ARTICLE IN PRESS

Information Processing and Management xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

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

Field independent probabilistic model for clustering multi-field documents

Shanfeng Zhu [a,b,*], Ichigaku Takigawa [c], Jia Zeng [d], Hiroshi Mamitsuka [c]

[a] Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
[b] School of Computer Science, Fudan University, 220 Handan Road, Shanghai 200433, China
[c] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[d] Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong

ARTICLE INFO

Article history:
Received 13 May 2008
Received in revised form 28 December 2008
Accepted 21 March 2009
Available online xxxx

Keywords:
Document clustering
Finite mixture model
Multivariate Bernoulli model
Multinomial model
Field independent clustering model

ABSTRACT

We propose a new finite mixture model for clustering multiple-field documents, such as scientific literature with distinct fields: title, abstract, keywords, main text and references. This probabilistic model, which we call field independent clustering model (FICM), incorporates the distinct word distributions of each field to integrate the discriminative abilities of each field as well as to select the most suitable component probabilistic model for each field. We evaluated the performance of FICM by applying it to the problem of clustering three-field (title, abstract and MeSH) biomedical documents from TREC 2004 and 2005 Genomics tracks, and two-field (title and abstract) news reports from Reuters-21578. Experimental results showed that FICM outperformed the classical multinomial model and the multivariate Bernoulli model, being at a statistically significant level for all the three collections. These results indicate that FICM outperformed widely-used probabilistic models for document clustering by considering the characteristics of each field. We further showed that the component model, which is consistent with the nature of the corresponding field, achieved a better performance and considering the diversity of model setting also gave a further performance improvement. An extended abstract of parts of the work presented in this paper has appeared in Zhu et al. [Zhu, S., Takigawa, I., Zhang, S., & Mamitsuka, H. (2007). A probabilistic model for clustering text documents with multiple fields. In *Proceedings of the 29th European conference on information retrieval, ECIR 2007. Lecture notes in computer science (Vol. 4425, pp. 331–342)].*

1. Introduction

Document clustering uses unsupervised learning algorithms to group documents into different topics, and performs exploratory analysis on text collections (Ribeiro-Neto, 1999). Based on the structure of the final solution, clustering methods are divided into *partitional* (*flat*) clustering and *hierarchical* clustering (Jain, Murty, & Flynn, 1999). Similarity-based (discriminative) methods and model-based (generative) methods have been two major strategies for learning clusters (Zhong & Ghosh, 2003). The similarity-based methods, such as the classical *k*-means algorithm, measure the similarity between document pairs, and group similar documents into the same cluster. In document clustering, each document is usually represented by a vector of weighted selected words according to vector space model, and the similarity between two documents is calculated by Euclidean distance or the cosine of the angle between two corresponding vectors. On the other

* Corresponding author. Address: School of Computer Science, Fudan University, 220 Handan Road, Shanghai 200433, China. Tel.: +86 21 65643786; fax: +86 21 65654253.
E-mail addresses: zhushanfeng@gmail.com (S. Zhu), takigawa@kuicr.kyoto-u.ac.jp (I. Takigawa), j.zeng@ieee.org (J. Zeng), mami@kuicr.kyoto-u.ac.jp (H. Mamitsuka).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.03.005

hand, model-based methods can generate the documents in the same cluster by an identical model, which is usually specified *a priori* according to the characteristic of the data set. Without explicitly computing similarity between each document pair, the model-based methods usually have a less computational complexity compared to the similarity-based methods. Additionally they provide an intuitive explanation for each cluster through the corresponding model (Zhong & Ghosh, 2003).

Various model-based clustering algorithms have already been proposed to tackle the problem of clustering high dimensional and very sparse text documents. Multivariate Bernoulli model (BM) (McCallum & Nigam, 1998) and multinomial model (MM) (McCallum & Nigam, 1998) are the two most popular models. Zhong and Ghosh (2003) proposed a unified framework for probabilistic model-based clustering, which divides a model-based approach into, a model re-estimation step and data re-assignment step. In their recent study (Zhong & Ghosh, 2005), they compared the performance of three different model-based clustering algorithms (BM, MM and von Miser–Fisher model Banerjee, Dhillon, Ghosh, & Sra, 2003) with different strategies of assigning documents to clusters. They found that MM and von Miser–Fisher outperformed BM in clustering 15 text datasets. In addition, soft and deterministic annealing (DA) based assignments achieved better performance than hard assignment in most of datasets. Meila and Heckerman (2001) also compared soft assignment with hard assignment in multinomial model clustering text documents, and obtained similar results. Rigouste, Cappé, and Yvon (2007) have proposed multinomial mixture models to cluster documents, where each document is generated by a mixture of several MMs.

In this paper, we are particularly interested in probabilistic model-based partitional clustering algorithms. To the best of our knowledge, existing model-based algorithms treat each document as an integrated object. However, the document is usually composed of several distinct fields in reality. One typical example is the multiple-field scientific document constituted by the title, abstract, keywords, main text, and references. Although each field has a common objective of presenting the document's topic, it plays a different role, and thus has different distributions over various vocabularies. For example, the title is usually very short with around ten words summarizing the topic of the document, while the abstract is much longer with about 100–200 words briefly describing the motivation of the work, the proposed solution and the experimental result. Another typical example is the news report, which includes two important fields: title and body. Similar to the scientific document, the title in the news report summarizes the topic, and the body gives a more detailed description. Moreover, the first paragraph of the body most likely outlines what has happened in the news.

Here we propose a new probabilistic model referred to as field independent clustering model (FICM) to explicitly handle each field in a document separately. FICM is an extension of a finite mixture model to consider the distinct word distribution in each field. It then allows us to integrate all of them together. Given the cluster that a document belongs to, FICM assumes that each field of this document is conditionally independent. The generative model for each field can be different according to characteristics of data. FICM outperforms other classical model-based methods for the following reasons. First, it integrates the discriminative ability of each field by modeling each field separately. Second, we can further select the most suitable model for each field, and thus strengthen the performance of FICM. The basis of FICM is the conditional independence assumption of each field. This type of conditional independence has been widely employed and has achieved great successes in information retrieval and machine learning (Domingos & Pazzani, 1996; Lewis, 1998). For instance, the classical MM and BMs are also based on this kind of assumption, where, given the cluster the document belongs to, each word in the document is generated, being conditionally independent. We stress that the corresponding models work very well in practice, although this kind of assumption may not follow reality exactly. This is because that the conditional independence assumption may change the posterior probabilities of each cluster, but the cluster with the maximum posterior probability is often unchanged.

We conducted extensive and thorough experiments by applying FICM to clustering both scientific documents and news reports. We focused on clustering biomedical literature (Jensen, Saric, & Bork, 2006), which has been gaining more and more attention. MEDLINE (Wheeler et al., 2005) is the largest biomedical literature database for biomedical text mining. For each document (record) in MEDLINE, many distinct fields are provided, such as title, authors, institution, source, MeSH (Medical Subject Headings) and abstract, among which title, abstract and MeSH are the most informative. MeSH (Nelson, Schopen, Savage, Schulman, & Arluk, 2004) is a controlled vocabulary thesaurus defined by the National Library of Medicine for indexing documents in the MEDLINE database. It includes a set of description terms organized in a hierarchical structure. To make a reliable evaluation, we have built 100 datasets randomly generated from TREC 2004 and 2005 Genomics track, respectively, which makes use of 10-years MEDLINE records as the corpus. In addition, we also created 100 news report datasets based on Reuters-21578 to evaluate the performance of FICM.

We first compared the performance of FICM with those of BM and MM in the experiment because of their wide usage. In the simplest case, the direct extension of BM and MM by applying the same model (BM or MM) for all fields in FICM has made a significant improvement over the original models in the majority of cases, being statistically significant in some cases. For example, over the TREC 2004 Genomics data, the direct extension of MM by FICM outperformed MM in 63 out of total 100 datasets, and 15 of them were statistically significant at the 95% confidence level. From this result, we found that the significant improvement in performance comes from the integration of the discriminative ability of each field. We can then improve the performance of FICM further by assigning a suitable component model to each field. For example, over the TREC 2004 Genomics data, FICM outperformed MM in 98 out of total 100 datasets, and 78 of them were statistically significant at the 95% confidence level by assigning MM to the abstract field and BM to the title and the MeSH fields. We further investigated the component model assignment problem to achieve the best clustering performance, by exploring all eight possible combinations of component assignments in FICM for clustering TREC 2004 and TREC 2005 Genomics data.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

We propose a basic strategy for this problem:"to assign the best model to each field, keeping the diversity of each component model", and experimental results demonstrated the effectiveness of this strategy. Finally, we explore the effect of a method, which we call 'Field Weighting', where a field is weighted more by assuming that the words in the field appear more often. Experimental results showed that weighting the title field moderately further improves the clustering performance of FICM.

The remainder of the paper is organized as follows. In Section 2, we propose the FICM for clustering multiple-field documents. We also discuss the superiority and optimal setting of component model of FICM. Section 3 briefly describes the three data collections, TREC Genomics track 2004, 2005 and Reuters-21578, as well as the procedure of generating 300 test datasets from these three collections. We summarize the evaluation criteria and experimental procedures in Section 4. Section 5 presents the detailed experimental results, which demonstrates the superiority of FICM and the effectiveness of component model selection strategy. Section 6 draws conclusions and envisions future works.

2. Field independent clustering model (FICM)

In this section, we first briefly introduce the classical BM and MM in turn, and then describe the proposed FICM in detail. In particular, we discuss the superiority as well as the time and space complexities of FICM.

To fix notation, let D be a set of documents, and d be a document in D. Let Z be the set of classes (topics) of D, z be a class in Z, and K be the number of the clusters in the dataset. C is the set of fields, and c is a specific field in C, which could be a title, abstract or MeSH in this work. Let D_c be a set of documents only considering field c, and d_c be field c (e.g. title) of a document d. We denote W as the set of words appearing in D, W_c as the set of words appearing in D_c, and w as a word. Let $N_{w,d}$ be the frequency of word w appearing in document d, and N_{w,d_c} be the frequency of w appearing in d_c. Let $B_{w,d}$ be 1 if $N_{w,d} > 0$, otherwise 0. Let B_{w,d_c} be 1 if $N_{w,d_c} > 0$, otherwise 0.

2.1. Mixture model for document clustering

All three models, BM, MM and the proposed FICM, belong to one important category of generative models: mixture models. The basic assumption for generative models for document clustering is that each document d in D is independently and identically generated by the model (Duda, Hart, & Stork, 2001). That is, the probability of generating document dataset D is the product of the probability of generating every document d in D:

$$p(D) = \prod_{d \in D} p(d)$$

Moreover, the mixture model assumes that, each document is generated by a finite mixture distribution of the form $p(d) = \sum_{i=1}^{K} \pi_i p(d; \theta_i)$, where π_i is the prior probability of cluster i, K is the number of components and $p(d; \theta_i)$ is the probability of generating d in cluster i, which depends on a parameter vector θ_i. Using the notation defined above, the probability of generating document d can also be written as

$$p(d) = \sum_{z \in Z} p(z) p(d|z)$$

The specific model structure of $p(d|z)$ could be further hypothesized, for example, as a multivariate Bernoulli distribution or multinomial distribution. Then the basic goal of the mixture model is to use all documents in the dataset to estimate the model parameters, which are usually learned by maximum-likelihood (ML) estimation and the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). With these model parameters, we can easily calculate the cluster membership of each document $p(z|d)$. The differences between the BM, MM and FICM models lie in the detailed hypothesis employed to generate a document, i.e., the specific probability structure of $p(d|z)$.

2.2. Multivariate Bernoulli model (BM)

In BM, each document is represented by a binary vector, which denotes the presence or absence of each word in the document. The basic assumption is that, given the cluster that a document belongs to, the occurrence of each word in the document is assumed to be conditionally independent. For a document in class z, the probability of word w appearing in the document is $p(w|z)$, while the probability for the absence of w in the document is $1 - p(w|z)$. We then try to maximize the log-likelihood of generating the whole set of documents D. The ML estimators of this model can be obtained by the following EM algorithm, which repeats the E- and M-steps alternatively until some stopping condition is satisfied. In the M-step, we employ a Laplacian prior to avoid zero probabilities, which is a form of *maximum a posteriori* (MAP) parameter estimation (DeGroot, 1970).

Probabilistic structure:

$$L(D) = \sum_{d \in D} \log p(d) = \sum_{d \in D} \log \left(\sum_{z \in Z} p(z) p(d|z) \right) = \sum_{d \in D} \log \left(\sum_{z \in Z} \left(p(z) \prod_{w \in d} p(w|z)^{B_{w,d}} (1 - p(w|z))^{1 - B_{w,d}} \right) \right)$$

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

E-step:

$$p(z|d) \propto p(z)p(d|z) = p(z)\prod_{w\in d}\left(p(w|z)^{B_{w,d}}(1-p(w|z))^{1-B_{w,d}}\right)$$

M-step: (with Laplacian smoothing)

$$p(z) \propto \sum_{d\in D} p(z|d)$$

$$p(w|z) = \frac{1+\sum_{d\in D}p(z|d)\cdot B_{w,d}}{2+\sum_{d\in D}p(z|d)}$$

2.3. Multinomial model (MM)

In contrast to BM, MM assumes that, given the cluster that a document belongs to, every occurrence of each word in the document is assumed to be conditionally independent. Given a cluster z, it generates each word in a document independently with constraint $\sum_{w\in W}p(w|z) = 1$. The probabilistic structure of multinomial model and its corresponding E- and M-steps are shown below.

Probabilistic structure:

$$L(D) = \sum_{d\in D}\log p(d) = \sum_{d\in D}\log\left(\sum_{z\in Z}p(z)p(d|z)\right) = \sum_{d\in D}\log\left(\sum_{z\in Z}\left(p(z)\prod_{w\in d}p(w|z)^{N_{w,d}}\right)\right)$$

E-step:

$$p(z|d) \propto p(z)p(d|z) = p(z)\prod_{w\in d}p(w|z)^{N_{w,d}}$$

M-step: (with Laplacian smoothing)

$$p(z) \propto \sum_{d\in D} p(z|d)$$

$$p(w|z) = \frac{1+\sum_{d\in D}p(z|d)\cdot N_{w,d}}{|W|+\sum_{w'\in W}\sum_{d\in D}p(z|d)\cdot N_{w',d}}$$

2.4. Field independent clustering model (FICM)

In spite of different assumptions for document representation and generation, both BM and MM treat the occurrence of each word at the document level rather than the field level. Conversely, the basic idea of FICM is that, given the cluster to which a document belongs, each component field of a document is conditionally independently generated. Although in practice a document may not fully obey this rule, this kind of independence assumption has been widely used successfully in machine learning and information retrieval (Domingos & Pazzani, 1996; Lewis, 1998). In fact, this kind of assumption is also employed in the classical BM and MMs, which assume that, given the cluster to which a document belongs, each word in the document is conditionally independently generated by the underlying probabilistic models. Under the assumption that each field is generated independently, the probability of generating a document d is given as follows:

$$p(d) = \sum_{z\in Z}p(z)p(d|z) = \sum_{z\in Z}p(z)\prod_{c\in C}p(d_c|z)$$

Here $p(d_c|z)$ is the probability of generating the field c of document d given the underlying cluster z. In the simplest case of having only two clusters z_1, z_2, the likelihood-ratio LR of assigning d into z_1 rather than z_2 can be calculated by the following formula:

$$LR = \frac{p(z_1|d)}{p(z_2|d)} = \frac{p(z_1)}{p(z_2)}\cdot\frac{p(d|z_1)}{p(d|z_2)} = \frac{p(z_1)}{p(z_2)}\cdot\prod_{c\in C}\frac{p(d_c|z_1)}{p(d_c|z_2)} \quad (1)$$

Since $p(z_1)/p(z_2)$ is the ratio of the prior distribution of clusters z_1, z_2 in the dataset, LR is determined by $\prod_{c\in C}(p(d_c|z_1)/p(d_c|z_2))$, which is the product of the discriminative ability of each field. If we assume each cluster has the same prior distribution, LR will be a direct integration of the discrimination value of each field. Thus the strength of FICM relies on the integration of the clustering ability of each field, which can be further improved by choosing a good probabilistic model for each field. We can see that FICM is more like a framework, whose implementation depends on the probabilistic model used for each field. In this work, the component models are constrained into BM and MM. However, it can be easily extended to other probabilistic models. Let C_b be the set of fields modeled by the BM, and C_m be the set of fields modeled by the MM. We can derive the probabilistic structure of FICM as below, and show the E- and M-steps of the EM algorithm to estimate the parameters of this model.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Probabilistic structure:

$$L(D) = \sum_{d\in D} \log p(d) = \sum_{d\in D} \log\left(\sum_{z\in Z} p(z)p(d|z)\right) = \sum_{d\in D} \log\left(\sum_{z\in Z}\left(p(z)\prod_{c\in C_b} p(d_c|z) \times \prod_{c'\in C_m} p(d_{c'}|z)\right)\right)$$

$$= \sum_{d\in D} \log\left(\sum_{z\in Z}\left(p(z)\prod_{c\in C_b}\prod_{w\in d_c} p(w|z,c)^{B_{w,d_c}}(1-p(w|z,c))^{1-B_{w,d_c}} \times \prod_{c'\in C_m}\prod_{w\in d_{c'}} (p(w|z,c')^{N_{w,d_{c'}}})\right)\right)$$

E-step:

$$p(z|d) \propto p(z)p(d|z) = p(z)\prod_{c\in C_b} p(d_c|z) \times \prod_{c'\in C_m} p(d'_c|z)$$

$$= p(z)\prod_{c\in C_b}\prod_{w\in d_c}\left(p(w|z,c)^{B_{w,d_c}}(1-p(w|z,c))^{1-B_{w,d_c}}\right)\prod_{c'\in C_m}\prod_{w\in d_{c'}} (p(w|z,c')^{N_{w,d_{c'}}})$$

M-step: (with Laplacian smoothing)

$$p(z) \propto \sum_{d\in D} p(z|d)$$

$$p(w|z,c) = \frac{1+\sum_{d\in D} p(z|d)\cdot B_{w,d_c}}{2+\sum_{d\in D} p(z|d)} \quad \text{if } c\in C_b$$

$$p(w|z,c') = \frac{1+\sum_{d\in D} p(z|d)\cdot N_{w,d_{c'}}}{|W_{c'}|+\sum_{w'\in W}\sum_{d\in D} p(z|d)\cdot N_{w',d_{c'}}} \quad \text{if } c'\in C_m$$

In the simplest case, we can use the same probabilistic model for all fields in FICM, such as $C_b = C$ or $C_m = C$, which is called as Field Independent Clustering Bernoulli Model (FICBM) or Field Independent Clustering Multinomial Model (FICMM), respectively.

2.5. Superiority of FICM

As shown in Eq. (1), FICM assumes that each field has a certain degree of discriminative ability, and the overall clustering performance could be improved by integrating each field's discriminative information. It is analogous to ensemble learning, which combines a set of individually trained classifiers for improving the overall performance on novel examples (Polikar, 2006). Ensemble learning has been successfully applied in many classification problems, and recently has also been used for clustering problems (Ghosh, 2002). Although both FICM and ensemble learning rely on the component learner (model) to achieve a better performance, the fundamental difference is that FICM directly integrates the discriminative ability of each component in the framework of the generative model for clustering, while ensemble learning computes the final clustering result by aggregating a set of different clustering results, which are obtained by each component learner, respectively. Dietterich (2000) analyzed the success of ensemble learning from the statistical, computational and representational point of view, and Hansen and Salamon (1990) summarized that the necessary and sufficient condition for the superiority of an ensemble of classifiers over any single classifier is that these component classifiers are accurate and diverse. Similarly, we think that these two conditions for the component model also hold for FICM for achieving the better clustering performance. Each component model should accurately describe document topic in order to bring certain benefits to the final clustering result. Additionally, diversity of component models can reduce the bias caused by a single component model, because different component models may assign the same document to different clusters. Therefore, we have the basic principle of the component model design in FICM: "Assign the best model to each field, and in the meanwhile maintain the diversity of each component model".

2.6. A 'Field Weighting' extension: Weighting fields differently

To further improve the performance of FICM, we weigh (scale) each field by assuming that the words in each field appear more often. This kind of approach has also been employed by other researchers. Recently Li, Xu, and Zhang (2007) have studied the problem of clustering blog documents of the World Wide Web (WWW). Each blog document consists of three components: the title, body and comments of the blog pages. They used the k-means clustering algorithm to cluster blog documents with assigning different weights to different components. Their experimental results indicated that assigning a large weight value to the blog comments produced a better clustering solution. In their method, each blog document was still deemed as a vector, and was a weighted sum of all three vectors, representing the title, body and comments, respectively. On the other hand, FICM models keeps fields being conditionally independent, and integrates them as a whole in a probabilistic framework. In spite of significant difference, the idea of assigning different weights to different fields can be also applied in FICM.

For a field d_c and a cluster z, if the occurrence of each word in d_c is multiplied by λ_c, which can be deemed as weighting this field by λ_c, then the probability of generating this weighted field by MM will be $p(d_c|z)^{\lambda_c}$. In terms of BM, we can assume

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 1
Time and space complexities.

Model	Time	Space
BM	$O(\|Z\| \cdot \|D\| \cdot \|W\|)$	$O(\|D\| \cdot \|W\|)$
MM	$O(\|Z\| \cdot \|D\| \cdot \|W\|)$	$O(\|D\| \cdot \|W\|)$
FICM	$O(\|C\| \cdot \|Z\| \cdot \|D\| \cdot \|W\|)$	$O(\|C\| \cdot \|D\| \cdot \|W\|)$

that field d_c happens λ_c times, instead of only once, then the probability of generating these λ_c fields by BM will be $p(d_c|z)^{\lambda_c}$, too. With this approach, the probability of generating d in FICM will be as follows:

$$p(d) = \sum_{z \in Z} p(z)p(d|z) = \sum_{z \in Z} p(z) \prod_{c \in C} p(d_c|z)^{\lambda_c}$$

Then we can explore the effect of assigning different λ_c to d_c, which we call 'Field Weighting'.

2.7. Time and space complexities

Table 1 summarizes the time and space complexities of BM, MM and FICM. The most time-consuming part of BM is computing $p(z|d)$ and $p(w|z)$, which is $O(|Z| \cdot |D| \cdot |W|)$. As for the space complexities, besides the space for input dataset D, we only need to keep space for $p(z), p(z|d)$ and $p(w|z)$. The space complexity is then upper-bounded by $|D| \cdot |W| + |Z| \cdot |D| + |Z| \cdot |W|$. Normally, $|Z|$ is far smaller than $|D|$ and $|W|$, and so it can be simplified as $O(|D| \cdot |W|)$. We can easily see that MM has the same time and space complexities as BM. For FICM, the most time-consuming part is computing $p(z|d)$ and $p(w|z, c)$. Since $|W_c| \leqslant |W|$, the time complexity of FICM is upper-bounded by $O(|C| \cdot |Z| \cdot |D| \cdot |W|)$. Regarding the space complexities of FICM, we need to keep each field of the D and the space for saving $p(z), p(z|d)$, as well as $p(w|z, c)$, which is upper-bounded by $O(|C| \cdot |D| \cdot |W|)$. Although FICM has slightly higher time and space complexities than BM and MM with considering each field explicitly, it can take advantage of integrating discrimination ability of each field to improve the clustering performance.

3. Datasets

To examine the performance of the proposed FICM, we have built evaluation datasets from three collections:

(1) TREC Genomics track 2004 and 2005[1] (Hersh et al., 2004; Hersh, Cohen, Bhupatiraju, Johnson, & Hearst, 2005);
(2) Reuters-21578 news collection.[2]

The Genomics track of the TREC[3] conference provides a testbed and benchmark for comparing different information retrieval systems for biomedical documents. There are totally 4,591,008 documents (MEDLINE records from year 1994 to 2003) in the TREC Genomics track corpus for 2004 and 2005. In the 2004 track, biologists defined 50 relatively independent topics, while in the 2005 track, biologists must additionally follow a semantic template to define 50 topics. In both tracks, biologists manually assessed the relevance of retrieved records, and obtained a set of reliably relevant documents for each topic. The Reuters-21578 news collection contains 21,578 news stories which appeared on the Reuters newswire in 1987. There are totally 135 topics, and one document may belong to multiple topics.

In the Genomics track 2004 and Genomics track 2005, we extracted not only the whole text of each record, but also three distinct fields (title, abstract and MeSH) of each record. In Reuters-21578, we used the first paragraph of the news body as a distinct field, because it usually gives the summary of the news analogous to the abstract in a scientific document. Thus we extracted two distinct fields for clustering, title and abstract (the first paragraph of the news body). In all three sources, we did not consider those topics associated with only nine or fewer documents, and removed those documents with empty fields or relevant to more than one topic, and finally retained 39 topics (4400 documents) in the Genomics track 2004, 24 topics (2317 documents) in the Genomics track 2005 and 42 topics (8574 documents) in the Reuters-21578, respectively. In the Reuters-21578, the two largest topics, "earn" and "acq", have respectively 3735 and 2125 documents, which are much larger than the other 40 topics which have at most 355 relevant documents. To avoid the dominance of these two topics in the dataset, we only retained all other 40 topics (2714 documents) for evaluation.

From these filtered topics, we have built two sets of datasets based on the TREC Genomics tracks of 2004 and 2005, and one set of datasets based on the Reuters-21578, which are called as the Genomics2004, Genomics2005 and Reuters1987 collections, respectively. For a robust comparison, each collection includes 100 datasets, which were generated from the corresponding source by randomly choosing three or more (no more than 12) topics. With a specific number of topics, we built 10 different datasets. Compared with other randomly generated document datasets from MEDLINE (Yoo & Hu,

[1] http://ir.ohsu.edu/genomics/.
[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/.
[3] http://trec.nist.gov/.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 2
Summary of statistical characteristics of Genomics2004, Genomics2005 and Reuters1987. The 'min of all 100', 'mean of all 100' and 'max of all 100' refer to the minimum, average and maximum value of each feature, respectively, out of all 100 datasets in each collection.

Collection	Data	N_d	W	K	N_l	Balance	N_t	N_a	N_m	W_t	W_a	W_m
Genomics2004	T200412a	1687	4288	12	165.3	0.0281	9.1	113.9	40.4	942	3951	1230
Genomics2004	min of all 100	133	879	3	138.3	0.0216	6	93.4	33.2	97	774	251
Genomics2004	max of all 100	1960	4630	12	175.9	0.5625	9.8	120.3	48	1044	4290	1381
Genomics2004	mean of all 100	860.2	2812.3	7.5	161	0.1056	8.3	109.3	40.9	543.7	2597.1	790.1
Genomics2005	T200512a	1163	3153	12	155	0.029	8.8	106.2	38.4	662	2906	890
Genomics2005	min of all 100	71	570	3	132.8	0.0211	5.6	92.8	28.3	60	485	158
Genomics2005	max of all 100	1469	3442	12	165.1	0.6667	8.9	111.3	41.1	727	3168	950
Genomics2005	mean of all 100	690.7	2214.4	7.5	148.9	0.0711	8.0	103	35.3	400.6	2025.3	608.8
Reuters1987	R198712a	714	915	12	19	0.0338	4	14.3	–	322	858	–
Reuters1987	min of all 100	52	116	3	12	0.0282	2.3	8.6	–	2.6	107	–
Reuters1987	max of all 100	1232	1280	12	19.6	0.619	4.4	14.6	–	529	1198	–
Reuters1987	mean of all 100	500.8	622.3	7.5	18.0	0.1002	3.8	13.4	–	223.8	575	–

2006), Genomics2004 and Genomics2005 are of high quality, where the topic of each document was assessed manually by the biologists. In the pre-processing step, we removed stop words, carried out case folding and tokenized the documents using the Porter's stemming algorithm (Porter, 1980). Similar to Zhong and Ghosh (2005), we eliminated any (stemmed) word that appears in less than three documents. The same procedure was also applied to the words in each field.

Table 2 summarizes the statistical characteristics of Genomics2004, Genomics2005, and Reuters1987, where N_d is the number of documents, W is the number of distinct words (tokens), K is the number of classes, N_l is the average number of words in each document, balance is the size ratio of the smallest class to the largest class, $N_t(N_a, N_m)$ is the average number of words in the title (abstract, MeSH) field, and $W_t(W_a, W_m)$ is the number of distinct words in the title (abstract, MeSH) field. Each dataset in Genomics2004 and Genomics2005 is named by combining an initial alphabet "T", the year, the number of topics, and the order of the dataset. For example, "T200412a" represents the first dataset with twelve topics generated from the TREC Genomics track 2004. Similarly, each dataset in Reuters1987 is named by combining the initial alphabet "R", the year, the number of topics, and the order of this dataset. Table 2 shows that Genomics2004 varies greatly in some important characteristics: the number of documents varies from 133 to 1960, the number of classes from 3 to 12, the balance from 0.022 to 0.563 and the number of distinct words in each dataset from 879 to 4630. Compared with Genomics2004, Genomics2005 is slightly smaller but also has large variations. In contrast, Reuters1987 is the smallest in terms of the average length and the average number of documents in each dataset. The number of documents ranges from 52 to 1232, the number of distinct words from 116 to 1280, and the balance from 0.0282 to 0.619. Overall, the diversity of Genomics2004, Genomics2005 and Reuters1987 makes them very suitable for comparing different clustering algorithms.

4. Evaluation criteria and experiment design

In this section, we describe the criterion based on the normalized mutual information to evaluate the proposed FICM. Furthermore, we discuss how to compare the performance of different models using our new performance representation form, *S-Pair*.

4.1. Normalized mutual information (NMI)

We used external measures as evaluation criteria in document clustering. External measures evaluate the clustering result using the correct (true) class labels of the dataset, which is not provided during the clustering processes. Well-known external measures include purity, average entropy, F-measure and mutual information. Ghosh (2003) compared these external measures, and showed that mutual information is a superior measure over other external measures. Both purity and average entropy favor large number of clusters, while F-measure is biased towards coarser clusterings. Therefore, we used the normalized mutual information (NMI) to compare the performance of FICM with the other models. Normalized mutual information (NMI) is calculated by the following formula:

$$\mathrm{NMI} = \frac{I(X;Y)}{\sqrt{H(X)\cdot H(Y)}}$$

where X and Y are the predicted clusters and the correct class labels, respectively, $I(X;Y)$ is the mutual information between X and Y, and $H(X)$ and $H(Y)$ are the entropy of X and Y, respectively. In practice, Zhong and Ghosh (2003, 2005) proposed a sample estimate to calculate the NMI,

$$\mathrm{NMI} = \frac{\sum_{h,l} n_{h,l} \log\left(\frac{n\cdot n_{h,l}}{n_h n_l}\right)}{\sqrt{\left(\sum_h n_h \log\frac{n_h}{n}\right)\left(\sum_l n_l \log\frac{n_l}{n}\right)}} \tag{2}$$

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

where n is the total number of documents in the whole collection, n_h is the number of documents in class h (standard), n_l is the number of documents in cluster l (predicted), and $n_{h,l}$ is the number of documents in both class h and cluster l. The NMI value ranges from zero to one, where an NMI value of zero means that the result is equal to random partitioning, and an NMI value close to one means that the result is almost identical to the true class labeling.

4.2. Model comparison: S-Pair

In Genomics2004 and Genomics2005, each document consists of three fields, title, abstract and MeSH, while in Reuters1987, each document consists of two fields, title and abstract. When the BM is applied on the documents with title field only, abstract field only, MeSH field only and all fields, we denote these models as *B-title, B-abstract, B-mesh* and *B-whole*, respectively. When the MM is used, the corresponding models are denoted as *M-title, M-abstract, M-mesh* and *M-whole*, respectively. As described in Section 2, FICM is a framework where the specific implementation depends on the component models assigned to each field. Since we focus on two popular probabilistic models, BM and MM, there are $2^3 = 8$ possible implementations of FICM for clustering datasets in Genomics2004 and Genomics2005, and there are $2^2 = 4$ possible implementations of FICM for clustering datasets in Reuters1987. To make a distinction, each specific implementation of FICM is denoted by a combination of an initial alphabet *F*, the character '-' and a number of alphabets, where each representing a model used in this field. The character '*b*' and '*m*', represent the BM and the MM, respectively. For example, '*F-mmm*' stands for a specific implementation of FICM, FICMM, where all three fields are modeled by MM. Table 3 lists the abbreviations of some models and their corresponding meanings.

To make a fair comparison, the same stop criterion for the EM algorithm was adopted for all the models: the relative change of the maximum log-likelihood in two consecutive iteration is less than 0.001% or the number of iterations of the EM algorithms exceeds 30. In addition, the number of classes, K, was given as *a priori* parameter in all the experiments. Moreover, to reduce the possible bias caused by a random initial partition, each experiment was carried out 100 times, and the means, standard deviations and the paired t-test were used to compare different models. The comparison of two different models, θ_1 and θ_2, is carried on clustering 100 datasets of Genomics2004, Genomics2005, and Reuters1987. The number of the datasets that θ_1 outperforms θ_2 is denoted by $N_{\theta_1>\theta_2}$, in which the number of those datasets with statistically significant improvement at the 95% confidence level is denoted by $N_{\theta_1>\theta_2,+}$. Similarly, the number of datasets that θ_2 outperforms θ_1 is denoted by $N_{\theta_2>\theta_1}$, in which the number of datasets with statistically significant improvement at the 95% confidence level is denoted by $N_{\theta_2>\theta_1,+}$. Thus for comparing θ_1 with θ_2 on each collection, we can obtain a pair of two numbers $\left(+\frac{N_{\theta_1>\theta_2,+}}{N_{\theta_1>\theta_2}}, -\frac{N_{\theta_2>\theta_1,+}}{N_{\theta_2>\theta_1}}\right)$. Here we call it *Significant Pair*, for short *S-Pair*, of comparing θ_1 with θ_2. Please note $N_{\theta_1>\theta_2} + N_{\theta_2>\theta_1} = 100$ because there are 100 datasets in each collection. We can say that θ_1 outperforms θ_2 ($\theta_1 > \theta_2$) if and only if $N_{\theta_1>\theta_2} > N_{\theta_2>\theta_1}$ and $N_{\theta_1>\theta_2,+} \geqslant N_{\theta_2>\theta_1,+}$.

To show the integrated power of FICM, we divided the datasets in two types in the following manner: We first call the datasets on which FICM outperforms the classical model (BM or MM on the entire text) *superior datasets* and the other datasets are called *neutral datsets*. If a component model can also obtain a better clustering result over superior datasets than over neutral datasets, we may say that the performance of FICM can be achieved by the good performance of this component model. To check this hypothesis, we focused on two simplest implementations of FICM: FICBM and FICMM. For example, for FICBM, we used the performance of *B-whole* (BM over the entire text) as the baseline, and the power of each component model was represented by a relative measure, the ratio of the clustering performance of the component model to the baseline, which we call the *relative discriminative ability* (RDA) of this field hereafter. That is, for the title field, $\text{RDA}_{\text{title}}$ can be computed as $\frac{\text{NMI}_{B\text{-}title}}{\text{NMI}_{B\text{-}whole}}$. Similarly, for this example, we can compute $\text{RDA}_{\text{abstract}}$ and RDA_{MeSH} for the abstract field and MeSH. We can

Table 3
Model abbreviations.

Abbreviation	Meaning
BM	Multivariate Bernoulli model
MM	Multinomial model
FICM	Field independent clustering model
FICBM	Applying BM to all fields in FICM
FICMM	Applying MM to all fields in FICM
B-title	Using BM on title only
B-abstract	Using BM on abstract only
B-mesh	Using BM on MeSH only
B-whole	Using BM on whole text
M-title	Using MM on title only
M-abstract	Using MM on abstract only
M-mesh	Using MM on MeSH only
M-whole	Using MM on whole text
F-bmb	Applying BM to the first and third field, and MM to the second field in FICM
F-bbb	Applying BM to all three fields in FICM
F-mb	Applying MM to the first field, and BM to the second field in FICM

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

then compute the total RDA over all fields as $\mathrm{RDA} = \frac{\mathrm{RDA}_{\mathrm{title}} + \mathrm{RDA}_{\mathrm{abstract}} + \mathrm{RDA}_{\mathrm{MeSH}}}{3}$ for the case of three fields: title, abstract and MeSH. This is possible for FICMM by using *M-whole* as the baseline. Similarly, for the case of only two fields, title and abstract, this RDA is given as $\mathrm{RDA} = \frac{\mathrm{RDA}_{\mathrm{title}} + \mathrm{RDA}_{\mathrm{abstract}}}{2}$. We denote the RDA for superior datasets and neutral datasets by $\mathrm{RDA}^{\mathrm{sup}}$ and $\mathrm{RDA}^{\mathrm{neu}}$, respectively.

5. Experimental results

Experimental results include two parts. In the first part, we compared BM and MM with their direct extension, FICBM and FICMM. We found that FICBM (FICMM) outperformed *B-whole* (*M-whole*) consistently in all three collections, which demonstrated the effectiveness of FICM, even in the simplest configuration. In the second part, we further examined different combinations of component models in FICM, and obtained some insights into the component field configurations.

5.1. FICBM and FICMM

5.1.1. Comparison of B-title, B-abstract, B-mesh, B-whole and FICBM

We compared the performance of the BM-based clustering models. In Genomics2004 and Genomics2005, FICBM is denoted by *F-bbb* since it consists of three different fields, while in Reuters1987 with two different fields, FICBM is denoted by *F-bb*. Table 4 illustrates the performance of different models in terms of NMI in Genomics2004, Genomics2005, and Reuters1987. For each collection, it presented the result on one example dataset in the collection, as well as the average result over all 100 datasets. For the example dataset (or the average of all datasets), we used bold face to highlight the model with the highest NMI. Experimental results showed that FICBM achieved the highest average NMI (0.756, 0.723 and 0.497) in all three collections, respectively. Table 5 presented a comparison of these models by the paired *t*-test for indicating the statistical significance of improvement. The experimental results showed that FICBM outperformed *B-whole* consistently for all three collections, with S-Pair (+41/80, −0/20) in Genomics2004, S-Pair (+63/95, −1/5) in Genomics2005 and S-Pair (+73/89, −1/11) in Reuters1987.

5.1.2. Comparison of M-title, M-abstract, M-mesh, M-whole and FICMM

We compared the performance of MM-based clustering models and FICMM: *F-mmm* (or *F-mm* for Reuters1987). Table 6 presents the experimental result of these models on one example dataset and the average of all 100 datasets for Genomics2004, Genomics2005 and Reuters1987. We can see that FICMM achieved higher NMI than *M-whole* in all three collections. For example, with Genomics2004, FICMM achieved the highest NMI of 0.740, which was followed by *M-whole* of 0.736. In addition, interestingly, for Genomics2005 and Reuters1987, *M-title* (using MM on the title only) obtained the highest average NMI out of all five models, which suggests that only a few distinguished words in the title are effective for document clustering. We further compared these models in terms of the significance of improvement by paired *t*-test. Table 7 gives the results for each collection. We found that FICMM outperformed *M-whole* consistently in all three collections, with S-Pair (+15/63, −0/37) in Genomics2004, S-Pair (+12/67, −3/33) in Genomics2005 and S-Pair (+69/94, −0/6) in Reuters1987.

5.1.3. The superiority of FICBM over B-whole and FICMM over M-whole relies on the integration of discriminative ability of each component model

In the above two rounds of experiments, the direct extension of BM and MM by FICM (FICBM and FICMM) outperformed the corresponding original models (*B-whole* and *M-whole*) significantly, and the improvement of FICBM over *B-whole* was

Table 4
Performance of *B-title*, *B-abstract*, *B-mesh*, *B-whole* and *FICBM* (*F-bbb* for Genomics2004 and Genomics2005 or *F-bb* for Reuters1987) in terms of *NMI* (*mean ± standard deviation*).

Collection	Data	*B-title*	*B-abstract*	*B-mesh*	*B-whole*	*F-bbb* (*F-bb*)
Genomics2004	T200412a	.630 ± .03	.734 ± .04	.739 ± .03	.756 ± .04	**.782 ± .04**
Genomics2004	Mean of all 100	.628 ± .06	.723 ± .07	.715 ± .05	.743 ± .06	**.756 ± .06**
Genomics2005	T200512a	.671 ± .05	.697 ± .05	.634 ± .03	.706 ± .04	**.745 ± .04**
Genomics2005	Mean of all 100	.671 ± .06	.693 ± .07	.656 ± .05	.697 ± .07	**.723 ± .07**
Reuters1987	R198712a	.300 ± .07	.359 ± .06	–	.357 ± .05	**.369 ± .07**
Reuters1987	Mean of all 100	.446 ± .07	.466 ± .07	–	.460 ± .07	**.497 ± .08**

Table 5
The comparison of *B-title*, *B-abstract*, *B-mesh*, *B-whole* and *FICBM*(*F-bbb* in Genomics2004 and Genomics2005 or *F-bb* in Reuters1987) in terms of S-Pair.

Collection	*B-whole* > *B-title*	*B-whole* > *B-abstract*	*B-whole* > *B-mesh*	*F-bbb* (*F-bb*) > *B-whole*
Genomics2004	(+92/95, −2/ 5)	(+53/80, −6/20)	(+61/74, 13/26)	(+41/80, −0/20)
Genomics2005	(+53/68, −25/32)	(+22/56, −13/44)	(+75/84, −9/16)	(+63/95, −1/ 5)
Reuters1987	(+57/70, −25/30)	(+ 2/36, −11/64)	–	(+73/89, −1/11)

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 6
Performance of *M-title*, *M-abstract*, *M-mesh*, *M-whole* and *FICMM* (*F-mmm* in Genomics2004 and Genomics2005 or *F-mm* in Reuters1987) in terms of *NMI* (*mean ± standard deviation*).

Collection	Data	*M-title*	*M-abstract*	*M-mesh*	*M-whole*	*F-mmm* (*F-mm*)
Genomics2004	T200412a	.740 ± .03	.781 ± .03	.716 ± .03	.790 ± .03	**.795 ± .04**
Genomics2004	Mean of all 100	.709 ± .05	.719 ± .06	.676 ± .05	.736 ± .06	**.740 ± .06**
Genomics2005	T200512a	.755 ± .02	.754 ± .03	.603 ± .03	.758 ± .03	**.758 ± .03**
Genomics2005	Mean of all 100	**.723 ± .05**	.702 ± .06	.588 ± .05	.702 ± .06	.705 ± .06
Reuters1987	R198712a	.413 ± .04	.423 ± .03	–	.407 ± .03	**.430 ± .03**
Reuters1987	Mean of all 100	**.493 ± .05**	.482 ± .06	–	.473 ± .06	.492 ± .06

Table 7
The comparison of *M-title*, *M-abstract*, *M-mesh*, *M-whole* and *FICMM*(*F-mmm* in Genomics2004 and Genomics2005 or *F-mm* in Reuters1987) in terms of S-Pair.

Collection	*M-whole* > *M-title*	*M-whole* > *M-abstract*	*M-whole* > *M-mesh*	*F-mmm* (*F-mm*) > *M-whole*
Genomics2004	(+64/76, −15/24)	(+61/89, −0/11)	(+89/94, −1/6)	(+15/63, −0/37)
Genomics2005	(+17/26, −57/74)	(+13/50, −10/50)	(+95/97, −2/3)	(+12/67, −3/33)
Reuters1987	(+23/33, −50/67)	(+ 8/32, −35/68)	–	(+69/94, −0/ 6)

especially remarkable. As discussed in Section 4, we can compare RDA^{sup} of FICBM (and FICMM) with RDA^{neu} of that. Through this comparison, we can check if there is a significant difference between the superior and neutral by paired *t*-test. Note that we used RDA of the three field case for Genomics2004 and Genomics2005, while that of the only two field case for Reuters1987. As shown in Table 8, the upper three rows are by FICBM and the lower three rows are by FICMM. In all six cases, the RDA^{sup} was higher than RDA^{neu}. Moreover, in four of them, the difference was statistically significant at the 95 % confidence level (*p*-value less than 0.05), and one of them obtained a *p*-value of 0.081, which was close to 0.05. The only exception was the case of FICMM on Genomics2005 with a *p*-value of 0.43. The reason may be due to the dependence of different fields and the unbalanced contribution of each field in the FICMM. In contrast to BM, MM considers all occurrences of every word, which means that longer fields, such as abstracts, contributed much more than shorter fields, such as titles, in the performance of FICMM. Over all, we believe that the significant performance improvement of FICBM over BM on the entire text (*B-whole*) and FICMM over MM on the entire text (*M-whole*) comes from the integration of discriminative ability of each component model.

5.2. The combination of different component models in FICM

As discussed in Section 2, to make good use of FICM, the component model should be accurate and diverse, which requires us to assign the best model to each field and to also maintain a diversity of component models. To examine this strategy, we shall explore the effect of different combination of component models in FICM. Since Reuters1987 is too simple with only two fields and four possible combinations in FICM, we focus on exploring different combinations of component models on clustering datasets in the data collections Genomics2004 and Genomics2005, which consists of three fields and eight possible combinations. First, we identified which model is more suitable, BM or MM, for each independent field. Second, we checked if the performance of FICM could be improved by configuring a better model to one field with fixed model settings on the other two fields. Finally, we discussed the optimal configuration of FICM compared with classical BM and MM.

5.2.1. Identifying the best model for each field independently

To identify the best model for each field, it is necessary to know both the important characteristics of each field and the strengths (and weaknesses) of each model. Alternatively, we may carry out some preliminary experiments on a small training dataset to elucidate an appropriate model for each field. In Genomics2004 and Genomics2005, considering the distinct

Table 8
The comparison of relative discriminative ability (RDA) of FICBM (FICMM) between superior datasets and neutral datasets in Genomics2004, Genomics2005 and Reuters1987.

Collection	Model	RDA^{sup}(mean ± standard deviation)	RDA^{neu}(mean ± standard deviation)	*p*-Value
Genomics2004	FICBM	.944 ± .04	.914 ± .04	**6.51e−004**
Genomics2005	FICBM	.982 ± .05	.943 ± .04	**2.09e−004**
Reuters1987	FICBM	1.011 ± .11	.928 ± .11	**2.80e−003**
Genomics2004	FICMM	.971 ± .03	.947 ± .04	**4.20e−002**
Genomics2005	FICMM	.966 ± .03	.957 ± .04	0.430
Reuters1987	FICMM	1.042 ± .05	1.023 ± .05	0.081

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

features of the MeSH and title fields, we believe that the best model for MeSH and the title field would be BM and MM, with the following reasons. The MeSH terms are originally organized in a hierarchal structure for indicating the document theme. Here, both BM and MM treat them flatly without considering this hierarchical information. Some general terms, such as "human", "protein" and "genetics", would appear in the MeSH field very frequently, which actually brings little information for clustering. The situation will become even worse for MM since it favors frequent terms for clustering. Conversely, by considering binary occurrences only, BM would be more robust against this situation. On the other hand, for the title field, words that appear usually are very informative and highly related for expressing the document topic. In this situation, MM, in which the probability for generating each distinct word sums to 1, adds some constraints to those relevant informative words, and would be more appropriate than BM since the latter deals with the presence/absence of each word independently. To examine these hypotheses, we compared the performance of MM and BM on the title and MeSH fields using a paired *t*-test. As shown in Table 9, the assignment of these models is justified through the experimental results for both collections. For example, for Genomics2004, *B-mesh* outperformed *M-mesh* with S-Pair (+81/94, −5/6), and *M-title* outperformed *B-title* with S-Pair (+89/92, −5/8).

Table 9
The comparison of *B-title* with *M-title*, *B-abstract* with *M-abstract* and *B-mesh* with *M-mesh* in terms of S-Pair.

Data	*M-title* > *B-title*	*B-abstract* > *M-abstract* $K < 8$	*M-abstract* > *B-abstract* $K \geqslant 8$	*B-mesh* > *M-mesh*
Genomics2004	(+89/92, −5/ 8)	(+35/43, −4/ 7)	(+35/37, −10/13)	(+81/94, −5/ 6)
Genomics2005	(+73/85, −6/15)	(+24/29, −12/21)	(+34/40, −7/10)	(+97/97, −1/3)

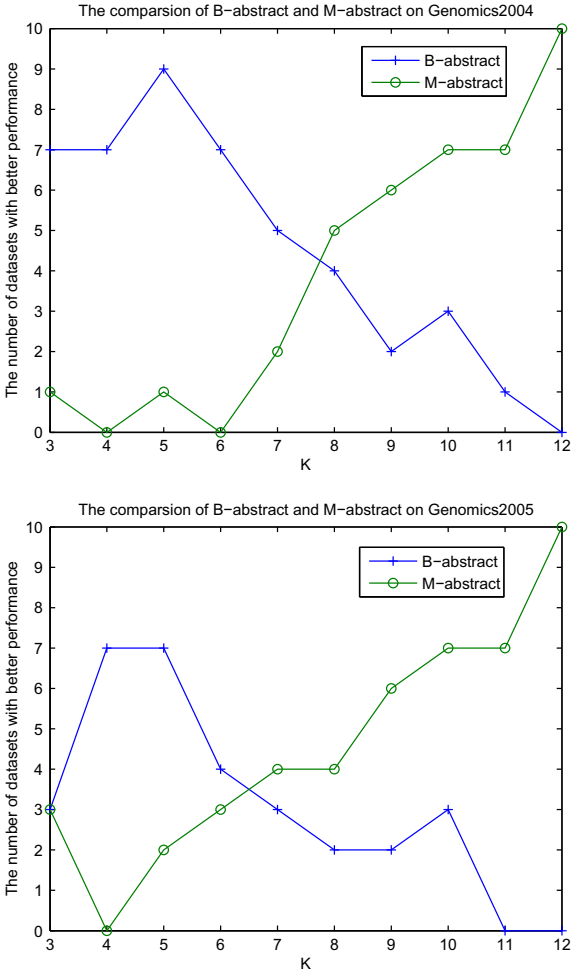

Fig. 1. Comparison of *B-abstract* with *M-abstract* for Genomics2004 and Genomics2005: the number of datasets with *K* topics where one model outperforms another statistically significant.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

For the abstract field, we speculate that BM would be more suitable on small datasets with few topics, while MM outperforms multivariate models definitely on large datasets, which comes from McCallum and Nigam'findings (1998). They compared MM with BM for naive Bayes text classification, and found that BM performed very well on datasets with a limited vocabulary, while MM usually performed better on datasets with a larger vocabulary size. This finding is also consistent with our experiments which compared *B-abstract* with *M-abstract*. As illustrated in Fig. 1, *B-abstract* usually outperformed *M-abstract* on datasets with small number of topics, while *M-abstract* outperformed *B-abstract* on datasets with more topics. For example, for both Genomics2004 and Genomics2005, *M-abstract* outperformed *B-abstract* statistically significantly for all 10 datasets with 12 topics. The crossover of *B-abstract* and *M-abstract* with similar performance happens when the number of topics in the dataset is around 8. And thus, by dividing the datasets into two groups: $K \geqslant 8$ or $K < 8$, we summarized the comparison of *B-abstract* and *M-abstract* in Table 9.

Over all, in Genomics2004 and Genomics2005, the best model for MeSH and title fields would be BM and MM, respectively. And for the abstract field, BM is more suitable than MM when the number of topics in the datasets is small ($K < 8$). Otherwise, MM is more suitable.

5.2.2. Replacing the model setting for only one field in FICM

We compared the performance of different models with paired *t*-test by changing the model for only one component field and fixing the models for the other two fields. The hypothesis is that assigning a better model to a field in FICM will improve the clustering performance.

As shown in Table 10, changing the component model from MM to BM for the MeSH field improved the clustering performance significantly for all combinations. For example, for Genomics2005, *F-mmb* outperformed *F-mmm* with S-Pair (+67/96, −0/4), and *F-bbb* outperformed *F-bbm* with S-Pair (+54/94, −2/6). Moreover, Table 11 shows consistently that replacing MM with BM for the abstract would increase clustering performance when the number of topics is less than 8. The clustering performance decreased when the number of topics is equal to or larger than 8. For example, for Genomics2004, *F-bbb* outperformed *F-bmb* with S-Pair (+34/43, −2/7) when $K < 8$ as shown in Table 11, while *F-bmb* outperformed *F-bbb* with S-Pair (+37/41, −6/9) when $K \geqslant 8$ in Table 12. These experimental results are highly consistent with our hypothesis that applying a better component model will improve the clustering performance of FICM.

Compared with the abstract and MeSH, the effect from changing the model for the title is much weaker and more complicated as illustrated in Table 13. For example, for Genomics2005, *F-mbm* differed slightly from *F-bbm* with S-Pair (+5/53, −4/47), and *F-mbb* differed slightly from *F-bbb* with S-Pair (+1/52, −1/48). This indicates that model selection for the abstract or MeSH is more important than model selection for the title with respect to the performance of FICM. A main reason may be that the title is much shorter compared with the other two fields, so that the performance of FICM is dominated by the component model setting in abstract and MeSH. Another reason would be the relative high correlation between the title and the abstract. One popular measure for correlation is the cosine of the angle between two corresponding vectors (cosine similarity). For each dataset in the Genomics2004 and Genomics2005 collections, the cosine similarities between any two fields for both MM and BM are computed, and averaged in Table 14. We displayed the highest correlated pair in every combination of data collection and model in boldface. We see that in all four cases, the title–abstract pair always obtained the largest similarity, which was especially significant in MM. For example, in Genomics2004, the average cosine similarity between the title and the abstract in MM was 0.46, while the average cosine similarity between the title and the MeSH was

Table 10
Changing model settings on the MeSH field in FICM.

Data	*F-bbb* > *F-bbm*	*F-mbb* > *F-mbm*	*F-bmb* > *F-bmm*	*F-mmb* > *F-mmm*
Genomics2004	(+30/74, −0/26)	(+27/78, −2/22)	(+38/91, −0/9)	(+47/90, −0/10)
Genomics2005	(+54/94, −2/6)	(+51/90, −0/10)	(+46/97, −0/3)	(+67/96, −0/4)

Table 11
Changing model settings on the abstract field in FICM when $K < 8$.

Data	*F-bbb* > *F-bmb*	*F-bbm* > *F-bmm*	*F-mbm* > *F-mmm*	*F-mbb* > *F-mmb*
Genomics2004	(+34/43, −2/7)	(+31/43, −3/7)	(+29/44, −2/6)	(+21/31, −9/19)
Genomics2005	(+25/33, −9/17)	(+18/30, −10/20)	(+21/30, −11/20)	(+27/42, −1/8)

Table 12
Changing model setting on the abstract field in FICM when $K \geqslant 8$.

Data	*F-bmb* > *F-bbb*	*F-bmm* > *F-bbm*	*F-mmm* > *F-mbm*	*F-mmb* > *F-mbb*
Genomics2004	(+37/41, −6/9)	(+32/41, −6/9)	(+21/34, −9/16)	(+38/44, −4/6)
Genomics2005	(+40/44, −2/6)	(+37/44, −3/6)	(+27/34, −8/16)	(+31/39, −5/11)

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 13
Changing model setting on the title field in FICM.

Data	*F-mbm* > *F-bbm*	*F-mbb* > *F-bbb*	*F-bmm* > *F-mmm*	*F-bmb* > *F-mmb*
Genomics2004	(+9/65, −4/35)	(+8/60, −1/40)	(+15/74, −0/26)	(+13/76, −3/24)
Genomics2005	(+5/53, −4/47)	(+1/52, −3/48)	(+35/82, −0/18)	(+13/68, −0/32)

Table 14
The average cosine similarity (*mean* ± *standard deviation*) between any two fields over all datasets in Genomics2004 and Genomics2005.

Data	Model	Title–abstract	Title–MeSH	Abstract–MeSH
Genomics2004	BM	**.30 ± .01**	.26 ± .02	.25 ± .01
Genomics2004	MM	**.46 ± .02**	.23 ± .03	.25 ± .01
Genomics2005	BM	**.31 ± .01**	.27 ± .03	.24 ± .02
Genomics2005	MM	**.46 ± .03**	.23 ± .03	.24 ± .02

0.23, and the average cosine similarity between the abstract and the MeSH was 0.25. In this case, assigning different models to the title and the abstract, respectively, will usually obtain good clustering results, which confirms the importance of the diversity among component models. For example, as shown in Table 13, for Genomics2004, *F-bmb* outperformed *F-mmb* with S-Pair (+13/76, −3/24), and *F-bmm* outperformed *F-mmm* with S-Pair (+15/74, −0/26), and for Genomics2005 *F-bmb* outperformed *F-mmb* with S-Pair (+13/68, −0/32), and *F-bmm* outperformed *F-mmm* with S-Pair (+35/82, −0/18).

In short, assigning a better model to a field in FICM usually improves the clustering performance, with exception on the title field, which may due to its short length and relative high correlation with the abstract.

5.2.3. Discussion on the best model configuration in FICM

As discussed in Section 2, FICM should work very well if each component model is accurate and diverse. According to this principle, we attempted to determine the best configurations of FICM for clustering datasets in Genomics2004 and Genomics2005. The basic idea is to assign the best model to each field with the constraint of keeping the diversity of each component model. Additionally, for Genomics2004 and Genomics2005, we also found that the effect of model selection on the title field was much weaker than the other two fields. Here, we first compared different configuration of FICM against the two classical models: BM (*B-whole*) and MM (*M-whole*). In this case, when compared with *B-whole*, the component setting on the abstract field in FICM is fixed to BM, and when compared with *M-whole*, the component setting on the abstract field in FICM is fixed to MM. In addition, by dividing the datasets into two groups, $K \geqslant 8$ and $K < 8$, we compared all eight possible configurations in FICM to find the best combination.

In the first case, since the model setting for abstract is already fixed, we only need to set the models for the other two fields: title and MeSH. As we discussed before, the best models for MeSH and title are BM and MM, respectively. Considering the high correlation between title and abstract, we should assign different models to each field. So, we assigned BM to MeSH, and the complement model of abstract to title. Under these settings, we conducted a comparison experiment for BM by all possible three combinations of *B-whole*, *F-bbb* and *F-mbb*. Table 15 shows S-Pairs obtained by these combinations. For MM, a similar comparison experiment was done by using *M-whole*, *F-mmm* and *F-bmb*. Table 16 presented S-Pairs obtained by all three combinations. We found that clustering performance could be improved, especially in the case of MM. For example, for Genomics2004, *F-mmm* outperformed *M-whole* slightly with S-Pair (+15/63, −0/37), while *F-bmb* outperformed both *M-whole* and *F-mmm* remarkably with S-Pairs (+78/98, −0/2) and (+58/93, −0/7), respectively. This result is totally true of Genomics2005.

In addition, by dividing the datasets into two categories, $K \geqslant 8$ and $K < 8$, we determined the best configuration of FICM. In the former case, the best model would be *F-bmb*, while in the latter case, two models, *F-mbb* and *F-bbb*, with very close performances, outperformed all other models significantly. As shown in Table 17, when $K \geqslant 8$, *F-bmb* performed slightly better than *F-mmb*, which outperformed all other models significantly for both Genomics2004 and Genomics2005. When $K < 8$, *F-mbb* outperformed all other models significantly for both Genomics2004 and Genomics2005, except *F-bbb* with a performance similar to that of *F-mbb*. The close performance of *F-bbb* and *F-mbb* reflects the weak effect of model setting in the title field. Over all, the experimental results further validate the effectiveness of our strategy of assigning the best component model to each field while keeping diversity in FICM to achieve good performance.

Table 15
The comparison of *B-whole* with *F-bbb*, *B-whole* with *F-mbb* and *F-bbb* with *F-mbb* in terms of S-Pair.

Data	*F-bbb* > *B-whole*	*F-mbb* > *B-whole*	*F-mbb* > *F-bbb*
Genomics2004	(+41/80, −0/20	(+49/86, −0/14)	(+8/60, −1/40)
Genomics2005	(+63/95, −1/5)	(+59/94, −0/6)	(+1/52, −3/48)

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 16
The comparison of *M-whole* with *F-mmm*, *M-whole* with *F-bmb* and *F-mmm* with *F-mbb* in terms of S-Pair

Data	*F-mmm* > *M-whole*	*F-bmb* > *M-whole*	*F-bmb* > *F-mmm*
Genomics2004	(+15/63, −0/37)	(+78/98, −0/2)	(+58/93, −0/7)
Genomics2005	(+12/67, −3/33)	(+69/93, −0/7)	(+72/98, −0/2)

Table 17
The comparison between *F-mbb* (*F-bbb*) with others when $K < 8$, and the comparison between *F-bmb* (*F-mmb*) with others when $K \geqslant 8$ over all datasets in Genomics2004 and Genomics2005.

Best model	Collection	*F-bbm*	*F-bmm*	*F-mbm*	*F-mmm*	*F-bbb*	*F-bmb*	*F-mbb*	*F-mmb*
F-mbb	Genomics2004	$(+\frac{18}{41}, -\frac{0}{9})$	$(+\frac{37}{48}, -\frac{0}{2})$	$(+\frac{18}{42}, -\frac{0}{8})$	$(+\frac{37}{46}, -\frac{0}{4})$	$(+\frac{1}{22}, -\frac{1}{28})$	$(+\frac{27}{42}, -\frac{1}{8})$	–	$(+\frac{27}{44}, -\frac{1}{6})$
($K < 8$)	Genomics2005	$(+\frac{25}{48}, -\frac{0}{2})$	$(+\frac{29}{44}, -\frac{2}{6})$	$(+\frac{29}{47}, -\frac{0}{3})$	$(+\frac{34}{45}, -\frac{1}{5})$	$(+\frac{1}{23}, -\frac{2}{27})$	$(+\frac{21}{31}, -\frac{9}{19})$	–	$(+\frac{23}{34}, -\frac{9}{16})$
F-bbb	Genomics2004	$(+\frac{21}{41}, -\frac{0}{9})$	$(+\frac{35}{48}, -\frac{0}{2})$	$(+\frac{20}{46}, -\frac{1}{4})$	$(+\frac{36}{48}, -\frac{1}{2})$	–	$(+\frac{34}{43}, -\frac{2}{7})$	$(+\frac{1}{28}, -\frac{1}{22})$	$(+\frac{29}{41}, -\frac{2}{9})$
($K < 8$)	Genomics2005	$(+\frac{28}{48}, -\frac{1}{2})$	$(+\frac{33}{41}, -\frac{1}{9})$	$(+\frac{32}{49}, -\frac{1}{1})$	$(+\frac{34}{43}, -\frac{1}{7})$	–	$(+\frac{25}{33}, -\frac{9}{17})$	$(+\frac{2}{27}, -\frac{1}{23})$	$(+\frac{25}{36}, -\frac{8}{14})$
F-bmb	Genomics2004	$(+\frac{41}{44}, -\frac{5}{6})$	$(+\frac{24}{49}, -\frac{0}{1})$	$(+\frac{36}{43}, -\frac{4}{7})$	$(+\frac{43}{49}, -\frac{0}{1})$	$(+\frac{37}{41}, -\frac{6}{9})$	–	$(+\frac{31}{39}, -\frac{5}{11})$	$(+\frac{8}{41}, -\frac{0}{9})$
($K \geqslant 8$)	Genomics2005	$(+\frac{46}{48}, -\frac{2}{2})$	$(+\frac{30}{50}, -\frac{0}{0})$	$(+\frac{46}{48}, -\frac{0}{2})$	$(+\frac{48}{50}, -\frac{0}{0})$	$(+\frac{40}{44}, -\frac{2}{6})$	–	$(+\frac{38}{44}, -\frac{4}{6})$	$(+\frac{9}{35}, -\frac{0}{15})$
F-mmb	Genomics2004	$(+\frac{38}{42}, -\frac{5}{8})$	$(+\frac{9}{43}, -\frac{0}{7})$	$(+\frac{32}{41}, -\frac{6}{9})$	$(+\frac{36}{47}, -\frac{0}{3})$	$(+\frac{31}{41}, -\frac{8}{9})$	$(+\frac{0}{9}, -\frac{8}{41})$	$(+\frac{28}{37}, -\frac{6}{13})$	–
($K \geqslant 8$)	Genomics2005	$(+\frac{39}{46}, -\frac{2}{7})$	$(+\frac{18}{42}, -\frac{0}{8})$	$(+\frac{40}{47}, -\frac{0}{3})$	$(+\frac{44}{50}, -\frac{0}{0})$	$(+\frac{37}{41}, -\frac{6}{9})$	$(+\frac{0}{15}, -\frac{9}{35})$	$(+\frac{34}{41}, -\frac{5}{9})$	–

5.3. The effect of field weighting: assigning different weights to different fields in FICM

To further explore the capability of FICM, we can assign different weights to different fields. Here we focus on two best configurations: *F-bmb* and *F-mbb*. We use a vector $\lambda = (\lambda_t, \lambda_a, \lambda_m)$ to represent the weights assigned to fields, where λ_t, λ_a and λ_m are the weights for the title, abstract and MeSH fields, respectively. For each field, we examined four different λ values: 3, 5, 8 and 10. The experimental results of FICM on *F-bmb* and *F-mbb* with field weighting, are shown in Tables 18 and 19, respectively. For example, in Table 18, we first present the performance of original *F-bmb* without field weighting ($\lambda = (1,1,1)$), and then the performance of *F-bmb* with field weighting and the corresponding S-Pair between the original model and the one with field weighting. We highlighted the configuration in boldface that outperformed the original model significantly. In spite of examining the effect of field weighting on two different collections, Genomics2004 and Genomics2005, and two different configurations: *F-bmb*and *F-mbb*, we were able to observe almost the same tendency from the experiment results. That is, weighting the title field moderately was able to improve the clustering performance significantly. For example, on the Genomics2005 collection, with ($\lambda = (3,1,1)$), the average NMI of *F-bmb* was improved from 0.724 to 0.735 with S-Pair (+39/80, −1/20). The highest improvement was observed when λ_t was set to 3 or 5. Overweighing the title field seems impair the clustering performance. In fact, if we set $\lambda = (10,1,1)$ in the above example, the average NMI slided from 0.735 ($\lambda = (3,1,1)$) to 0.727. On the other hand, adding more weights to the abstract and MeSH fields reduced the clustering performance in almost all cases. For example, on the Genomics2005 collection, with ($\lambda = (1,3,1)$), the average NMI of *F-bmb* was reduced from 0.724 to 0.701 with S-Pair (+0/12, −45/88). Overall, the performance of FICM can be further enhanced when we apply the Field Weighting extension to the suitable fields.

Table 18
The effect of changing weights on F-mbb in Genomics2004 and Genomics2005.

λ	Genomics2004		Genomics2005	
	NMI	S-Pair	NMI	S-Pair
(1,1,1)	.759 ± .06		.724 ± .07	
(3,1,1)	**.766 ± .06**	(+22/73, −2/27)	**.735 ± .07**	(+39/80, −1/20)
(5,1,1)	**.765 ± .06**	(+32/65, −7/35)	**.736 ± .06**	(+49/75, −5/25)
(8,1,1)	.757 ± .06	(+27/49, −23/51)	**.732 ± .06**	(+49/69, −15/31)
(10,1,1)	.750 ± .06	(+19/40, −32/60)	**.727 ± .06**	(+40/60, −20/40)
(1,3,1)	.737 ± .07	(+0/5, −61/95)	.701 ± .07	(+0/12, −45/88)
(1,5,1)	.727 ± .07	(+0/2, −74/98)	.692 ± .07	(+1/9, −66/91)
(1,8,1)	.719 ± .07	(+0/1, −85/99)	.687 ± .07	(+0/3, −72/97)
(1,10,1)	.716 ± .07	(+0/2, −86/98)	.682 ± .07	(+1/3, −79/97)
(1,1,3)	.751 ± .06	(+2/32, −24/68)	.706 ± .06	(+2/13, −48/87)
(1,1,5)	.739 ± .06	(+2/15, −48/85)	.689 ± .06	(+1/6, −74/94)
(1,1,8)	.728 ± .06	(+1/6, −66/94)	.675 ± .06	(+0/4, −88/96)
(1,1,10)	.723 ± .06	(+1/5, −76/95)	.667 ± .06	(+0/2, −92/98)

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Table 19
The effect of changing weights on F-bmb in Genomics2004 and Genomics2005.

λ	Genomics2004		Genomics2005	
	NMI	S-Pair	NMI	S-Pair
(1,1,1)	.756 ± .06		.731 ± .06	
(3,1,1)	**.764 ± .06**	(+21/76, −1/24)	**.738 ± .06**	(+21/74, −0/26)
(5,1,1)	**.763 ± .06**	(+24/63, −4/37)	**.742 ± .06**	(+32/80, −0/20)
(8,1,1)	.755 ± .06	(+18/48, −27/52)	**.740 ± .06**	(+37/71, −7/29)
(10,1,1)	.751 ± .06	(+11/39, −33/61)	**.735 ± .06**	(+26/58, −13/42)
(1,3,1)	.728 ± .06	(+0/1, −77/99)	.706 ± .06	(+0/5, −75/95)
(1,5,1)	.716 ± .06	(+0/0, −92/100)	.696 ± .06	(+0/2, −85/98)
(1,8,1)	.710 ± .06	(+0/0, −92/100)	.691 ± .06	(+0/1, −87/99)
(1,10,1)	.708 ± .06	(+0/1, −91/99)	.689 ± .06	(+0/1, −88/99)
(1,1,3)	**.761 ± .06**	(+11/61, −3/39)	.724 ± .06	(+7/32, −28/68)
(1,1,5)	.755 ± .06	(+8/43, −24/57)	.714 ± .06	(+5/20, −56/80)
(1,1,8)	.746 ± .06	(+5/31, −43/69)	.700 ± .06	(+4/13, −72/87)
(1,1,10)	.739 ± .06	(+5/23, −52/77)	.690 ± .06	(+3/9, −80/91)

6. Conclusions

We have presented a probabilistic model, FICM, for clustering multi-field text documents. The advantage of FICM comes from the integration of the discriminative ability of each field and the power of choosing the most suitable generative model for each field. In order to achieve good performance, the component models of FICM should be accurate and diverse. We have experimentally shown that a direct extension of the classical BM and MMs by FICM, FICBM and FICMM, are able to achieve a better performance, and by configuring each field with a suitable model, we obtain much better clustering results. The component model setting is practical when we have some prior knowledge on the fields, such as BM for MeSH field in our work, which can improve the performance significantly. Alternately, we can carry out some preliminary experiments on the dataset to determine the suitable model for each field. In addition, we introduced a 'Field Weighting' extension, which assigns different weights to different fields in FICM, and found that it can further improve the clustering performance significantly. Over all, we emphasize that our idea of selecting the best model for each field and of integrating them independently can be applied to other documents with multiple fields, meaning that FICM is capable of improving the performance of clustering documents in other applications.

In our experiments, the number of documents in each dataset ranges from 52 to 1960. We would like to examine the performance of FICM on some larger datasets in the future. Techniques for determining the number of topics in the dataset (Cheung, 2005) is likely to be incorporated into FICM. Finally, we hope that FICM will be applied to another clustering problem where it also has multiple components that can be modeled separately and integrated.

Acknowledgements

The authors would like to thank anonymous reviewers for their helpful comments and advice. Shanfeng Zhu and Hiroshi Mamitsuka have been supported in part by BIRD of Japan Science and Technology Agency (JST), and Shanghai Key Lab of Intelligent Information Processing, Fudan University. In addition, this project has been partly supported by Startup Fund of Fudan University, the Shanghai Committee of Science and Technology, China (Grant Nos. 08DZ2271800 and 09DZ2272800) and The State Key Lab of Bio-Organic & Natural Products Chemistry, CAS.

References

Banerjee, A., Dhillon, I., Ghosh, J., & Sra, S. (2003). Generative model-based clustering of directional data. In *The proceedings of the SIGKDD'03, Washington DC, USA, August 24–27, 2003* (pp. 19–28).

Cheung, Y. (2005). On rival penalization controlled competitive learning for clustering with automatic cluster number selection. *IEEE Transactions on Knowledge and Data Engineering, 17*(11), 1583–1588.

DeGroot, M. (1970). *Optimal statistical decisions*. McGraw-Hill.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–38.

Dietterich, T. (2000). Ensemble methods in machine learning. In *First international workshop on multiple classifier system, MCS2000, Cagliari, Italy, June 21–23, 2000. Lecture notes in computer science* (Vol. 1857, pp. 1–15).

Domingos, P., & Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Proceedings of the thirteenth international conference on machine learning (ICML'96), Bari, Italy, July 3–6, 1996* (pp. 105–112). Morgan Kaufmann.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Ghosh, J. (2002). Multiclassifier systems: Back to future. In *Third international workshop on multiple classifier system, MCS2002, Cagliari, Italy, June 24-26, 2002. Lecture notes in computer science* (Vol. 2364, pp. 1–15).

Ghosh, J. 2003. Scalable clustering methods for data mining. In Nong Ye (Ed.). *Handbook of data mining* (pp. 247–277). Mahwah, NJ: Lawrence Erlbaum Associates.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005

Hansen, L., & Salamon, P. (1990). Neural network ensembles. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 12*, 993–1001.

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., & Kraemer, D.F. (2004). Trec 2004 genomics track overview. In E. M. Voorhees, & L. P. Buckland (Eds.). *The proceedings of the thirteenth text retrieval conference (TREC 2004), Gaithersburg, Maryland, November 16–19, 2004*.

Hersh, W. R., Cohen, A., Bhupatiraju, R. T., Johnson, P., & Hearst, M. (2005). Trec 2005 genomics track overview. In E. M. Voorhees, & L. P. Buckland (Eds.). *The proceedings of the thirteenth text retrieval conference (TREC 2004), Gaithersburg, Maryland, November 15–18, 2005*.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jensen, L., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics, 7*(2), 119–129.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European conference on machine learning (ECML-98), Chemnitz, Germany, April 21–23, 1998. Lecture notes in computer science* (Vol. 1398, pp. 4–15).

Li, B., Xu, S., & Zhang, J. (2007). Enhancing clustering blog documents by utilizing author/reader comments. In *Proceedings of the 45th annual southeast regional conference, 2007, Winston-Salem, North Carolina, USA, March 23–24, 2007* (pp. 94–99). ACM.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI workshop on learning for text categorization* (pp. 41–48).

Meila, M., & Heckerman, D. (2001). An experimental comparison of model-based clustering methods. *Machine Learning, 42*(1/2), 9–29.

Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J.-L., & Arluk, N. (2004). The mesh translation maintenance system: Structure, interface design, and implementation. In M. E. A. Fieschi (Ed.), *Proceedings of the 11th world congress on medical informatics; 2004 September 7–11; San Francisco, CA* (pp. 67–69). Amsterdam: IOS Press.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuit and System Magazine*, 21–45.

Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Ribeiro-Neto, R. B.-Y. B. (1999). *Modern information retrieval*. New York: Addison Wesley.

Rigouste, L., Cappé, O., & Yvon, F. (2007). Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management, 43*(5), 1260–1280.

Wheeler, D. et al (2005). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, D39–D45.

Yoo, I., & Hu, X. (2006). A comprehensive comparison study of document clustering for a biomedical digital library medline. In G. Marchionini et al. (Eds.). *ACM/IEEE joint conference on digital libraries, JCDL 2006* (pp. 220–229).

Zhong, S., & Ghosh, J. (2003). A unified framework for model-based clustering. *Journal of Machine Learning Research, 4*, 1001–1037.

Zhong, S., & Ghosh, J. (2005). Generative model-based document clustering: A comparative study. *Knowledge and Information Systems, 8*(3), 374–384.

Please cite this article in press as: Zhu, S., et al. Field independent probabilistic model for clustering multi-field documents. *Information Processing and Management* (2009), doi:10.1016/j.ipm.2009.03.005