Theoretical and Applied Informatics
ISSN 1896-5334
Vol. 25 (2013), no. 3-4
pp. 239 – 250
DOI: 10.2478/thai-2013-0017

Classification of text documents by using expanded terms in Latent Semantic Analysis

BOŻENA ŚMIAŁKOWSKA, MARCIN GIBERT

Faculty of Computer Science
West Pomeranian University of Technology
ul. Żołnierska 49, Szczecin, Poland
mgibert@wi.zut.edu.pl, bsmialkowska@wi.zut.edu.pl

Received 31 July 2013, Revised 9 September 2013, Accepted 13 October 2013

Abstract: In this article attention is paid to improving the quality of text document classification. The common techniques of analysis of text documents used in classification are shown and the weakness of these methods are stressed. Discussed here is the integration of quantitative and qualitative methods, which is increasing the quality of classification. In the proposed approach the expanded terms, obtained by using information patterns are used in the Latent Semantic Analysis. Finally empirical research is presented and based upon the quality measures of the text document classification, the effectiveness of the proposed approach is proved.

Keywords: text classification, information extraction, Latent Semantic Analysis, information retrieval, text representation

1. Introduction

Classification is one of the main methods used in applications for the analysis of text documents [1]. This type of application is most popular among companies, institutions and by private persons. Examples are e-mail programs and search engines on the Internet. In order to develop such an application a lot of attention is paid to the quality of the classification [2].

The aim of classification is to allocate a text document to one or more classes which have been defined previously [3]. These classes are defined by a set of pattern text documents, categorized by the expert. Therefore the classification is based on calculation of the similarity of a classifiable text document to the pattern text documents from each category [4]. Finally the text document is classified to the category, which has the highest number of similar text documents.

There are two types of methods of text document analysis, which are used in the classification. These are quantitative and qualitative methods [5]. The first type of these methods is based on the Vector Space Model (VSM) [6]. In this approach text

documents are represented by a vector whose elements are the frequencies of terms, in common case words. In this approach the similarity of two vectors is calculated by using cosine measure [6].

The advanced solution of classification based upon the Vector Space Model is Latent Semantic Analysis [7]. In this case different words, which have a similar meaning are merged and replaced by the hidden semantic structures. Obtained structures are the new elements of the vector, which is representing the text document. Therefore this method achieves in many cases a better quality of classification than common Vector Space Model methods. However, all previously described methods, are focused in general on the structure of the text documents. This causes a considerable limit of text document analysis and therefore the quality of the classification is insufficient in many cases.

A better solution to this problem can be the use of qualitative methods of a text classification. These methods focus on the meaning of the text document. In these methods the expanded information extracted from the text documents with handmade rules are used to classify a text document [8]. The expanded information is obtained by using information patterns defined by the expert [9]. This kind of classification is characterized by high quality of the classification, but the design of a complete model of rules for a large amount of information is difficult and very time consuming [8].

Therefore this paper proposes an alternative solution to increase the quality of the classification by using integration of the previously mentioned methods. In this approach the expanded information extracted from a text document are the new terms in the Latent Semantic Analysis.

2. Expanded terms obtained by using information patterns

In order to obtain expanded terms the information patterns are defined by the expert. These patterns are special structures, which contain some area of knowledge referred to in text documents [9]. This knowledge is used to identify the important information for the classification process of the text documents. For study of the stages of developing information patterns an example concerning symptoms of anaemia and stress have been presented. In the first step, two types of information differentiating the specific symptoms of stress and anaemia have been chosen. For example for identification of anaemia it is necessary to define information like: “yellow skin”, “difficulty concentrating”, “rapid heartbeat” etc. and for identification of stress: “upset stomach”, “chest pain”, “low energy” etc. Next, the important words for identifying the specific information have been categorised. A list of exemplary categories is:

<category name=”part”>leg, skin, heart, hand, head…</category>
<category name=”symptom”>cramp, tingle, ulcer, pain… </category>
<category name=”colour”>pale, red, yellow, black…</category>
<category name=”measure”>strong, difficult, short, rapid…</category>

Where: <category>…</category> - tags of start and close of category definition; name=”…”– name of category; leg, skin, heart…- words from each category.

In order to build the information patterns the relations between previously described categories have been defined by using XML notation:

```
<pattern name=”symptom” level=”1”>
        <plus>
                <category name=”part”> </category>
                <option>
                        <category name=”symptom”> </category>
                        <category name=”colour”> </category>
                </option>
        </plus>
</ pattern >

< pattern name=”measure” level=”2”>
        <plus>
                <option>
                        <category name=” colour,”> </category>
                        <category name=” measure”> </category>
                </option>
                < pattern name=” symptom”> </ pattern >
        </plus>
 </ pattern>
```

Where: < pattern >…</ pattern > - tags of start and close of pattern definition; name=”…”– name of pattern; level=”…” – level of nesting of patterns; element – category or pattern; element* - zero, one or more elements; [element 1, element 2] – alternative of two elements; element 1+ element 2– link of two elements.
Based upon previously defined information patterns, the expanded terms from the classifiable text document and pattern text documents from each category have been extracted. Exemplary terms extracted from text documents are presented below:

1) Term “yellow skin” obtained by using information pattern:

```
<pattern name=”symptom” >
        <category name=”part”>skin</category>
        <category name=”colour”>yellow</category>
</ pattern >
```

2) Term “strong head pain” obtained by using information pattern:

```
<pattern name=”measure” >
        <category name=”measure”>strong</category>
```

```
        <pattern name=”symptom” >
                <category name=”symptom”>pain</category>
                <category name=”part”>head</category>
        </ pattern >
</ pattern >
```

Where between tags <category…></category> is put identified word from the category.

In the first step of extraction the correctness of expectation of all the potential patterns, in which there is a certain category, is validated. Next, after the identification of the patterns the relevant terms are extracted. The exact specificity of the extraction algorithm is presented in literature [9].

3. Classification by using Latent Semantic Analysis

The expanded terms extracted from text documents are used in Latent Semantic Analysis. In the first step the function of importance, which determines the information value of the term by giving the proper weight, is used [10]. A weight can either be local or global. This weight determines the impact of a term, within, respectively, the text document and the text corpus. An example of the function of importance, which gives global weight, is the inverse document frequency - IDF. The IDF is calculated in formula (1) [10].

$$IDF = \log(\frac{n}{df_i}) + 1 \tag{1}$$

Where i – the term from 1 to m, m - maximum index of the term, n – the total number of text documents, df_i (document frequency) - the number of text documents, where there is a term with index i. In this case weight is higher for terms which occur less in different text documents. The less occurring terms are more important discriminants of text documents from each category and therefore they have much greater influence on the classification process.

In the next step, based upon an obtained set of terms, the vectors which represent a text document have been built. The elements of these vectors are the weights of terms calculated by the function of importance.

The main goal of Latent Semantic Analysis is the reduction of vectors dimensionality. In this case terms from different text documents, which have a similar meaning, but extracted from different information patterns, are merged and replaced by the hidden semantic structures, which are the new elements of the vectors [11]. Therefore texts documents thematically similar, but made up of different terms, have a high similarity measure and therefore the quality of classification is also improved. The second reason of using LSA is that the reduced size of the vectors increases the efficiency of the classification system.

In practice the reduction of vectors dimensionality is realised by Singular Value Decomposition (SVD) for the sparse matrix of terms and text documents, which is built by linking all vectors [7]. This method selects the optimal projection for a given

amount of semantic structures, since they correspond to the greatest diversity of vector elements. A sparse matrix of terms and text documents is decomposed according to the formula (2) [11].

$$A_{mxn} = U_{mxd} S_{dxd} (V_{nxd})^T \tag{2}$$

Where m – the total number of terms, n – the total number of text documents, d=min(m,n), U_{mxd}– the matrix of terms, S_{dxd}– the matrix of singular values, V_{nxd}– the matrix of text documents. In order to reduce the dimensionality of a matrix to a certain number of detected semantic structures between terms, the matrix of singular values S and the matrix of text documents V are reduced to k columns, where k <m. The value of k is chosen based on singular value analysis using specific rules, described in literature [9]. The matrix of terms, which are new semantic structures and text documents are calculated by multiplying the reduced matrix S and $(V_{nxd})^T$.

In the next step, the similarity of the classifiable text document to pattern text documents, belonging to each category, is calculated. The similarity is calculated using the selected measures. The most commonly used one is the cosine measure, calculated according to formula (3) [9].

$$\cos(p,d) = \frac{\sum_{i=1}^{k} p_i d_i}{\sqrt{\sum_{i=1}^{k} p_i^2} \sqrt{\sum_{i=1}^{k} d_i^2}} \tag{3}$$

Where p –classifiable text document, d – pattern text document, p_i – the value of the element i in the vector of classifiable text, d_i – the value of the element i in the vector of a pattern text document. In this case the similarity measure shows how close a classifiable text document is with every pattern text document. If this measure is low it means that text documents include terms with similar meaning.

In the last step, based upon a calculated similarity measure for every pair of classifiable text document and pattern text document, the category in which there is the greatest number of similar text documents is selected. The greatest number of similar text documents is calculated based on a text documents form category which havea higher similarity than the level of similarity established by the expert. Therefore KNN (K-Nearest Neighbour) , one of the most popular algorithms has been used [12]. The stages of this algorithm are:

1) The pattern text document with similarity to classifiable text document higher than similarity level established by the expert is selected.
2) Membership of selected text document for each category is checked.
3) The text is classified into the category with higher number of relevant text documents.

4. Empirical research based on the average precision and recall

The testing research relies on the basic measures which characterize the classification quality. These measures are precision and recall [2]. Recall is expressed by formula (4), and average precision by formula (5) [13]. Table 1 gives the specific results of the classification system for defining these quality measures.

Category	**Expert allocates YES**	**Expert allocates NO**	**Overall**
System allocates YES	TP_i	FP_i	m_i
System allocates NO	FN_i	TN_i	$N-m_i$
Together	n_i	$N-n_i$	N

Tab 1. The specificity of the results of the classification system.

Where TP_i(True Positive) – text documents properly considered to be relevant, FP_i(False Positive) – text documents falsely considered to be relevant,
FN_i(False Negative) – text documents falsely considered to be non-relevant,
TN_i(False Positive) – text documents properly considered to be non-relevant,
n_i– number of all relevant text documents allocated by the expert,
m_i – number of all relevant text documents allocated by the system.

$$r(i) = \frac{TP_i}{n_i} \tag{4}$$

where: r(i) – recall on i-level.

$$p_{avgj} = \frac{\sum_{1}^{k} p_j}{k},\ j = 1,2...k,\ p_j(i) = \frac{TP_i}{m_i} \tag{5}$$

where: TP_i(True Positive) – text documents properly considered to be relevant, m_i – number of all relevant text documents allocated by the system, i – level of recall,
j – number of pattern text document.

In this research the average precision of the classification is calculated at different levels of recall for four variants of integration of text analysis methods. These variants are:

1) classification based upon Vector Space Model with terms built on the singular words,
2) classification based upon Latent Semantic Analysis with terms built on the singular words,
3) classification based upon Vector Space Model with expanded terms built on information patterns,
4) classification based upon Latent Semantic Analysis with expanded terms built on information patterns.

The test task was to classify a text document into one of two categories for the three, different sets of classifiable and pattern text documents; see Table 2. The first category was related to the symptoms of anaemia and the second to the symptoms of stress. The four different pattern text documents describing the symptoms are assigned to each category. Prior to testing, the classifiable text document has been assessed by an expert to the second category.

For the first variant all words from text documents have been changed to a basic form [9]. Next, the low-value words have been removed by using stop-list. Next, for calculating the weight of the terms, the singular words in this case, the function of importance - IDF has been used. The degree of similarity of the classifiable text document to pattern text documents by using the cosine measure, has been calculated. The calculation results are shown in Table 2.

Similarity	**Category 1**				**Category 2**			
	Pattern text doc. 1	**Pattern text doc. 2**	**Pattern text doc. 3**	**Pattern text doc. 4**	**Pattern text doc. 5**	**Pattern text doc. 6**	**Pattern text doc. 7**	**Pattern text doc. 8**
Classifiable text 1	0,2412	0,0958	0,1097	0,0819	0,1416	0,0381	0	0,0668
Classifiable text 2	0,0842	0,1436	0,0293	0,1561	0,1225	0,076	0,0661	0,2414
Classifiable text 3	0,0375	0,1604	0,367	0,117	0,0931	0,0413	0	0,1214

Tab 2. The similarity of text documents calculated for the terms based on single words.

Each single row in Table 2 is a set of classifiable and pattern text documents. In each set there is one classifiable text document, for which the similarity level (cosine measure) to every four pattern text documents from each category is calculated.
On the four possible levels of recall corresponding to the number of pattern text documents from each category, the average precision for three different sets of classifiable and pattern text documents has been calculated. The average precision is expressed by formula (6) [14]:

$$p_{avgj} = \frac{\sum_{1}^{k} p_j}{k},\ j = 1,2...k,\ p_j(i) = \frac{TP_i}{m_i} \tag{6}$$

where: TP_i(True Positive) – text documents properly considered to be relevant, m_i – number of all relevant text documents allocated by the system, i – level of recall,
j – number of set of classifiable and pattern text documents. For the calculation of precision on i-level of recall the rankings of similarity of the classifiable text document to pattern text documents has been built. An example of ranking for the second level of recall and for the first set of classifiable and pattern text documents is shown in Table 3.

Position	Text	Similarity	Category
1	1	0,2412	1
2	6	0,1416	2
3	7	0,1097	1

Tab 3. The similarity ranking for the second level of recall.

The calculation of the precision for every pattern text document on the different level of recall and also average precision for all levels of recall is shown in Table 4.

Level of recall	Precision			Average precision
	Pattern text 1	Pattern text 2	Pattern text 3	
25%	100%	50%	100%	**83,33%**
50%	66%	66%	100%	**77,33%**
75%	75%	60%	75%	**70%**
100%	80%	50%	57%	**62,33%**

Tab 4. The average precision for terms based on single words.

The highest average precision achieves 83,33% for the first level of recall, which amounts to 25%. For the maximum level of recall - 100% - is achieved 62,33% of the average precision.

In the second variant the terms based upon singular words are used in Latent Semantic Analysis. In this case singular words are merged and replaced by the hidden semantic structures, which are the new terms in the classification process. The sparse matrix of all terms and text documents is reduced to only four new terms, which are identified hidden semantic structures. The similarity of text documents based upon their new representation and subsequently the average precision for each set of classifiable and pattern text documents have been calculated.

In the third variant of the classification, the information patterns for each category have been prepared. The expanded terms from all text documents have been extracted. These expanded terms are the new terms in the Vector Space Model used in the text document classification. For each term the IDF weight has been obtained and the similarity of text documents and average precision for every set of text documents have been calculated.

In the last variant proposed by the authors, the expanded terms extracted from the text documents have been used in Latent Semantic Analysis. The matrix of all expanded terms and text documents has been reduced with the rule of proportion and the text document similarity has been calculated. Then the average precision for each level of recall has been obtained. The total calculation of the average precision for all variants of the text document classification is presented in Table 5.

Level of recall	Average precision			
	Classification Variant I – terms base on singular words in VSM	Classification Variant II – terms base on singular words in LSA	Classification Variant III – expanded terms in VSM	Classification Variant IV – expanded terms in LSA
25%	83,33%	100%	100%	100%
50%	77,33%	77,33%	83,33%	100%
75%	70%	65%	78,33%	83,33%
100%	62,33%	57,66%	52,33%	60%

Tab 5. The average precision for all variants of text document classification.

The visualisation of the results is presented in Fig. 1.

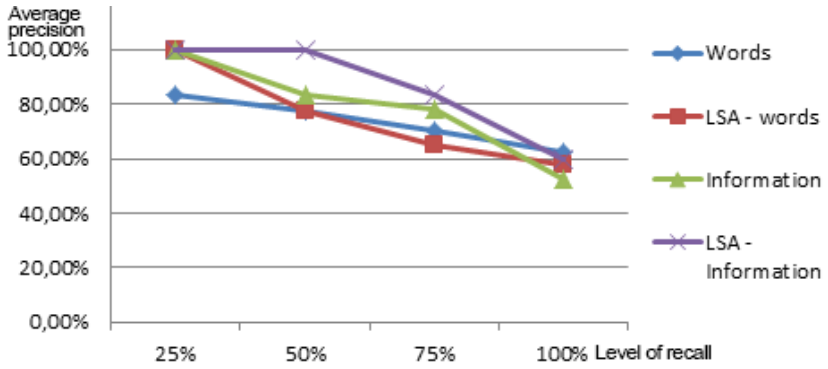

Fig 1. The average precision of all variants of text document classification.

In Table 5 and in Fig. 1 the advantage of the classification by using expanded terms in the Latent Semantic Analysis is made visible. Almost on every level of recall the classification by using expanded terms in Latent Semantic Analysis is better than other methods. Only in this variant on the first two levels of recall the average precision of 100% was maintained. There is also a significant difference between the classification by using expanded terms in Vector Space Model (third variant of classification) and the first two variants of classifications using terms based on single words.

6. Conclusions

Based upon the empirical research it can be concluded that a classification based upon the expanded terms obtained in this way and used in the Latent Semantic Analysis gives a higher quality of classification in comparison with a classification based upon terms with single words. Also the reduction of the size of the sparse matrix of terms and text documents by using Latent Semantic Analysis increases the efficiency of the classification system.

This model of text document classification is used in a situation, which on the one hand, requires precise information extraction from a text document, while on the other

hand the construction of classification rules is time-consuming or difficult to achieve. The two most popular examples of text classification found in literature, in which the described solution may be used, concern the diagnosis of diseases based on a patient's description of symptoms, and a classification of emails.

By analysing the results it should be kept in mind that the quality measures of classification are based only on precision and recall. In the given example classification applied only to two categories and a small amount of pattern text documents. For larger collections of text documents, the precision of 100% is very difficult to obtain. The concept of Latent Semantic Analysis for the text document representation using the extracted information can reasonably be used for classification of large sets of text documents. The condition of this task is to maintain a relatively small area of knowledge, for which the information patterns are designed.

References

[1] Jackson P. i Moulinier I., *Natural Language Processing for Online Applications: Text Retrieval, Extraction, and Categorization*. John Benjamins Publishing, 2002.

[2] Lewis D. D., *Representation quality in text classification: an introduction and experiment*, w Proceedings of the workshop on Speech and Natural Language, Stroudsburg, PA, USA, 1990, pp. 288–295.

[3] Sebastiani F., *Machine learning in automated text categorization*, ACM Comput Surv, t. 34, nr 1, ss. 1–47, mar. 2002.

[4] Metzler D., Dumais S., i Meek C., *Similarity Measures for Short Segments of Text*, in Advances in Information Retrieval, G. Amati, C. Carpineto, i G. Romano, Red. Springer Berlin Heidelberg, 2007, pp. 16–27.

[5] Stefano Ferilli M. B., *Combining Qualitative and Quantitative Keyword Extraction Methods with Document Layout Analysis*", pp. 22–33, 2009.

[6] Salton G., Wong A., i Yang C. S., *A vector space model for automatic indexing*, Commun ACM, t. 18, nr 11, ss. 613–620, lis. 1975.

[7] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., i Harshman R., *Indexing by latent semantic analysis*, J. Am. Soc. Inf. Sci., t. 41, nr 6, ss. 391–407, 1990.

[8] Hayes P. J. i Weinstein S. P., *CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories*, in Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence, 1991, ss. 49–64.

[9] Lubaszewski W., *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*. AGH Uczelniane Wydawnictwa Naukowo-Dydaktyczne, 2009.

[10] Xia T. i Chai Y., *An improvement to TF-IDF: Term Distribution based Term Weight Algorithm*, J. Softw., t. 6, nr 3, mar. 2011.

[11] Landauer T., Foltz P., i Laham D., *An Introduction to Latent Semantic Analysis*, Discourse Process., nr 25, ss. 259–284, 1998.

[12] Guo G., Wang H., Bell D., Bi Y., i Greer K., *Using kNN model for automatic text categorization*, Soft Comput., t. 10, nr 5, ss. 423–430, mar. 2006.

[13] Raghavan V., Bollmann P., i Jung G. S., *A critical investigation of recall and precision as measures of retrieval system performance*, ACM Trans Inf Syst, t. 7, nr 3, ss. 205–229, lip. 1989.

[14] Berry M. W., Kogan J., i SIAM International Conference on Data Mining, *Text mining applications and theory*. Chichester, U.K.: Wiley, 2010.

Klasyfikacja dokumentów tekstowych przy użyciu rozbudowanych wyrażeń w niejawnej analizie semantycznej

W artykule skoncentrowano się na poprawie jakości klasyfikacji dokumentów tekstowych. Zostały przybliżone najpopularniejsze techniki analizy dokumentów tekstowych wykorzystywanych w klasyfikacji. Zwrócono uwagę na słabe stronnych opisanych technik. Omówiono możliwość integracji metod ilościowych i jakościowych analizy tekstu i jej wpływ na poprawę jakości klasyfikacji. Zaproponowano rozwiązanie, w którym rozbudowane wyrażenia otrzymane za pomocą wzorców informacyjnych są wykorzystywane w niejawnej analizie semantycznej. Ostatecznie w oparciu o miary jakości klasyfikacji dokumentów tekstowych zaprezentowano wyniki badań testowych, które potwierdzają skuteczność zaproponowanego rozwiązania.