

# Effect of Document Representation on the Performance of Medical Document Classification

Fathi H. Saad<sup>1</sup>, B. de la Iglesia<sup>1</sup>, and G. D. Bell<sup>2</sup>

<sup>1</sup>School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

<sup>2</sup>Endoscopy Unit, Norfolk and Norwich University Hospital, Colney Lane, Norwich NR4 7UY

**Abstract**— Text classification in the medical domain is a real world problem with wide applicability. This paper investigates extensively the effect of text representation approaches on the performance of medical document classification. To accomplish this objective, we evaluated seven different approaches to represent real word medical documents. The text representation approaches investigated in this paper are basic word representation (bag-of-words), key-phrases, collocation extracted from preprocessed text, collocation extracted from post-processed text, single-word-nouns, combination of single-word-noun and adjectives and combination of single-word-noun, adjective and verbs. A set of experiments was conducted to make comprehensive evaluation of the effects of these representation approaches using real world medical documents by measuring the classification performance. We measured classification performance with information retrieval metrics; precision, recall, F-measure and accuracy. Our experimental results show that bag-of-words representation outperforms all other representation approaches. In addition, careful use of selected features improve the classification performance.

## I. INTRODUCTION

Electronic medical data is often presented either fully or partially in the form of free-text (e.g. medical reports attached to patients' records). Medical text documents represent a significant source of clinical data, especially data that are not available in coded electronic form. Such documents contain important information about patients, disease progression and management, but are difficult to analyse with conventional data mining techniques due to their unstructured or semi-structured nature. Medical staff may have a number of interesting and highly clinically relevant questions that can be asked of such data, but do not have a readily available automated method for reading, categorising and analysing what might amount to hundreds or even thousands of electronic patients' reports.

The Gastroenterology unit in Norfolk and Norwich University Hospital had just such a problem as they collected electronic reports on thousands of colonoscopy procedures, but could not give answer to simple questions, such as the percentage of successful colonoscopies undertaken. Colonoscopy refers to the

passage of the colonoscope from the lowest part (anus and rectum) right around the colon to the caecum and in some cases into the terminal ileum via the ileo-caecal valve. This constitutes a complete or total examination. The aim of colonoscopy is to check for medical problems such as bleeding, colon cancer, polyps, colitis, etc [6]. After each colonoscopy procedure, the endoscopist at the NNUH would generate a detailed report about the current status of the examined part of the body and the result of the procedure itself using the Endoscribe database. There was a section of the report for free text and general comment.

The information contained in the Endoscribe generated report was extremely valuable for clinical purposes but difficult to handle electronically due to the lack of structure. Colonoscopy can be classified as successful or unsuccessful depending on whether the endoscopists had or had not been able to examine the whole or only part of the colon. The reasons for an incomplete examination such as poor bowel preparation, an obstructing cancer, patient intolerance etc also needs to be monitored as does polyp detection rates. Classifying colonoscopy procedure reports into categories such as successful or unsuccessful intubation to the caecum or polyp detected (yes/no) are document classification tasks.

This is a common problem in medical data. For example, a recent study distinguishing planned and unplanned readmissions shows that coded medical information alone was not sufficient for classifying admissions, and that information in text reports significantly improved the classification [1].

Text classification is the process of labelling unstructured text with categories from a predefined set. The main approach to text classification is based on machine learning, where a general inductive process automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents. In document classification, the classification process consists of four steps.

*Step 1:* dividing the documents set into train and test documents set.

*Step 2:* transforming the content of the document into a set of selected features that is suitable for classification methods to use. The way in which text is represented

(selected features) has a strong impact on the performance of text classification systems [2], [28], [35]. That means, a good representation should allow documents to be efficiently classified with high accuracy. The features could be words, phrases, names or other linguistic structures. In this step, the number of features could also be reduced by removing stop words and/or by stemming.

*Step 3:* training the classifier in order to build the classification model. In this step, the features are usually weighted in order to indicate their importance within the document and their distribution among the documents. Then, it is necessary to adopt a model to view the documents and their features. There are various models to view the documents and their features such as vector-space models [21], similarity-space models [22] or graph-based models [19]. The last task of the third step is to train the classification algorithm using labelled and unlabelled document sets to build the classifier or the classification model.

*Step 4:* applying the built classifier in order to evaluate the performance of the classification model, if sufficient, the model can be applied to the characteristics of new documents.

The document representation step, is concerned with selecting a *good* set of features, i.e. one that will give improved classification performance. A naïve approach would be to use all features in the documents thus avoiding the feature selection problem. This may be problematic for a number of reasons, for example, the features are not equally useful and weak features may deteriorate the classifier's performance. Also, using all features may result in a less efficient classifier. There are a number of general guiding principles for choosing the features to represent a document:

- 1) Irrelevant aspects of a document should be ignored;
- 2) Aspects of a document which are common across many unrelated documents should be ignored;
- 3) Aspects of a documents which are likely to reflect relevance should be retained;
- 4) Information about a document which requires expensive processing should be ignored.

The problem of text classification within the medical domain is both an important and challenging one. Because the medical documents are contextually rich and grow rapidly, manual categorization is necessarily time consuming. Developing tools that automate or semi-automate the process of classifying medical documents has drawn great research interest.

It is known that the performance of a generated classifier depends greatly on the actual representation of the text to be classified [35]. In domain-specific text

classification such as medical domain, finding the optimal representation approach that leads to the best classification performance is non trivial. The main goal of this paper is therefore to answer the following question: which linguistic components of a medical document should be represented for the task of documents classification?

In this paper, we investigate the effects of different text representation approaches on the classification performance. The classification task here is to classify colonoscopy procedures reports into successful and unsuccessful classes. To conduct the set of experiments for the evaluation, we use a text classification system that supports different representation approaches. Four classification performance measures are used to measure the classification performance for the seven approaches.

## II. RELATED WORKS

Different representation approaches were evaluated in many studies. The bag-of-words (which is described with the other approaches in the following section) is the common approach taken by many researches to represent the text. Some studies [3], [27], [29], [30] concluded that effectiveness would not be improved if representations more sophisticated than bag-of-words are used. However, this kind of representation is not recommended by some other studies [24], [34]. Retrieval experiments conducted in [32] reported that the phrase-based representation was superior to bag-of- words representation for vector space model. The authors did not test their representation for machine learning algorithms used for text classification. Encouraging results were obtained in [7] when documents were represented using phrases as features. Collocation representation is considered in some studies [8], [9], [10] and applied in some real-world text mining applications [11], [12], [14]. Combination of some selected single words, namely combining nouns, verbs, adjectives and adverbs, were studied in [13]. The use of medical domain knowledge and a natural language processor to extract medical concepts to be used as an alternative to bag-of-words is investigated in [31].

Examination of the literature suggests bag-of-words are strongly recommended. Disparity of opinion over the best approach with some others, but not all strongly recommend bag-of-words. However, many of these have not been tested for the representation of medical text for classification task, the only exception is [31]. Given the reported importance of document representation on the classification performance [35] and the importance of the real world problem being addressed, we set out to comprehensively evaluate the options available.

### III. REPRESENTATION APPROACHES

We comprehensively evaluate seven representation approaches in this paper. The first representation is based on free text alone. The following three representations are based on using natural language processing which required tokenising, part-of-speech tagging (POS) and phrases/collocations identification and extraction. The last three representations are based on limited natural language processing which required only tokenising and tagging features.

The first two representations were included in our evaluation because of the strong recommendation of many studies. The effectiveness and efficiency of representing text as collocations is showed in many studies and used in some real world text applications. In addition, they have never been evaluated to represent medical text. The last three representations are introduced, as they have not been tested on a medical domain. The seven approaches are briefly described below:

1) *Bag-of-words*: The most frequently used method to represent text is *bag-of-words* representation where all words from the set of documents are taken and no ordering of words or any structure of text is used [3]. Each distinct word corresponds to a feature with a weight as its value that is correlated to the number of times the word occurs in the document. The main advantages of this representation are simple and easy.

2) *Bag-of-phrases*: The bag-of-words representation approach has two problems. First, since each distinct word corresponds to a feature, this increases the dimensionality of feature space which is known to have a negative effect on the classification performance. For example, *distal descending colon* correspond to one feature using bag-of-phrases approach and correspond to three features in bag-of-words approach. The second problem is that information is lost due to feature splits; we may lose stronger features. In addition, the percentage of long phrases, such as *distal descending colon*, in medical vocabulary is very high. This high prevalence of phrases represents a problem for medical text classification.

There are two well-known methods to extract phrases: statistical and syntactic methods [5], [34]. In the statistical method, if a pattern occurs often, it is probably a phrase. The phrases are selected by corpus term frequency or document frequency. There are many approaches e.g., based on word co-occurrences, mutual information, partial parsing techniques. The statistical method is very fast and reasonably accurate. The syntactic method uses part of speech taggers and the extracted phrases are called syntactic phrases. First we

assign POS tags to terms using POS tagger (there are many approaches for POS tagging such as probabilistic, rule-based...etc.) Then we match phrases by POS patterns. The method is reasonably fast but slower than the statistical method. However, it is more accurate. We use syntactic phrases [5] for representation because syntactic phrases are often better indicators of content than statistical phrases [34].

3) *Collocations from original text*: We used a *Perl* program to extract collocations from original text. The maximum length of the extracted collocation is four words. The collocations that do not occur more than two times were removed. Neither stop words removal nor stemming was used to extract the collocations. We called this kind of representation *Collo\_Orig*, which means extracting collocation from original documents.

4) *Collocations from words*: In this case the stop words were removed before extracting the collocations. This kind of representation is called *Collo\_Words*.

To guarantee accurate selection and extraction of the last three approaches, a part of speech tagger program is needed to assign tags to each word. For this purpose, five part of speech taggers were evaluated. We used *TreeTagger* [6] which achieved the best accuracy. The last three approaches are:

5) *Single\_Word\_Nouns* (NN): In this approach, only singular (e.g., Colon) and plural (e.g., Polyps) single word nouns were extracted. We named this approach NN.

6) *NN\_Adjectives* (NN\_JJ): In addition to NN, we extracted adjectives to form this approach which we named it NN\_JJ. Only adjective (e.g., large), comparative adjective (e.g., larger) and superlative (e.g., largest) adjective were extracted.

7) *NN\_Adjective\_Verbs* (NN\_JJ\_VV): All kinds of verbs were extracted in addition to NN\_JJ to form this approaches. This approach was named NN\_JJ\_VV. Examples of extracted verbs biopsied, taken, bleeding, remains...etc

The details of the techniques used to extract the last six approaches in addition to the result of the comparison of six POS taggers were omitted due to space limitation.

### IV. EXPERIMENTAL SETUP

#### A. Document set

For these experiments we used real world medical documents collected from the Gastroenterology unit in Norfolk and Norwich University Hospital. These documents contain information on colonoscopy procedures including preparation of the bowel, features of the colon identified in examination, abnormalities found

during examination with their description, patient's reaction to the procedure, etc.

The procedure can be classified as successful or unsuccessful depending on what the clinicians claim they have been able to examine and the reasons for any limited examinations. The number of documents in this collection is 4,876. 25% of these documents were selected using *1-in-4 include* sampling strategy to be used as test documents. The rest (75%) were used to create training sets. 120 documents from the positive class were selected as the positive set from the training set. The rest of the documents were used as the unlabelled set. Table 1 below shows the number of documents in each class for the two data sets as classified by expert in the domain field and we used them as "gold" standard.

TABLE 1  
STATISTICS SHOWS THE NUMBER OF DOCUMENTS IN EACH CLASS FOR THE TWO DATA SETS

	Train dataset (NS)	Test dataset (TS)
All Docs.	3,657 (75%)	1,219 (25%)
Successful	3,178 (86.9% of NS)	1,042 (85.5% of TS)
Failed	359 (9.8% of NS)	177 (14.5% of TS)
Positive set	120 (3.3% of NS)	0

### B. Documents Pre-Processing

Not all the words in the documents are important, so they may degrade the classifier's performance. In addition, representing small set of documents that may have hundreds of different words using *bag-of words* will generate a huge feature space and thus affect the classification performance negatively. To solve these problems, methods to reduce the feature space dimension are needed. We used three methods:

- 1) As a result of consulting an expert in the domain field, we removed unhelpful sentences from the documents such as "Informed consent was obtained with the benefits, risks and alternatives for the procedure explained", which is found in all reports;
- 2) We have removed stop words from all data sets using stop-lists containing common words such as "the", "a", "an";
- 3) We stemmed the words using Porter's suffix-stripping algorithm [3]. Words are considered the same if they share the same stem.

### C. Performance Measure

Four evaluation measures were used to evaluate the performance of the classifier for each approach: *recall*, *precision*, *F-measure* and *accuracy* [23]. *Precision* is the percentage of correctly identified positive documents over those classified as positive (Equation 2). *Recall* is the percentage of correctly identified positive documents

over all positive documents (Equation 1). *Accuracy* is the ratio of correct classification for the overall document set (Equation 4). The F-measure has been proposed to balance recall and precision by giving them equal weights (Equation 3). Therefore, for the evaluation of text classifiers, precision and recall need to be used in conjunction with the F-measure and/or accuracy.

To calculate these measures, we first found the values of the following parameters for each classifier:

TP: number of the documents correctly assigned to the positive class

FP: number of the documents incorrectly assigned to the positive class

FN: number of the documents incorrectly rejected from the positive class

TN: number of the documents correctly rejected from the positive class

Then, we calculated precision and recall for each of the classifiers with the following formulas:

$$Precision (P) = TP / (TP + FP) \quad (1)$$

$$Recall (R) = TP / (TP + FN) \quad (2)$$

$$F-Measure = 2 * P * R / (P + R) \quad (3)$$

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

### D. Classifying Documents

The classification technique used in these set of experiments is called *partially supervised classification* which is effective and computationally efficient [20]. This technique is different from the traditional classifications techniques. In traditional binary classification, a text classifier is built using a classification learning algorithm and a set of labelled (often manually) documents [25]. This approach, called supervised learning [26], requires considerable effort to manually label a large number of training examples for the two classes (positive and negative). The effort is larger for multi-class problems. On the other hand, partially supervised classification alleviates some labour-intensive effort as it is based on the use of a large set of unlabelled documents and a small set of labelled documents for the class of interest (positive class). No documents labelling is required for the negative class, as negative documents will be identified automatically from the set of unlabelled documents.

The technique we used, was developed based on study reported in [33] which shows that using both labelled and unlabelled documents is better than using the small labelled set alone. Using this approach, we build a two-class or binary classifiers with only positive and unlabelled examples and no negative examples. This

approach is based on two-step strategy. The first step identifies a set of reliable negative documents from the unlabelled set to create the negative class. The second step builds a set of classifiers by iteratively applying a classification algorithm and then selects a good classifier from the set of generated classifiers. The tool used in this paper which applies this approach is called *Spy-Expectation Maximization* (S-EM) [20]. The two steps of S-EM are described as follow:

Step 1: A technique called *Spy* is introduced to select reliable negative *RN* documents from unlabelled set *U*. It first randomly selects a set *S* of positive documents from *P* and puts them in *U*. Then the documents in *S* act as spies documents from the positive set to the unlabelled set *U*. The spies behave similarly to the unknown positive documents in *U*. Hence, they allow the algorithm to infer the behavior of the unknown positive documents in *U*. It then runs EM algorithm using the set *P-S* as positive and the set *U-S* as negative. EM basically runs Naïve Bayesian (NB) twice. After EM completes, the resulting classifier uses the probabilities assigned to the documents in *S* to decide a probability threshold *th* to identify possible negative documents in *U* to produce the set *RN*.

Step 2: building a set of classifiers using EM algorithm. Basically, EM iteratively runs NB to revise the probabilistic label of each document in set  $Q = U - RN$ . Since each iteration of EM produces a NB classifier, S-EM has a mechanism to select a good classifier from the set of classifiers produced by NB. The EM algorithm is shown bellow:

- 1) Each document in *P* is assigned the class label 1;
- 2) Each document in *RN* is assigned the class label -1;
- 3) Each document  $d \in Q (= U - RN)$  is not assigned any label initially. At the end of the first iteration of EM, it will be assigned a probabilistic label,  $Pr(1|d)$ . In subsequent iterations, the set *Q* will participate in EM with its newly assigned probabilistic classes.
- 4) Run the EM algorithm using the document sets, *P*, *RN* and *Q* until it converges.

## V. EXPERIMENTAL METHODOLOGY

Seven approaches were included in our experiments and explained in the document representation subsection. All the features in each one of the seven approaches are weighted. There are many weighting feature schemes to indicate their importance within the document and their distribution among the documents such as binary [15], term frequency – inverse document frequency (*tf-idf*) [4], [16], term frequency – inverse document frequency using logistic function (*tf-idf(ls)*) [17] and entropy [18]. In our experiments the features were weighted using the most

well known method, *tf-idf*. Once the features weights are computed, the *Viewing Model* should be used to view the documents with their features' weight. In our experiments we viewed the documents using the most commonly used model which is vector-space model [21]. In this model, each document has an associated vector, which is expressed as a term-by-document-matrix. Each entry in the matrix represents the importance (measured using *tf-idf* scheme) of a particular feature in a certain document. When the vector-space matrix is constructed, the classification algorithm is applied to build the classification model (the classifier) using the train data set (positive and unlabelled sets). Finally, the classifier is used to classify the test set to measure the classification performance.

## VI. RESULTS AND ANALYSIS

A primary concern of ours was to experimentally evaluate the effect of using various approaches for the purpose of document representation on medical document classification performance. Based on the definitions of recall and precision, the highest value of either of them does not mean good overall performance. The classifier that did not assign any document to the positive class could have a perfect precision but low recall. Conversely, if the classifier assigns all documents to the positive class it may give 100% recall but unacceptably low precision. This can be clearly noted from Table 3 in the case of NN and bag-of-phrases, where both approaches produced the highest precision results 98.37 and 94.07 respectively, but in contrast they produced the lowest recall results 68.36 and 62.71. For this reason the comparison will be based on the F-measure and accuracy results. Table 3 shows the recall, precision, F-measure and accuracy results obtained by the classification algorithm using the seven approaches under investigation. There are many observations can be made by analyzing Table 3. The big difference in the F-measure and accuracy results based on the highest and lowest percentages obtained (F-measure: 85.88%-67.46% = 18.42%, accuracy: 95.98%-88.84% = 7.14%) is emphasising that the performance of the classifier is greatly depends on the actual representation of the text to be classified. The bag-of-words approach outperforms all other approaches in terms of F-measure and Accuracy results. This coincides with the results obtained in [27] where the authors include bag-of-words, phrases, clusters of words and clusters of phrases in their evaluation. The second approach that performs well is collocations extracted after removal of stop words. The approach that performed the worst in terms of F-measure and Accuracy is the collocations extracted from the original text. The

second worst performer is the bag-of-phrases. While collocations extracted from original text and phrases are less ambiguous than individual words, they are not good features to represent medical documents. Two of our proposed features, namely *NN* and *NN\_JJ* perform equally in terms of accuracy, and they obtained very competitive Accuracy results.

TABLE 3  
RECALL, PRECISION, F-MEASURE AND ACCURACY RESULTS OBTAINED VIA THE CLASSIFIER FOR SEVEN APPROACHES FOR DIFFERENT DOCUMENT REPRESENTATIONS

Approach	Recall	Precision	F-Measure	Accuracy
Bag-of-Words	84.18	87.65	85.88	95.98
Collo_Orig.	79.66	58.51	67.46	88.84
Collo_Words	85.31	82.97	84.12	95.32
Phrases	62.71	94.07	75.25	94.01
NN	68.36	98.37	80.67	95.24
NN_JJ	76.84	83.95	80.24	94.50
NN_JJ_VV	83.05	84.	83.52	95.24

Using the approach that achieved the best results in terms of F-measure and Accuracy, the bag-of-words, another set of experiments was conducted to attempt to improve the document classification accuracy by applying feature reduction method. As shown in table 2, the final total number of distinct features in the collection is 2,636. The frequencies of these features vary from the highest frequency 7,111 to the lowest frequency 1. 1,124 of these features occurred only once. The same experimental methodology used above was repeated with a reduced feature set using *bag-of-words* approach. In each case, only the  $\gamma$  top features according to their frequency are selected to build the classifier. The four values of  $\gamma$  used are 100, 200, 300 and 500. Table 4 shows the achieved accuracy and F-measure values for these sets or experiments.

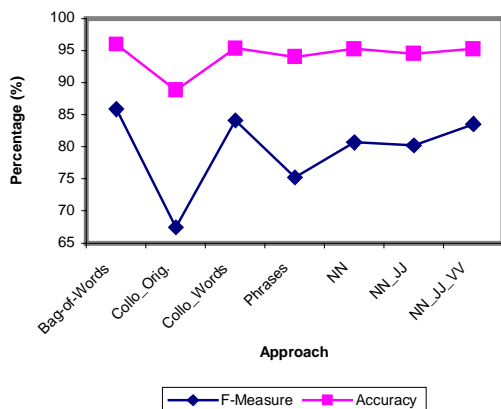


Fig. 1. F-measure and accuracy results obtained via the classifier for seven t approaches for document representation

TABLE 4  
F-MEASURE AND ACCURACY RESULTS OBTAINED BY THE CLASSIFICATION ALGORITHM FOR ALL, TOP 100, TOP 200, TOP 300 AND TOP 500 FEATURES FOR DOCUMENT REPRESENTATION

	Recall	Precision	F-measure	Accuracy
All features	84.18	87.65	85.88	95.98
$\gamma =$ Top 100 features	82.02	66.36	73.37	91.31
$\gamma =$ Top 200 features	85.30	88.49	86.87	96.20
$\gamma =$ Top 300 features	84.27	86.71	85.47	95.82
$\gamma =$ Top 500 features	84.75	87.72	86.21	96.06

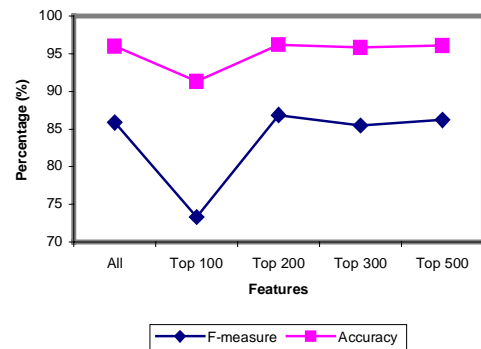


Fig. 2. F-measure and accuracy results obtained by the classification algorithm for all, top 100, top 200, top 300 and top 500 features for document representation

Using the top 100 features significantly degraded the classification accuracy. This indicates that a set of 100 top features is too small to represent the collection of documents. The results obtained using the top 200 features slightly improve the classification accuracy. Larger feature sets ( $\gamma=300$  and 500) did not provide significantly improved results and in some cases produced slightly worse results. The main observations from the last set of experiments are:

- 1) Selecting a reduced set of features to represent the documents can improve the performance of the classifier as measured by the F-measure and Accuracy;
- 2) A very reduced feature set may affect the classification performance;
- 3) Finding a sufficient set of features can improve the accuracy while also increasing efficiency, but it may require some experimentation.

## VII. CONCLUSION

In this paper, seven approaches to document representation are experimentally studied and evaluated in order to determine the most appropriate representation to improve the classification performance. Our experimental results show that bag-of-words outperforms all other approaches. This experimentation is performed using real-world medical documents. It was found that the choice of features used to represent the document

impacts the classification accuracy to a large extent. Practically, we managed to reach 96.2% classification accuracy.

### VIII. FUTURE WORK

We are already working to develop the approach further. This includes investigating other approaches such as noun-phrases. Evaluating other feature selection approaches such as mutual information and information gain instead of using the term frequency. The introduction of additional features to improve the classification performance. We are going to test the approach validity by investigating other clinical problems such as identifying polyps using different medical documents.

### REFERENCES

- [1] M. P. Kossovsky, F. P. Sarasin, F. Bolla, M. Gaspoz, and F. Borst, "Distinction between planned and unplanned readmissions following Discharge from a Department of Internal Medicine". *Methods Inf. Med.* 38(2): 140-3, 1999.
- [2] T. Menon, L. H. Tong, S. Sathyankeerthi, and A. Brombacher, "Automated Text Classification for Fast Feedback – Investigating the effects of Document Representation". *Lecture Notes in Computer Science*. Springer-Verlag. ISSN 0302-9743. Vol. 2774, pp. 1008-1014, 2003
- [3] H. Benbrahim, and M. A. Barmer, "Neighborhood Exploitation in Hypertext Categorization", In *Research and Development in Intelligent Systems XXI*, Springer-Verlag, 2005.
- [4] M. F. Porter, "An algorithm for suffix stripping", *Program; automated library and information systems*, 14(3), pp. 130-137, 1980.
- [5] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning, "Domain-specific keyphrase extraction". In: *Proceedings of the Sixteenth Int. Joint Conference on Artificial Intelligence*. Morgan Kaufmann. San Francisco, CA, 1999
- [6] H. Schmid, "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of the Conference on New Methods in Language Processing*. Manchester, 1994.
- [7] O. Zamir, and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration". In *Proceedings of the 21st Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
- [8] A. L. Gorin, B. A. Parker, R. M. Sachs, and G. Wilpon, "How May I Help You". In *Proceedings of the Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, pp 57-60, 1996
- [9] I. Dagan, L. Lee, and F. C. N. Pereira, "Similarity based methods for word sense disambiguation". In *Proceedings of Eighth Conference of the European Chapter of the Association for Computational Linguistics*. New Jersey, pp 56-63, 1997
- [10] J. Chu-Carroll, and B. Carpenter, "Vector-based Natural Language Call Routing", *Computational Linguistic Journal*, Vol. 25 #3, 1999.
- [11] S. J. Cox, and B. Shahshahani, "Improved techniques for automatic call-routing". In *Institute of Acoustics Workshop on Innovation in Speech Processing*, 2001.
- [12] V. N. Polyakov, and V. V. Sinitin, "Rubryx: Technology of Text Classification Using Lexical Meaning Based Approach", in *Proc. of Int. Conference Speech and Computer*. SPECOM-2003. Moscow, MSLU, pp. 137-143, 2003. Available at [www.sowsoft.com/rubryx/](http://www.sowsoft.com/rubryx/)
- [13] J. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, "Effectiveness of Simple Linguistic Processing in Automatic Sentiment Classification of Product Reviews". In *proceedings of the Eighth Int. Knowledge Organisation and the Global Information Society Conference*. London, 2004
- [14] V. N. Polyakov, and V. V. Sinitin, "Method Automatic Classification of Web-resource by Patterns" in *Text Processing and Cognitive Technologies*. Paper Collection. Issue 6. Kazan, Otechestvo, pp. 120-126, 2001.
- [15] G. Salton, and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval", *Information Processing and Management*. 24(5), pp. 513-523, 1988.
- [16] G. Salton, and M. McGill, "Introduction to Modern Information Retrieval". McGraw-Hill. 1983.
- [17] D. Pyle, "Data Preparation for Data Mining", Morgan Kaufman, pp. 257-258, 1999.
- [18] S. T. Dumais, "Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval", Technical Report TM-ARH-017527, Bellcore, 1990.
- [19] Y. Zhao, and G. Karypis, "Criterion functions for document clustering: Experiments and analysis". Technical Report TR #01-40, Department of Computer Science, University of Minnesota, Minneapolis, MN, Feb 2002.
- [20] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially Supervised Classification of Text Documents". *Proceedings of the Nineteenth Int. Conference on Machine Learning (ICML-2002)*, Sydney, Australia. 2002.
- [21] G. Salton, "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer". Addison-Wesley, 1989.
- [22] G. Karypis, "CLUTO: A Clustering Toolkit". Technical Report: #02-017. University of Minnesota, Department of Computer Science, 2003.
- [23] Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization". *Journal of Information Retrieval*. Vol. 1 #1/2. Kluwer, pp 68-90, 1999.
- [24] S. Finch, "Partial orders for document representation: a new methodology for combining document features", in *Proceedings of the 18th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, pp. 264-272, 1995.
- [25] G. Cong, W. S. Lee, H. Wu, and B. Liu, "Semi-supervised Text Classification Using Partitioned EM". 11<sup>th</sup> Int. Conference on Database Systems for Advanced Applications (DASFAA), pp 482-493, 2004.
- [26] Y. Yang, and X. Liu, "Are-examination of text categorisation methods". In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42-49, 1999.
- [27] D. D. Lewis, "Feature Selection and Feature Extraction for Text Categorization". In *Proceedings of the Speech and Natural Language Workshop*, Harriman, Morgan Kaufmann, pp. 212-217, 1992.
- [28] D. D. Lewis, "Representation Quality in Text Classification: An Introduction and Experiment". In *Proceedings of the Speech and Natural Language Workshop*, Hidden Valley, Morgan Kaufmann, pp. 288-295, 1990.
- [29] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms for text categorization". In *Proceedings of CIKM*. Bethesda, MD, 1998.
- [30] D. D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task". In *Proceedings of SIGIR*. 1992.
- [31] A. Wilcox, G. Hripesak, and Friedman, C.: "Using Domain Knowledge Sources to Improve Classification of Text Medical Reports". In *Proceedings of ACM SIGKDD Workshop on Text Mining*. 2000.
- [32] W. Mao, and W. W. Chu, "Free-text Medical Document Retrieval via Phrase-based Vector Space Model". In *Proceedings of AMIA*. 2002.
- [33] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Learning to classify text from labelled and unlabelled documents". AAAI-98 (pp. 792-799). Madison, US: AAAI Press, Menlo Park, US. 1998.
- [34] T. Strzalkowski, "Document Representation in Natural Language Text Retrieval". Available on <http://www.inf.ed.ac.uk/teaching/courses/tts/papers/H94-1072.pdf>
- [35] D. D. Lewis, "Text Representation for Intelligent Text Retrieval: A Classification-Oriented View", In Paul S. Jacobs' editors, *Text-based Intelligent Systems: Current Research and Practical in Information Extraction and Retrieval*, Lawrence Erlbaum Associates, p. 179-97, 1992.