project on $\mathrm{R}_s$. Recall from Section 3.1 that we keep a record about the subscripts in MOPP but use the simplified MOPP for the learning and inference process. After the learning and inference process, we need to *reverse* map the learnt latent values back to the plate that they originally belong to. This process is crucial because our goal is to retrieve the topic proportions of each participant ($\theta_c$) and the topic assignment of the investigated project ($\mathrm{Z}_{gp}$). The topic proportions in $\theta_c$ based on the investigated project's topic will become the score of each participant in that project. An end-to-end pseudo-code that summarizes the steps to score each participant's expertise to projects and mine the outlier participants is shown in Algorithm 2.

**Algorithm 2** Mining outlier participants algorithm

**Input**: 1) Investigated projects' $L_2$-normalized unit vectors, 2) Each participant from the investigated projects with their $L_2$-normalized unit vectors, 3) $O$: number of outlier participants

**Output**: *topO* : List of Top-$O$ [<outlier participants, his project, his score>] outlier participants

**Steps**:

```
 1: Let $lPart \leftarrow \emptyset$, $topO \leftarrow \emptyset$
 2: Translate MOPP to simplified MOPP (Sect. 3.1)
 3: Do the learning and inference process (Sect. 3.2)
 4: Reverse map the learnt latent values to the respective projects. (Sect. 3.3)
 5: for every investigated project $p \in \mathrm{P}_g$
 6:   Get its topic assignment $\mathrm{Z}_{gp}$
 7:     for every participant $c \in \mathrm{C}^{\mathrm{P}_g}$
 8:         Get the topic proportions $\theta_c$
 9:         Get his score, $partScore \leftarrow \theta_c[\mathrm{Z}_{gp}]$
10:         $lPart \leftarrow <c, p, partScore>$
11: Sort $lPart$ in an ascending order based on $partScore$ values
12: $topO \leftarrow$ take the participants in the first Top-$O$ lowest scores in $lPart$
13: Output $topO$
```

We translate MOPP to simplified MOPP (Line 2) following our description in Sect. 3.1. Line 3 refers to the learning and inference process, which is summarized in the E-step and M-step (Sect. 3.2). Line 4: after we learn the model, we translate back the simplified MOPP to MOPP. Lines 5–10: for every project $p$ in $\mathrm{P}_g$, we infer its topic assignment ($\mathrm{Z}_{gp}$) and score each participant based on his topic proportion in $\mathrm{Z}_{gp}$. Lines 11–13: we label *topO* lowest scores participants as outlier participants.

## 4 Experiments

In this section we present our experiments to evaluate the performance of our proposed model. The proposed model is implemented in Matlab and we conduct the experiments on a machine with Intel® Core(TM) Duo CPU T6400 @2.00 GHz, 1.75 GB of RAM.

### 4.1 Baseline Methods

For our baseline methods, we use a cosine similarity test and a latent topic model Latent Dirichlet Allocation (LDA). We specifically measure the cosine similarity between the TF-IDF vectors of each participant and his/her current project. We form the TF-IDF vector of each participant from the words of all his/her previous projects, while the TF-IDF vector for a project is extracted from the words in the title. The cosine similarity is defined as follows:

$$cos(vec_1, vec_2) = \frac{vec_1 \cdot vec_2}{|vec_1||vec_2|} \tag{8}$$

The original LDA model (Sect. 2.1) is not suitable to be used directly for our purpose. We introduce the modified LDA for our purpose in Fig. 4. Following the translation for the simplified MOPP in Sect. 3.1, we keep a record of the subscripts. We perform the reverse mapping after the learning and inference process. The remaining steps are the same as step 5-13 in Algorithm 2. We set $\eta = 0.1$ and $\alpha = 50/\text{K}$, where K is the number of topics [15]. Because LDA returns topic mixture for each project and each participant, we use (9) to score the participants. We sort the scores in an ascending order to list the outlier participants, where $t_p$ is the topic mixture of project $p$ and $s_{i,p}$ is the topic mixture of a participant $i$ in project $p$.

$$< t_p \cdot s_{i,p} > \tag{9}$$

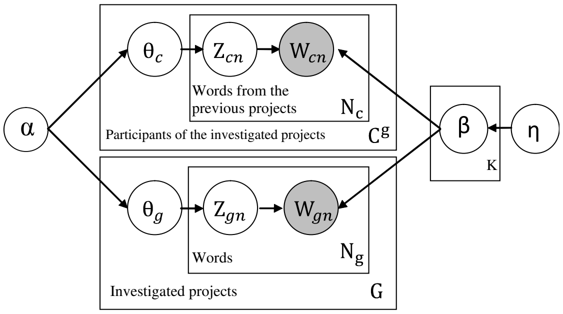


**Fig. 4.** The modified LDA for mining outlier participants

### 4.2 Semi Synthetic and Synthetic Data sets

Our experiments are divided into two parts: using semi-synthetic and synthetic data sets. For the semi-synthetic data set, we use data from the Arxiv HEP-TH (high energy physics theory) network[3]. This data set was originally released as a part of 2003 KDD Cup. We analyze the publications in year 2003 and extract the authors. We then take a number of authors from DBLP[4] and their publications. These DBLP authors will act as the outlier participants. For all publications extracted from HEP-TH and DBLP, we use words from the title. We then inject the authors from DBLP to the HEP-TH randomly. We have 1,212 HEP-TH authors in 554 projects. We vary the number of outlier participants and the results are the average values over ten trials. To evaluate the performance of our proposed model and the baseline methods, we use precision, recall and the F1 score. We show the results in Fig. 5.

[3] http://snap.stanford.edu/data/cit-HepTh.html

[4] www.informatik.uni-trier.de/ ley/db/

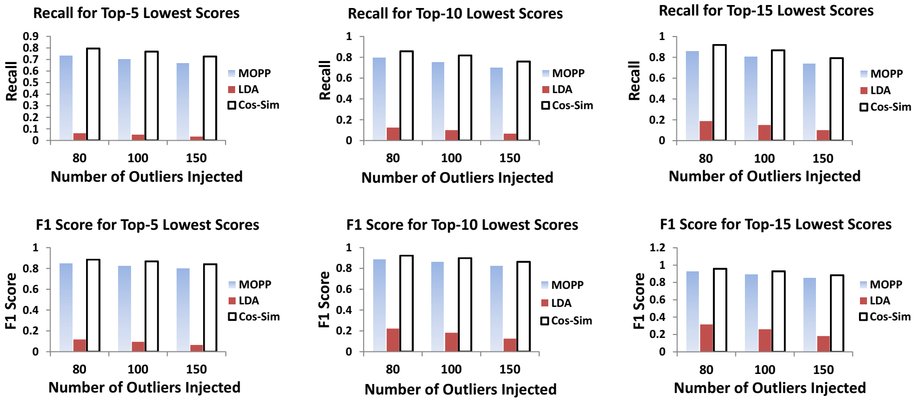


**Fig. 5.** Recall and F1 score for the first Top-5, 10 and 15 lowest scores in MOPP, LDA, and cosine similarity (Cos-Sim)

Figure 5 shows that LDA gives the lowest performance in detecting outlier participants (both the recall and F1 score are very low), while cosine similarity seems to be slightly better than MOPP. This is intuitive because the nature of the data set itself (HEP-TH and DBLP) is almost well-separated (the words in computer science are less likely to appear in the physics publications)[5].

We use the Normalized Mutual Information (NMI) measure to compare the performance of LDA and MOPP in reconstructing the underlying label distribution in the data set. NMI [17] is used to evaluate the clustering result [18, 19]. NMI is defined as follows:

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{10}$$

where $X$ represents cluster assignments, $Y$ represents true labels on the data set, $I$ and $H$ represent mutual information and marginal entropy. Table 3 presents the result and shows that MOPP gives a better cluster quality than LDA does.

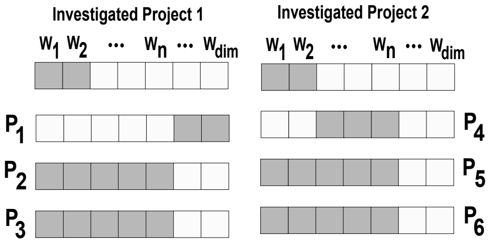
**Table 3.** NMI score: MOPP vs. LDA. The best values are highlighted, where NMI score close to 0 represents bad clustering quality and NMI = 1 for perfect clustering quality

| NMI | Number of Outliers Injected | | |
|---|---|---|---|
| | 80 | 100 | 150 |
| LDA | 0.308 | 0.404 | 0.626 |
| MOPP | **0.803** | **0.811** | **0.815** |

[5] In our initial experiments, we also included a classical outlier detection method Local Outlier Factor (LOF) [16]. Unfortunately LOF fails to detect outlier participants in any settings so we do not include the result here.

In the second part of the experiment, we form a small synthetic data set that represents a scenario illustrated in Fig. 6(a). This scenario is often observed in the real word that the extracted words from the participants may not appear in the investigated projects' extracted words. For example in Fig. 6(a) the word $W_1$ and $W_2$ appear in the investigated project 2, but do not appear in the previous projects of participant $P_4$. Because cosine similarity compares directly between TF-IDF vector of a participant and an investigated project, obviously $P_1$ and $P_4$ in Fig. 6(a) will be marked as outlier participants in project 1 and 2 respectively. However if we analyze in the topic space, the *true* outlier participant should be $P_1$, because $P_4$'s words are likely to share same topic with the investigated project 2 (through words in $P_5$ and $P_6$). We generate the synthetic data set by first generating words with random occurrence for the investigated projects. We then generate words for the participants with also random occurrence[6]. We generate three types of participant: "normal", "spurious-outlier", and "true-outlier" participants. $P_2$, $P_3$, $P_5$ and $P_6$ are examples of the normal participants, $P_4$ is an example of the spurious-outlier participant, and $P_1$ is an example of the true-outlier participant. We randomly assign all the participants to the investigated projects and inject the true-outlier participants.

The table in Fig. 6(b) shows that MOPP outperforms both cosine similarity and LDA[7] in this scenario. As we can observe MOPP returns all the true-outlier participants and all true-outlier participants have the lowest score. The cosine similarity returns all the true-outlier participants together with the spurious-outlier participants (low precision and high recall score). On the other hand, LDA correctly returns the true-outlier participants (high precision score), but misses many true outliers (low recall score).



(a)

| | MOPP | Cos-Sim | LDA |
|---|---|---|---|
| Precision | **1.0** | 0.7 | 1.0 |
| Recall | **1.0** | 1.0 | 0.05 |
| F1 Score | **1.0** | 0.8 | 0.09 |

(b)

**Fig. 6.** (a) Scenario used in the synthetic experiment: *dim* is the dimensionality of vocabulary, $W$ represents word, $P$ represents participant, shaded box of $W_n$ represents a certain number of appearances of word $W_n$ and unshaded box of $W_n$ means word $W_n$ does not appear (b) Precision, recall and F1 score of MOPP, cosine similarity (Cos-Sim) and LDA

[6] To fit the scenario, we generate words with occurrence $> 0$

[7] For MOPP and cosine similarity, we consider results from the lowest returned score, while for LDA we take the results from the Top-5 lowest returned scores.

**Running time:** MOPP **vs. LDA.** In this section we compare the running time of LDA and MOPP w.r.t. the dimension of data. We form synthetic data sets with various dimensions: 100, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, and 5,000. We then run LDA and MOPP with various number (5, 10, 15 and 20) of topics/clusters. Figure 7 presents the results of the running time for 1,000 iterations of the respective model. It is clear that MOPP scales linearly as the dimensionality of the data set increases. As the number of topics and dimension keep increasing, MOPP can run under 500 seconds for 1,000 iterations or less than half second per iteration.

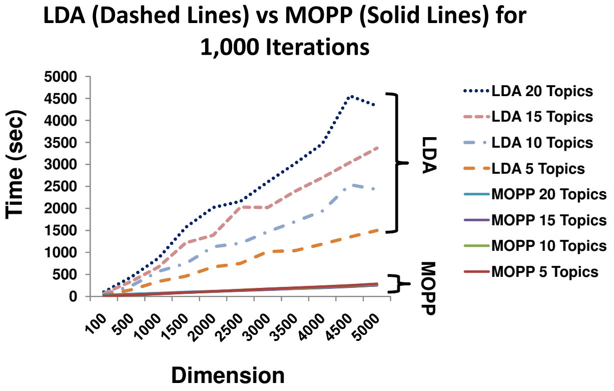


**Fig. 7.** Running time: MOPP vs. LDA for synthetic data set with various dimensions

### 4.3 Real Data set

**Experimental Settings.** We use a subset of DBLP to evaluate the performance of MOPP. In a bibliographic setting, a publication represents a project and an author represents a participant in our proposed model[8].

We take projects from year 2005 to 2011 of four conferences that represent four main research fields, i.e. VLDB (databases), SIGKDD (data mining), SIGIR (information retrieval), and NIPS (machine learning). We analyze the title of each project. We remove words which are *too common* or *too rare*[9] from our analysis. We have 141,999 projects in total.

We only consider projects which have at least two participants who have more than nine previous[10] projects. We assume that a participant who has at least ten previous projects in his/her profile has already "matured" his/her research direction. We refer these filtered projects as the *investigated projects*.

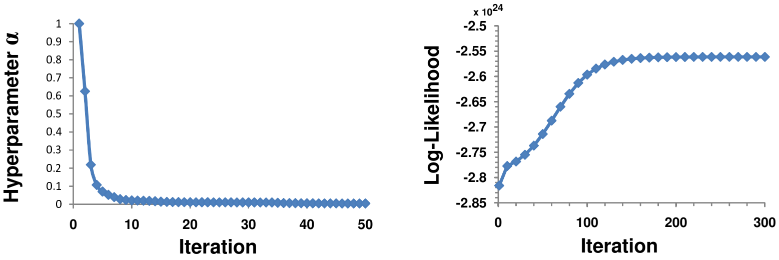
[8] Henceforth we use the term a project for a publication and a participant for an author.

[9] We determine words that appear less than 100 times are too rare and more than 100,000 times are too often.

[10] This includes projects before year 2005 as well.

At the end of process, we have 792 investigated projects, 1,424 participants and the dimensionality of data (size of vocabulary) is 2,108. In average, each participant has 28.71 previous projects before he/she joins in the investigated project. The number of topics K that we use is five, i.e. four main research topics and one for the other), $\kappa$ for MOPP = 2,500, and $\alpha$ for MOPP is initialized to 1 to represent a non-informative prior [20, 21].

**Convergence Rate.** We now show our empirical verification that the variational EM of MOPP is able to converge. The convergence rate of $\alpha$ and variational EM for DBLP data set are shown in Fig. 8(a) and Fig. 8(b) respectively.



(a) (b)

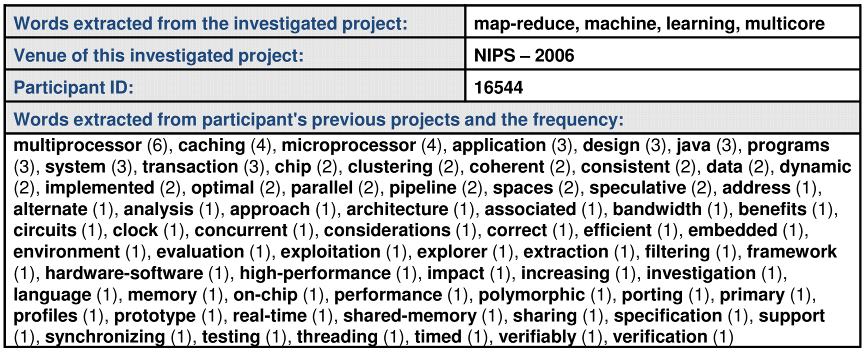
**Fig. 8.** DBLP: Rate of convergence of (a) $\alpha$ value (b) variational EM

**Experimental Results.** From MOPP and the baseline methods, we aim to mine those participants who have the lowest scores. We examine these three cases for our analysis and present the results in the following paragraphs (Fig. 9, 10 and 11):

***Case 1:*** Participant who has the lowest score from the cosine similarity only
***Case 2:*** Participant who has the lowest score from LDA only
***Case 3:*** Participant who has the lowest $\theta_c$ from MOPP only

***Case 1.*** The cosine similarity measures the correlation between two vectors. We focus on the participants who have cosine similarity with the investigated project equal to 0. Figure 9 shows the extracted words from an investigated project and previous projects of a participant (*ID 16544*)[11]. The cosine similarity of this participant's TF-IDF vector and the investigated project's TF-IDF vector is zero. This participant is marked as an outlier participant by cosine similarity. However, we can observe that the words from his previous projects (*multiprocessor*, *microprocessor*, *chip*, *pipeline*, *architecture*) have a strong relation with the the word from the investigated project (i.e. *multicore*). The score from LDA is 0.23 and from MOPP is 0.047. Both of these scores are not the lowest scores in the respective models.

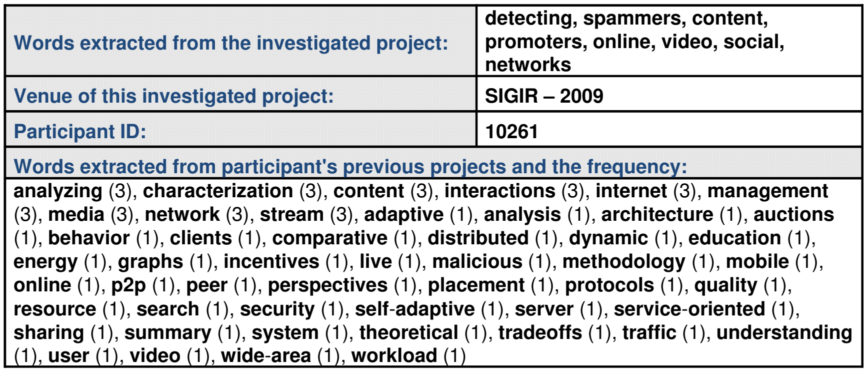
[11] We use an anonymous ID for a participant



| | |
|---|---|
| Words extracted from the investigated project: | map-reduce, machine, learning, multicore |
| Venue of this investigated project: | NIPS – 2006 |
| Participant ID: | 16544 |
| Words extracted from participant's previous projects and the frequency: | |
| **multiprocessor** (6), **caching** (4), **microprocessor** (4), **application** (3), **design** (3), **java** (3), **programs** (3), **system** (3), **transaction** (3), **chip** (2), **clustering** (2), **coherent** (2), **consistent** (2), **data** (2), **dynamic** (2), **implemented** (2), **optimal** (2), **parallel** (2), **pipeline** (2), **spaces** (2), **speculative** (2), **address** (1), **alternate** (1), **analysis** (1), **approach** (1), **architecture** (1), **associated** (1), **bandwidth** (1), **benefits** (1), **circuits** (1), **clock** (1), **concurrent** (1), **considerations** (1), **correct** (1), **efficient** (1), **embedded** (1), **environment** (1), **evaluation** (1), **exploitation** (1), **explorer** (1), **extraction** (1), **filtering** (1), **framework** (1), **hardware-software** (1), **high-performance** (1), **impact** (1), **increasing** (1), **investigation** (1), **language** (1), **memory** (1), **on-chip** (1), **performance** (1), **polymorphic** (1), **porting** (1), **primary** (1), **profiles** (1), **prototype** (1), **real-time** (1), **shared-memory** (1), **sharing** (1), **specification** (1), **support** (1), **synchronizing** (1), **testing** (1), **threading** (1), **timed** (1), **verifiably** (1), **verification** (1) | |

**Fig. 9.** Case 1: participant with the lowest score from cosine similarity only (Cos-Sim score: 0). Highlight the weakness of cosine similarity at the word level.

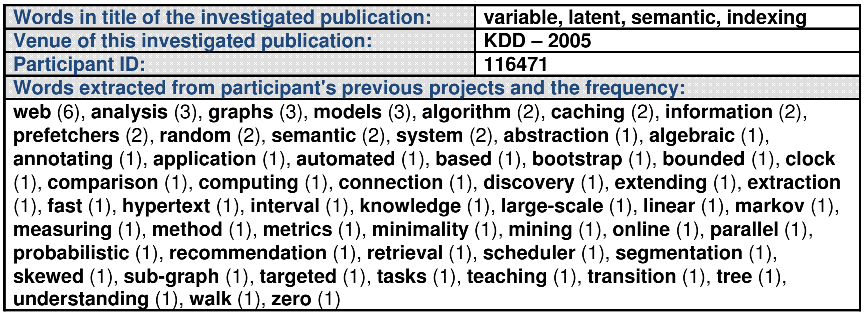
***Case 2.*** In case 2, we present a participant (*ID 10261*) which LDA marks as an outlier participant (Figure 10). However, as we can see here too that the words from his previous projects (*internet, network, malicious, online*) are related to the extracted words in the investigated project (*spammer, online, social, networks*). The scores from LDA, cosine-similarity and MOPP are 0.095, 0.097, and 0.167 respectively.



| | |
|---|---|
| Words extracted from the investigated project: | detecting, spammers, content, promoters, online, video, social, networks |
| Venue of this investigated project: | SIGIR – 2009 |
| Participant ID: | 10261 |
| Words extracted from participant's previous projects and the frequency: | |
| **analyzing** (3), **characterization** (3), **content** (3), **interactions** (3), **internet** (3), **management** (3), **media** (3), **network** (3), **stream** (3), **adaptive** (1), **analysis** (1), **architecture** (1), **auctions** (1), **behavior** (1), **clients** (1), **comparative** (1), **distributed** (1), **dynamic** (1), **education** (1), **energy** (1), **graphs** (1), **incentives** (1), **live** (1), **malicious** (1), **methodology** (1), **mobile** (1), **online** (1), **p2p** (1), **peer** (1), **perspectives** (1), **placement** (1), **protocols** (1), **quality** (1), **resource** (1), **search** (1), **security** (1), **self-adaptive** (1), **server** (1), **service-oriented** (1), **sharing** (1), **summary** (1), **system** (1), **theoretical** (1), **tradeoffs** (1), **traffic** (1), **understanding** (1), **user** (1), **video** (1), **wide-area** (1), **workload** (1) | |

**Fig. 10.** Case 2: participant with the lowest score from LDA only (LDA score: 0.095).

***Case 3.*** The last case that we consider is when MOPP gives the lowest score. Figure 11 presents the results for participant *ID 116471*. From the words extracted from this participant's previous projects, it seems that this participant focus more on web graph analysis. However, the words from the investigated project suggest that this project presents method in text analysis algorithms. This participant has a score 0.197 from LDA and 0.0828 from cosine similarity.



| | |
|---|---|
| Words in title of the investigated publication: | variable, latent, semantic, indexing |
| Venue of this investigated publication: | KDD – 2005 |
| Participant ID: | 116471 |
| Words extracted from participant's previous projects and the frequency: | |
| **web** (6), **analysis** (3), **graphs** (3), **models** (3), **algorithm** (2), **caching** (2), **information** (2), **prefetchers** (2), **random** (2), **semantic** (2), **system** (2), **abstraction** (1), **algebraic** (1), **annotating** (1), **application** (1), **automated** (1), **based** (1), **bootstrap** (1), **bounded** (1), **clock** (1), **comparison** (1), **computing** (1), **connection** (1), **discovery** (1), **extending** (1), **extraction** (1), **fast** (1), **hypertext** (1), **interval** (1), **knowledge** (1), **large-scale** (1), **linear** (1), **markov** (1), **measuring** (1), **method** (1), **metrics** (1), **minimality** (1), **mining** (1), **online** (1), **parallel** (1), **probabilistic** (1), **recommendation** (1), **retrieval** (1), **scheduler** (1), **segmentation** (1), **skewed** (1), **sub-graph** (1), **targeted** (1), **tasks** (1), **teaching** (1), **transition** (1), **tree** (1), **understanding** (1), **walk** (1), **zero** (1) | |

**Fig. 11.** Case 3: participant with the lowest score from MOPP only (MOPP score: 0).

## 5 Conclusion

In this paper, we introduce the **M**ining **O**utlier **P**articipants from **P**rojects (MOPP) model, to address the problem of scoring and ranking participant's expertise to their projects. Each participant is scored based on their project working history to the project's topic. Participants who have the lowest scores are marked as outlier participants, which means that these participants have different topic interests compare to the projects that they are working on. MOPP incorporates the structure and nature of hierarchical generative model and directional distribution, the von Mises-Fisher distribution. Experiments on semi-synthetic and synthetic data sets show that MOPP outperforms baseline methods. We also present the result from real data set extracted from DBLP. The proposed model consistently gives more meaningful and semantically correct results from the bibliographic network DBLP. For future work, we would like to extend the model to non-parametric model and compare its performance to other non-parametric topic models. We also plan to implement MOPP in different domains.

## References

1. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys **41** (2009)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research (JMLR) **3** (2003) 993–1022
3. Mardia, K.: Statistical of directional data (with discussion). Journal of the Royal Statistical Society **37**(3) (1975) 390
4. Mardia, K., Jupp, P.: Directional Statistics. John Wiley and Sons, Ltd. (2000)
5. Fisher, N., Lewis, T., Embleton, B.: Statistical Analysis of Spherical Data. Cambridge University Press (1987)
6. Zhong, S., Ghosh, J.: Generative model-based document clustering: A comparative study. Knowledge and Information Systems **8**(3) (2005) 374–384
7. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research (JMLR) **6** (2005) 1345–1382

8. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Generative model-based clustering of directional data. In: ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). (2003)
9. Banerjee, A., Ghosh, J.: Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres. Proceedings International Joint Conference on Neural Networks **15**(3) (2002) 15901595
10. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Machine Learning **42**(1) (2001) 143175
11. Reisinger, J., Waters, A., Silverthorn, B., Mooney, R.J.: Spherical topic models. In: International Conference on Machine Learning (ICML). (2010)
12. Ide, T., Kashima, H.: Eigenspace-based anomaly detection in computer systems. In: ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). (2004)
13. Fujimaki, R., Yairi, T., Machida, K.: An approach to spacecraft anomaly detection problem using kernel feature space. In: ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). (2005)
14. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In: ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). (2008)
15. Griffith, T., Steyvers, M.: Finding scientific topics. PNAS **101** (2004) 5228–5235
16. Breunig, M., Kriegel, H., Ng, R., Sander, J.: Lof: Identifying density-based local outliers. In: IEEE International Conference on Data Mining (ICDM). (2000)
17. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research (JMLR) **3** (2002) 583–617
18. Zhong, S., Ghosh, J.: Generative model-based document clustering: a comparative study. Knowledge and Information Systems **8** (2005) 374–384
19. Hu, X., Zhang, X., Lu, C., Park, E., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: ACM Conference on Knowledge Discovery and Data Mining (SIGKDD). (2009)
20. DeGroot, M.H.: Probability and Statistics. Addison-Wesley, 2nd edition (1986)
21. Xu, Z., Ke, Y., Wang, Y., Cheng, H., Cheng, J.: A model-based approach to attributed graph clustering. In: SIGMOD. (2012)