

An enhanced text classifier for automatic document classification

Wijewickrema, P. K. C. M¹. and Gamage, R. C. G²

Abstract

Automatic classification has become an important research area due to the exponential growth of digital content in the modern world. Evidently, manual classification of documents is very painstaking and labor-intensive task. It takes much time to organize a collection of documents according to the subject area. This research has developed a computer programme that can automatically classifying a given text document. Therefore, the user gets correct classification results just after feeding the document to the new system. For the process of classification, we use a new algorithm developed by enhancing basic form of an existing text classifier called tf-idf. The results were obtained for classification accuracy of the new text classification algorithm. They were compared with the results obtained for the basic tf-idf classifier. The research revealed that, the newly developed classifier algorithm can obtain better classification accuracy than the basic tf-idf classifier.

Keywords: Automatic classification, Text classification, tf-idf weight function

Introduction

Due to the continued growth of information in both printed and electronic formats, it is becoming extremely difficult to organize text materials in a proper way. Therefore, it is a usual practice of following a standard classification scheme for the purpose of organizing bibliographic materials suitably. Generally, the classification schemes play an outstanding role since they facilitate subject access as well as locating of the available bibliographic

¹ Assistant Librarian, Sabaragamuwa University of Sri Lanka. Sri Lanka. Email: manju@sab.ac.lk

² Senior Assistant Librarian, University of Moratuwa, Sri Lanka. Email: ruga@uom.lk

materials in a library or any other repository. Although the importance of these classification systems attain such an exceeding significance, the overall work of manual classification is very tricky. Hence, one may feel the output of the process will not so worth relative to the contribution as the classifier has to pay a plenty of time and full attention throughout this work. Contrast to the printed documents, the volume of electronic information is rapidly increased due to the higher usage of Internet and Web content available in the modern world. As a result, the difficulty of classifying electronic documents has begun to make huge troubles than classifying the printed documents. Thus, there is an inevitable need for a tool that can classify electronic text documents.

For this purpose, automatic text classification systems have been introduced to automatically assign the electronic documents into appropriately pre-defined categories. Moreover, many classification algorithms have been proposed to accomplish this task and they are known as text classifiers. In general, these algorithms determine how far a given test document¹ matches with the pre-defined categories. Different text classifiers have their own methodologies to match the most related subject category to the given test document. For example, probabilistic methods, distance learning methods, support vector machines, genetic algorithms, hidden Markov models, decision tree methods, decision rule methods, regression methods, neural network methods (Sebastiani, 2002; Tao, Ling, & Cheng, 2005) and tf-idf (term frequency-inverse document frequency) based methods (Abbas, Smaïli, & Berkani, 2010; Tao *et al*, 2005) use different concepts to classify the given documents.

In our study, we mainly focus to enhance the basic form of the tf-idf classifier as it has some limitations. In general, the tf-idf weight function (Salton, Wong, & Yang, 1975) has the ability of numerically representing the importance of a particular word to an entire

¹ Document that we need to classify.

document. For this, it uses the term frequency¹ and the number of documents where the considered term presence at least once. This importance increases proportionally to the number of times a word appears in the document and offset by the number of documents where the particular term appears. However, since the basic form of the tf-idf algorithm categorizes a document based on single key term (highest frequency term) of the document, the results may not be very accurate. Because some of the times, the subject area of a document may not be determined by the highest frequently occurring term. Some other term may have less number of occurrences, yet being a key term of the document. Therefore, it is worthier to consider the effect of a sufficient number of key terms that have higher frequencies than determining the subject based only on the highest frequency term. Moreover, this algorithm does not give any importance for the additional subject-relevant key terms appearing in the test document; if the same term appears among the key terms of a document in the training set². But if the algorithm is able to determine the subject of a document based on more than one keyword and how far they are important (relevant) to that subject in the test document, then it may give a more accurate classification than basic form of the tf-idf algorithm.

In this study, we show that the proposed algorithm more successfully classifies text documents than basic form of the tf-idf algorithm does. Success of the new classifier than the conventional one was evaluated in terms of precision, recall and F_1 measures. All three average values obtained for them were higher in case of the new classifier than the values obtained for the basic tf-idf classifier. Moreover, a pre-prepared computer program known as Lucene API (Application Programming Interface) was used to implement the new algorithm and construct the automatic text classifier. Even though

¹ Number of times a distinct word occurs in a document.

² A collection of pre-classified documents that is used to select the most relevant subject category to the input document.

Lucene has been designed for the purpose of using as an information retrieval search engine, its flexibility has been exploited by this study to adopt the system as a text classifier as well.

In addition to the introduction part, we have organized this paper as follows.

Section 2 of the paper represents an account of some related works to text classification.

We discuss the details of the methodology in section 3. The experimental results of the research are discussed in the section 4. Finally, we have concluded the paper in section 5.

Related work

The area of text classification and classification algorithms have been studying extensively for many years. As a result, a wide range of supervised learning¹ methods has been used in this area.

Using a considerable number of text classifiers can be seen with the development of Internet applications. Dumais and Chen (2000) give an account of one of these kinds of initiatives called hierarchical approach. This study explores the use of hierarchical structure for classifying a vast amount of heterogeneous collection of Web documents. For this purpose, it uses Support Vector Machine (SVM) learning model which was not previously used for hierarchical problems. A probabilistic description-oriented approach (Gövert, Lalmas, & Fuhr, 1999) has also been reported for categorizing of Web content. Here, probabilistic indexing is used and documents are categorized using the k –nearest neighbor (kNN) classifier by giving it a probabilistic interpretation. However, this study has considered the features specific to Web documents as well as standard features of text documents.

¹ Machine is given a pre-defined example documents and the goal of the machine is to learn to produce the matching output for the input document.

Calvo, Lee, & Li (2004) applied automatic Naïve Bayes classifier on news stories from Reuters RCV1 corpus and to another with over 41,000 Web sites. It performed flat multi-label classification using two distinct thresholding strategies called score-based and rank-based. Billsus and Pazzani (1999) report another news classification effort which especially focused on the difference between user's long-term and short-term interests with their dynamic nature.

A classification system for incoming e-mail messages has been introduced by Manco, Masciari, Ruffolo, & Tagarelli (2002). This innovation is based on clustering algorithms for data classification which are extracted from e-mails in an unsupervised way. However, as it does not use a supervised learning method, authors themselves found some difficulties in their work. Crawford, Kay, & McCreath (2001) explores the i-ems¹ (Intelligent Electronic Mail Sorter) project which has developed initiatives to build a system to assist users in managing electronic mail. In order to classify e-mails, it uses learning rules methodology. This system is not a completely automatic one as it does not automatically arrange messages into archive folders. Moreover, Pantel and Lin (1998) report a method known as SpamCop to filter junk emails. This is an example of the use of text classification in addressing the problem of email spam.

Meanwhile, a quite a few projects have been launched to automatically classifying documents based on the classification schemes such as Dewey Decimal Classification (DDC) and Universal Decimal Classification (UDC). According to Toth (2002) and Golub (2006), there have been developed numerous automatic subject classification systems with the extensive idea of categorizing Web documents. Among them, the Nordic WAIS/WWW (Wide Area Information Server/World Wide Web) Project and GERHARD (German Harvest Automated Retrieval and Directory) project were based on the UDC

¹ A project aimed to investigate the automatic induction of e-mail filtering rules.

system while the Online Computer Library Center's (OCLC) project Scorpion built tools for automatic subject recognition, using DDC.

Methodology

Processing

Two major pre-processing phases have been followed in this research. First, the removal of stop words has been taken place to reduce less significant words of the text. Secondly, stemming process is carried out to reduce the number of index terms with the same root. Lucene's default stop words list has been used to remove the stop words. It covers a wide area of stop words that are known in general. However, it is further enriched by comparing with some other available stop word lists.^{1 2} The figure 1 shows a part of the stop words list that has been used in this research.

<i>a</i>	<i>be</i>	<i>everybody</i>
<i>able</i>	<i>because</i>	<i>everyone</i>

Figure 1: Portion of the stop words list used

Whenever it is required, this system allows the user to update the stop words list in order to improve the performances of the system.

Stemming is the next step of pre-processing. For this purpose, Porter's stemming algorithm is used as it is the most commonly used algorithm for word stemming in English language (Smirnov, 2008; Zhou, Smalheiser, & Yu, 2006). A sample of the stemmed words can be given as in the figure 2 along with their original terms.

¹ http://meta.wikimedia.org/wiki/MySQL_4.0.20_stop_word_list

² <http://www.textfixer.com/resources/common-english-words.txt>

Word	Stem
<i>Administered</i>	<i>administ</i>
<i>Clairvoyance</i>	<i>clairvoy</i>
<i>Define</i>	<i>defin</i>

Figure 2: Sample of words and their corresponding stems

Then the frequencies of the remaining words are counted after the removing stop words and stemming are completed. Hence, a list of the remaining terms and their frequencies are generated as the output of this pre-processing phase. This pre-processing scenario can be shown as in the figure 3.

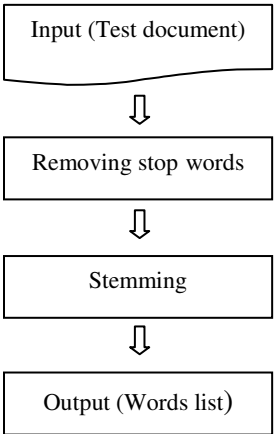


Figure 3: Pre-processing

A considerable amount of text classification approaches have been using term frequencies for selecting the most important words of a text as they have realized that the frequently occurring terms have a close relationship with the subject stream of a document (Khan, Baharudin, Lee, & Khan, 2010; Song, Lim, Kang, & Lee, 2005). This implies that, based on a collection of high frequency terms, one can suggest one or more subject streams where the document can belong. Therefore, after having the output of

the initial stage, now it is easy to recognize and extract such kind of highly related terms using the following process.

Feature Selection

One of the greatest difficulties in text categorization comes with deciding the factors which determine the relationship between the input document and the training documents¹. To determine these relationships; the most important words of the test document is compared with the amount of presence of similar words in each document of the training set. This study has used the tf-idf values of the considered terms for selecting these keywords. Since the current study is taking one test document at a time, the 'idf' value remains constant and only term frequencies are considered. Then the most frequently occurring terms are selected as the most critical terms in determining the relevancies between the test document and the training documents. After determining the critical candidates, it is required to decide how many of them should be considered. This process of determining the proper number and terms which most appropriately describe the subject area of the test document is known as feature selection. In our study, the size of the feature space (the sample of most critical terms) is experimentally determined as follows.

First, we have implemented the new text classifier algorithm (given by the equation (4)) using the Lucene search engine library. Then, 47 electronic test documents have been used for the classification. These documents fall within the subject range of class numbers 110 to 139 of the DDC (edition 21) scheme. Furthermore, these documents were selected without having any detail knowledge of the specific content.

After that they were properly classified by an experienced subject classifier. At this stage, the classifier carefully examined their core content. Then again, the same set of documents was classified using the new automatic classifier. This time, 47 test documents

¹ Pre-classified documents in the training set.

were classified based on 1 to 6 numbers of keywords. That means each document is classified six times by keeping the dimension of the feature space as 1, 2, 3, 4, 5 and 6 for each¹. Here, each document of the test collection was tested for the classification accuracy based on 385 training documents.

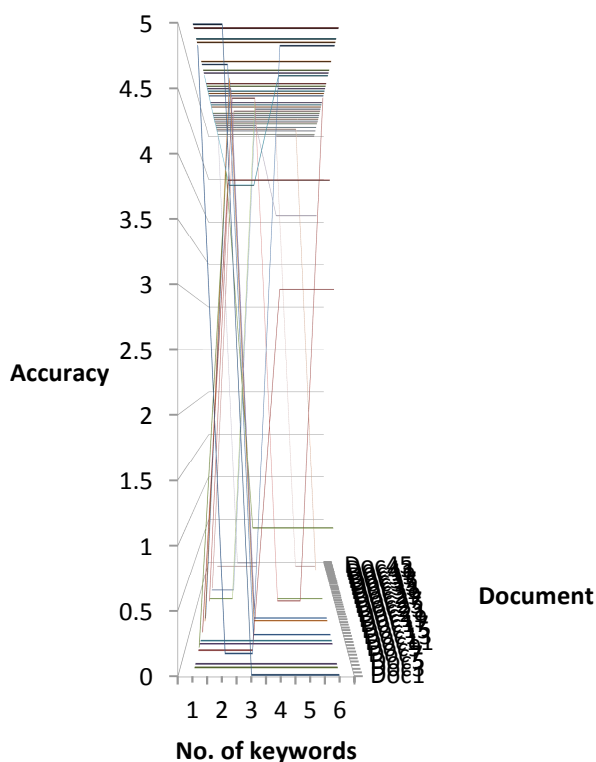


Figure 4: Variation of the classification accuracy with respect to the number of keywords

¹ Since the computation time increases with the number of keywords being considered, it has been limited to 6 terms.

Variation of the accuracy of classification against the number of keywords can be given as in the figure 4. Here, accuracy level 5 was given for an exact matching (i.e. if the subject of the test document is exactly matched with the classification results) while the minimum zero was given for inaccurate classification results. Moreover, accuracy levels 4, 3, and 2 were given for one step super/sub, two steps super/sub and three steps super/sub classification to the exact subject of the test document respectively. In addition, accuracy level 1 was given for the same level of subjects (or their sub classes).

By examining the figure 4, one can conclude that the accuracy of classification considerably goes down with one and three number of keywords. However, it has given some increased accuracy with respect to two keywords. Yet, there are a lot of variations around two and it is not so safe to select this number. With respect to four and more keywords, good and steady classification results have been given. This characteristic is common for most of the test documents. For example, figure 5 shows that the accuracy of classification is increased after considering four keywords for given two individual documents of the test sample.

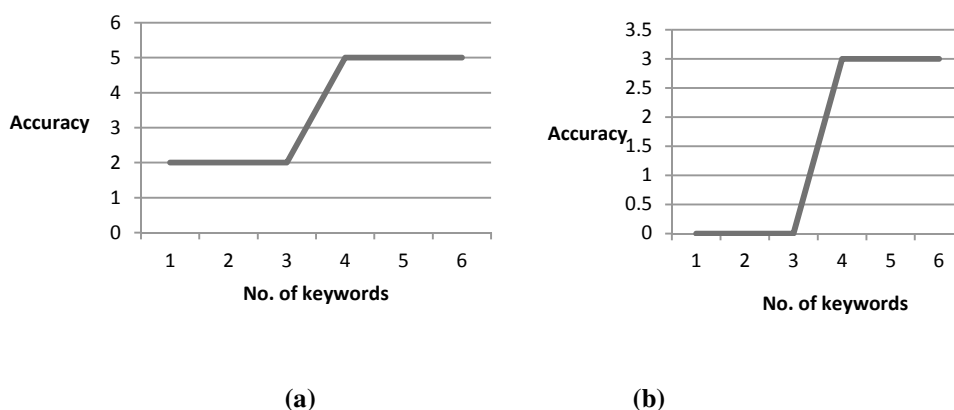


Figure 5: Variation of the classification accuracy of two documents with respect to the number of keywords

Accordingly, in this study, the feature space has been limited to a fixed value of four elements. This threshold value is the minimum value which starts to give the most accurate results for the test documents. As a result of selecting the most appropriate minimum value, now the system is able to do the classification process with least effort and taking less time. Furthermore, the proposed method is more beneficial as it uses a fixed and pre-determined size for the dimension of the feature space. These two facts further reduce the computational time that has to necessarily spend in automatic feature selection.

Scoring

In this study, a new text classifier has been developed using an existing term frequency weight function called the tf-idf weight function. This function has the ability to assign a value for a document based on a few factors. They are; the term frequency, the total number of terms in that document, the number of documents that a particular word occurs in the collection and the number of total documents in the corpus. Accordingly, the tf-idf weight function can also be considered as a basic type of classifier.

This basic tf-idf function is given by the equation (1).

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \quad (1)$$

Where,

$(tf - idf)_{i,j}$: term frequency-inverse document frequency

$tf_{i,j}$: term frequency of the term t_i in the document d_j .

idf_i : inverse document frequency of the term t_i in the collection.

Here, $tf_{i,j}$ and idf_i values can be defined as follows.

$$tf_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \quad (2)$$

Where,

$f_{i,j}$: number of occurrences of the considered term t_i in the document d_j .

$\sum_k f_{k,j}$: sum of the number of occurrences of all terms in document d_j .

and,

$$idf_i = 1 + \ln \left(\frac{|N|}{|\{n: t_i \in N\}| + 1} \right) \quad (3)$$

Where,

$\ln \left(\frac{|N|}{|\{n: t_i \in N\}| + 1} \right)$ is the natural logarithm value of $\left(\frac{|N|}{|\{n: t_i \in N\}| + 1} \right)$.

$|N|$: total number of documents in the collection.

$|\{n: t_i \in N\}|$: number of documents where the term t_i appears.

N denotes the entire collection of documents.

But before using this tf-idf weight as a basic tf-idf classifier, it has to be further developed to find out solutions for the following flaws.

1. This basic form considers tf-idf value only regarding a single keyword. Therefore, it is possible to incorrectly classify the documents when the term with the highest frequency does not correctly imply the subject of a document. Moreover, there can be certain subjects which cannot be decided using only a single key term.
2. Since this function considers only a single key term, it will not be able to consider the importance of the terms with equally higher frequencies.

In order to overcome these two difficulties, the study has recommended the following resolutions. The new text classifier algorithm is built based on these explanations.

1. Instead of considering a single key term, this research focuses on more than one key term.
2. As the threshold value is expanded into four, the text classifier has to be enhanced to consider the importance of up to four key words from the test document and the same from the training set. Moreover, it is also necessary to consider the level of importance of those terms in determining the subject stream of the test document. In order to do this, it is essential to numerically represent how far they gain significance within the test document.

To achieve these goals, we have developed the following new algorithm (4).

$$\text{Document Score} = \{(w_{1,D}) \times (tf - idf)_{1,d}\} + \{(w_{2,D}) \times (tf - idf)_{2,d}\} + \{(w_{3,D}) \times (tf - idf)_{3,d}\} + \{(w_{4,D}) \times (tf - idf)_{4,d}\} \quad (4)$$

Where,

$$\text{Weight of the term } t_i \text{ in test document } D: (w_{i,D}) = \frac{(tf - idf)_{i,D}}{\sqrt{\sum_{k=1}^4 (tf - idf)_{k,D}^2}} \quad (5)$$

Training Sample

Training sample or the training set is a collection of documents where the training documents are stored. By using them, one can compare and determine the subject of an external document. This can be done by evaluating the similarities or relevancies between the test document and the training documents in the training set. In this study, we have used 385 training documents within the domain of philosophy related subjects. Only text documents have been selected since this study focuses only on textual information. Documents for the training set have been selected from the Wikipedia

online encyclopedia, Stanford encyclopedia of philosophy, Google directory and also from the subject gateway of Bulletin Board for Libraries. As the selected documents are not pre-classified they were further examined and classified by experienced subject classifiers according to the DDC scheme. One of the important facts noticeable here is, more than one training document have been selected from the same subject. However, their contents are not similar to each other. Hence, there is more possibility to select the most relevant training document for the given test document as there are multiple documents from the same subject stream.

Results

For the evaluation, 58 test documents belonging to 32 distinct subjects from the selected domain (DDC subject class 110 to 139) were chosen. After selection, they were specifically classified by an experienced subject classifier. These test documents were again classified automatically by using the new classifier and the basic tf-idf classifier. As the training set, 385 pre-classified text documents were used. Finally, the results were evaluated based on the precision, recall and F_1 measures.

In text classification precision can be defined as the fraction of retrieved categories that are relevant.

$$\text{Precision} = \frac{\text{Categories found and correct}}{\text{Total categories found}} \quad (6)$$

Recall is the fraction of relevant categories that are retrieved.

$$\text{Recall} = \frac{\text{Categories found and correct}}{\text{Total categories correct}} \quad (7)$$

A single measure that combines precision and recall is the F_1 measure.

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

The obtained results are given by the table 1 for each subject.

Table 1: Precision, Recall, & F_1 measures obtained for the new algorithm & basic tf-idf classifier

<i>Subject</i>	<i>Precision</i>		<i>Recall</i>		<i>F₁</i>	
	New classifier	Basic tf-idf	New classifier	Basic tf-idf	New classifier	Basic tf-idf
Apparitions	0.40	0.20	0.50	0.25	0.44	0.22
Aries	0.60	0.20	0.75	0.25	0.67	0.22
Attributes-Faculties	0.71	0.60	0.71	0.60	0.71	0.59
Axiology	0.60	0.60	0.75	0.75	0.67	0.66
Causation	0.80	1.00	0.80	1.00	0.80	1.00
Cosmology	0.80	0.60	0.80	0.60	0.80	0.60
Epistemology	0.50	0.40	0.50	0.40	0.50	0.40
Evil Spirits	0.71	0.60	0.71	0.60	0.71	0.59
Feng Shui	1.00	1.00	1.00	1.00	1.00	1.00
Geomancy	0.20	0.20	0.25	0.25	0.22	0.22
Leo	0.80	0.80	1.00	1.00	0.89	0.88
Libra	0.60	0	0.60	0	0.60	0
Love	0.80	0.80	0.80	0.80	0.80	0.80
Mind	0.80	0.40	0.80	0.40	0.80	0.40
Ontology	1.00	0.60	0.83	0.49	0.91	0.54
Other Religion	0.71	0.60	0.71	0.60	0.70	0.59
Palmistry	0.80	0.60	1.00	0.75	0.89	0.66
Phrenology	1.00	1.00	1.00	1.00	1.00	1.00
Pisces	0.60	1.00	0.60	1.00	0.60	1.00
Poltergeists	1.00	0.86	0.71	0.61	0.83	0.72
Precognition	0.80	0.70	0.80	0.70	0.80	0.70
Psychic Phenomena	0.60	0.80	0.50	0.66	0.54	0.72
Psycho Kinesis	1.00	0.33	1.00	0.33	1.00	0.33
Reincarnation	0.80	0.66	0.67	0.55	0.73	0.60
Space	0.80	0.80	0.67	0.66	0.73	0.72
Specific Mediumistic Phenomena	0.20	0.20	0.17	0.16	0.18	0.18

Spells-Curses-Charms	0.47	0.26	0.33	0.19	0.39	0.22
Spiritualism	0.60	0.60	0.75	0.75	0.67	0.66
Taurus	0.40	0.40	0.50	0.50	0.44	0.44
Teleology	0.80	0.80	1.00	1.00	0.89	0.85
Telepathy	1.00	0.40	0.83	0.33	0.91	0.36
Time	0.90	1.00	0.75	0.83	0.82	0.90

In order to determine the precision and recall values; it was considered only five topmost results given by the system. For example, when the value for precision was calculated, the score 1.00 (or 100%) was given when all the five topmost labels that were retrieved matched with the subject of the document which was tested. Furthermore, when there were more than one test document from the same subject the average values were considered. After calculating the precision and recall values; the F_1 values for the same set of test documents has been computed. These values are also given in the table 1.

Table 2: Average values of Precision, Recall, & F_1 measures

Performance Measures	New classifier	Basic tf-idf classifier
Average Precision	0.7127	0.5942
Average Recall	0.7124	0.5953
Average F_1 value	0.7076	0.5862

To depict the way the system classifying documents for both classifiers are shown in the Annexure of the paper. The example shows the classification results obtained for the documents with the subjects Telepathy in psychic phenomena and Mind in philosophy.

According to the values obtained for precision and recall, it is possible to draw the precision-recall curve given in the figure 6. In fact, this graph has been drawn according to

the precision and recall values obtained for the DDC class Spells-Curses-Charms and based on the new classification algorithm. One can notice that, it has the inherent saw-tooth shape of a general Precision-Recall graph.

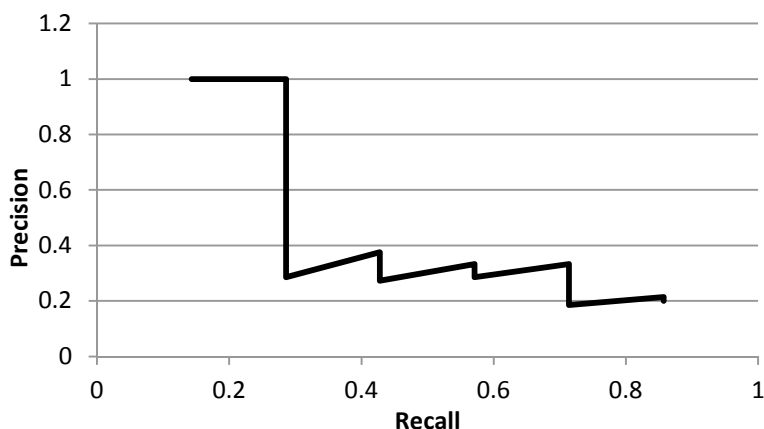


Figure 6: Precision-recall graph for the class Spells-Curses-Charms

Conclusions

This study has developed a new form of text classifier using an existing weight algorithm. The best classification results were given when it was considered topmost four keywords of the test documents. Therefore, four keywords can be considered as the most suited feature space size for the new algorithm.

Then the classification capabilities of the new classifier and the basic tf-idf classifier were evaluated based on the precision, recall and F_1 measures. According to the results, the new system gains 0.7127 average precision, 0.7124 average recall and 0.7076 average F_1 value. Hence, we can conclude that, out of all the retrieved subject categories, 71.27% are related while 71.24% of the relevant categories are retrieved by the new system out of all the relevant subjects in the training set.

The basic tf-idf classifier has gained 0.5942 for average precision. The values 0.5953 and 0.5862 have been obtained for average recall and average F_1 respectively. Considering

these factors, first we can conclude that, the new text classifier is able to classify documents much accurately than the basic tf-idf classifier does. Secondly, it is obvious that, the highest frequency term is not the only factor that correctly determines the subject of a document. But quite a few other high frequency terms as well. Moreover, this depends on the nature of the text classifier.

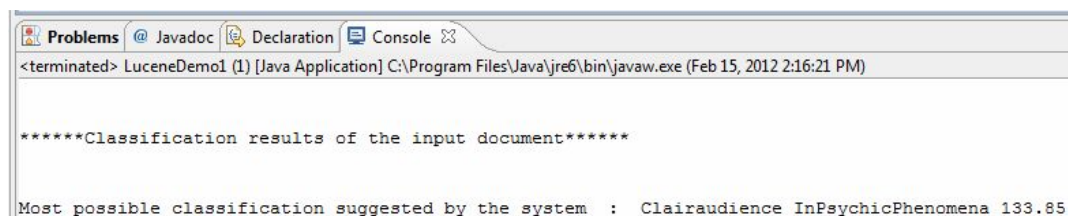
References

- Abbas, M., Smaïli, K., & Berkani, D. (2010). Efficiency of TR-Classifer versus TFIDF. *2010 First International Conference on Integrated Intelligent Computing*, Retrieved from http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5571449 (DOI: 10.1109/ICIIC.2010.60).
- Billsus, D., & Pazzani, M. J. (1999). A Hybrid User Model for News Story Classification. *Proceedings of the 7th International Conference on User Modeling*, Retrieved from <http://www.cs.usask.ca/UM99/Proc/billsus.pdf>
- Calvo, R. A., Lee, J. M., & Li, X. (2004). Managing Content with Automatic Document Classification. *Journal of Digital Information*, 5(2). Retrieved from <http://journals.tdl.org/jodi/article/viewFile/135/133>
- Crawford, E., Kay, J., & McCreath, E. (2001). Automatic Induction of Rules for e-mail Classification. *Proceedings of the 6th Australian Document Computing Symposium*, Retrieved from <http://cs.anu.edu.au/~Eric.McCreath/papers/adcs2001.pdf>
- Dumais, S., & Chen, H. (2000). Hierarchical Classification of Web Content. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Retrieved from <http://research.microsoft.com/en-us/um/people/sdumais/sigir00.pdf>
- Golub, K. (2006). Automated Subject Classification of Textual Web Pages, Based on a Controlled Vocabulary: Challenges and Recommendations. *New Review of Hypermedia and Multimedia*, 12(1), 11-27. Retrieved from <http://homes.ukoln.ac.uk/~kg249/publ/Hypermedia2006.pdf>
- Gövert, N., Lalmas, M., & Fuhr, N. (1999). A probabilistic description-oriented approach for categorizing Web documents. *Proceedings of the 8th International Conference on Information and Knowledge Management*. Retrieved from http://www.is.inf.uni-due.de/bib/pdf/ir/Goevert_etal:99.pdf

- Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1), 4-20. Retrieved from <http://ojs.academypublisher.com/index.php/jait/article/view/01010420>
- Manco, G., Masciari, E., Ruffolo, M., & Tagarelli, A. (2002). Towards An Adaptive Mail Classifier. *AIIA* 2002, Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.18.8794&rep=rep1&type=pdf>
- Pantel, P. & Lin, D. (1998). SpamCop: A Spam Classification & Organization Program. *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, Retrieved from <http://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-017.pdf>
- Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620. Retrieved from http://www.cs.uiuc.edu/class/fa05/cs511/Spring05/other_papers/p613-salton.pdf
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47. Retrieved from <http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf>
- Smirnov, I. (2008). Overview of Stemming Algorithms. Retrieved from <http://the-smirnovs.org/info/stemming.pdf>
- Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005). Automatic Classification of Web Pages based on the Concept of Domain Ontology. *Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05)*, 645-651. Retrieved from <http://portal.acm.org/citation.cfm?id=1122212>
- Tao, Z. Y., Ling, G., & Cheng, W. Y. (2005). An Improved TF-IDF Approach for Text Classification. *Journal of Zhejiang University SCIENCE*, 6A(1), 49-55. Retrieved from <http://www.zju.edu.cn/jzus/2005/A0501/A050108.pdf>
- Toth, E. (2002). Innovative Solutions in Automatic Classification: A Brief Summary. *Libri*, 52(1), 48-53. Retrieved from <http://www.librijournal.org/pdf/2002-1pp48-53.pdf>
- Zhou, W., Smalheiser, N. R., & Yu, C. (2006). A Tutorial on Information Retrieval: Basic Terms and Concepts. *Journal of Biomedical Discovery and Collaboration*, 1(2). Retrieved from <http://www.biomedcentral.com/content/pdf/1747-5333-1-2.pdf>

Annexure

Figure A1 and Figure A2 show the classification results obtained for a test document belongs to the subject Telepathy in psychic phenomena. The same document was classified twice using the basic tf-idf and the new classifier. Here one can notice that how far the classification accuracy has been increased after implementing the new classifier.

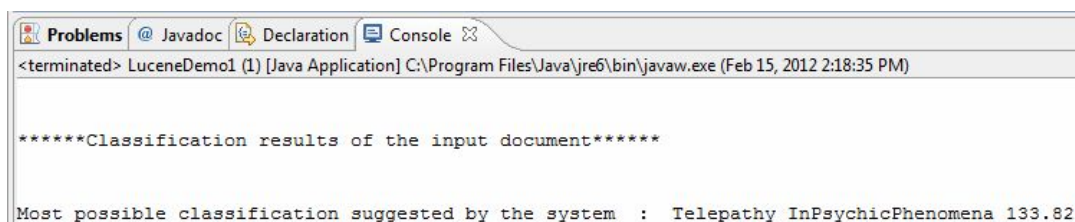


```
Problems @ Javadoc Declaration Console
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:16:21 PM)

*****Classification results of the input document*****

Most possible classification suggested by the system : Clairaudience_InPsychicPhenomena_133.85
```

Figure A1: Classification results obtained for the basic tf-idf classifier



```
Problems @ Javadoc Declaration Console
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:18:35 PM)

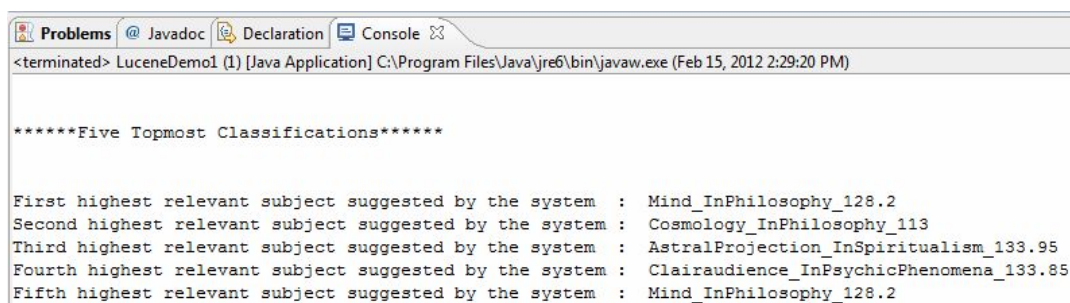
*****Classification results of the input document*****

Most possible classification suggested by the system : Telepathy_InPsychicPhenomena_133.82
```

Figure A2: Classification results obtained for the new classifier

Following figures show the classification results obtained by the system for some test documents.

Figure A3 and Figure A4 give the five topmost results for the basic tf-idf classifier and the new classifier. In these two cases, we have input a document which belongs to the subject area of Mind in philosophy.

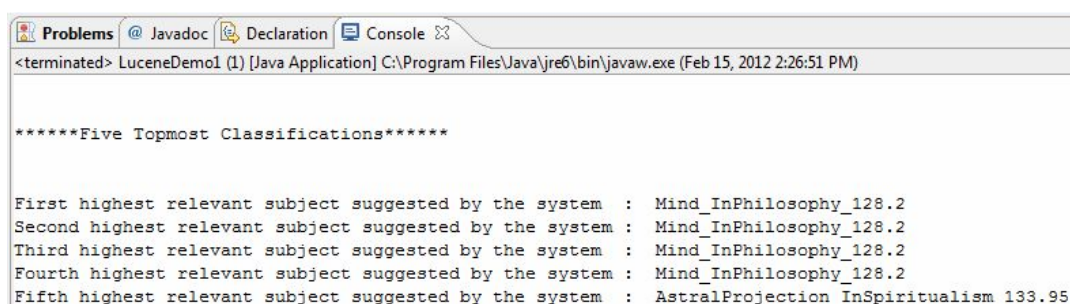


```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:29:20 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Second highest relevant subject suggested by the system : Cosmology_InPhilosophy_113
Third highest relevant subject suggested by the system : AstralProjection_InSpiritualism_133.95
Fourth highest relevant subject suggested by the system : Clairaudience_InPsychicPhenomena_133.85
Fifth highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
```

Figure A3: Five topmost results for the basic tf-idf classifier



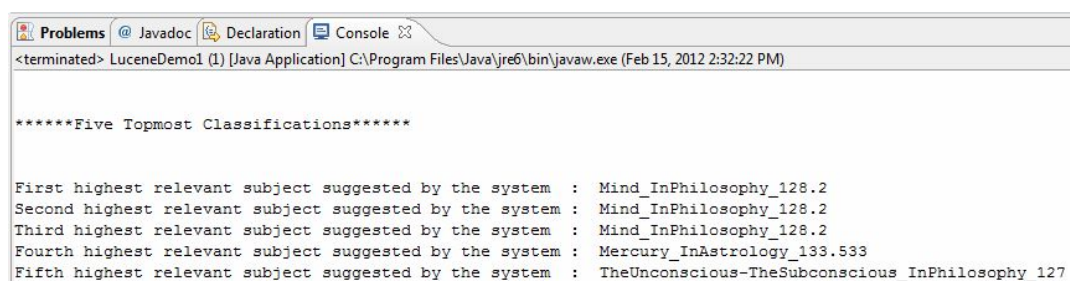
```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:26:51 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Second highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Third highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Fourth highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Fifth highest relevant subject suggested by the system : AstralProjection_InSpiritualism_133.95
```

Figure A4: Five topmost results for the new classifier

Figure A5 and Figure A6 show the five topmost results obtained for the basic tf-idf classifier and the new classifier. In these cases, we have input a test document which belongs to the subject area of Telepathy in psychic phenomena.

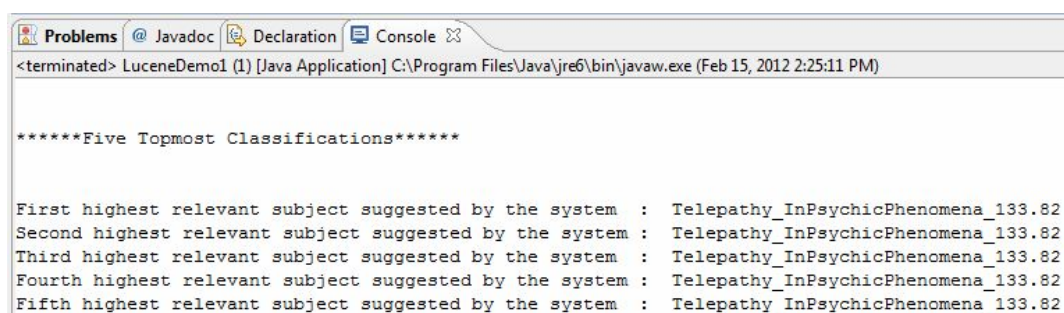


```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:32:22 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Second highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Third highest relevant subject suggested by the system : Mind_InPhilosophy_128.2
Fourth highest relevant subject suggested by the system : Mercury_InAstrology_133.533
Fifth highest relevant subject suggested by the system : TheUnconscious-TheSubconscious_InPhilosophy_127
```

Figure A5: Five topmost results for the basic tf-idf classifier



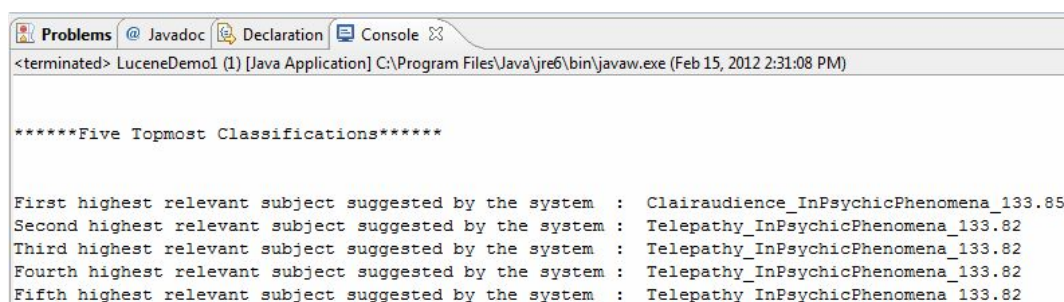
```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:25:11 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Second highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Third highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fourth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fifth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
```

Figure A6: Five topmost results for the new classifier

Again another test document which falls into the subject area of Telepathy in psychic phenomena was classified separately using the two classifiers. Five topmost results are given by the Figure A7 and Figure A8.

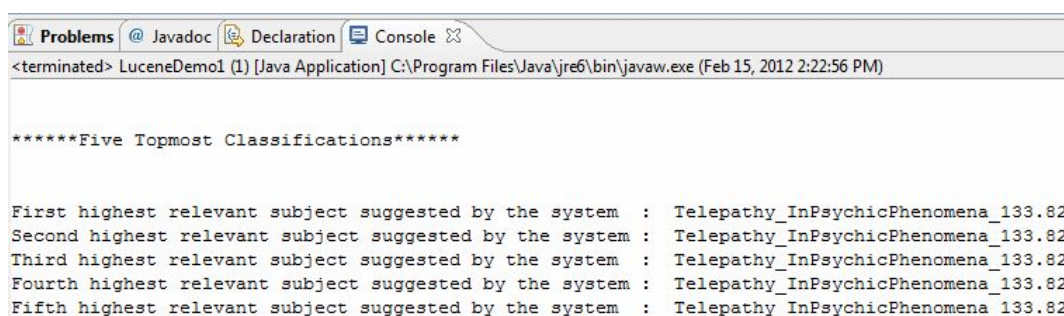


```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:31:08 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Clairaudience_InPsychicPhenomena_133.85
Second highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Third highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fourth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fifth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
```

Figure A7: Five topmost results for the basic tf-idf classifier



```
<terminated> LuceneDemo1 (1) [Java Application] C:\Program Files\Java\jre6\bin\javaw.exe (Feb 15, 2012 2:22:56 PM)

*****Five Topmost Classifications*****

First highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Second highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Third highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fourth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
Fifth highest relevant subject suggested by the system : Telepathy_InPsychicPhenomena_133.82
```

Figure A8: Five topmost results for the new classifier