A Hybrid Statistical Data Pre-processing Approach for Language-Independent Text Classification

Yanbo J. Wang[1], Frans Coenen[2], and Robert Sanderson[2]

[1] Information Management Center, China Minsheng Banking Corp., Ltd.
Room 606, Building No. 8, 1 Zhongguancun Nandajie,
100873 Beijing, China
wangyanbo@cmbc.com.cn
[2] Department of Computer Science, University of Liverpool,
Ashton Building, Ashton Street, Liverpool, L69 3BX, UK
{Coenen,Azaroth}@liverpool.ac.uk

Abstract. Data pre-processing is an important topic in Text Classification (TC). It aims to convert the original textual data in a data-mining-ready structure, where the most significant text-features that serve to differentiate between text-categories are identified. Broadly speaking, textual data pre-processing techniques can be divided into three groups: (i) linguistic, (ii) statistical, and (iii) hybrid (i) & (ii). With regard to language-independent TC, our study relates to the statistical aspect only. The nature of textual data pre-processing includes: Document-base Representation (DR) and Feature Selection (FS). In this paper, we propose a hybrid statistical FS approach that integrates two existing (statistical FS) techniques, DIAAF (Darmstadt Indexing Approach Association Factor) and GSSC (Galavotti·Sebastiani·Simi Coefficient). Our proposed approach is presented under a statistical "bag of phrases" DR setting. The experimental results, based on the well-established associative text classification approach, demonstrate that our proposed technique outperforms existing mechanisms with respect to the accuracy of classification.

Keywords: Associative Classification, Data Pre-processing, Document-base Representation, Feature Selection, (Language-independent) Text Classification.

1 Introduction

Text mining is a promising topic of current research in data mining and knowledge discovery. It aims to extract various types of hidden, interesting, previously unknown and potentially useful knowledge from sets of collected textual data. In a natural language context, a given textual dataset is usually refined to produce a document-base, i.e. a set of electronic documents that typically consists of thousands of documents, where each document may contain hundreds of words. One important aspect of text mining is Text Classification (TC) – "*the task of assigning one or more predefined categories to natural language text documents, based on their contents*" [10]. Broadly speaking, TC studies can be separated into two divisions: *single-label* that assigns exactly one pre-defined category to each "unseen" document; and

R. Huang et al. (Eds.): ADMA 2009, LNAI 5678, pp. 338–349, 2009.

multi-label that assigns one or more pre-defined category to each “unseen” document. With regard to single-label TC, two distinct approaches can be identified: *Binary* TC which in particular assigns either a pre-defined category or the complement of this category to each “unseen” document; and *multi-class* TC which simultaneously deals with all given categories and assigns the most appropriate category to each “unseen” document. This paper is concerned with the single-label multi-class TC approach.

Text mining requires the given document-base to be first pre-processed, where the (unstructured) original textual data is converted in a (structured) data-mining-ready format, and the most significant text-features that serve to differentiate between text categories are identified. Thus the entire process of TC, in general, can be identified as textual data (document-base) *pre-processing* plus traditional *classification*. Broadly speaking, textual data pre-processing techniques can be divided into three groups: (i) linguistic, (ii) statistical, and (iii) hybrid (i) & (ii). Both the linguistic and the hybrid aspects pre-process document-bases depending on the rules and/or regularities in semantics, syntax and/or lexicology of languages. Such techniques are designed with particular languages and styles of language as the target, and involve deep linguistic analysis. For the purpose of building a language-independent text classifier that can be applied to cross-lingual, multi-lingual and/or unknown lingual textual data collections, this paper is only concerned with the statistical aspect of textual data pre-processing.

In [17] the nature of textual data pre-processing is characterized as: *Document-base Representation* (DR) which designs an application oriented data model to precisely interpret a given document-base in an explicit and structured manner; and *Feature Selection* (FS) which extracts the most significant information (text-features) from the given document-base. In DR the Vector Space Model (VSM) [20] is considered appropriate for many text mining applications, especially when dealing with TC problems. The VSM is usually presented in a binary format, where “*each coordinate of a document vector is zero (when the corresponding attribute is absent) or unity (when the corresponding attribute is present)*” [14]. In TC, two common DR approaches that are used to define VSM are the “bag of words” and the “bag of phrases”. The motivation for the latter approach is that phrases seem to carry more contextual and/or syntactic information than single words. In [22] Scheffer and Wrobel argue that the “bag of words” representation does not distinguish between “*I have no objections, thanks*” and “*No thanks, I have objections*”, where the “bag of phrases” approach seems to deal with this kind of situation better. Hence the experimental work in this paper is designed with respect to the “bag of phrases” DR setting.

In theory, the textual attributes of a document can include every text-feature (word or phrase) that might be expected to occur in a given document-base. However, this is computationally unrealistic, so it requires some FS mechanism (during the pre-processing phase) to identify the *key* text-features that will be useful for a particular text mining application, such as TC. In the past, a number of approaches have been proposed for TC, under the heading of statistical FS. Two major ones are the Darmstadt Indexing Approach Association Factor (DIAAF) and the Galavotti·Sebastiani·Simi Coefficient (GSSC). Other existing methods include: Relevancy Score (RS), Mutual Information (MI), etc.

Classification Rule Mining (CRM) is a well-established area in data mining and knowledge discovery for identifying hidden classification rules from a given class-database (i.e. usually a relational data table with a set of pre-defined class labels), the objective being to build a (rule based) classifier to categorize "unseen" data records. It should be noted that CRM refers to the rule based approach of the traditional classification problem. Approaches that are parallel to CRM include: probabilistic classification, support vector machine based classification, neural network based classification, etc. One CRM implementation mechanism is to employ association rule mining [1] methods to identify the desired classification rules, i.e. associative classification [2]. Coenen *et al.* [5] and Shidara *et al.* [24] indicate that results presented in the studies of [15, 16, 28] show that in many cases associative classification offers greater classification accuracy than other classification approaches, such as C4.5 [19] and RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [7].

In the past decade, associative classification has been proposed for application in TC (e.g. [6, 29]). In [3] Antonie and Zaïane argue: an associative text classifier "*is fast during both training and categorization phases*", especially when handling large document-bases; and such classifiers "*can be read, understood and modified by humans*". In comparison, TC techniques other than the rule based, i.e. probabilistic based, support vector machine based, neural network based, etc., do not present the classifier in a human readable fashion, so that users do not see why the classification predictions have been made. Given the advantages offered by associative classification with respect to TC, this approach is adopted in our study to support the investigation of statistical (textual) data pre-processing for language-independent TC.

In this paper, we propose a statistical FS approach, which combines the ideas of DIAAF and GSSC mechanisms, namely Hybrid DIAAF/GSSC. The evaluation of Hybrid DIAAF/GSSC, under a statistical "bag of phrases" DR setting, is conducted using the TFPC (Total From Partial Classification) [5] associative classification algorithm; although any other associative classifier generator could equally well have been used. The experimental results demonstrate that Hybrid DIAAF/GSSC based textual data pre-processing approach outperforms alternative techniques with respect to the accuracy of classification. This in turn improves the performance of language-independent TC. The rest of this paper is organized as follows. In section 2, we describe the statistical "bag of phrases" DR approach. In section 3, we review the DIAAF and the GSSC statistical FS mechanisms. The Hybrid DIAAF/GSSC approach is proposed in section 4. Section 5 presents the experimental results. Finally our conclusions and open issues for further research are provided in section 6.

2 Statistical "Bag of Phrases" Document-Base Representation

In the "bag of phrases" DR approach, each element in a document vector represents a phrase describing an ordered combination of words appearing contiguously in sequence. Preliminarily, some definitions with regard to the statistical aspect are given as follows.

- **Words:** Words in a document-base are defined as continuous sequences of alphabetic characters delimited by non-alphabetic characters, e.g. punctuation marks, white space and numbers.
- **Noise Words (N):** Common and rare words are collectively defined to be *noise* words in a document-base. Note that noise words can be identified by their *support* value, i.e. the percentage of documents in the training dataset in which the word appears.
- **Upper Noise Words:** Common (upper noise) words are words with a support value above a user-supplied Upper Noise Threshold (UNT).
- **Lower Noise Words:** Rare (lower noise) words are words with a support value below a user-supplied Lower Noise Threshold (LNT).
- **Potential Significant Words:** A potential significant word, also referred to as a *key* word, is a non-noise whose *contribution* value exceeds some user-specified threshold *G*. The contribution value of a word is a measure of the extent to which the word serves to differentiate between classes and can be calculated in a number of ways (noted as various statistical FS mechanisms).
- **Significant Words (G):** The first *K* words (i.e. the first *k* words for each pre-defined class) that are selected from the ordered list of potential significant words (in a descending manner based on their contribution value) are defined to be significant words.
- **Ordinary Words (O):** Other non-noise words that have not been selected as significant words.
- **Stop Marks (S):** Not actual words but six key punctuation marks (, . : ; ! and ?). All other non-alphabetic characters are ignored.

In [6] the authors (based on the above definitions) propose a statistical "bag of phrases" (DR) approach for TC, namely DelSNcontGO: phrases are Delimited by stop marks (S) and/or noise words (N), and (as phrase contents) made up of sequences of one or more significant words (G) and ordinary words (O); sequences of ordinary words delimited by stop marks and/or noise words that do not include at least one significant word (in the contents) are ignored. The experimental results presented in [6] show that DelSNcontGO performs well with respect to the accuracy of classification. In this paper, this statistical "bag of phrases" DR approach will be further concerned in the section of experimental results.

3 Statistical Feature Selection

Statistical FS techniques automatically compute a weighting score for each text-feature in a document. A significant text-feature can be identified when its weighting score exceeds a user-defined weighting threshold. Methods under this heading do not involve linguistic analysis but focus on some document-base statistics. With regard to TC, the common intuitions of various methods here can be described as: (i) the more times a text-feature appears in a class the more relevant it is to this particular class; and (ii) the more times a text-feature appears across the document-base in documents of all classes the worse it is at discriminating between the classes.

A number of mechanisms have been proposed in statistical FS. Two major statistical models can be identified: Darmstadt Indexing Approach Association Factor (DIAAF) and Galavotti·Sebastiani·Simi Coefficient (GSSC).

DIAAF: The Darmstadt Indexing Approach (DIA) [11] was originally "*developed for automatic indexing with a prescribed indexing vocabulary*" [12]. In a machine learning context, Sebastiani [23] argues that this approach "*considers properties (of terms, documents, categories, or pairwise relationships among these) as basic dimensions of the learning space*". Examples of the properties include the length of a document, the frequency of occurrence between a text-feature and a class, etc. One of the pair-wise relationships considered is the term-category relationship, noted as the DIA Association Factor (DIAAF) [23], which can be applied to select significant text-features for TC problems. The calculation of the DIAAF score, and reported in [10], can be specified in probabilistic form using:

$$diaaf\_score(u_h, C_i) = \mathrm{P}(C_i \mid u_h)\ ,$$

where u_h represents a text-feature in a given document-base $Đ$ ($Đ = \{D_1, D_2, \ldots, D_{m-1}, D_m\}$), and C_i represents a set of documents (in $Đ$) labeled with a particular text-class. The DIAAF weighting score expresses the proportion of a feature's occurrence in the given class divided by a feature's document-base occurrence.

GSSC: The GSS (Galavotti·Sebastiani·Simi) Coefficient defined in [13] represents the core calculation as well as a simplified variant of both the Chi-square Statistics (χ^2) and the Correlation Coefficient (CC) statistical FS mechanisms. In [27, 30], the authors state: (i) the well-established χ^2 statistic can be applied to measure the lack of independence between a term u_h and a pre-defined class C_i; and (ii) if the feature/term and the class are independent, the calculated χ^2 score has a natural value 0. In [18] Ng *et al.* introduce CC as a refined variant of χ^2 to generate a better set of key/significant features and improve the performance of the χ^2 approach. Ng *et al.* argue that "*words that come from the irrelevant texts or are highly indicative of non-membership in*" a class C_i are not as useful; and indicate that CC "*selects exactly those words that are highly indicative of membership in a category, whereas the χ^2 metric will not only pick out this set of words but also those words that are indicative of non-membership in the category*". In [13] Galavotti *et al.* provide an explanation of the rationale to further refine the CC approach, and demonstrate that this very simple approach (GSSC) can produce a comparable performance to the χ^2 metric. The GSSC is defined in probabilistic form using:

$$gssc\_score(u_h, C_i) = \mathrm{P}(u_h, C_i) \times \mathrm{P}(\neg u_h, \neg C_i) - \mathrm{P}(u_h, \neg C_i) \times \mathrm{P}(\neg u_h, C_i)\ ,$$

where $\neg u_h$ represents a document that does not involve the feature u_h, and $\neg C_i$ ($Đ - C_i$) represents the set of documents labeled with the complement of the pre-defined class C_i.

Existing statistical FS techniques other than DIAAF and GSSC include: Mutual Information (MI), Relevancy Score (RS), etc. In this section, we further provide a brief review of MI and RS. Note that both MI and RS are referenced in the evaluation section of this paper (section 5).

MI: Early work on Mutual Information (MI) can be found in [4, 9]. This statistical model is used to determine whether a genuine association exists between two text-features or not. In TC investigations, MI has been broadly employed in a variety of approaches to select the most significant text-features that serve to classify documents. The calculation of the MI score between a text-feature u_h and a pre-defined text-class C_i, as reported in [10], is achieved in probabilistic form using:

$$mi\_score(u_h, C_i) = log(\mathrm{P}(u_h \mid C_i) \mathbin{/} \mathrm{P}(u_h)) \,.$$

This score expresses the proportion (in a logarithmic terms) of the probability with which the feature occurs in documents of the given class divided by the probability with which the feature occurs in the document-base.

RS: The initial concept of Relevancy Score (RS) was introduced by Salton and Buckley [21] as relevancy weight. It aims to measure how "unbalanced" a text-feature (term) u_h is across documents in a document-base $Đ$ with and without a particular text-class C_i. Salton and Buckley define a term's relevancy weight as: "*the proportion of relevant documents in which a term occurs divided by the proportion of nonrelevant items in which the term occurs*". In [26] the idea of RS was proposed based on relevancy weight with the objective of selecting significant text-features in $Đ$ for the TC application. A term's relevancy score can be defined as: the number of relevant (the target text-class associated) documents in which a term occurs divided by the number of non-relevant documents in which a term occurs. Fragoudis *et al.* [10] and Sebastiani [23] show that the RS score can be calculated in probabilistic form using:

$$relevancy\_score(u_h, C_i) = log((\mathrm{P}(u_h \mid C_i) + d) \mathbin{/} (\mathrm{P}(u_h \mid \neg C_i) + d)) \,,$$

where d is a constant damping factor. In [26] the value of d was initialized as 1/6. For the simplicity, we choose 0 as the value of d in our study.

4 Proposed Statistical Feature Selection

With regard to language-independent TC, in this section, we introduce a new statistical FS technique. In the previous section, two statistical FS techniques DIAAF and GSSC were presented in detail. The newly proposed mechanism is a variant of the original GSSC approach that makes use of the DIAAF approach, namely Hybrid DIAAF/GSSC.

Recall that the probabilistic formula for calculating the DIAAF score is given by:

$$diaaf\_score(u_h, C_i) = \mathrm{P}(C_i \mid u_h) \,.$$

Recall that the probabilistic formula for GSSC is:

$$gssc\_score(u_h, C_i) = \mathrm{P}(u_h, C_i) \times \mathrm{P}(\neg u_h, \neg C_i) - \mathrm{P}(u_h, \neg C_i) \times \mathrm{P}(\neg u_h, C_i) \,.$$

Substituting each of the four probabilistic components in GSSC by its DIAAF related function, a DIAAF based formula is derived in a GSSC fashion:

$$diaaf\text{-}gssc\_score(u_h, C_i) = \mathrm{P}(C_i \mid u_h) \times \mathrm{P}(\neg C_i \mid \neg u_h) - \mathrm{P}(\neg C_i \mid u_h) \times \mathrm{P}(C_i \mid \neg u_h) \,.$$

An example of Hybrid DIAAF/GSSC score calculation is provided in Table 1. Given a document-base *Ð* containing 100 documents equally divided into 4 classes (i.e. 25 per class), and assuming that text-feature (word) u_h appears in 30 of the documents, then the Hybrid DIAAF/GSSC score per class can be calculated as shown in the Table.

Table 1. Hybrid DIAAF/GSSC score calculation

Class	# docs per class	# docs with u_h per class	# docs without u_h per class	# docs with u_h in other classes	# docs without u_h in other classes	# docs with u_h in *Ð*	# docs without u_h in *Ð*	Hybrid DIAAF/ GSSC Score
1	25	15	10	15	60	30	70	0.357
2	25	10	15	20	55	30	70	0.119
3	25	5	20	25	50	30	70	-0.119
4	25	0	25	30	45	30	70	-0.357

The algorithm for identifying significant text-features (i.e. *key* words in the current context, with regard to sections 2 – Potential Significant Words) in *Ð*, based on Hybrid DIAAF/GSSC, is given as follows:

```
Algorithm: Key Word Identification – Hybrid DIAAF/GSSC
Input: (a) A document-base Ð (the training part, where the noise words have been removed);
       (b) A user-defined significance threshold $G$;
Output: A set of identified key words $S_{KW}$;
Begin Algorithm:
  $S_{KW}$ ← an empty set for holding the identified key words in Ð;
  $C$ ← catch the set of pre-defined text-classes within Ð;
  $W_{GLO}$ ← read Ð to create a global word set, where the word document-base support $supp_{GLO}$ is associated with each word $u_h$ in $W_{GLO}$;
  for each $C_i \in C$ do
      $W_{LOC}$ ← read documents that reference $C_i$ to create a local word set, where the local support $supp_{LOC}$ is associated with each word $u_h$ in $W_{LOC}$;
      for each word $u_h \in W_{LOC}$ do
```

```
            contribution ← $(u_h.supp_{LOC}$ / $u_h.supp_{GLO})$ ×
    $(((|Đ|$ - $|C_i|)$
                    -$(u_h.supp_{GLO}$ - $u_h.supp_{LOC}))$ / $(|Đ|$ -
            $u_h.supp_{GLO}))$ -
                    $((u_h.supp_{GLO}$ - $u_h.supp_{LOC})$ / $u_h.supp_{GLO})$ ×
                    $((|C_i|$ - $u_h.supp_{LOC})$ / $(|Đ|$ - $u_h.supp_{GLO}))$;
            if (contribution $\geq G$) then
                    add $u_h$ into $S_{KW}$;
        end for
end for
return ($S_{KW}$);
End Algorithm
```

The intuition behind the Hybrid DIAAF/GSSC approach is:

1. The contribution of term u_h for class C_i tends to be high if the ratio of the class based term support to the document-base term support is high,
2. The contribution of term u_h for class C_i tends to be high if the ratio of the class-complement based term support of non-appearance to the document-base term support of non-appearance is high,
3. The contribution of term u_h for class C_i tends to be high if the ratio of the class-complement based term support to the document-base term support is low, and
4. The contribution of term u_h for class C_i tends to be high if the ratio of the class based term support of non-appearance to the document-base term support of non-appearance is low.

5 Experimental Results

This section presents an evaluation of the proposed statistical FS approach, using three well-known text collections (i.e. Usenet Articles, Reuters-21578 and MedLine-OHSUMED). The aim of this evaluation is to assess the approach with respect to the accuracy of classification in statistical "bag of phrases" DR setting. All evaluations described in this section were conducted using the TFPC[1] associative classification algorithm; although any other classifier generator could equally well have been used. All algorithms involved in the evaluation were implemented using the standard Java programming language. The experiments were run on a 1.87 GHz Intel(R) Core(TM)2 CPU with 2.00 GB of RAM running under the Windows Command Processor. For the experiments four individual document-bases (textual datasets) were used. Each was prepared/extracted (as a subset) from one of the above mentioned text collections. The preparation of Usenet Articles ("20 Newsgroups") based document-bases adopted the approach of Deng *et al.* [8], where the entire collection was randomly split into two document-bases covering 10 classes each: 20NG.D10000.C10 and 20NG.D9997.C10. The preparation of Reuters-21578 and the MedLine-OHSUMED document-bases recalled the idea of Wang *et al.* [25], where the Reuters.D6643.C8 and OHSUMED.D6855.C10 document-bases were generated.

[1] TFPC software may be obtained from
http://www.csc.liv.ac.uk/~frans/KDD/Software/Apriori-TFPC/aprioriTFPC.html

The experiments reported below were designed to evaluate the proposed Hybrid DIAAF/GSSC FS approach, in comparison with alternative mechanisms (i.e. DIAAF, GSSC, MI, and RS), with regard to the DelSNcontGO statistical "bag of phrases" DR approach. Accuracy figures, describing the proportion of correctly classified "unseen" documents, were obtained using the Ten-fold Cross Validation (TCV). A support threshold value of 0.1% and a Lower Noise Threshold (LNT) value of 0.2% were used, as suggested in [6]. A confidence threshold value of 50% was used (as proposed in the published evaluations of a number of associative classification studies [5, 15, 28]). The Upper Noise Threshold (UNT) value was set to 20%. The parameter *K* (maximum number of selected final significant words) was chosen to be 1,000. Note that the value of *K* was changed to be 900 instead of 1,000 for OHSUMED.D6855.C10. The reason to decrease the value of *K* here was that 1,000 selected final significant words generated more than 2^{15} significant phrases; and, for reasons of computational efficiency, the TFPC associative classifier limits the total number of identified attributes[2] (significant phrases) to 2^{15}. To ensure that there are enough candidate final significant (potential significant) words to be selected for each category, the *G* parameter was given a minimal value (*almost zero*) so that the *G* parameter could be ignored.

Table 2. Classification accuracy – comparison of the five statistical FS techniques in the statistical "bag of phrases" DR setting

	DIAAF	**GSSC**	**RS**	**MI**	**Hybrid DIAAF/GSSC**
20NG.D10000.C10	76.36	0	76.36	76.36	**76.43**
20NG.D9997.C10	81.45	0	81.45	81.45	**81.62**
Reuters.D6643.C8	87.57	0	87.79	87.79	**88.23**
OHSUMED.D6855.C10	78.83	0	79.64	79.53	**79.74**
Average Accuracy	81.05	0	81.31	81.28	**81.51**
# of Best Accuracies	0	0	0	0	**4**

The results presented in Table 2 are the classification accuracy values (obtained by different statistical FS mechanisms in the DelSNcontGO statistical "bag of phrases" DR setting), based on the 4 extracted/prepared document-bases. From Table 2 it can be seen that the proposed Hybrid DIAAF/GSSC mechanism outperforms other alternative approaches:

1. The number of instances of best classification accuracies obtained throughout the 4 document-bases can be ranked in order as: Hybrid DIAAF/GSSC (all cases), and DIAAF, GSSC, RS and MI (none of any case), which demonstrates the stability of Hybrid DIAAF/GSSC's good performance;

[2] The TFPC algorithm stores attributes as a signed short integer.

2. The average accuracy of classification throughout the 4 document-bases can be ranked in order as: Hybrid DIAAF/GSSC (81.51%), RS (81.31%), MI (81.28%), DIAAF (81.05%), and GSSC (0%), which shows the overall advantage of the proposed mechanism; and
3. The column of GSSC is shown with value '0' for all the records. The reason of this is that when applying the GSSC feature selection technique, with the TFPC associative text classifier, too many rules were generated thus causing computational difficulty and consequently no results were obtained.

6 Conclusions

This paper is concerned with an investigation of statistical feature selection for (single-label multi-class) language-independent text classification. A description of the statistical document-base representation in terms of "bag of phrases" was provided in section 2. Both the DIAAF and GSSC statistical FS approaches were reviewed in section 3. A new statistical FS technique (Hybrid DIAAF/GSSC) was consequently introduced in section 4, which integrates the ideas of DIAAF and GSSC. From the experimental results, it can be seen that the proposed Hybrid DIAAF/GSSC approach outperforms existing mechanisms regarding the DelSNcontGO (statistical) "bag of phrases" DR setting and the TFPC associative text classification. This in turn improves the performance of language-independent text classification.

The results presented in this paper corroborate that the traditional text classification problem can be solved, with good classification accuracy, in a language-independent manner. Further research is suggested to identify the improved statistical textual data pre-processing approach in terms of (statistical) document-base representation and (statistical) feature selection, and improve the performance of language-independent text classification.

Acknowledgment. The authors would like to thank Professor Paul Leng of the Department of Computer Science at the University of Liverpool, and Dr. Jiongyu Li and Fan Li of the Information Management Center in the China Minsheng Banking Corp., Ltd. for their support with respect to the work described here.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, May 1993, pp. 207–216. ACM Press, New York (1993)
2. Ali, K., Manganaris, S., Srikant, R.: Partial Classification using Association Rules. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, CA, USA, August 1997, pp. 115–118. AAAI Press, Menlo Park (1997)
3. Antonie, M.-L., Zaïane, O.R.: Text Document Categorization by Term Association. In: Proceedings of the 2002 IEEE International Conference on Data Mining, Maebashi City, Japan, December 2002, pp. 19–26. IEEE Computer Society, Los Alamitos (2002)

4. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. In: Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, Vancouver, BC, Canada, pp. 76–83. Association for Computational Linguistics (1989)
5. Coenen, F., Leng, P., Zhang, L.: Threshold Tuning for Improved Classification Association Rule Mining. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS, vol. 3518, pp. 216–225. Springer, Heidelberg (2005)
6. Coenen, F., Leng, P., Sanderson, R., Wang, Y.J.: Statistical Identification of Key Phrases for Text Classification. In: Proceedings of the 5th International Conference on Machine Learning and Data Mining, Leipzig, Germany, July 2007, pp. 838–853. Springer, Heidelberg (2007)
7. Cohen, W.W.: Fast Effective Rule Induction. In: Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, July 1995, pp. 115–123. Morgan Kaufmann, San Francisco (1995)
8. Deng, Z.-H., Tang, S.-W., Yang, D.-Q., Zhang, M., Wu, X.-B., Yang, M.: Two Odds-radio-based Text Classification Algorithms. In: Proceedings of the Third International Conference on Web Information Systems Engineering Workshop, Singapore, December 2002, pp. 223–231. IEEE Computer Society, Los Alamitos (2002)
9. Fano, R.M.: Transmission of Information – A Statistical Theory of Communication. The MIT Press, Cambridge (1961)
10. Fragoudis, D., Meretaskis, D., Likothanassis, S.: Best Terms: An Efficient Feature-Selection Algorithm for Text Categorization. Knowledge and Information Systems 8(1), 16–33 (2005)
11. Fuhr, N.: Models for Retrieval with Probabilistic Indexing. Information Processing and Management 25(1), 55–72 (1989)
12. Fuhr, N., Buckley, C.: A Probabilistic Learning Approach for Document Indexing. ACM Transactions on Information System 9(3), 223–248 (1991)
13. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. In: Borbinha, J.L., Baker, T. (eds.) ECDL 2000. LNCS, vol. 1923, pp. 59–68. Springer, Heidelberg (2000)
14. Kobayashi, M., Aono, M.: Vector Space Models for Search and Cluster Mining. In: Berry, M.W. (ed.) Survey of Text Mining – Clustering, Classification, and Retrieval, pp. 103–122. Springer, New York (2004)
15. Li, W., Han, J., Pei, J.: CMAR: Accurate and Efficient Classification based on Multiple Class-association Rules. In: Proceedings of the 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, November-December 2001, pp. 369–376. IEEE Computer Society Press, Los Alamitos (2001)
16. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, August 1998, pp. 80–86. AAAI Press, Menlo Park (1998)
17. Mladenic, D.: Text-learning and Related Intelligent Agents: A survey. IEEE Intelligent Systems 14(4), 44–54 (1999)
18. Ng, H.T., Goh, W.B., Low, K.L.: Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, July 1997, pp. 67–73. ACM Press, New York (1997)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco (1993)
20. Salton, G., Wong, A., Yang, C.S.: A Vector Space Model for Automatic Indexing. Information Retrieval and Language Processing 18(11), 613–620 (1975)

21. Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing & Management 24(5), 513–523 (1988)
22. Scheffer, T., Wrobel, S.: Text Classification beyond the Bag-of-words Representation. In: Proceedings of the Workshop on Text Learning, held at the Nineteenth International Conference on Machine Learning, Sydney, Australia (2002)
23. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)
24. Shidara, Y., Nakamura, A., Kudo, M.: CCIC: Consistent Common Itemsets Classifier. In: Proceedings of the 5th International Conference on Machine Learning and Data Mining, Leipzig, Germany, July 2007, pp. 490–498. Springer, Heidelberg (2007)
25. Wang, Y.J., Sanderson, R., Coenen, F., Leng, P.H.: Document-Base Extraction for Single-Label Text Classification. In: Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, Turin, Italy, September 2008, pp. 357–367. Springer, Heidelberg (2008)
26. Wiener, E., Pedersen, J.O., Weigend, A.S.: A Neural Network Approach to Topic Spotting. In: Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, USA, April 1995, pp. 317–332 (1995)
27. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning, Nashville, TN, USA, July 1997, pp. 412–420. Morgan Kaufmann Publishers, San Francisco (1997)
28. Yin, X., Han, J.: CPAR: Classification based on Predictive Association Rules. In: Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, May 2003, pp. 331–335. SIAM, Philadelphia (2003)
29. Yoon, Y., Lee, G.G.: Practical Application of Associative Classifier for Document Classification. In: Lee, G.G., Yamada, A., Meng, H., Myaeng, S.-H. (eds.) AIRS 2005. LNCS, vol. 3689, pp. 467–478. Springer, Heidelberg (2005)
30. Zheng, Z., Srihari, R.: Optimally Combining Positive and Negative Features for Text Categorization. In: Proceedings of the 2003 ICML Workshop on Learning from Imbalanced Data Sets II, Washington, DC, USA (2003)