A Supervised Parameter Estimation Method of LDA

Liu Zhenyan[1,2,3,4], Meng Dan[3], Wang Weiping[3], and Zhang Chunxia[4]

[1] Institute of Computing Technology, Chinese Academy of Sciences, BeiJing, 100190, China
[2] University of Chinese Academy of Sciences, BeiJing, 100049, China
[3] Institute of Information Engineering, Chinese Academy of Sciences, BeiJing, 100093, China
[4] School of Software, Beijing Institute of Technology, BeiJing, 100081, China
zhenyanliu@bit.edu.cn

Abstract. Latent Dirichlet Allocation (LDA) probabilistic topic model is a very effective dimension-reduction tool which can automatically extract latent topics and dedicate to text representation in a lower-dimensional semantic topic space. But the original LDA and its most variants are unsupervised without reference to category label of the documents in the training corpus. And most of them view the terms in vocabulary as equally important, but the weight of each term is different, especially for a skewed corpus in which there are many more samples of some categories than others. As a result, we propose a supervised parameter estimation method based on category and document information which can estimate the parameters of LDA according to term weight. The comparative experiments show that the proposed method is superior for the skewed text classification, which can largely improve the recall and precision of the minority category.

Keywords: LDA, parameter estimation, Gibbs sampling, skewed text classification, term weighting.

1 Introduction

Probabilistic topic model are receiving extensive attention in text mining, information retrieval, natural language processing and so on. Latent Dirichlet Allocation (LDA) proposed by Blei et al. [1] is one of the most notable and most successful probabilistic topic models for unsupervised and supervised learning. Especially for the text classification problem, LDA is a very effective dimension-reduction tool which can automatically extract latent topics and dedicate to text representation in a lower-dimensional semantic topic space.

In text classification, LDA is commonly unsupervised because the parameters of LDA are estimated without reference to category label of the documents in the training corpus. And the terms (or words) in LDA vocabulary are viewed as equally important, but the category discriminating of each term is different, especially for a skewed corpus in which there are many more samples of some categories than others. In other words, for the skewed corpus the importance of a term not only depends on the relationship between it and all categories, but also the sample size of each category. However,

R. Cheng et al. (Eds.): APWeb 2015, LNCS 9313, pp. 401–410, 2015.
DOI: 10.1007/978-3-319-25255-1_33

LDA ignores the valuable information, that is, its two default assumptions are that both the training corpus is balanced and each terms in vocabulary is equally important. Undoubtedly this will cause suboptimal categorization peformance for the skewed text classification.

To address this shortcoming, this paper will propose a novel parameter estimation method based on category and document information which can estimate the parameters of LDA according to the weight of terms. The rest of this paper is organized as follows. Section 2 will first briefly review the related work. Section 3 will analyze LDA model and its classical Gibbs Sampling parameter estimation method. Based on the analysis in section 3, a surpervised parameter estimation method will be presented in section 4, which can especially cope with the skewed text classification problem. In section 5, we will present our comparative experiments of this new method. Finally, section 6 will give the conclusions.

2 Related Work

The LDA model is still a newcomer of topic model, which is in a relatively early stage of development up to now, and most variants of LDA focus on three research directions: parameters extension, context introduction, and orienting special task [2]. However, a few researchers pay attention to term weight in LDA. In fact, the original LDA [1] didn't mention how to build vocabulary, and in subsequent sLDA (supervised Latent Dirichlet Allocation) [3] the vocabulary was chosen by TF-IDF which computed the weight of a term using the product of the term frequency (TF) and the inverse document frequency (IDF) [4]. Madsen et al. [5] proposed Dirichlet Compound Multinomial model using TF-IDF for term weighting. Similarly, Reisinger et al. [6] also made use of TF-IDF to compute term weight in their Spherical topic model.

Moreover, Zhang et al. [7] proposed a weighted LDA model in which the weight of a term is computed based on Gauss function.Wilson et al. [8] extended the LDA model by accommodate the Pointwise Mutual Information to compute term weight.The two extended LDA model aimed to reduce the negative effect of the high frequency terms for topic distribution and incorporated term weight to parameter estimation of LDA.

However, term weight computed by the above methods can only reflect the document information, not the category information. As a result, especially for a skewed corpus, most terms chosen by them may be come from a majority category, which will tend to degrade the performance of classifier directly. But at present few scholars focus on using LDA model for dimensionality reduction of the skewed corpus.

In order to tackle the skewed text classification problem, currently one of the most popular solution pursue to improve traditional term selection method, which are Document Frequency (DF), Information Gain (IG) and so on, or traditional term weighting method, such as TFIDF. For example, Wu et al. [9] proposed a novel term selection method based on category DF, Xu et al. [10] introduced Inverse Document Frequency (ICF) into IG and so on, and Zhang et al. [11] converted TF-IDF to

TF-IDF-IG. Their experiments showed these methods can largely increase the precision and recall of the minority category. Inspired by these methods, this paper will propose a new term weighting method based on TFIDF, and accommodate it to LDA for dimensionality reduction of the skewed corpus.

3 LDA and Gibbs Sampling

For text classification, LDA is a very effective dimension-reduction tool which can automatically mine topics hidden in the documents, where each topic can be viewed as a collection of correlative words, thus each document can be represented using the latent topics.

The basic idea of LDA can be thought as be origined from Latent Semantic Indexing (LSI)[12],which uses the co-occurrences of words to capture the latent semantic associations of words and constructs a lower-dimension latent semantic feature space. This derivation of LSI aims at dimensionality reducion, however there isn't a clear conception of "topic" in LSI. The mathematical basis of LSI is linear algebra, not probability theory. So LSI is not a probabilistic topic model, but lays the foundation for probabilistic topic model.

After LSI, an alternative to LSI named probabilistic LSI (pLSI) was introduced by Hoffmann. The basic idea of pLSI is a document is a mixture of topics and a topic is a mixture of words. In pLSI the concept "topic" appreares clearly, thus pLSI is regarded as the actual origin of probabilistic topic model. However, the two parameters of pLSI —the topic distributions for each document and the word distributions for each topic — don't be treated as random variables. For this reason, pLSI is not a complete probabilistic topic model.

From Bayesian School's opinion, every parameter should be random variable and every random varialble should follow a prior distribution. So pLSI model is extended by treating the two parameters of pLSI as random variables and introducing Dirichlet prior on them. This new extended model is LDA in which the parameters of model are estimated by Bayesian method. The graphical model of LDA is depicted in Fig. 1.

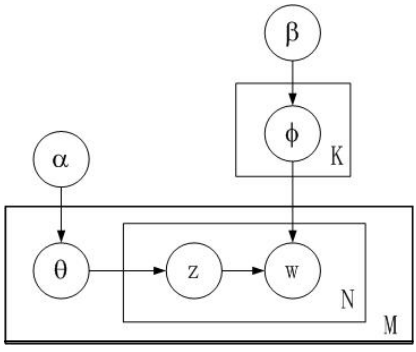

Fig. 1. Graphical model representation of LDA

Where w refers to the observed word in a document which contains N words, z refers to a latent topic, θ refers to the topic distribution for each document, ϕ refers to the word distribution for each topic, α and β are hyperparameters for Dirichlet prior distribution over both θ and ϕ respectively, M is the size of a corpus, N is the length of a document, and K is the number of latent topics in the corpus.

The generative process for a corpus under the LDA model is as follows.

1. Choose $\phi_k \sim$ Dirichlet (β), k∈[1, K]
2. For each document *m*∈[1, M]
   (a) Choose $\theta_m \sim$ Dirichlet (α)
   (b) For the nth word in document *m*, n∈[1, N_m]
   Choose a topic $z_{m,n} \sim$ Multinomial (θ_m)
   Choose a word $w_{m,n} \sim$ Multinomial ($\phi_{z_{m,n}}$)

That is, to make a new document, at first LDA chooses ϕ_k (k∈[1, K]) where $\phi_{i,j} = p(w_i|z_i=j)$ refers to the probability that the *j*th topic is sampled for the *i*th word, then for each document *m*, chooses θ_m where $\theta_i = p(z_i=j)$ refers to the probability of word w_i under topic *j*, after that, for each word in the current document, chooses a topic $z_{m,n}$, and draws a word $w_{m,n}$ from that topic $z_{m,n}$.

In such LDA model, the probability of a word w_i within a document is:

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j)\, P(z_i = j) \quad (1)$$

Furthermore, for a corpus consists of M documents and K latent topics, let $\phi = \{\phi_k\}_{k=1}^{K}$ refer to the multinomial distribution over words for each topic, and let $\theta = \{\theta_m\}_{m=1}^{M}$ refer to the multinomial distribution over topics for each document. Based on this, both ϕ and θ are the main objectives of LDA inference where ϕ represents a K×W (W is the size of the vocabulary) matrix and θ represents a M×K matrix. The parameters ϕ and θ indicate which words are important for which topic and which topics are important for a particular document, respectively.

Unfortunately, it is intractable to learn the parameters ϕ and θ directly.Instead of directly estimating them, another approach is to directly estimate the posterior distribution over z (the assignment of words to topics). A typical implement of this approach is Gibbs Sampling proposed by Griffiths et al. [13], a specific form of Markov Chain Monte Carlo (MCMC) that refers to a set of approximate iterative techniques for obtaining samples from complex distributions.

Gibbs Sampling simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables where each subset is conditioned on the value of distribution. The sampling is done sequentially and proceeds until the sampled values approximate the target distribution [14]. In Gibbs Sampling method, parameters do not be estimated directly, but be apporximated using posterior estimation of z.

Gibbs Sampler for LDA needs to compute the pobability of a topic being assigned to a word, given all other topic assignments to all other words. For an observed word w_i ($w_i \in \mathbf{w}$) in document d_i, according to Bayesian's rule, the conditional posterior distribution for $z_i = k$ (k ∈[1,K]) is given by

$$P(z_i = k|\mathbf{z}_{\neg i}, \mathbf{w}) \propto P(w_i|z_i = k, \mathbf{z}_{\neg i}, \mathbf{w}_{\neg i})P(z_i = k|\mathbf{z}_{\neg i})$$
$$= \frac{n_{\neg i,k}^{(w_i)} + \beta}{n_{\neg i,k}^{(\cdot)} + W\beta} \cdot \frac{n_{\neg i,k}^{(d_i)} + \alpha}{n_{\neg i,\cdot}^{(d_i)} + K\alpha} \quad (2)$$

Where $\mathbf{z}_{\neg i}$ means all topic assignment except z_i, $\mathbf{w}_{\neg i}$ means all words except w_i in the vocabulary. $n_{\neg i,k}^{(w_i)}$ is the number of times of word w_i assigned to topic k except the current assignment, $n_{\neg i,k}^{(\cdot)}$ is the total number of words assigned to topic k except the current assignment, $n_{\neg i,k}^{(d_i)}$ is the number of words from document d_i assigned to topic k except the current assignment, and $n_{\neg i,\cdot}^{(d_i)}$ is the number of words in document d_i, not including the current word w_i.

Gibbs Sampler starts by assigning each word to a random topic index in [1…K], and then assign a new topic index for every word during each iteration of Gibbs Sampling. After a burn-in period of a few hundred interations, Gibbs Sampler can reach its converged state and two matrices ϕ and θ are estimated from all topic assignment as follows.

$$\phi_{k,w_i} = \frac{n_{\neg i,k}^{(w_i)} + \beta}{n_{\neg i,k}^{(\cdot)} + W\beta} \qquad \theta_{d_i,k} = \frac{n_{\neg i,k}^{(d_i)} + \alpha}{n_{\neg i,\cdot}^{(d_i)} + K\alpha} \quad (3)$$

Here, we can see each term (or word) is equally important in calcuating the conditional posterior distribution for $z_i = k$. However, in the skewed text classification, the important of terms should be especially distinguished, or else which will largely degrade classification performance. That is, the traditional LDA model ignoring term weight must make many mistakes when classifying skewed documents. In order to overcome this limitation of LDA, we will propose an excellent term weighting method to compute term weight, which will be used to estimate the parameters of LDA.

4 A Supervised Parameter Estimation Method of LDA

The IDF factor of the traditional TFIDF is used to indicate the category discriminating power of a term, who believes that the fewer documents a term occurs in, the more discriminating power the term contributes to text classification. However, a term occurred in many documents from a category should be viewed as a strong feature, while a term occurred in fewer documents from some different categories should be viewed as a weak feature. The term weight computed by TFIDF can only reflect the document difference, not the category difference. As a result, TFIDF must be improved based on the category difference and the document difference.

Firstly, for the skewed corpus, the absolute category document frequency of term t, which is the number of documents from a category that have at least one occurrence of term t, cannot accurately measure its category discriminating power. For example, the document frequency of term t is 90 in a major category that contains 1000 instances, and the document frequency of term t is 90 in a minor category that contains

100 instances. Thus term t is more useful to identify this minor category. Therefore, a term that occurs in a minor category should be more valuable than in a major category in case of the same document occurrence number in each category. We will use Relative Category Document Frequency Difference (R-CDFD) to measure the difference of documents contain term t between category c_i and its complement category $\bar{c}_i$. The corresponding formula is given by

$$R\text{-}CDFD(t, c_i) = P(t|c_i) - P(t|\bar{c}_i) \tag{4}$$

Where $P(t|c_i)$ is the conditional probability of term t occurrence given category c_i, $P(t|\bar{c}_i)$ is the conditional probability of term t occurrence given category $\bar{c}_i$. $P(t|c_i) = \frac{D_t \cap D_{c_i}}{D_{c_i}}$, here D_t denotes the number of documents that have at least one occurrence of term t, D_{c_i} denotes the number of documents that belong to category c_i. $P(t|c_i) = \frac{D_t \cap D_{\bar{c}_i}}{D_{\bar{c}_i}}$, here $D_{\bar{c}_i}$ denotes the number of documents that belong to category $\bar{c}_i$.

Secondly, in order to give a higher score to a term occurred in a minor category, the category distribution should be taken into account. The lower the probability of the category contains term t, the higher weight term t will achieve. Moreover, another important factor is the relation between term and category which can be measured by the conditional probability of a category given that term t occurred. And then the higher this conditional probability is, the higher weight term t will achieve.

The above three factors, i.e., R-CDFD, the category distribution, the relation between term and category, can characterize respectively a profile of term weight, so the three factors should be integrated to compute term weight. Hence, an integrated factor named as Relative Category Difference (RCD) is constructed, which contains the above three sub-factors, and the corresponding formula is as follows.

$$\begin{aligned} RCD(t) &= \sum_{i=1}^{|C|} |RCDFD(t, c_i)| \lg \frac{1+P(c_i|t)}{P(c_i)} \\ &= \sum_{i=1}^{|C|} |P(t|c_i) - P(t|\bar{c}_i)| \lg \frac{1+P(c_i|t)}{P(c_i)} \end{aligned} \tag{5}$$

Where $|C|$ denotes the total numbers of categories in the corpus, D denotes the total number of documents in the corpus, $P(c_i|t) = \frac{D_{c_i} \cap D_t}{D_t}$ is the conditional probability of category c_i given term t occurred, and $P(c_i) = \frac{D_{c_i}}{D}$ is the probability of category c_i, here D_{C_i} denotes the number of documents that belongs to category c_i and D denotes the total number of documents in the corpus.

Then, the RCD is incorporated to replace the IDF of TFIDF. In LDA the new TF-RCD term weighting schema will be used to choose the vocabulary, and estimate parameters which replace the term frequency in Eq.3 and Eq.4 with the sum of term weight as follows.

$$\phi_{k,w_i} = \frac{wS_{\neg i,k}^{(w_i)} + \beta}{wS_{\neg i,k}^{(\cdot)} + W\beta} \qquad \theta_{d_i,k} = \frac{wS_{\neg i,k}^{(d_i)} + \alpha}{wS_{\neg i,\cdot}^{(d_i)} + K\alpha} \tag{6}$$

Where $wS_{\neg i,k}^{(w_i)}$ is the weighted sum of word w_i assigned to topic k except the current assignment, $wS_{\neg i,k}^{(\cdot)}$ is the weighted sum of words assigned to topic k except the current assignment, $wS_{\neg i,k}^{(d_i)}$ is the weighted sum of words from document d_i assigned to topic k except the current assignment, and $wS_{\neg i,\cdot}^{(d_i)}$ is the weighted sum of words in document d_i, not including the current word w_i.

5 Experiment

In order to verify the new parameter estimation method of LDA, we construct experiments focused on a comparison of TFIDF and TF-RCD in LDA. We run experiments on a subset of WebKB dataset from Ana [17], which have been pre-processed that includes tokenization and stop word removal. The experiment dataset contains 4,199 documents with four categories: "project", "course", "faculty" and "student", which is a skewed corpus that 504 documents belong to "project" category, 930 to "course", 1,124 to "faculty", and 1,641 to "student". 84% of all distinct words are observed in "student" category, 80% in "faculty", 68% in "course", and 64% in "project". Fig.2 gives the category distribution and term distribution of the WebKB dataset used in our experiment.

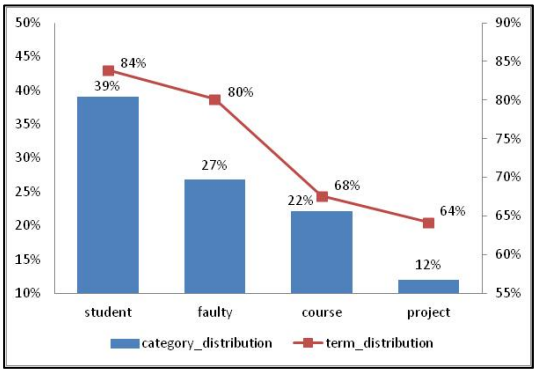

Fig. 2. Category distribution and term distribution in WebKB

On this skewed experiment dataset LDA model is trained. A 5000-term vocabulary of LDA is chosen by TFIDF or TF-RCD. And term weight is incorporated into Gibbs Sampling to assign a proper topic for the term. Then the documents are represented in latent topic space drawn by LDA. We build SVM (Support Vector Machine) classifier with LIBSVM development kit [18], in which linear kernel function is used. The reason for using SVM is that SVM has a better performance than other classification methods in text classification since it is based on the structural risk minimization principle.

Commonly the evaluation metrics for the skewed text classification are macro-averaged precision, macro-averaged recall, macro-averaged F1 [19]. Since

macro-averaged scores are averaged values over the number of categories, and then the performance of classifier is not dominated by major categories. Let P be the precision, R be recall, and *m* denotes the total number of categories, then macro-averaged precision is $\frac{1}{m}\sum_{i=1}^{m} P_i$, macro-averaged recall is $\frac{1}{m}\sum_{i=1}^{m} R_i$, macro-averaged F1 is $\frac{1}{m}\sum_{i=1}^{m} F1_i$, where F1 is $\frac{2PR}{P+R}$.

Five-fold cross-validation is performed on the experiment dataset. For this purpose, the corpus is initially partitioned into five folds. In each experiment, four fold's data are used to train while one fold's data are used to test. The average of five experiments results is reported in Table1.

Table 1. The F1 scores comparison of four schemes

	TF-RCD(P)	TF-RCD(V)	TF-IDF(P)	TF-IDF(V)
student	0.9699	0.9649	0.9650	0.9550
faculty	0.9443	0.9085	0.8695	0.8436
course	0.9396	0.9050	0.8595	0.8335
project	0.9091	0.8049	0.7647	0.7595
macro_ave_F1	0.9407	0.8958	0.8647	0.8479

In Table1, TF-RCD(P) denotes using TF-RCD in both parameter estimation and vocabulary choosing of LDA, TF-RCD(V) denotes only using TF-RCD in vocabulary choosing, TF-IDF(P) denotes using TF-IDF in both parameter estimation and vocabulary choosing , and TF-IDF (V) denotes only using TF-IDF in vocabulary choosing.

The macro-averaged F1 score of TF-RCD(V), compared with TF-IDF (V), is just improved about 3%, and then we can draw a conclusion that term weight only used to build vocabulary will make a little benefit for the performance of the skewed text classifier. But if term weight doesn't pay enough attention to minor category, though term weight is used for parameter estimation, it can't also largely improve the performance of the skewed text classifier. This conclusion can be drawn from the comparison of TF-IDF (V) and TF-IDF(P).

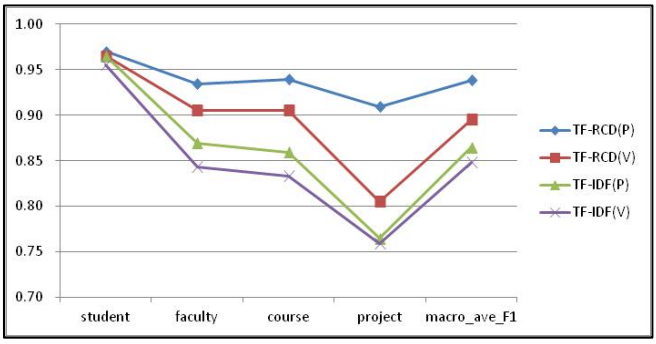

Fig. 3. The classifier performance comparison of four schemes

From Table1 we can see that the macro-averaged F1 score of TF-RCD(P), compared with TF-RCD(V), TF-IDF (P) and TF-IDF (V), is the highest and the minority categories benefit most significantly. Fig.3 gets further insights about the comparison of TF-RCD(P), TF-RCD(V), TF-IDF(P), and TF-IDF (V) with chart form. As can be seen from Fig.3, the use of TF-RCD in vocabulary choosing and parameter estimation can greatly improve the whole performance of the skewed text classifier.

6 Conclusion

TF-RCD is a superior term weighting method especially for skewed text categorization. The term weight computed by TF-RCD can not only reflect the document difference but also the category difference, while TFIDF can only reflect the document difference. The RCD of TF-RCD integrate three important factors, i.e., the relative category document frequency difference, the category distribution, the relation between term and category, can devote to measure the category discriminating power of a term.

As a result, TF-RCD can fairly choose more discriminative terms from every category to build vocabulary for LDA, and the term weights computed by TF-RCD are incorporated into parameter estimation to mine latent topics in skewed corpus. The comparative experiments show that the supervised parameter estimation method is superior for the skewed text classification, which can largely improve the recall and precision of rare category.

Acknowledgements. This work was financially supported by National Natural Science Foundation of China (61272361), also supported by Key Project of National Defense Basic Research Program of China (B11201320), National HeGaoJi Key Project (2013ZX01039 -002-001-001), National High-Tech Research and Development Program of China (2012AA011002).

References

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Machine Learning Research 3(3), 993–1022 (2003)
2. Xu, G., Wang, H.: The Development of Topic Models in Natural Language Processing. Chinese Journal of Computers 34(8), 1423–1436 (2011) (in Chinese)
3. Blei, D., McAuliffe, J.: Supervised topic models. Advances in Neural Information Processing Systems 20, 121–128 (2008)
4. Salton, G., Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval. Journal of Information Processing & Management 24(5), 513–523 (1988)
5. Madsen, R., Kauchak, D., Elkan, C.: Modeling word burstiness using the dirichlet distribution. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 545–552 (2005)
6. Reisinger, J., Waters, A., Silverthorn, B., Mooney, R.: Spherical topic models. In: Proceedings of the 27th International Conference on Machine Learning, pp. 903–910 (2010)

7. Zhang, X., Zhou, X., Huang, H., et al.: An improved LDA Topic Model. Journal of Beijing Jiaotong University 34(2), 111–114 (2010) (in Chinese)
8. Wilson, A., Chew, P.: Term weighting schemes for latent dirichlet allocation. In: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 465–473 (2010)
9. Wu, D., Zhang, Y., Yin, F., Li, M.: Feature Selection Based on Class Distritution Difference and VPRS for Text Classification. Journal of Electronics & Information Technology 29(12), 2880–2884 (2007) (in Chinese)
10. Xu, Y., Li, J., Wang, B., Sun, C., Zhang, S.: A Study of Feature Selection for Text Categorization on Imbalanced Data. Journal of Computer Research and Development 44(suppl.), 58–62 (2007) (in Chinese)
11. Zhang, A., Jing, H., Wang, B., Xu, Y.: Research on Effects of Term Weighting Factors for Text Categorization. Journal of Chinese Information Processing 24(3), 97–104 (2010) (in Chinese)
12. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990)
13. Heinrich, G.: Parameter estimation for text analysis. Technical Note Version 2.9. http://www.arbylon.net/publications/text-est2.pdf (2009)
14. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis. Erlbaum, Hillsdale (2007)
15. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: Proceedings of the 14th International Conference on Machine Learning, pp. 170–178 (1997)
16. Mladenic, D., Grobelnk, M.: Feature selection for unbalanced class distribution and Naïve Bayes. In: Proceeding of the 16th International Conference Machine Learning, pp. 258–267 (1999)
17. http://web.ist.utl.pt/~acardoso/datasets/
18. http://www.csie.ntu.edu.tw/~cjlin/
19. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2010)