IFLA WLIC 2013
SINGAPORE

Submitted on: June 15, 2013

An Ontology Based Fully Automatic Document Classification System Using an Existing Semi-Automatic System

CHAAMINDA MANJULA WIJEWICKREMA
Main Library, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.
manju@sab.ac.lk

RUWAN GAMAGE
National Institute of Library and Information Sciences, University of Colombo, Colombo, Sri Lanka.
mailruga@gmail.com

Abstract:

Automatic classification of documents has become an important research area due to the exponential growth of digital content and because manual or semi-automatic organization is not effective. On one hand, manual and semi-automatic classification is very painstaking and labor-intensive. On the other hand, misclassifications due to vagueness of the documents and classification schemes are inevitable in these two methods.

Hence, the current study sought to shed a light on these issues. This research proposes an automated system that can completely classify a given text document by minimizing the vocabulary ambiguities. One of our previous studies has developed a semi-automatic system for document classification and here we propose to extend it furthermore to obtain a fully automatic document classification system.

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

1 INTRODUCTION

The amount of information available in both printed and electronic formats has increased dramatically in recent years. This speedy growth of information in the modern world leads people to organize text materials in easy to access ways. Moreover, standard library classification schemes are used to accomplish this goal. However, some flaws that are

inherent in natural languages may create severe issues in the process of classification. In particular, the vocabulary ambiguities which exist in natural languages may increase the inaccuracy of classification. For example, the homonym 'ontology' in computer science might also be classified under 'ontology' in philosophy. This, in turn, affects the likelihood of a document being found. Time spent and effort put into organizing materials is also high in traditional methods of classification. In addition, the volume of digital documents also increases rapidly, demanding easy and affordable mechanised methods.

Automatic Text Classification (ATC) is a field of study with research on automatically categorising digital documents into pre-defined classes. In fact, the major factor behind the ATC is known as the classification algorithm or the text classifier. Support vector machines, probabilistic methods, generic algorithms, distance learning methods, hidden Markov models, decision tree methods, regression methods, decision rule methods, neural network methods (Sebastiani, 2002; Tao, Ling, & Cheng, 2005) and tf-idf (term frequency-inverse document frequency) based methods (Abbas, Smaili, & Berkani, 2010; Tao, Ling, & Cheng, 2005) are the frequently used classification algorithms in the process of text classification.

In this study, we attempt to further enhance a tf-idf based semi-automatic (hybrid) text classification system (termed HTCS in this paper) previously introduced by the authors (Wijewickrema & Gamage, 2012a, 2012b). The basic tf-idf function used there has been enhanced here. In order to reduce vocabulary ambiguities, a domain-ontology has been used.

Although HTCS gives better results than the manual method, it has some limitations because of the semi-automatic nature of the process. As manual intervention for classification has not been fully eliminated from the process, the final classification still depends on human decisions. On the one hand, the semi-automatic system selects a few candidate classes, and the human classifier still has to select the most suitable class from among them. On the other hand, a fully automatic method can further minimise time and labour spent on a given task.

2 LITERATURE REVIEW

The existence of the ambiguities of natural languages has been widely discussed in both linguistics and natural language processing (Richardson & Smeaton, 1995). In 1985, Princeton University started the WordNet project (Miller et al., 1990; Morato et al, 2004) to develop a lexical resource for the English language. In addition, many other electronic lexical resources have been developed (Best, Nathan, & Lebiere, 2010; Prevot, Borgo, & Oltramari, 2005; Valitutti, Strapparava & Stock, 2004).

Various attempts to use lexical resources (or ontologies) for automatic classification are also evident. Prabowo et al. (2002) and Song et al. (2005) report ontology based systems for classification of Web pages. The former used ontologies based on the Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC) schemes. However, according to Song et al. (2005), Prabowo et al.'s (2002) work is less productive as it is not adaptive to creating a sophisticated classification. Song et al. (2005) developed their ontology semi-automatically in the domain of economy. However, as this ontology was not so descriptive, one cannot expect highly accurate results for classification.

In addition to the Web page classification, ontologies are also applied in classification of emails and news in digital format. Taghva et al. (2003) formulated an ontology based classification system for emails. Tenenboim, Shapira, and Shoval (2008) report an ontology based classification system for electronic newspapers.

3 METHODOLOGY

The proposed system is mainly based on three crucial classification stages.

3.1 First Stage of Classification

3.1.1 Stemming and Removing Stop Words

The first stage of the system concerns the stemming to limit the index terms to the root terms and elimination of stop words to reduce the less significant words from the input document. After these two events, the refined document (now limited to a list of words) is taken as the output of the process. Using this output, the first few highly frequent terms are selected as the key terms to determine the discipline of the document in focus (input document).

3.1.2 Training Set

Training set is a collection of documents where the training documents are stored. One can use them to compare with the input document to determine the subject of it. This can be done by evaluating the similarities between previously selected key terms of the input document and the terms in the training documents. In our study, we used 385 training documents within the domain of philosophy. These have been classified by experienced subject classifiers according to the DDC scheme.

3.1.3 Text Classification Algorithm

In general, the text classification algorithms numerically determine to what extent a given document relates to a particular subject. In one of our previous studies (Wijewickrema & Gamage, 2012b), a new text classifier algorithm was developed using an existing term frequency weight function called the tf-idf weight function (Salton & Buckley, 1988). The same classifier was used here as well. It compares the similarities between the input document and each of the training documents. Hence, the discipline of the training document which obtains the highest numerical value is determined as the discipline of the input document. The new classifier uses the formula given below.

$$Document\ Score = \sum_{i=1}^{4} \frac{(tf-idf)_{i,D}}{\sqrt{\sum_{k=1}^{4}(tf-idf)_{k,D}{}^{2}}} \times (tf-idf)_{i,d} \qquad (1)$$

where, $(tf-idf)_{i,D}$ and $(tf-idf)_{i,d}$ represent the tf-idf weighting for the term t_i in the input document D and the training document d respectively. Here, we limit the summation only up to four as we consider only the first four highest frequency terms to determine the subject of the input document.

However, $(tf-idf)_{i,j}$ is defined as follows:

$$(tf-idf)_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}} \times \left[1 + \ln\left(\frac{|N|}{|n_i| + 1}\right)\right] \qquad (2)$$

where, $f_{i,j}$ gives the number of occurrences of the keyword t_i in the text of document d_j, N is the total number of documents in the collection and n_i is the number of documents in which the keyword t_i appears. In general, tf-idf weight function has the ability to assign numerical values for documents based on a few factors. They include the term frequency, the total number of terms in that document, the number of documents in which 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 classifier. This basic form has some

drawbacks, such as, consideration of only a single keyword for classification and not considering the importance of terms with equally higher frequencies. However, our new classifier is able to wipe out these difficulties to some extent.

3.2 Second Stage of Classification

In the second stage, the domain ontology was used to reduce the vocabulary ambiguities which may affect during the intermediate process of classification. Furthermore, we developed a computer programme to direct the broad classification results of the first stage to the ontology. Therefore, at the end of the second stage of the classification process, we get candidate classes for the given input document.

3.2.1 Ontology

Elimination of vocabulary ambiguities in classifying text documents is one of the main objectives of this research. Therefore, it is intended to be achieved with the use of a well-structured ontology. In order to build the ontology, we used two sources. Ontology was first constructed using the DDC scheme and then enriched using the Sears list. Our ontology is limited to the domain of philosophy and paranormal phenomena related subjects. They belong from classes 110 to 139 of the 21st edition of the DDC scheme. This limitation has been made due to the practical difficulties of building a large knowledge base. Therefore, the study is intended to classify documents only within that range. However, the approach maps the semantic relationships whenever there exist associations among the terms within the domain and the terms which are outside of the domain. Furthermore, a particular concern of the new ontology is to develop such relationships as like synonyms, near synonyms, hypernyms, and hyponyms.

For example, figure 1 shows the way that the term ‘cosmology’ is being presented in the ontology.

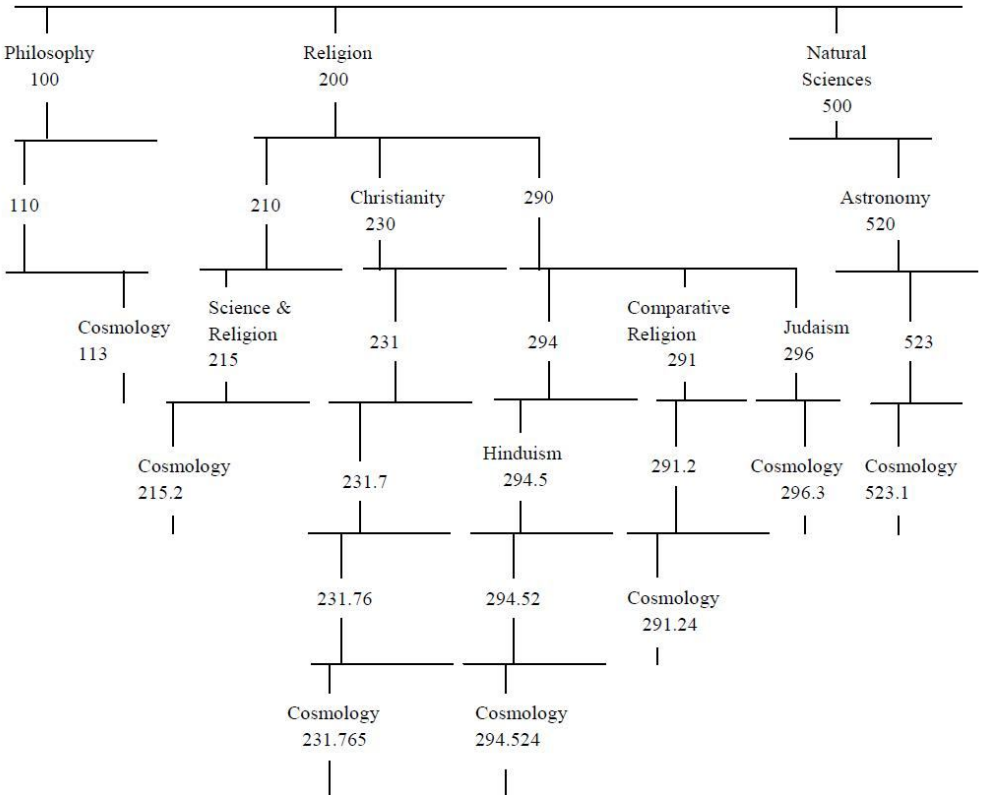

Figure 1: Structural arrangement of the term 'cosmology' in the domain ontology

As figure 1 illustrates, the term 'cosmology' appears in more than one place of the ontology and it builds relationships among them appropriately. So that, once we query the term 'cosmology' from the ontology, it retrieves the following possibilities with their descriptions.

Astronomy-Cosmology_InPhilosophyOfReligion_215.2
Cosmology_InAstronomy_523.1
Cosmology_InCreation_InChristianity_231.765
Cosmology_InHinduism_294.524
Cosmology_InPhilosophy_113
Creation-Cosmology_InReligion_291.24
Theology-Ethics-ViewsOfSocialIssues_InJudaism_296.3

3.3 Third Stage of Classification

In the final stage, another strategy has been followed to filter all the possible classification results that are given by the system after the second stage so that the user gets only one suggestion as the class number. As in figure 2, it compares the terms given at the end of the second filtration with the terms in the original input document. Therefore, the discipline which obtains the highest matching score during the process can be considered as the most suitable subject for the given input document.

The major stages of the process can be shown as in the figure 2.

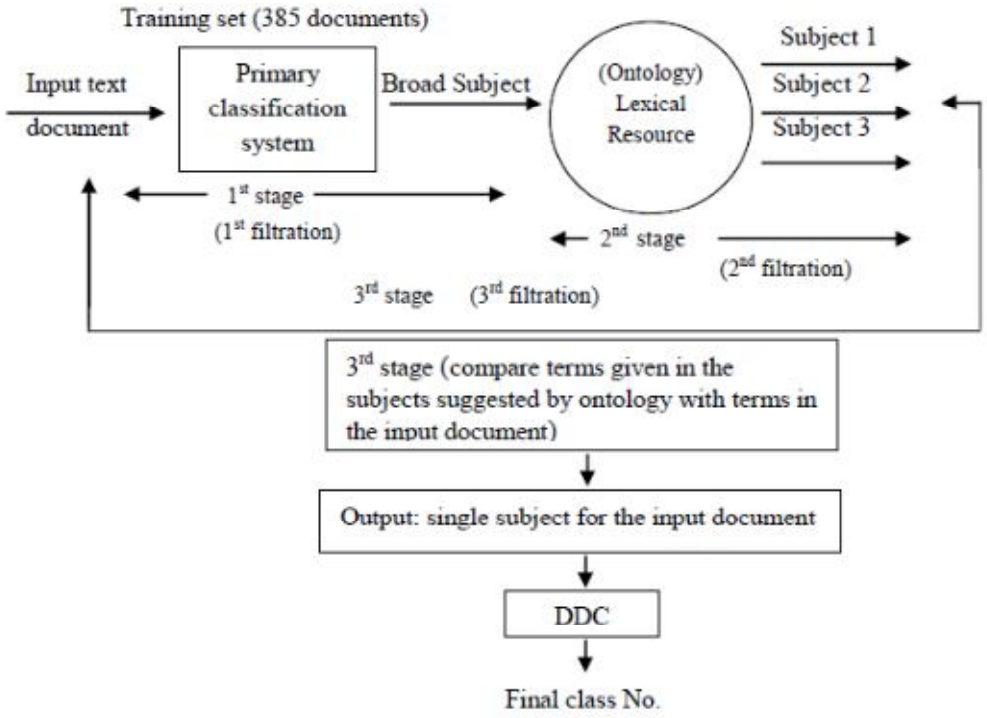

Figure 2: All stages of the fully automatic classification system.

3.4 Tools Used to Implement the System

In the first stage, we used Lucene API (Application Programming Interface) to choose the most relevant broad subject appropriate to the given input document. To initiate the second stage, the selected subject label was sent to a pre-built OWL ontology. This ontology was built using the Protégé ontology editor. After developing the ontology in OWL format, it was combined with the Lucene API using Protégé-OWL API. Here, the Protégé-OWL API was used as a tool to retrieve desired information from the OWL ontology file. As the output of the second stage, the system produced several candidate classes. We used this output as the input of the third stage to select one subject which is most relevant to the input document. For this purpose, Lucene API was used again.

The use of each tool and the way each has been positioned in the entire system can be shown as in the figure 3.

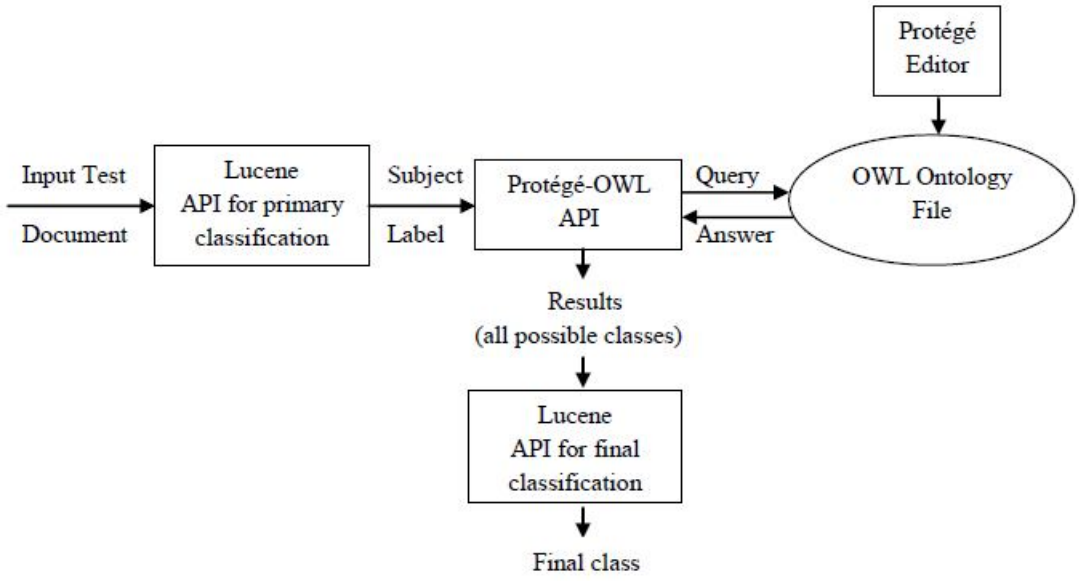

Figure 3: Implementing the system using tools

4 Results

In our previous study (Wijewickrema and Gamage, 2012a), a set of 58 input documents were classified according to the DDC system by an experienced subject classifier. Moreover, the classifier was asked to read the documents and to give all the possible subjects/disciplines and the one most preferred. Then the same set of documents was classified by the same subject classifier with the aid of the semi-automatic system. Again the classifier was instructed to select the most relevant subject out of all the possibilities given by the system. Following this sequence, we obtained the subject of the input document by manual and semi-automatic means at the end of the second filtration. As the input documents were pre-classified, the results were compared for accuracy.

For example, the classification results obtained after the second filtration can be shown as in figure 4. This input document belongs to the subject of 'Other Religions'.

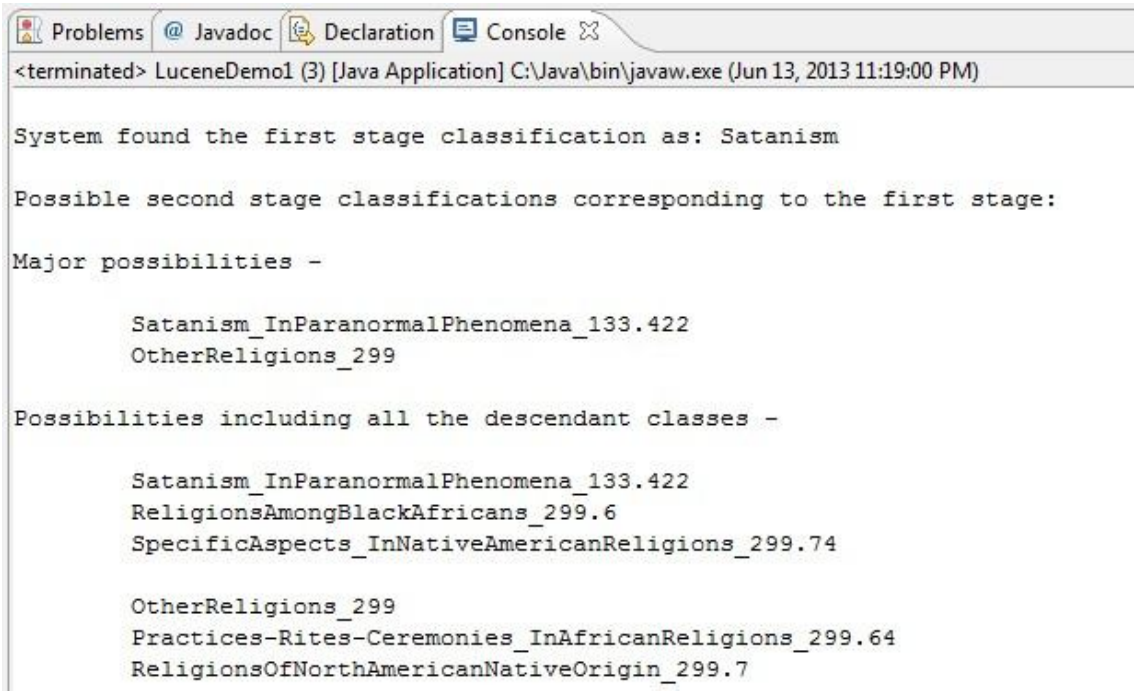

Figure 4: Classification results obtained after the second filtration

The same set of 58 input documents was again used to obtain the results of the current study. They were classified by the system automatically; the system gave only one option as the subject best matching the input document. Results were obtained in two ways. First, we compared the accuracy of classification for manual, semi-automatic and fully-automatic methods. Secondly, the relationship between vagueness of the documents and the inaccuracy of classification was measured. These results are also obtained for all three types of manual, semi-automatic and fully-automatic classification methods.

The classification result obtained by the new system for the same document which belongs to "Other Religions" is given in figure 5.

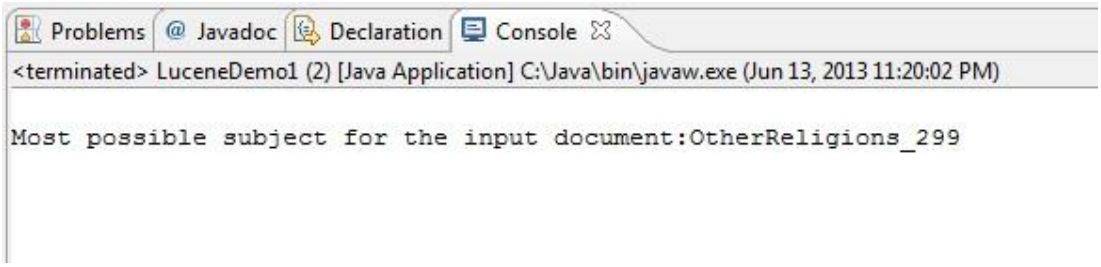

Figure 5: Classification results obtained by the fully-automatic system

4.1 Results Obtained for Classification Accuracy

Table 1 shows the results that were obtained to compare the accuracy of classification with respect to manual, semi-automatic and fully-automatic classification methods. The first column of the table gives the number of documents from each subject that was selected to test the accuracy of classification. Second, third and fourth columns show the correct number of classifications that were obtained using the respective classification method.

Table 1: Correct manual, hybrid and automatic classifications

Subject	Total Documents	Correct Manual Classifications	Correct Hybrid Classifications	Correct Fully Automatic Classifications
Apparitions	1	0	1	1
Aries	1	1	1	1
Attributes-Faculties	1	0	0	0
Axiology	1	1	1	1
Causation	1	0	1	1
Cosmology	4	2	3	0
Epistemology	2	1	2	0
Evil Spirits	1	0	0	0
Feng Shui	1	0	0	1
Geomancy	1	1	1	1
Leo	1	1	1	1
Libra	1	1	1	1
Love	1	1	1	1
Mind	1	0	0	1
Ontology	4	3	3	4
Other Religion	1	1	1	1
Palmistry	2	2	2	2
Phrenology	2	2	2	2
Pisces	1	1	1	1
Poltergeists	3	3	2	2
Precognition	2	2	2	2
Psychic Phenomena	2	2	2	0
Psycho Kinesis	3	2	2	3
Reincarnation	3	2	2	2
Space	1	1	1	1
Specific Mediumistic Phenomena	3	2	0	1
Spells-Curses-Charms	4	3	3	0

Spiritualism	1	0	1	1
Taurus	2	2	2	2
Teleology	1	1	1	1
Telepathy	3	3	2	2
Time	2	2	2	2

According to the results in Table 1, five distinct documents have been more accurately classified by the fully-automated system than by the human classifier with the assistance of the automated system. These documents belong to the subjects of Feng Shui, Mind, Ontology, Psycho kinesis and Specific mediumistic phenomena. On the other hand, there are ten documents that were more correctly classified by the hybrid system than by the fully-automated system. They belong to Cosmology, Epistemology, Psychic phenomena, and Spells-Curses-Charms. However, subject wise, out of all the subjects that were used, 3.125% were more correctly classified by the fully-automated system than by the hybrid method.

Figure 6 gives a comparison of the percentages of correct semi-automatic classification with the percentages of correct fully-automatic classification.

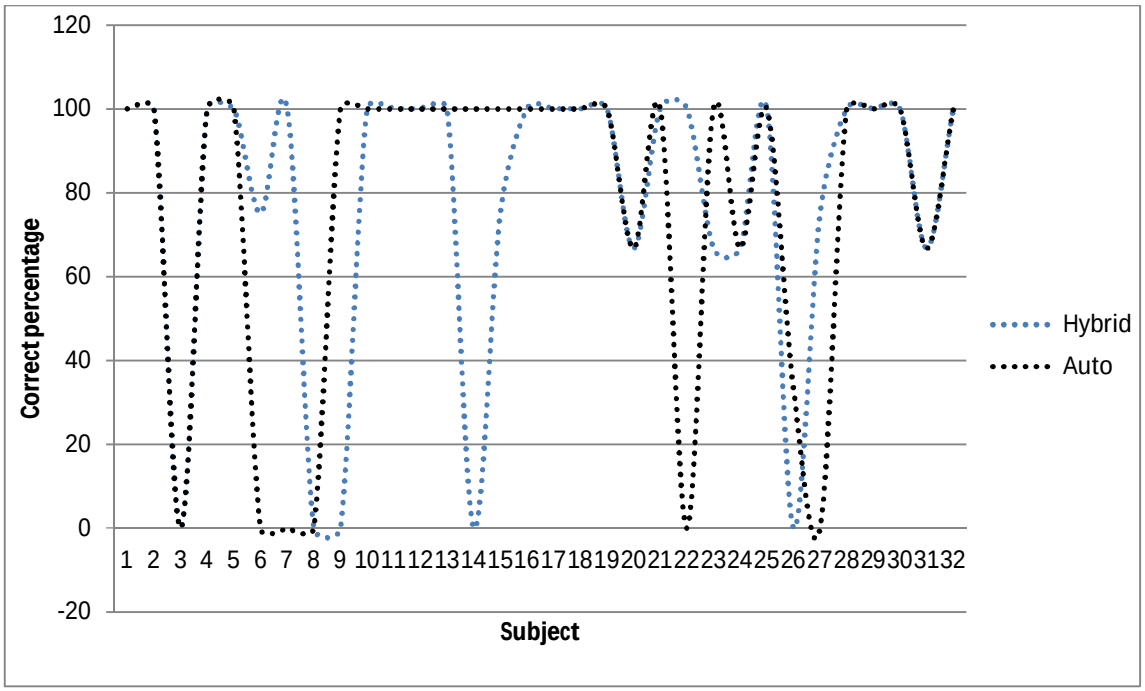

Figure 6: Percentages of correct semi-automatic and fully-automatic classifications

Here, one can notice that there is not much difference in the deviation of the average number of correct classifications of the two curves in figure 6. Although there is an improvement of the correctness of the new system, it does not appear extensively in figure 6 as the difference is not so significant.

4.2 Results Obtained for Document Vagueness and Misclassifications

The same set of documents was analysed to check whether there is a relationship between the vagueness of the content and the three methods of classification. In addition, the classifier was asked to mention whether the documents were vague or not. Vague documents were marked as 1 while others were marked as 0. After the classification, they were also marked. If a document was incorrectly classified, it was marked as 1 while 0 was given for the correct classifications. The obtained results are given in the table 2.

Table 2: Document vagueness and inaccurate classifications done by manual, hybrid and fully-automatic methods

Subject	Vagueness	Incorrect Manual Classification	Incorrect Hybrid Classification	Incorrect Fully Automatic Classification
Apparitions	1	1	0	0
Aries	0	0	0	0
Attributes Faculties	1	1	1	1
Axiology	0	0	0	0
Causation	1	1	0	0
Cosmology	1	1	0	1
	1	0	0	1
	1	1	1	1
	1	0	0	1
Epistemology	1	1	0	1
	0	0	0	1
Evil Spirits	1	1	1	1
Feng Shui	0	1	1	0
Geomancy	0	0	0	0
Leo	0	0	0	0
Libra	0	0	0	0
Love	1	0	0	0
Mind	1	1	1	0
Ontology	1	0	0	0
	1	1	1	0
	1	0	0	0
	1	0	0	0
Other Religion	1	0	0	0
Palmistry	0	0	0	0
	0	0	0	0
Phrenology	0	0	0	0
	0	0	0	0
Pisces	0	0	0	0
Poltergeists	0	0	0	0
	0	0	1	1
	0	0	0	0
Precognition	0	0	0	0
	0	0	0	0
Psychic Phenomena	1	0	0	1
	1	0	0	1
Psycho Kinesis	0	0	0	0
	0	0	0	0
	1	1	1	0

Reincarnation	0	0	1	1
	1	1	0	0
	0	0	0	0
Space	1	0	0	0
Specific Mediumistic Phenomena	1	0	1	0
	1	0	1	1
	1	1	1	1
Spells-Curses-Charms	0	0	0	1
	0	0	0	1
	0	0	0	1
	1	1	1	1
Spiritualism	1	1	0	0
Taurus	0	0	0	0
	0	0	0	0
Teleology	0	0	0	0
Telepathy	0	0	1	1
	0	0	0	0
	0	0	0	0
Time	1	0	0	0
	1	0	0	0

Analysis of the above results has been done using binary logistic regression. Here, the odds ratio has been used to determine the existence of a relationship between the vagueness of the documents and the incorrect classifications obtained by manual, semi-automatic and fully-automatic methods. In fact, the odds ratio is used to measure the size of the effect, describing the strength of association between two binary data values.

Odds ratio for the vagueness of documents and incorrect classification of documents done manually was 29.00, while the same ratio for the vagueness of documents and incorrect classifications done by the semi-automatic method was 3.61. Furthermore, the odds ratio for the vagueness of documents and incorrect classification of documents carried out by the fully-automatic system was 2.46. Hence, the odds ratio obtained for the effect of vagueness in the inaccurate classifications done automatically is the lowest.

5 Conclusions

Classification results obtained by manual, semi-automatic and fully-automatic methods were compared to determine the classification performances. Although the hybrid system shows accurate classifications for more documents than in the fully-automatic classification, 3.125% more subjects were correctly classified by the fully-automated system than by the hybrid system. The reason could be the misjudging some disciplines, such as Cosmology and Spells-Curses-Charms, by the new system. Therefore, whenever more input documents belong to these subjects, the number of misclassified documents also goes up. Apart from those special cases, the overall performance gives higher classification accuracy for the new system than the previous method. Moreover, this study further suggests that the vagueness of documents has a higher effect on semi-automatic classification, than on the fully-automatic method of classification.

References

Abbas, M., Smaïli, K., & Berkani, D. (2010). Efficiency of TR-Classifier versus TFIDF. *2010 First International Conference on Integrated Intelligent Computing*. doi: 10.1109/ICIIC.2010.60

Best, B. J., Nathan, G., & Lebiere, C. (2010). Extracting the Ontological Structure of OpenCyc for Reuse and Portability of Cognitive Models. *Proceedings of the 19th*

Conference on Behavior Representation in Modeling & Simulation (BRiMS 2010). Retrieved from http://www.adcogsys.com/pubs/Brims2010-best-gerhart-lebiere-opencyc.pdf

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*. Retrieved from http://courses.media.mit.edu/2002fall/mas962/MAS962/miller.pdf

Morato, J., Marzal, M. A., Llorens, J., & Moreiro, J. (2004). WordNet Applications. *Proceedings of the 2nd International Conference on Global WordNet*. Retrieved from http://www.fi.muni.cz/gwc2004/proc/105.pdf

Prabowo, R., Jackson, M., Burden, P., & Knoell, H. D. (2002). Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation. *Proceedings of the 3rd International Conference on Web Information Systems Engineering*. Retrieved from http://portal.acm.org/citation.cfm?id=674083

Prevot, L., Borgo, S., & Oltramari, A. (2005). Interfacing Ontologies and Lexical Resources. *Proceedings of OntoLex 2005*. Retrieved from http://www.loa-cnr.it/Papers/%5B22%5DprevotBorgoOltramari-3.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. doi: 10.1109/APSEC.2005.46

Tenenboim, L., Shapira, B., & Shoval, P. (2008). Ontology-based Classification of News in an Electronic Newspaper. *Proceedings of the International Conference on Intelligent Information and Engineering Systems*. Retrieved from http://www.foibg.com/ibs_isc/ibs-02/IBS-02-p12.pdf

Valitutti, A., Strapparava, C., & Stock, O. (2004). Developing Affective Lexical Resources. *PsychNology Journal*, 2(1), 61-83. Retrieved from http://www.psychnology.org/File/PSYCHNOLOGY_JOURNAL_2_1_VALITUTTI.pdf

Wijewickrema, P. K. C. M. & Gamage, R. C. G. (2012). Automatic Document Classification Using a Domain Ontology, *Proceedings of the 09th National Conference on Library and Information Science (NACLIS 2012)*, ISSN 978-955-9075-17-2, 85-107.

Wijewickrema P. K. C. M. & Gamage, R. C. G. (2012). An enhanced text classifier for automatic document classification, *Journal of the University Librarians' Association of Sri Lanka, 16* (2), ISSN 1391-4081, 138-159.