IOP Conf. Series: Journal of Physics: Conf. Series **953** (2018) 012047 doi:10.1088/1742-6596/953/1/012047

Feature extraction for document text using Latent Dirichlet Allocation

P M Prihatini[1], I K Suryawan[1] and IN Mandia[2]
[1] Electrical Engineering Department, Politeknik Negeri Bali, Kampus Bukit Jimbaran, Kuta, Badung 80361 Bali Indonesia
[2] Accounting Department, Politeknik Negeri Bali, Kampus Bukit Jimbaran, Kuta, Badung 80361 Bali Indonesia

manikprihatini@pnb.ac.id

Abstract. Feature extraction is one of stages in the information retrieval system that used to extract the unique feature values of a text document. The process of feature extraction can be done by several methods, one of which is Latent Dirichlet Allocation. However, researches related to text feature extraction using Latent Dirichlet Allocation method are rarely found for Indonesian text. Therefore, through this research, a text feature extraction will be implemented for Indonesian text. The research method consists of data acquisition, text pre-processing, initialization, topic sampling and evaluation. The evaluation is done by comparing Precision, Recall and F-Measure value between Latent Dirichlet Allocation and Term Frequency Inverse Document Frequency KMeans which commonly used for feature extraction. The evaluation results show that Precision, Recall and F-Measure value of Latent Dirichlet Allocation method is higher than Term Frequency Inverse Document Frequency KMeans method. This shows that Latent Dirichlet Allocation method is able to extract features and cluster Indonesian text better than Term Frequency Inverse Document Frequency KMeans method.

1. Introduction

Information retrieval through digital text media requires several stages to produce information that meets user requirement. The first step is converting digital text from its original source into filtered text units to produce useful text units for further processing. The next stage is extracting the meaning contained within each unit of text so that the text unit can represent the feature of user information requirement. The last step is comparing the result of text unit representation to user information requirement to get the highest similarity value, so as be the most information required by user.

Each stage of information retrieval has vital function and is interdependent with one another. If text units that have been filtered at the text processing stage are not the best text units, surely the result of the representation at the extraction stage can not represent the criteria of the user information requirement. So that, the process of comparison between extraction results and user information requirement will generates some documents that are unable to satisfy the user. Moreover, information retrieval results are not only useful to this needs only, but also very useful for other needs such as automatic document clustering and summarization. Therefore, all three stages in the process of information retrieval are important to be examined.

In the previous research, the phase of digital text processing from its original source has been successfully implemented for the Indonesian text [1]. This research has processed digital news documents using stopword dictionary, root word dictionary and the rules of affixes deletion. Text

documents that have been parsed into text units are filtered using stopwords dictionary. The filtered text units have stemmed using a Porter-like Stemmer algorithm based on the rules of affixes deletion and root word dictionary. The testing process for stemming results has produced precision and recall value which are quite good but still to be optimized.

Text unit extraction can be done automatically by utilizing several methods, such as the frequency of word occurrence in documents (TF-IDF) that have been commonly used. Zhao used TF-IDF to represents each document in corpora and implemented it to graph regularized with data reconstruction for text clustering [2]. Tutkan used average TF-IDF score from each term in each document for classification [3]. Noh used TF-IDF, frequency, variance and K-Means clustering for document analysis [4]. Ceci used TF-IDF in presenting a word in the sentence for document image summarization [5].

As information retrieval technology advances, text extraction today tends to be based on topics because a unit of text is not only capable of representing only one meaning, but can represent several meanings in different contexts. One of the topic based methods is Latent Dirichlet Allocation (LDA) that was introduced by Blei [6]. LDA is very interesting to be examined with several inference methods, such as Variational Tempering [7], Stochastic Variational Inference [8], Structured Stochastic Variational Inference [9] and Gibbs Sampling [10]. Many researchers have implemented LDA for feature extraction such as in Chinese reviews [11], friend recommendation system for social networks [12], predicting risk of credit in banking [13] and interpreting public sentiment on Twitter [14].

Text unit extraction by LDA requires an initialization process to determine the initial topic for each text unit. In the classic LDA, this is done using a multinonimal random function. The difference between LDA and other topic-based methods is LDA extracting text units into three levels. The deepest level is extracting the meaning of each text unit in each document by calculating its probability and defining a new topic for the text unit. This process is done continuously until it reaches a determined threshold value or convergent. The second level is determining topic distribution of each document based on the probability and topic of the converged text unit. The result is that each document can represent several topics with their respective probabilities. The outer level is determining the distribution of topics for the corpus (set of documents) based on document distribution topic. Thus, it can be concluded that LDA is not only useful for extracting the meaning of the text unit, but at the same time being able to make soft clustering for documents based on topic.

Researches on LDA had been done previously for Indonesian text. The first research was compared the performance of LDA using Mean Variational Inference and Gibbs Sampling reasoning algorithms [15]. The test results showed the Gibbs Sampling reasoning algorithm performs better on LDA than Mean Variational Inference for Indonesian text. The second research has developed the classical Gibbs Sampling algorithm by adding the concept of fuzzy logic [16]. The test results showed that Fuzzy Gibbs Sampling algorithm performs better on LDA than Gibbs Sampling classic. In these research, the extracting of text units for Indonesian was focused on the use of LDA methods and the tests was conducted only for the LDA method itself. Testing is not done by comparing it with non-topic based methods such as TF-IDF, so it is not known whether its performance is really better to apply to Indonesian text. Therefore, through this research, performance testing is done by comparing LDA method with TF-IDF method for Indonesian text with some improvement. The first research using 100 document while this research use 500 document. The second research used the initialization value of the frequency of word occurrence while this research used multinomial random values. In addition, considering that LDA can automatically perform document clustering, in this research the TF-IDF method is implemented with K-Means to cluster document text by the number of repetitions adjusted for the repetition achieved by the LDA.

2. Research Method

This research is carried out following the flow of information retrieval process to ensure interconnection between interacting processes, as shown in figure 1.

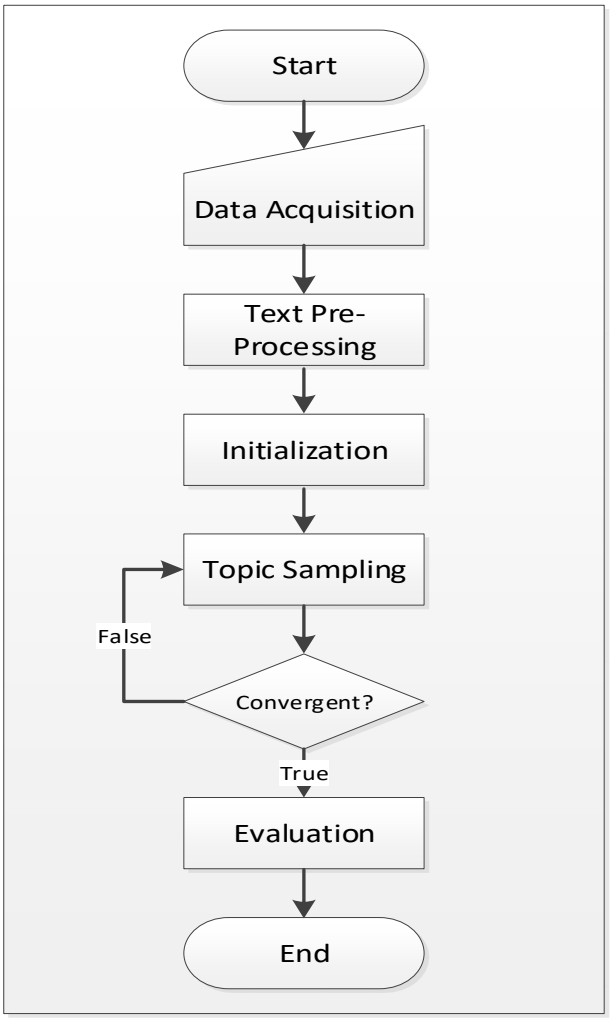

Figure 1. Research Method Flowchart.

2.1. Data Acquisition

Data acquisition stage is a process to collect data used in research. Research data in the form of news files obtained through the process of taking data online on Indonesian news media sites. Research data collected 500 news from 5 categories. Online data retrieval process is done by determining Indonesian news media sites as data source, selecting of news issuance time, selecting of news categories, copying the selected news and saving it into a text file.

2.2. Text Pre-Processing

The tokenization stage is a process to parse the text into the smallest unit such as paragraphs, sentences or words. This process is done on 500 news text files that have been collected. The tokenization process is done by parsing the text into a word, eliminating punctuation or symbols or numbers and saving it into text file.

Filtering stage is a process to eliminate meaningless words on text content based on stop words dictionary. This process is done on 500 tokenization files. The filtering process is done by matching each word with stop words dictionary. If it matches, the word is removed from the file. If it does not match, the word remains in the file. Then, saving the results into a text file.

Stemming stage is a process to get the root word by removing affixes that attaches to the word. This process is done on 500 filtering files. The stemming process is done by matching each word with

prefix rules, suffix rules, and root word dictionary. If it matches, remove the prefix or suffix on the word. Then, saving the result into a text file.

The re-filtering stage is a process to re-eliminate meaningless words on 500 stemming files. The re-filtering process is done in the same way as filtering.

2.3. Initialization

The initialization stage is a process to assign initial values and topics to text units. This process is done on 500 re-filtering files. The initialization process is done by calculating a random value for each word and assigning topic for each word with a multinomial distribution based on a random value, as in (1), where *K* is total number of topic, *n* is total number of word, *p* is a probability [17].

$$p(\vec{n}|\vec{p}) = \prod_{k=1}^{K} p_k^{n^{(k)}} = Mult(\vec{n}|\vec{p}, 1) \quad (1)$$

2.4. Topic Sampling

The topic sampling stage is the process of defining a new topic for each word in each text file. This process is done on 500 re-filtering files. The topic sampling process is done by decrementing the count of word-topic and document-topic matrix value. Then, calculating the probability value for each word, as in (2), where *nkt* is number of word *n* to topic *k*, *β* is constant parameter for topic, *nkm* is number of topic *k* for each word in a document, *α* is contanst parameter for word, *W* is total number of word in corpus, *V* is total number of unique word in corpus, *K* is total number of topic [17]. Then, determining new topic for each word with multinomial distribution based on the probability value. Then, incrementing the count of word-topic and document-topic matrix according to the new topic. All steps was repeating until it reach convergent condition, as in (3) where *N* is total number of unique word in corpus, *zi* is probability value for each word *n* [16].

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}) = \frac{n_{k,\neg i}^{(t)} + \beta}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + W\,\beta} \frac{n_{m,\neg i}^{(k)} + \alpha}{\sum_{k=1}^{K} n_{m,\neg i}^{k} + K\,\alpha} \quad (2)$$

$$\left\| \sum_{n=1}^{N} z_i - \sum_{n=1}^{N} z_{i-1} \right\| \leq \frac{\alpha}{\beta} \quad (3)$$

2.5. Evaluation

The evaluation stage is a process to test the results of LDA method implementation on Indonesian text. Evaluation was done by comparing the result of document clustering between LDA and TF-IDF K-Means method. The evaluation method used is Precision, Recall and F-Measure measurement metrics [18]. Precision *P* is calculated by dividing the number of retrieved relevant documents by the total number of retrieved documents, as in (4). Recall *R* is calculated by dividing the number of retrieved relevant documents by the total number of relevant documents, as in (5). F-Measure *F* is calculated from the mean weights of *P* and *R*, as in (6).

$$\boldsymbol{P} = \frac{\textbf{number of retrieved relevant documents}}{\textbf{total number of retrieved documents}} \quad (4)$$

$$\boldsymbol{R} = \frac{\textbf{number of retrieved relevant documents}}{\textbf{total number of relevant documents}} \quad (5)$$

$$\boldsymbol{F} = \frac{2 * \boldsymbol{P} * \boldsymbol{R}}{\boldsymbol{P} + \boldsymbol{R}} \quad (6)$$

3. Results and Discussions

3.1. Data Acquisition

The data acquisition stage generates a corpus that is a collection of 500 text files divided into 5 categories as in table 1.

Table 1. Corpus.

Category	Number of Files
News	101
Automotive	84
Sport	116
Technology	100
Business	99

3.2. Text Pre-Processing

The text pre-processing stage generates several files with the number of text units as in table 2. The data in table 2 shows that the original source text has been parsed into 306,870 text units, and then they filtered, stemmed and re-filtered, so that resulted 113,968 units of meaningful text for further processing.

Table 2. Text-Preprocessing File Results.

Step	Number of Text Units
Tokenization	306,870
Filtering	192,934
Stemming	119,022
Re-Filtering	113,968

3.3. Clustering with LDA Method

The text units that were generated at the text processing stage are extracted by the LDA method to obtain the unique feature value of each text unit. The LDA process in this research used the number of K topic is 5 topics, the parameter value of α is $50/K$ and the parameter value of β is 0.01 [17]. Convergent condition is achieved at the eighth iteration. The LDA extraction generates the highest, lowest, and average feature values as in table 3.

The LDA method also cluster 500 text files in the corpus automatically. The result of topics probability can be seen in figure 2. The x-axis on the graph shows the document number, while the y-axis shows the probability value of each document for the five given topics. The color on the graph shows the topic. From the graph, it is seen that clustering with LDA method directs distributed documents across all topics, as shown by cluster color distribution. This result is in accordance with the real condition that a document can be categorized in several topics as mentioned in the introduction.

Table 3. Result Values of LDA Method.

Description	Value of LDA
Minimum	0.000193
Maximum	0.999986
Average	0.560795

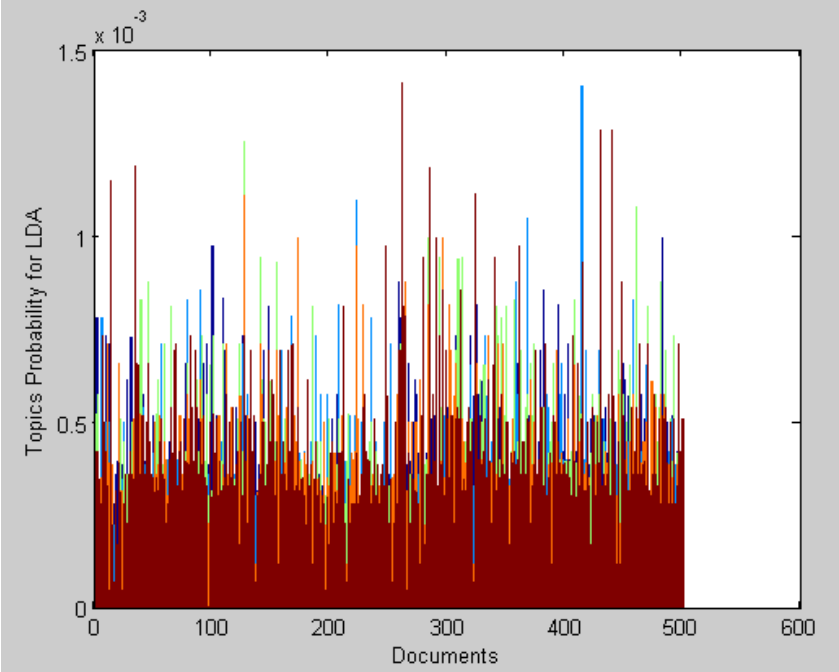

Figure 2. Topics Probability Result for LDA.

3.4. Clustering with TF IDF and K-Means Method

In this research, as comparison data for the LDA method, the text units were also extracted by the TF IDF method to obtain the unique feature value of each text unit. The TF IDF extraction generates the highest, lowest and average feature values as in table 4.

The extraction values are then used to cluster 500 text files in the corpus automatically using K-Means method with 8 repetitions according to the convergent conditions achieved by the LDA method. The result of topics probability can be seen in figure 3. The x axis on the graph shows the document number, while the y-axis shows the probability value of each document for five topics given. The color on the graph shows the topic. From the graph, it is seen that clustering with TFIDF-KMeans tends to lead a document for only one topic, as indicated by the blue, orange, and green cluster that was clear partitioned; while the other two groups are not so visible on the graph. This result has emphasized the weakness of the TFIDF method as mentioned in the introduction.

Table 4. Result Value of TF IDF Method.

Description	Value of TF IDF
Minimum	0.000008
Maximum	0.917112
Average	0.000825

3.5. Evaluation Results of Metric Measurement

Evaluation for the performance of the LDA and the TFIDF-KMeans methods in extracting feature values are calculated using P, R and F measurement metrics. The reference data used are news text file that has been grouped in the original source.

The evaluation results of the LDA methods are as in table 5. For news topic, the number of documents found is 118, while digital news sites manually classify 101 documents. This shows there is 17-news from other topics that clustered as news topic by the LDA method. Of the 118-news found, there are 90 news are relevant to manual classifying.

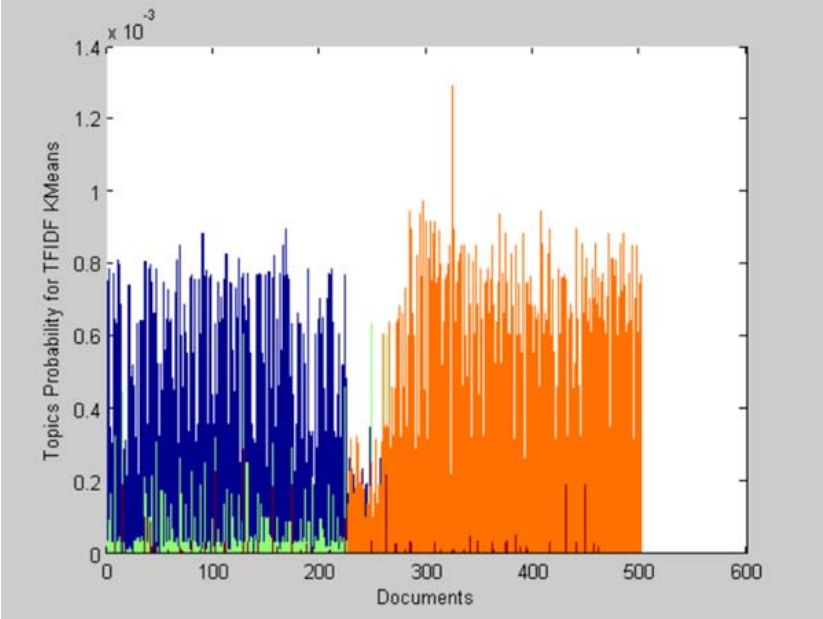

Figure 3. Topics Probability Result for TFIDF-KMeans.

For automotive topic, the number of documents found is 95, while digital news sites manually classify 84 documents. This shows there are 11 news from other topics that are clustered as automotive topic by the LDA method. Of the 95-news found, there are 77 news are relevant to manual classifying. For sport topic, the number of documents found is 115, while digital news sites manually classify 116 documents. This shows there is 1 news that are not clustered to sport topic by the LDA method. Of the 115-news found, there are 109 news are relevant to manual classifying. For technology topic, the number of documents found is 68, while digital news sites manually classify 100 documents. This shows there are 32 news that are not clustered to technology topic by the LDA method. Of the 68-news found, there are 85 news are relevant to manual classifying. This value exceeds the value of the document that was found due to the process of calculating the relevant documents found was directed to all topics. For business topic, the number of documents found is 104, while digital news sites manually classify 99 documents. This shows there are 5 news from other topics that are clustered as business topic by the LDA method. Of the 104-news found, there are 95 news are relevant to manual classifying.

Table 5. Evaluation Result for LDA.

Topic	Number of Relevant Items Retrieved	Number of Retrieved Items	Number of Relevant Items
News	90	118	101
Automotive	77	95	84
Sport	109	115	116
Technology	85	68	100
Business	95	104	99

The evaluation results of the TFIDF-KMeans methods are as in table 6. For news topic, the number of documents found is 237, while digital news sites manually classify 101 documents. This shows there are 136 news from other topics that are clustered as news topic by the TFIDF-KMeans method. Of the 237-news found, there are only 28 news are relevant to manual classifying. For automotive topic, there is no document found, while digital news sites manually classify 84 documents. This shows there are 84 news that are not clustered to automotive topic by the TFIDF-KMeans method. For sport topic, the number of documents found is 2, while digital news sites manually classify 116

documents. This shows there is 114 news that are not clustered to sport topic by the TFIDF-KMeans method. For technology topic, the number of documents found is 261, while digital news sites manually classify 100 documents. This shows there are 161 news from other topics that are clustered as technology topic by the TFIDF-KMeans method. Of the 261-news found, there are 100 news are relevant to manual classifying. For business topic, there is no document found, while digital news sites manually classify 99 documents. This shows there are 99 news that are not clustered to business topic by the TFIDF-KMeans method. The value of relevant documents retrieved for automotive, sport and business topic exceeds the value of the document that was found, because the process of calculating the relevant documents that was found directed to all topics.

Table 6. Evaluation Result for TFIDF-KMeans.

Topic	Number of Relevant Items Retrieved	Number of Retrieved Items	Number of Relevant Items
News	28	237	101
Automotive	2	0	84
Sport	114	2	116
Technology	100	261	100
Business	17	0	99

The average evaluation values of P, R and F for 500 text files of both methods are as in table 7, that were calculated based on the number of relevant documents found, the number of documents found and the number of documents relevant as described previously. The average evaluation values show the TFIDF- KMeans method has higher P than the LDA, but this is not a valid value because of greater than 1. This value was resulted because the value of relevant documents retrieved exceeds the value of the document that was found as mentioned previously. For R and F, the LDA method has higher than the TFIDF- KMeans. This is because the LDA method found the number of documents that relevant to the manual grouping by digital news site more than the TFIDF-KMeans method. So, it can conclude that the LDA method performs better than the TFIDF-KMeans method in extracting unique features of Indonesian text. In addition, the LDA method has a better ability to cluster Indonesian text than the TFIDF-KMeans method.

Table 7. Metric Evaluation Result.

Metrik	LDA	TFIDF-KMeans
Precision	0.9369	11.5003
Recall	0.9114	0.4911
F-Measure	0.9148	0.5304

4. Conclusion

The feature extraction in this research was implemented for 500 Indonesian text files using the LDA-based topics method. Evaluation was done for 5 topics, parameter value $\alpha=50/K$, parameter value $\beta=0.01$ and reached convergent condition at eighth iteration. The evaluation results show the average value of Precision, Recall and F-Measure are 0.9369, 0.9114 and 0.9148. For comparison, feature extraction is also done using TFIDF method and clustering by KMeans method with 8 iterations. The evaluation results show the average value of Precision, Recall and F-Measure are 11.5003, 0.4911 and 0.5304. From these evaluation results, it can be concluded that the LDA method performs better in extracting unique features and clustering Indonesian text files than the TFIDF-KMeans method. Therefore, the LDA method can be a better reference method for extracting features of Indonesian text.

5. **Acknowledgments**
This research was supported by grants of DIPA Politeknik Negeri Bali SP Dipa-042.01.2.401006/2017.

6. References

[1] Prihatini PM and Suryawan IK. 2016 *The 1st Int. Joint Conf. on Science and Technology (*Bali, Indonesia)
[2] Zhao Z, He X, Zhang L, Ng W and Zhuang Y Graph regularized feature selection with data reconstruction 2016 *IEEE Transactions on Knowledge and Data Engineering* **28** 689-700
[3] Tutkan M, Ganiz MC and Akyokuş S Helmholtz principle based supervised and unsupervised feature selection methods for text mining 2016 *Information Processing & Management* **52** 885-910
[4] Noh H, Jo Y and Lee S Keyword selection and processing strategy for applying text mining to patent analysis 2015 *Expert Systems with Applications* **42** 4348-60
[5] Ceci M, Loglisci C and Macchia L Ranking sentences for keyphrase extraction: a relational data mining approach 2014 *Procedia Computer Science* **38** 52-9
[6] Blei DM, Ng AY and Jordan MI Latent dirichlet allocation 2003 *Journal of Machine Learning Research* **3** 993-1022
[7] Mandt S, McInerney J, Abrol F, Ranganath R and Blei DM 2016 *19th Int. Conf. on Artificial Intelligence and Statistics* (Cadiz, Spain,JMLR: W&CP 41)
[8] Paisley J, Wang C, Blei DM and Jordan MI Nested hierarchical dirichlet processes 2015 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 256-70
[9] Hoffman MD and Blei DM 2015 *18th Int. Conf. on Artificial Intelligence and Statistics* (San Diego, CA, USA,JMLR: W&CP 38)
[10] Li Y, Zhou X, Sun Y and Zhang H Design and implementation of weibo sentiment analysis based on LDA and dependency parsing 2016 *China Communications* 91-105
[11] Xu H, Zhang F and Wang W Implicit feature identification in Chinese reviews using explicit topic mining model 2015 *Knowledge-Based Systems* **76** 166-75
[12] Wang Z, Liao J, Cao Q, Qi H and Wang Z Friendbook: a semantic-based friend recommendation system for social networks 2015 *IEEE Transactions on Mobile Computing* **14** 538-51
[13] Moro S, Cortez P and Rita P Business intelligence in banking: a literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation 2015 *Expert Systems with Applications* **42** 1314-24
[14] Tan S, Li Y, Sun H, Guan Z, Yan X, Bu J, Chen C and He X Interpreting the public sentiment variations on Twitter 2014 *IEEE Transactions on Knowledge and Data Engineering* **26** 1158-70
[15] Prihatini PM, Putra IKGD, Giriantari IAD and Sudarma M 2017 *Quality of Research (QIR) 2017* (Bali, Indonesia)
[16] Prihatini PM, Putra IKGD, Giriantari IAD and Sudarma M Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents 2017 *Contemporary Engineering Sciences* **10** 403-21
[17] Heinrich G 2008 *Parameter estimation for text analysis* (Germany: University of Leipzig)
[18] Manning CD, Raghavan P and Schutze H 2008 *An introduction to information retrieval* (England: Cambridge University Press)