Annals of Library and Information Studies
Vol. 61, December 2014, pp. 263-272

Impact of an ontology for automatic text classification

Chaaminda Manjula Wijewickrema

Senior Assistant Librarian, Main Library, Sabaragamuwa University of Sri Lanka, Belihuloya
E-mail: manju@sab.ac.lk

Received: 13 December 2013; Revised: 12 August 2014

The concept of ontologies has widely been used in various applications including email filtering and electronic news classification. It can be also used for the classification of digital documents in a library. Advancing the accuracy of classification is the main purpose of using ontologies for classification. Documents may be difficult to understand due to the vague terms used in the text. However, since ontologies represent the semantic relationships of the terms, they can be used to correctly identify the subject of a document. This study made an attempt to improve the classification accuracy of an automatic text classification system by using an ontology.

Classification results given by the automatic system with and without integrating the ontology were used to evaluate the impact of the ontology for automatic classification. Results showed that 32.76% more documents and 25% more subjects were correctly classified by the ontology based system than the system prior to use of ontology.

Keywords: Ontology; Automatic document classification; Text classification

Introduction

Automatic text (or document) classification is basically a machine learning task. In machine learning, the computers make and improve the decisions based on some data. Therefore, one can use machine learning task to automatically assign pre-defined categories to a document based on the contents of the document. As the assigned categories are pre-defined, this task is also known as a supervised machine learning task. Moreover, it uses mathematical algorithms to assign these categories to documents. Classification of digitized documents nowadays gains a higher significance due to the rapid growth of digital content. With respect to the growth, organizing them is a big challenge for efficient retrieval of relevant information. Therefore, finding and improving solutions for text classification has considerable importance.

Applications of automatic text classification cover a wide and diverse range of practical works like spam filtering[1,2], electronic news classification[3], email classification[4], and web page classification[5]. However, the higher significance of automatic text classification can be spoiled due to some existing practical constraints. Out of number of limitations, the existence of vocabulary ambiguities in natural languages can cause to reduce the accuracy of the classification results. For example, a document about 'bank' may give different meanings in different disciplines. Although a document in fact describes the edge of a river by the term 'bank', it can be misclassified under financial institutions. Therefore, the incidence of vague terms in natural language may cause to reduce the accuracy of the classification process. Types of vague terms in natural language appear in numerous forms. Homonyms are the terms that share the same spelling and the same pronunciation but have different meanings. Although the synonyms have different appearances, they bear exactly or nearly the same meaning. For example, as 'optics' and 'light' are synonymous, the classifier may struggle in classifying a document on optics if the classifier is not familiar with the term optics but only with the term light. Hence, synonyms also pose problems in the process of classification. Basically, the incidence of these two types of vocabulary ambiguities poses major challenges in classification.

This paper proposes a solution to reduce the number of misclassifications due to vocabulary ambiguities of the language used. This solution is based on a new ontology. As an ontology represents the relationships among the concepts and the descriptions of the concepts, it makes it easy to obtain and clearly understand all possible meanings of even a vague term. Hence, it can be used to select the most appropriate candidate out of number of other classification results. The selection of the suitable category is done by a newly developed automatic text classification system[6]. Also the proper integration of the ontology with the automatic text classifier furnishes the goal of obtaining an ontology based automatic text classification system. Hence, ultimately this research lays the foundation to increase the number of correctly classified documents as well as accurate placements of documents in a repository.

Review of literature

The precise meaning of a subject is very important in the field of document classification. A study[7] conducted to examine the concept of subject from a new viewpoint reveals that the results of subject classification depend on the behavior of the ontology structure of the subject. Moreover, the paper also emphasizes the significance of the proper use of an ontology in the classification process.

Ong and others[8] have developed an ontology based Web page classification system in order to improve the classification speed and the accuracy of the results. Unlike in the previous studies, the novel system does not consider the entire text in the Web page for classification. It uses a segmentation method that utilizes visual boundary of a region and matches the key terms within that region. Therefore, it can easily improve the classification speed. In addition to the speed, the system is able to identify and match the phrases efficiently. Moreover, it uses an ontology to resolve the problems arising with the words in a document that have similar meanings but which are not identical in their spelling. Although some classification systems fail to consider word disambiguation, the proposed system does not have that failure owing to the well constructed ontology. Therefore, it can improve the classification accuracy as well. However, one can dispute that it could help to increase the accuracy of classification, if the system considers the entire Web page for classification.

Prabowo and others[9] used ontologies for Web page classification with respect to the LCC (Library of Congress Classification) and DDC (Dewey Decimal Classification). Another similar kind of research[10] has been conducted for the classification of Web pages and they have identified some of the shortages of the Prabowo's and contributors' work. Song and others emphasized this as "the weakness of their approach is the fact that is not adaptive when users require more sophisticated classification, even if the approach follows the standard classifications". Here, the ontology was created semi-automatically in the domain of economy. Moreover, while building the ontology, concepts and their relational structure were defined through conference with specialists in the corresponding domain. Therefore, this system has not used any existing standard-like document (eg: existing classification scheme, encyclopedia, etc.) or resource to build the ontology.

Jain and Pareek[11]suggested an automatic topic identification system to select learning materials explaining a given topic. As a learning material may cover multiple topics related to multiple subject domains, it is difficult to classify it under single category. However, the proposed approach is able to identify the major topics covered by the material. A domain ontology has been developed by the study to retrieve the topics covered in the document. Similar kind of research has done by Rathore and Roy[12] to automatically identify the Web page topics. Another study on ontology based classification system for unstructured information was proposed by Burger and Stieger[13]. This system keeps the classification categories in the form of ontology classes and the metadata of files maps with these ontology classes to perform the classification process. Liu and others[14]proposed an ontology based classification method to apply in the field of agriculture. This system first search the product information agricultural field. It uses a learning algorithm to extract the topic knowledge of agricultural products. The results have shown that the system reports a high accuracy of the classification process.

Another study towards the personalized news classification[15] has been introduced to use semantic technologies to provide a number of on-line recommendation services for single and multiple users. The developed system called News@handwas aimed to use a controlled and structured vocabulary to define the news items and user profiles. These two are represented in terms of concepts in the domain ontologies. Then the news items are forwarded to the user based on the similarities between news text and user preferences and the semantic relationships between concepts.

Focusing on specific applications is one of the noticeable facts that is common in most of the above attempts. Therefore, investigating the classification problem in a more general view point adds a significant value to the field of automatic text classification. Moreover, among these solutions, it is difficult to find even a single ontology which is a combination of the DDC and the Sear's list. Hence, an appropriate combination of these two knowledge bases expands the ontology development using different combination of resources.

Objectives of the study

- To reveal the way that vocabulary ambiguities effect on the accuracy of document classification;
- To formulate semantic relationships for the terms in the selected domain to construct the ontology;
- To integrate a novel ontology with an existing text classifier; and
- To compare the classification performances of the ontology based automatic classification system and the automatic classification system prior to combine the ontology.

Methodology

Pre-processing

Pre-processing is one of the crucial parts of the automatic text classification. It sets the first step to classify a document. The process begins with removing less significant words of a document in a way that the most important terms can be used to determine the subject of the document. These less significant words are known as stop-words. Moreover, it removes frequently occurring functional words in the language of the text. For example, 'the', 'is', 'a', 'an', 'of', etc. are some common stop-words in English language. Then the stemming process is carried out to reduce the number of index terms with the same root. For example, the English word 'administered' is stemmed into 'administ'. The Porter's stemming algorithm has been used for this purpose as it is the most commonly used algorithm for word stemming in English language[16,17]. Therefore, removing stop-words and stemming make the test document (document need to be classified) suitable for classifying.

Feature selection

All possible terms in the collection of words after pre-processing do not have the same importance in deciding the subject of a document. Normally, when the frequency of a term goes down, its influence in deciding the subject also descends respectively. Therefore, it is worth selecting only a limited amount of keywords for deciding the subject. This process is known as feature selection. The selection criterion for the terms is based on their term frequencies while the size of the feature space (number of selected terms) is based on an experimental method. Forty seven text documents have been used for the experiment. The text classifier algorithm given by the equation (1) was implemented in the system prior to initiating the classification. Each of these 47 documents was classified six times by varying the size of the feature space from one to six. These features gained the first six highest frequencies in the test document. The experimental results proved that the system gives best results when the feature space is limited to four[6]. Hence, the first four highest frequency terms are chosen to decide the subject of the test document.

New text classifier

In general, text classifier decides how far a given document is related to a pre-defined category. These pre-defined categories are represented as example documents and the content of each of these documents are compared with the pre-selected features. The relevancies between the example documents and the feature space are determined by the text classifier algorithm as numerical values. The pre-defined category of the example document which gains the highest score is considered as the category (subject) of the test document. The algorithm which is used here was developed in a previous work done by Wijewickrema and Gamage[6] and the same is applied for this work as well.

Equation (1) represents the text classifier algorithm which is used for the research.

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

where, $(tf - idf)_{i,D}$ and $(tf - idf)_{i,d}$ represent the tf-idf weighting for the term t_i in the test document D and the training document d respectively. Here, we limit the summation only upto four as we limited the feature space into four. Moreover, $(tf - idf)_{i,j}$ is defined as follows:

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

where, $f_{i,j}$ gives the number of occurrence 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.

Example documents

Automatic document classification systems use a set of pre-defined documents to assign a category for the test document. The set of example documents (training set) facilitates the system to compare each of its training documents with the given features. Moreover, the category of the matching training document is assigned as the subject of the test document. Three hundred and eighty five text documents are used as the training set of this research. Documents for the training set have been selected from the Wikipedia online encyclopedia, Stanford encyclopedia of philosophy, Google directory and also from subject gateway of Bulletin Board for Libraries. These selected documents were further examined and classified by the experienced subject classifiers.

Ontology

Ancient Greek philosophy describes ontology as the theory of being, becoming, existence and reality. However, when it comes to computer science, the concept of ontology is defined as a formal, explicit specification of a shared conceptualization[18]. Therefore, ontologies represent the meanings of the concepts while they are giving relationships among them too. This leads to having all the available associations for a given concept. The associations among the concepts in an ontology are built through the semantic relations. Semantic relations appear in number of ways including synonyms, homonyms, hypernyms/hyponyms, meronyms and associatives. However, ontologies are used in NLP (Natural Language Processing) to resolve the issues that arise in natural languages due to their vocabulary ambiguities. Hence, we use this property to minimize the misclassifications which happen owing to vagueness of the document.

This research develops an ontology based on the DDC scheme. The Sears list is also used to further enrich the ontology. We impose restrictions upon the ontology to make it limited to a certain subject criteria. It includes only the subjects which belong to the DDC range 110 – 139. So rather than a universal ontology, this closes to a domain ontology. However, when there exist relationships from the subjects inside of the domain to the subjects outside, they are also mapped here.

As an example, Figure 1 shows a portion of the structure of ontology concerning the concept ‘Memory’. The subject notations have been assigned according to the DDC scheme.

As in Figure 2, first we input the test document to the primary classification system. The primary classification system includes the text classifier algorithm and the training set. Using these two, it decides the subject of the input document and the process up to this level is known as the first stage of classification. Then after establishing the ontology, the system directs the result (suggested subject) given by the text classifier algorithm into the ontology. Hence, it facilitates the ontology to find and present all the available associations to the input subject. Ultimately, even if the input document or the subject suggested by the text classifier is vague, the ontology makes it easier to understand by providing all the possible subjects and their descriptions incorporated with the input category. Moreover, this latter part of classification is called the second stage of classification.

This entire process is depicted in Figure 2.

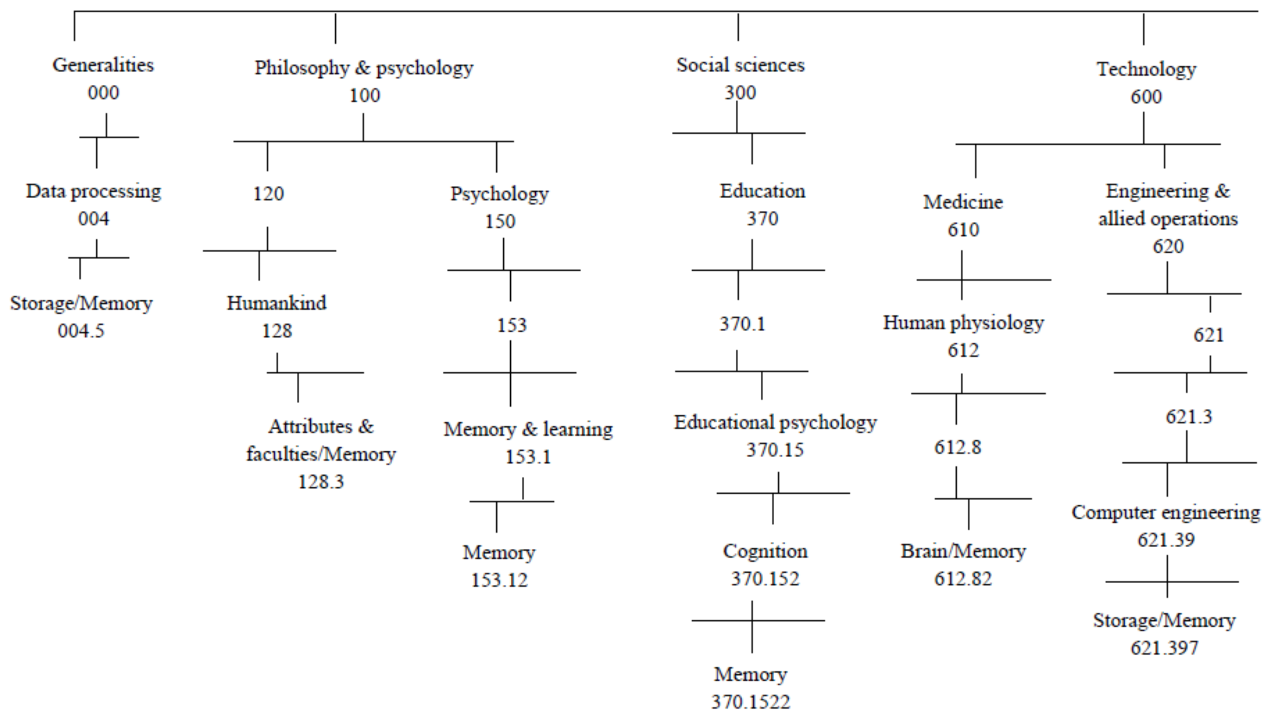

Fig. 1—Representation of the concepts

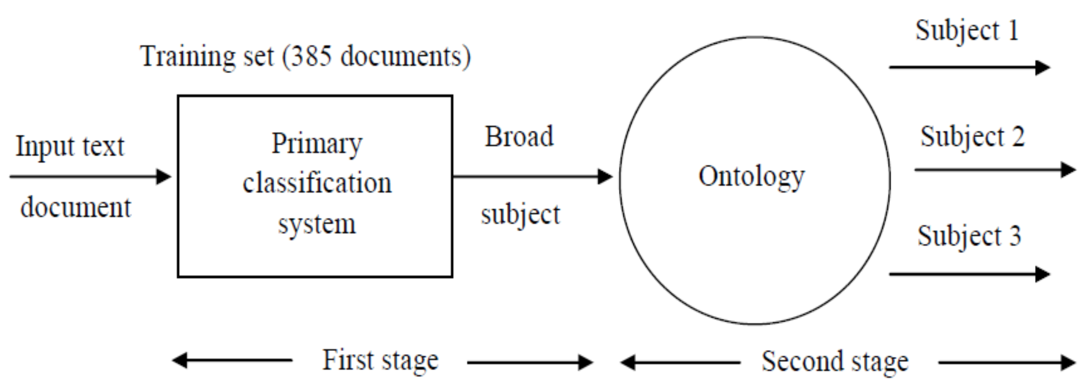

Fig. 2—Classification process

System implementation

We used three major software tools in order to implement the new system. Lucene API (Application Programming Interface) is used to complete the first stage of classification. In fact, Lucene search engine library is explicitly an API and not an application. Therefore, the user can change the programming codes and adopt it to accomplish the user needs. Here, it does the major tasks like pre-processing/index building, index searching, executing the algorithm, comparing results given by the algorithm, subject selection and direction. Then the Protégé ontology editor is used to implement the ontology. This ontology is built in OWL (Web Ontology Language) format. Protégé performs a set of actions that support the creation, visualization and manipulation of ontologies in different types of representation formats. In fact, using protégé one can describe the concepts and relationships among them.

Figure 3 shows a fraction of the ontology after establishing it using Protégé.

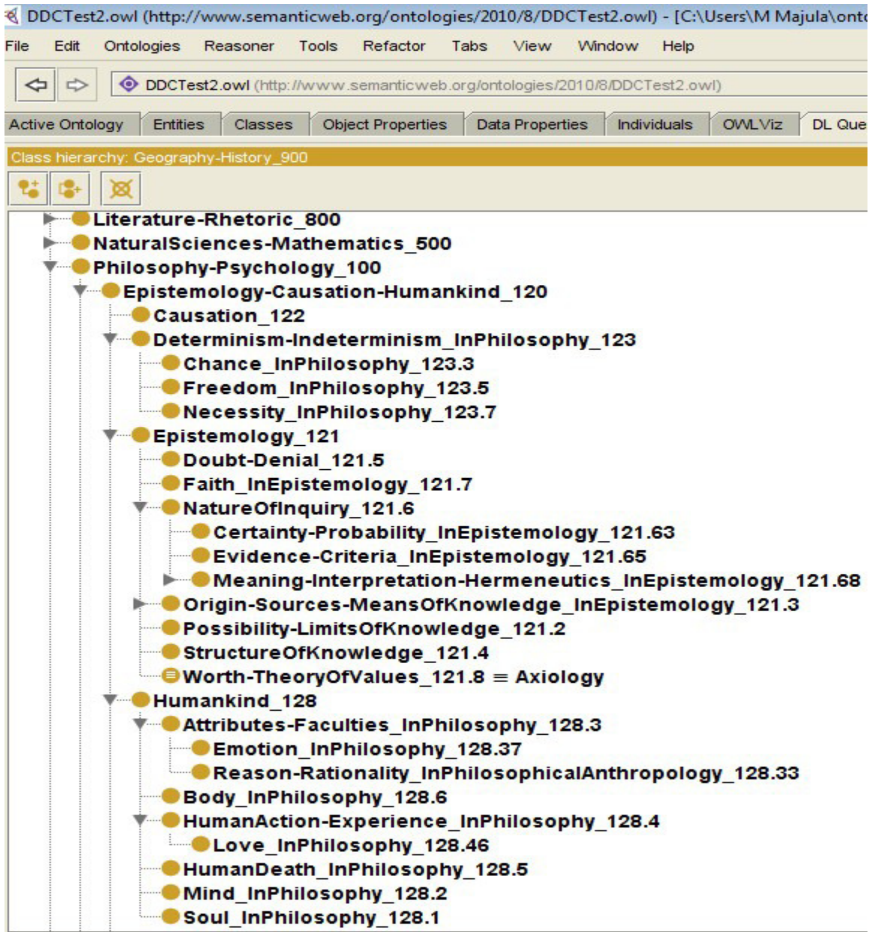

Fig. 3—Part of the ontology after implementing

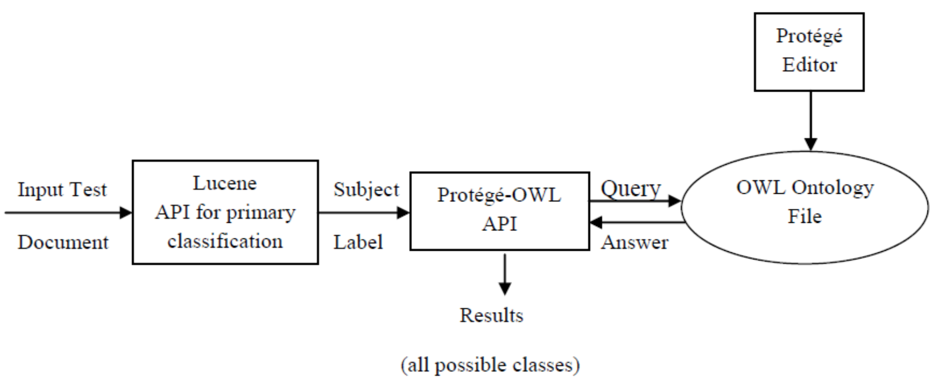

Fig. 4—System implementation using tools

Protégé-OWL API is used to query the ontology and to receive the feedback from it. It provides Java based programming codings to load and save OWL files to query and manipulate OWL data models and to carryout reasoning. Consequently, Protégé-OWL API can be considered as a bridge in between the OWL ontologies and the user who wants to query them.

Figure 4 shows the entire system implementation using each of these tools.

Classification results

Fifty eight test documents that belonged to 32 distinct subjects have been used for this evaluation. Evaluation process is carried out in two ways. First we checked the accuracy of classification results given after the first stage of classification. In that case, the system automatically selected the most relevant document to the input document from the training set. The subject of this most relevant document was given as the subject of the input document. Then we combined the ontology with the existing system and classified the same set of 58 documents. At this stage, the system automatically sent the result of the previous stage to the ontology and ontology further inquired all the possible relationships to the input. Sine, relationships ultimately led to concepts (i.e. categories), we found multiple subject categories as the possible and relevant subjects to the input document. To check the accuracy of classification, we looked for the presence of correct category from the results suggested by the system after the second stage of classification. However, here we did not exert any human intervention for selecting the most appropriate category from the suggestions. Instead, we simply looked for the presence of the most accurate category among them. If the correct category was available among them, then it was considered as a correct classification. Moreover, we have already developed this system furthermore[19] to explore some other classification properties of it. In order to obtain the results, both studies have been used the same set of test documents. Even though, the observations made and the results obtained were distinct from each other in both cases. For example, the previous study obtained results for correct manual, hybrid (after integrating ontology), and fully automatic (after integrating ontology) classification, the current study obtained results for correct fully automatic (without using ontology) classification and for the presence of correct category among the options (after using ontology).

Table 1 gives the number of correct classifications with and without using the ontology.

Above results shows that 53.45% of documents from all the evaluated documents were correctly classified by the system without using the ontology. However, the availability of correct classification of documents was 86.21% when the system was using the ontology. Moreover, in subject wise, 53.125% of subjects were correctly classified by the system without using the ontology while 78.125% subjects were correctly classified by the ontology based system for the same set of documents. Percentages for subject-wise calculations considered correct classification as only the subjects whose all the documents were classified correctly. Subjects like 'Evil Spirits', 'Leo', 'Love', 'Ontology', 'Other Religion', 'Space', 'Teleology', and 'Time' were correctly recognized by the ontology based system than the other one.

Figure 5 shows the variation of correct classifications obtained by the system without using the ontology and integrating the ontology.

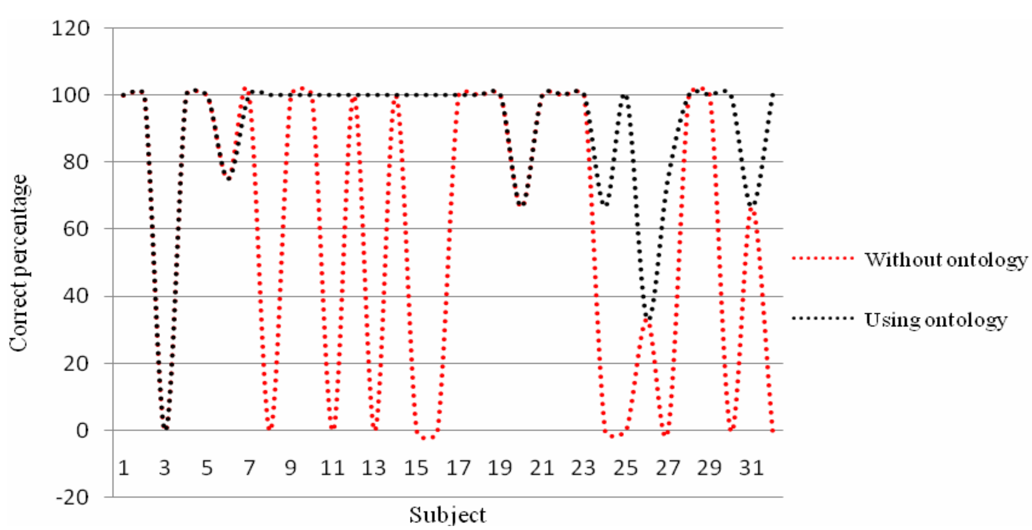

Fig. 5—Comparison of correct classifications with and without using the ontology

Table 1—Correct classification results obtained with and without using the ontology

Subject	Total documents	Correct classifications without ontology	Correct classifications using ontology
Apparitions	1	1	1
Aries	1	1	1
Attributes-Faculties	1	0	0
Axiology	1	1	1
Causation	1	1	1
Cosmology	4	1	3
Epistemology	2	2	2
Evil Spirits	1	0	1
FengShiui	1	1	1
Geomancy	1	1	1
Leo	1	0	1
Libra	1	1	1
Love	1	0	1
Mind	1	1	1
Ontology	4	0	4
Other Religion	1	0	1
Palmistry	2	2	2
Phrenology	2	2	2
Pisces	1	1	1
Poltergeists	3	2	2
Precognition	2	2	2
Psychic Phenomena	2	2	2
Psycho Kinesis	3	3	3
Reincarnation	3	0	2
Space	1	0	1
Specific Mediumistic Phenomena	3	1	1
Spells-Curses-Charms	4	0	3
Spiritualism	1	1	1
Taurus	2	2	2
Teleology	1	0	1
Telepathy	3	2	2
Time	2	0	2

According to the Figure 5, one can easily notice that there is a significant difference of the correct number of classifications given by the system without using the ontology and using the ontology.

As an example, Figure 6 shows the classification results obtained by the system with and without using the ontology for the same document.

In fact, the document which was classified in the above example belongs to the subject 'Other religions'. However, the system without using the ontology has been classified this document under 'Satanism'. Moreover, one can notice that the correct classification is available among the classification results given by the ontology based system. This was

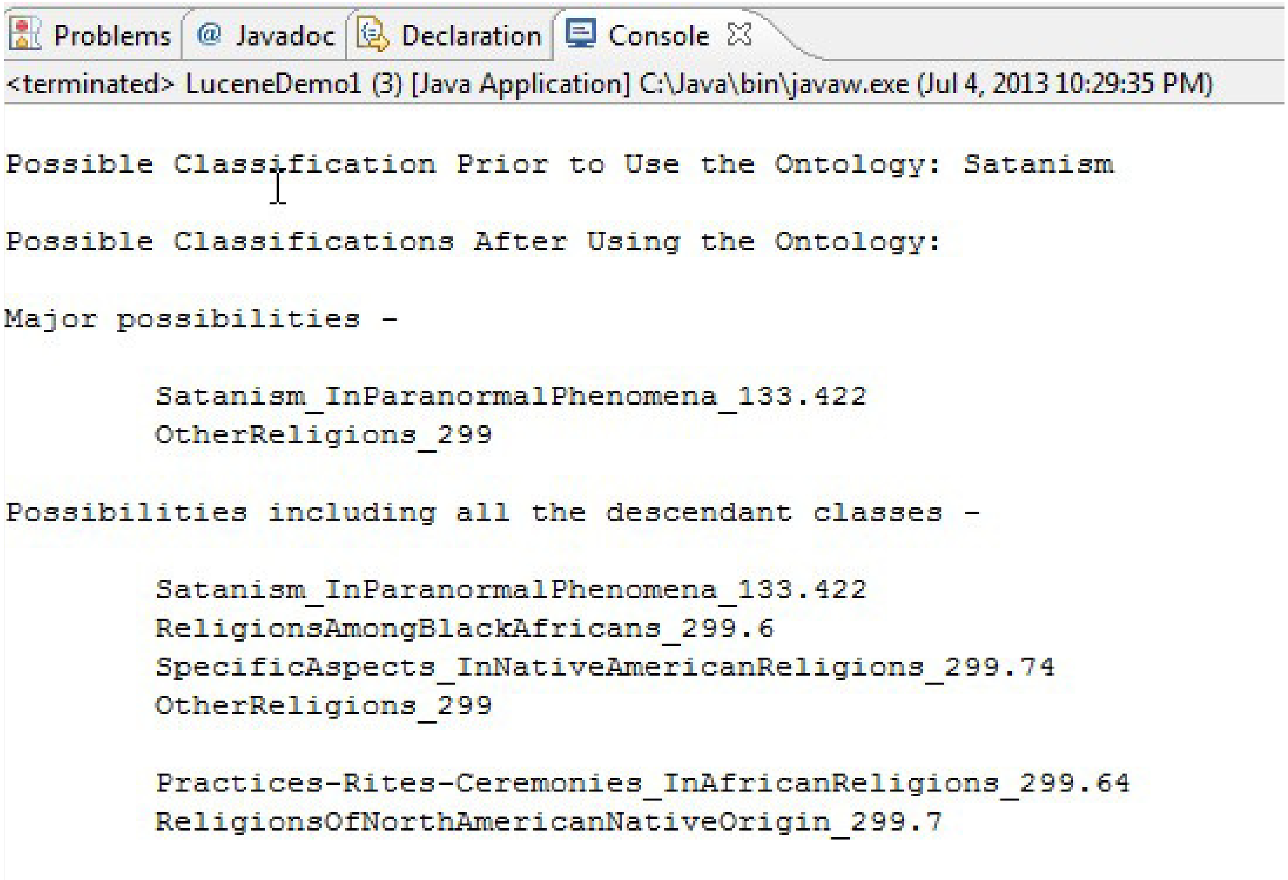

Fig. 6—Classification results given by the system with & without using the ontology

happened as the ontology represents all the possible associations available for the concept ‘Satanism’.

Conclusion

This research checked the classification accuracy of an automatic text classification system with and without integrating an ontology. The obtained results showed that 32.76% more documents were correctly classified by the ontology based system than the other. Furthermore, out of all the 32 subjects, 25% more subjects were correctly recognized by the system with ontology than the system which was not based on ontology. Hence, we can conclude that the integration of the ontology built here can be used to increase the classification accuracy of the automatic document classification system which is proposed by the research.

Although the ontology which is introduced by this research assists us to increase the accuracy of automatic classification, the final decision still has to be made manually. At this point, even if the correct category is present among the suggested results, the human classifier can select an inaccurate category from them. Therefore, the possibility of making an incorrect decision has not been fully eradicated yet. In order to overcome this problem, an extension[19] has been developed. Instead of giving multiple options to choose the correct one out of them, this extension is able to filter those options and select the most suitable one from them itself. However, the main difference between the two studies is that, while the current study attempts to find the effect of an ontology for classification accuracy of an automatic document classification system, the previous study[19] compared the classification accuracies among the manual, ontology based semi-automatic, and ontology based fully automatic systems. Hence, the former study did not reveal to which extent the ontology has been supported to increase the classification accuracy of the ontology based automated system than to the same system prior to use the ontology.

The future development of this study is towards the further increasing of the classification accuracy of the new system. Perhaps, this can be achieved in two different ways. On one hand, including more example documents in the training set would increase the classification accuracy. On the other hand, enhancement of the domain ontology can also improve the accuracy of the classification process.

References

1. Fumera G, Pillai I and Roli F, Spam filtering based on the analysis of text information embedded into images, *Journal of Machine Learning Research*, 7 (2006) 2699-2720.
2. Pantel P and Lin D, *SpamCop: A Spam Classification & Organization Program*, In Proceedings of the paper presented at the AAAI-98 Workshop on Learning for Text Categorization, Madison, Wisconsin, 26-27 July, 1998, p.26-27. Available at: http://www.aaai.org/Papers/Workshops/-1998/WS-98-05/WS98-05-017.pdf (Accessed on 31 July 2010).
3. Tenenboim L, Shapira B and Shoval P, *Ontology-based Classification of News in an Electronic Newspaper*, In Proceedings of the paper presented at theInternational Conference on Intelligent Information and Engineering Systems INFOS 2008, Verna, Bulgaria, 23 June-03 July 2008. Available at: http://www.foibg.com/ibs_isc/ibs-02/IBS-02-p12.pdf (Accessed on 02 April 2010).
4. Taghva K, Borsack J, Coombs J, Condit A,Lumos S and NartkerT,*Ontology based Classification of Email*, In Proceedings of the paper presented at the International Conference on Information Technology: Computers and Communications, Las Vegas, Nevada, 2003, p.194-198.
5. Dong H, Hussain F K and Chang E, *An Ontology-based Webpage Classification Approach for the Knowledge Grid Environment*, In Proceedings of the paper presented at the Fifth International Conference on Semantics, Knowledge and Grid, Zhuhai, China, 2009, p.120-127.
6. Wijewickrema P K C M and Gamage R C G, An enhanced text classifier for automatic document classification, *Journal of the University Librarians' Association of Sri Lanka,* 16 (2) (2012) 138-159.
7. Dutta B and Dutta C, Concept of 'Subject' in the context of library and information science from a new angle, *Annals of Library and Information Studies*, 60 (2) (2013) 78-87.
8. Ong W K, Hong J L, Fauzi F and Tan E X, *Ontological Based Webpage Classification*, In Proceedings of the paper presented at the 2012 International Conference on Information Retrieval & Knowledge Management, Kuala Lumpur, Malaysia, 2012, p.224-228.
9. Prabowo R, Jackson M, Burden P and Knoell H D, *Ontology-Based Automatic Classification for the Web Pages: Design, Implementation and Evaluation*, In Proceedings of the paper presented at the 3rd International Conference on Web Information Systems Engineering, Singapore, 2002, p.182-191.
10. Song M H, Lim S Y, Kang D J and Lee S J, *Automatic Classification of Web Pages based on the Concept of Domain Ontology*, In Proceedings of the paper presented at the 12th Asia-Pacific SoftwareEngineering Conference (APSEC'05), Taipei, Taiwan, 2005, p.645-651.
11. Jain S and Pareek J, *Automatic Topic(s) Identification from Learning Material: An Ontological Approach*, In Proceedings of the paper presented at the 2010 Second International Conference on Computer Engineering and Applications (ICCEA), Bali Island, Indonesia, 2010, p.358-362.
12. Rathore A S and Roy D, Ontology based web page topic identification, *International Journal of Computer Applications*, 85 (6) (2014) 35-40.
13. Burger S and Stieger B, *Ontology-based Classification of Unstructured Information*, In Proceedings of the paper presented at the Fifth International Conference on Digital Information Management (ICDIM), Thunder Bay, Canada, 2010, p.254-259.
14. Liu X, Duan X and Zhang H, *Application of Ontology in Classification of Agricultural Information*, In Proceedings of the paper presented at the 2012 IEEE Symposium on Robotics and Applications (ISRA), Kuala Lumpur, Malaysia, 2012, p.451-454.
15. CantadorI, Bellogin A and Castells P, *Ontology-based Personalised and Context-aware Recommendations of News Items*, In Proceedings of the paper presented at the WI-IAT '08 Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 2008, p.562-565.
16. Smirnov I, *Overview of Stemming Algorithms.*Available at: http://the-smirnovs.org/info/stemming.pdf (Accessed on 14 August 2010).
17. Zhou W, Smalheiser N R and Yu C, A tutorial on information retrieval: Basic terms and concepts, *Journal of Biomedical Discovery and Collaboration*, 1 (2), (2006). Available at: http://www.biomedcentral.com/content/pdf/1747-5333-1-2.pdf (Accessed on 17 January 2011).
18. Studer R, Benjamins V R and Fensel D, Knowledge Engineering: Principles and Methods, *Data & Knowledge Engineering*, 25(1-2) (1998) 161 – 197.
19. Wijewickrema P K C M and Gamage R C G, *An Ontology Based Fully Automatic Document Classification System Using an Existing Semi-Automatic System*, Paper presented at the IFLA World Library and Information Congress, 79th IFLA General Conference and Assembly, Suntec City, Singapore, 17-23 August 2013. Available at: http://library.-ifla.org/159/1/112-wijewickrema-en.pdf (Accessed on 07 August 2014).