The Categorization of News Articles with Informative Keywords

Taeho C. Jo
S/W business team, Samsung SDS
707-19, Yoksamdong, Kangnam Gu, Seoul, Korea
Phone: 82-2-3429-4532, FAX: 82-2-3429-3400
email: tcjo@sds.samsung.co.kr

Abstract - Text categorization means the process of assigning a category to the given document among categories predefined by users, as a function of text mining. It is necessary that the document be represented into structured form manipulated by computer for text categorization to perform this process. It is represented into a list of keywords by document indexing as structured form. Among keywords included in the document, not all reflect the contents of document. The keywords functioning grammatically, such as preposition, article, and conjunction, don't reflect its contents very little and appear in almost documents regardless of contents. Such keywords, what is called noises, are unnecessary for text categorization. These keywords should be excluded in the process of representing the document into the structured form. The others are called informative keyword and reflect its contents, after the keywords unnecessary for text categorization are removed. In advance of text categorization, two kinds of back data about keyword are necessary. One is called integrated back data and the other is categorical back data. The former provides the basis assigning substantial weight to each keyword for selecting informative keywords, and the latter does the basis assigning categorical weight to the document for categorizing it. Note that these back data are constructed as files or tables in a database. Therefore, the document is represented into a list of informative keywords based on the integrated back data, and the list is represented into a vector of categorical weights to each category in this paper. The category corresponding to the largest categorical weight will be assigned to the document.

1. Introduction

The capacity of storing data becomes enormous as the technology of computer hardware develops. So the amount of data is increasing exponentially, the information required by users becomes various. It is started the research for synthesizing the information, what is called knowledge, by analyzing the relations, patterns, and rules among the stored data, with the development of artificial intelligence techniques: neural networks, genetic algorithms, knowledge representations, and fuzzy theory [3][4]. This procedure is called data mining [1][2].

Text mining is the procedure of synthesizing the information by analyzing the relations, the patterns, and the rules among the textual data. Its functions are text summarization, text categorization, and text clustering [6][7][8][9]. Text summarization is the procedure to extract its partial content reflecting its whole contents automatically. Text categorization is the procedure of assigning a category to the text among categories predefined by users. Text clustering is the procedure of segmenting texts into several clusters, depending on the substantial relevance.

The content of this paper is restricted to text categorization. To do this, documents are classified in categories manually; assigning a category to a document by scanning one's contents. Therefore it costs very much time to categorize documents manually. It is necessary to make automatic this process of text categorization. Recently, it has been researched to make text classification automatic.

Joachims proposed the application of SVM (Super Vector Machine) to text categorization [11]. Koller and Sahami researched the hierarchical classification of texts with the minimal number of keyword [12]. Sahami, Hearst, and Saund proposed the combination of supervised learning model and unsupervised learning model to text categorization [13]. Larkey and Croft applied the combined model of k-Nearest Neighbor and Baysian classifier to the categorization of medical document [14]. Lewis, Schapire, Callan, and Papka proposed two linear learning models, Widrow-Hoff algorithm and Exponential Gradient algorithm, for text categorization [15]. Jo proposed the scheme of text categorization based on document indexing, instead of feature extraction [16]. In the scheme presented in [16], the categorical weights of the all keyword included in the document is computed and their weights are summed as the categorical weight of the given document. The category corresponding to the largest categorical weight is assigned to the given document. But not all keywords reflect the content of the text. There are two kinds of keywords. There exist the keywords reflecting the content of the document and such keywords are called informative keywords. The others are called functional keywords or noises only functioning grammatically, not reflecting its content. Such keywords are unnecessary to categorize the document. If all keywords are used to categorize it, its efficiency becomes very poor.

In this paper, document indexing is proposed as the alternative to the above scheme of representing a document into structured data. Document indexing means the process of representing a document into a list of keywords included in it. But not all keywords included in the document reflect its contents. The words functioning grammatically, such as article, conjunction, and preposition, reflect the contents of the document very little. Such keywords are unnecessary in the process of categorizing texts, and called noises. On contrary, the keywords, which reflect the content of the document strongly, are informative keywords and become very important contexts for text categorization. The process of

selecting only informative keywords from a list of keywords included in the documents is called keyword filtering. Note that back data should be constructed to assign the substantial weight to each keyword.

In the second section, the construction of the integrated back data and the categorical back data will be described. The integrated back data provides the basis to assign the substantial weight to each keyword and the categorical back data provides the basis to assign the categorical weight to each shared keyword. In the third section, the process of text categorization will be explained, and the result will be presented in the fifth section. In the sixth section, the meaning of the proposed scheme is discussed and the remaining task to improve the scheme of text categorization will be mentioned as conclusions.

2. Back Data

In this section, the construction of back data providing basis for selecting informative keywords and categorizing the documents will be described. There are two kinds of back data: integrated back data and categorical back data. The former provides the contexts to assign substantial weight, which means the degree of reflecting its contents. Informative keywords are selected based on the substantial weight. The latter does the contexts to assign categorical weights, which means the degree of representing its category. The document is categorized based on the categorical weights.

2.1. Integrated Back Data

Integrated back data is the back data providing the basis for assigning the substantial weight to each word, to select the informative keywords depending on it. The documents about all over the fields should be added almost uniformly to construct the integrated back data. The integrated back data can be stored as a file or a table in a particular database. And it includes records consisting of the fields: keyword and the accumulative number of documents including itself, among the total number of documents added to construct the integrated back data. Therefore, the integrated back data can be expressed as the following set of pairs.

$$IB = \{(k_1,\ nd_1),\ (k_2,\ nd_2),\ \ldots.\ ,\ (k_N,\ nd_N)\}$$

IB means the set represeting the integrated back data, and k_i is a keyword and nd_i is the accumulative number of documents including the keyword, k_i, in each pair belonging to the set IB. And the set, IB' is assumed to be the set containing the keywords included in the integrated back data and represented like the following this.

$$IB' = \{k_1,\ k_2,\ \ldots\ ,\ k_N\}$$

A document is represented into a set of keywords by indexing it. The elements of the set, D, are keywords included in the given document.

$$D = \{k_1',\ k_2',\ \ldots\ ,\ k_n'\}$$

For first, the initial back data is represented into the empty set of pairs like following this.

$$IB(0) = \varnothing,\ \ IB'(0) = \varnothing$$

The zero in the above parenthesis means the number of documents added to construct the integrated back data. And the integrated back data is empty initially, before any document is added.

If D is the set of keywords included in the first document added to the empty integrated back data, it is expressed like the following this.

$$IB(1) = \{(k_1',1),\ (k_2',1),\ \ldots.\ ,(k_n',1)\},$$
$$IB'(1) = D$$

And if the m documents are added, the integrated back data is assumed to be represented like the following this.

$$IB(m) = \{(k_1,\ nd_1),\ (k_2,\ nd_2),$$
$$\ldots.\ ,\ (k_N,\ nd_N)\}$$
$$IB'(m) = \{k_1,\ k_2,\ \ldots\ ,\ k_N\}$$

If a document is added and the set of keywords included in it is assumed to be the set, D, the element of the set, D, is assumed to be the keyword, k_i'. If there is the redundant keyword k_j to the keyword k_i', the pair included in the integrated back data is expressed like the following this.

$$\text{if}\ \ k_i' \in IB'(m)\ \text{and}\ k_i' = k_j,$$
$$\text{then}\ \ IB(m+1) = \{(k_1,\ nd_1),\ (k_2,\ nd_2),$$
$$\ldots.,(k_j,\ nd_j+1),\ldots.,(k_N,\ nd_N)\}$$

As expressed above, if there is the redundant keyword to the keyword, k_i' in the integrated back data, the accumulative number of documents including it is incremented by one.

On contrary, if there is no redundant keyword to k_i', the keyword is inserted in the integrated back data and the accumulative number of keywords corresponding to it is initialized by one.

$$\text{if}\ \ k_i' \notin IB'(m),$$
$$\text{then}\ IB(m+1) = \{(k_1,\ nd_1),\ (k_2,\ nd_2),$$
$$\ldots.,\ (k_N,\ nd_N),\ (k_i',\ 1)\}$$

Therefore, the integrated back data is contructed by iterating this process.

2.2. Categorical Back Data

While the integrated back data provides the basis of substantial weight as the measure of selecting informative keywords, categorical back data is the back data providing categorical back data as the measure of assigning a category to the document. To construct the integrated back data, documents are added with the regardless of their own categories. But the category of documents added to construct categorical back data is homogeneous. The number of the integrated back data is only one, while the number of the categorical back data is the number of the predefined categories.

The set of documents belonging to the category, a, is expressed into the following set, C_a.

$$C_a = \{D_1,\ D_2,\ \ldots\ ,\ D_m\}$$

The categorical back data corresponding to the category a, contains tuples, which includes the three attributes: keyword, the accumulative frequency, and the accumulative number of documents including itself. Therefore, it is expressed into the set of tuples, which includes the three attributes, like the following this.

$$CB_a = \{(k_{a1},\ nf_{a1}, nd_{a1}),\ (k_{a2},\ nf_{a2}, nd_{a2}),\ \ldots\ ,\ (k_{aN}, nf_{aN},\ nd_{aN})\}$$

And the set of keywords included in the categorical back data corresponding to the category a is expressed like the following set, CB'_a.

$$CB'_a = \{k_{a1},\ k_{a2},\ \ldots\ ,\ k_{aN}\}$$

Like the case of the integrated back data mentioned in the previous subsection, the categorical back data corresponding to the category, a is expressed into the empty sets to the both sets, CB_a and CB'_a, before any document is added.

$$CB_a(0) = \varnothing,\ \ \mathrm{C}B'_a(0) = \varnothing$$

The document belonging to the category a is expressed into the following set, D_h including the tuples, which includes the two attributes: keyword and its frequency within the document, D_h. Another set, D'_h is the set of keywords included in the given document.

$$D_h = \{(k'_{h1}, f_{h1}),\ (k'_{h2}, f_{h2}),\ \ldots\ ,\ (k'_{hn}, f_{hn})\}$$

$$D'_h = \{k'_{h1}, k'_{h2},\ \ldots\ ,\ k'_{hn}\}$$

In the above set, D_h, k'_{hi} means a keyword and f_{hi} does the frequency of the keyword k'_{hi} within the document.

The categorical back data corresponding to the category, a is expressed like the following this, when the first document is added.

$$CB_a(1) = \{(k'_{h1}, f_{h1}, 1), (k'_{h2}, f_{h2}, 1),\ \ldots\ , (k'_{hn}, f_{hn}, 1)\}$$

$$\mathrm{C}B'_a(1) = D'_h$$

It is assumed that the categorical back data corresponding to the category, a is expressed like the following this, when t documents are added.

$$CB_a(t) = \{(k_{a1}, nf_{a1}, nd_{a1}),\ (k_{a2}, nf_{a2}, nd_{a2}),\ \ldots\ ,\ (k_{aN}, nf_{aN}, nd_{aN})\}$$

$$CB'_a(t) = \{k_{a1},\ k_{a2},\ \ldots\ ,\ k_{aN}\}$$

If a document is added and there is the keyword k_{aj} in the categorical back data, redundant to k'_{hi} inlcuded in the document, it is expressed like the following this.

if $D_h \in C_a$ and $k'_{hi} \in CB'_a(t)$
and $k'_{hi} = k_{aj}$ then

$$CB_a(t+1) = \{(k_{a1}, nf_{a1}, nd_{a1}),\ (k_{a2}, nf_{a2}, nd_{a2}),\ \ldots, (k_{aj}, nf_{aj} + f_{hi}, nd_{a2} + 1),\ \ldots,\ (k_{aN}, nf_{aN}, nd_{aN})\}$$

In other words, if there is the keyword k_{aj} in the categorical back data corresponding to the category a, redundant to the keyword k'_{hi} included in the given document, its accumulative frequency incremented with its frequency within the document and the accumulative number of document including it is incremented with one.

If there is no redundant keyword in the categorical back data, it is expressed like the following this.

if $D_h \in C_a$ and $k'_{hi} \notin CB'_a(t)$, then

$$CB_a(t+1) = \{(k_{a1}, nf_{a1}, nd_{a1}),\ (k_{a2}, nf_{a2}, nd_{a2}),\ \ldots,\ (k_{aN}, nf_{aN}, nd_{aN}),\ (k'_{hi}, f_{hi}, 1)\}$$

3. Text Categorization

In this section, the process of text categorization will be described. Like the figure 1, a particular document is indexed to a list of keyword [17]. Through the integrated back data, informative keywords are selected from the list, and a category is assigned to the document based on the categorical back data.

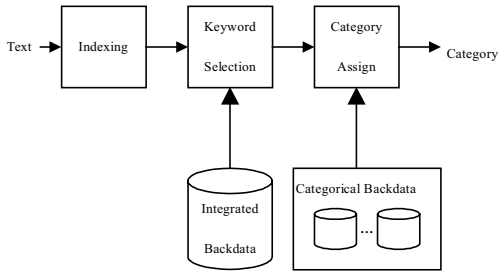

Figure 1. The Process of Text Categorization

In the above, a particular text is represented into a list of keywords by the process of document indexing. The substantial weight, the degree of reflecting the content of the text, is assigned to each keyword based on the integrated back data. These keywords are sorted in the descending order of the substantial weight. The keywords within the rank of their substantial weight are selected as informative keywords. The rank, the measurement of selecting informative keywords, is given as a parameter. Therefore, the text is represented into the list of informative keywords excluding functional keywords. To each informative keyword, categorical weight is assigned. The categorical weight to the text is computed by summing the categorical weight of the informative keywords. The category to the largest categorical weight is assigned to the text.

4. Experiment & Result

In this experiment, there are two groups of documents; one includes news articles in November 1998, the other includes news articles in January 1999. The results of text categorization with ranking keyword filtering are like the table 1.

Table 1. The Result with Ranking Keyword Filtering

	within the rank of 20	within the rank of 40
Nov 1998	76.7%	93.3%
Jan 1999	70.0%	83.3%

The results of that with critical keyword filtering are like the table 2.

Table 2. The Result with Critical Keyword Filtering

	over 1.0	over 0.5	over 0.1
Nov 1998	30.0%	70.0%	96.7%
Jan 1999	27.6%	73.3%	96.7%

The results of that with proportional keyword filtering are like the table 3.

Table 3. The Result with Proportional Keyword Filtering

	within the 10% of the whole	within the 30% of the whole
Nov 1998	80.0%	96.7%
Jan 1999	70.0%	96.7%

5. Conclusion

In this paper, it is proposed that the text be categorized with the informative keywords by removing the functional keywords, what is called noises. The schemes for selecting informative keywords are ranking selection, critical selection, and proportional selection. Through the experiment in the section 6, proportional selection of informative keywords presents the best performance in text categorization in the schemes proposed in this paper. In the proposed scheme, if the criterion of selecting informative keywords is very strict, the number of informative keywords is too small to classify the given text. The robustness of text categorization becomes very poor. On contrary, if the criterion of selecting informative keywords is very loose, the number of the selected informative keywords is so large that the efficiency of text categorization is poor. In other words, it costs unnecessary time to categorize the given text because even the functional keywords may be included in the list of informative keywords.

As a future work, it is necessary to select informative keywords for text categorization dynamically. The optimal criterion of selecting informative keywords is not fixed; very flexible depending on the size and the contents of the given text. In other words, the criterion of selecting informative keywords should be applied differently to each text.

6. Reference

[1] M.S. Chen, J. Han, P.S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transaction on Knowledge and Data Engineering Vol 8 No 6, pp886-883, 1996.

[2] K.M. Decker and S. Forcardi, "Technology Overview: A Report on Data Mining", Technical Report CSCS TR-95-02, Swiss Scientific Computing Center, 1995.

[3] N.J. Radcliffe and P.D. Surry, "Co-operation through Hierarchical Competition in Genetic Data Mining", Technical Report EPCC-TR-94-09, Edinbyrgh Parallel Computing Center, 1994.

[4] H. Lu, R. Setiono, and H. Liu, "Effective Data Mining using Neural Network", IEEE Transaction on Knowledge and Data Engineering Vol 8 No 6, pp957-961, 1996.

[5] V.I. Frants, J Shapiro, and V.G. Voi0skunskii, *Automated Information Retrieval: Theory and Methods*, Academic Press, 1997.

[6] M.A. Hearst, "Text Data Mining: Issues Techniques and the Relation to Information Access", http://www.sims.berkeley.edu/~hearst/talks/dm-talk, 1997.

[7] J. Eldredge, "Text Data Mining: an Overview", http://www.cs.columbia.edu/~radev/cs6998/class/cs6998-09-02/img001htm, 1997.

[8] A.D. Marwick, "Mining on Text Data", http://www.software.ibm.com/data/iminer/fortext/presentations/marwick/index.htm, 1998.

[9] R. Seiffert, "Text Mining and Retrieval: A Development View", http://www.software.ibm.com/data/iminer/fortext/presentations/seiffert/index.htm, 1998.

[10] D.D. Lewis, "Representation and learning in Information Retrieval", Dissertation of PhD, the Graduate School of the University of Massachusetts, 1992.

[11] T. Joachims, "Text Categorization with Support Vector Machines; learning with Many Relevant Features", LS-8 Report 23, Technical Report in University of Dortmund, 1997.

[12] D. Koller and M. Sahami, "Hierarchically Classifying Documents using very few Words", Proc. ICML 97, pp170-178, 1997.

[13] M. Sahami, M. Hearst, and E. Saund, "Applying the Multiple Case Mixture Model to Text Categorization", Proc. ICML 96, appearing, 1996.

[14] L.S. Larkey and W.B. Croft, "Combining Classifiers in Text Categorization", Proc. SIGIR 96, pp289-297 1996.

[15] D.D. Lewis, R.E.Schapire, J.P.Callan, and R.Papka, "Training Algorithm for Linear Text Classifier", pp298-315, The Proceedings of SIGIR 96, 1996.

[16] T.C. Jo, "News Article Classification based on Category Points from Keywords in Backdata", The Proc. CIMCA 99: Intelligent Image Processing, Data Analysis & Information Retrieval edited by M. Mohammadian, pp211-214, 1999.