Automatic Classification and Response of E-mails

Awatef Aloui
Multimedia, Information Systems and Advanced Computing Laboratory

Mahmoud Neji
Multimedia, Information Systems and Advanced Computing Laboratory

Abstract

With the fast development of Information Communication Technologies, the electronic mail has evolved as a convenient medium of communication and interactivity. In E-learning domain, tutors spend much time to process a large amount of e-mails. The focus of this paper is to propose a multi-agents system called EQASTO (E-mails Question Answering System using Text-mining and Ontological techniques) that is able to relieve the burden of e-mails processing. For this purpose, a combination of text-mining and ontological techniques will be useful to mine and classify semantically e-mails, fetch, generate, and send answers automatically to learners.

1. Introduction

In recent years, E-learning has been a newly growing field that helps learners to learn without time and distance barriers. The distance learning is based on the use of interactive technologies between learners and tutors.

With the ever-increasing use of Internet, e-mails have proved to be one of the main communication means between learners and tutors. But the exponential increase in the volume of e-mails can make the processing of e-mails tedious and time consuming for tutors. Therefore, the answer drafted by the tutor will be sent in few days, perhaps in few weeks and even nothing at all. This late or absence of answer has led to a serious problems, since they threaten one of the foremost objectives of E-learning, which is the development of the relations of collaboration and interactivity.

To solve this problem, an automatic processing of e-mails is very important in improving the tutor-learner interaction. For this purpose, this paper proposes a multi-agents system that combines text-mining and ontological techniques. It analyzes, classifies, and generates automatically e-mail responses. In this stage, our solution treats only e-mails in French language, but it can be adapted to other languages.

The remainder of the paper is organized as follows: section 2 provides a detailed literature survey of the related techniques of text-mining. Then, we provide an overview of our proposed methodology to analyze, classify, and answer e-mails in section 3. Finally, section 4 draws some conclusions and ideas for further research.

2. Related work

The search for relevant information was very important. Previously, the primarily difficulty consisted in having access to information. With the evolution and the technological development of data processing, mainly the speed and the reduction of the costs of storage and processing, a new field of research emerged, called Knowledge Discovery in Database where data-mining represents the essential and significant step that seeks to extract high level of knowledge from a low level of data.

The problem today, due to increasing use of the modern and unstructured forms of communication, is to filter among the profusion of information available those really exploitable. Thus, content analysis has matured into text-mining, literally excavation of textual data, which was created in the middle of 80 by the professor Don Swanson.

This section describes the main works in the literature that concern text-mining process in a general context of automatic processing of unstructured textual data. Then it provides several works which has focused on the automatic processing of electronic means of communication, especially e-mails.

2.1. Text-mining process

The text-mining, is also referred to as ‘’Text Data Mining’’ or ‘’Knowledge Discovery from textual databases’’, and can be defined as a new prospect for the analysis and the automatic processing for textual database allowing the discovery of knowledge [1]. It is an interdisciplinary area involving data-mining, statistics, information retrieval and natural language processing. Therefore, the process of the text-mining is similar to a traditional process of data-mining, its

characteristic lies in the specific steps of preparation of the data due to the semi-structured or unstructured nature of the text documents being processed.

Basically, two different stages are determined for this process.

The first one is the **pre-processing** that converts a textual data into a structured form, and process documents as bags of words using a linguistic [2] and semantic analysis [3].

Various studies related with pre-processing aims to obtain a set of representative terms called no-stop-word. For this purpose, a tokenization constitutes the departure of linguistic analysis; it converts the document into tokens or words (noun, verb, pronoun, article, conjunction and preposition) without understanding their meaning or relationships (i.e. out of context). Most of them are unimportant, called stop-words (white spaces, punctuation marks, pronoun, article, conjunction and preposition), so they are selected and taken out from document. Then, remaining words will be processed by a lemmatisation operation which consists in identifying the various inflections of a token (words brought back to their canonical form, adjectives brought back to the masculine), or variations of a verb (verbs brought back to infinitive). Later, a stemming algorithm replaces a set of words having a common morphological root by their stem. For instance, ‘‘learned, learning, learner’’ are all reduced to the stem “learn”. This operation reduces the number of words to process, and facilitates text-mining process. Lastly, a semantic analysis aims to extract key words that identify correctly the document using the criterion of high frequency weight of a term in the document. Currently three statistical methods [3] exist for extraction terms. Those methods are Boolean attributes, frequency attributes and N-gram attributes. Instead the statistical methods, semantic analysis can be realised through a semantic method. This method rests on an external element that can be static (replacing words by concepts) or dynamic (adding deduced information to the document).

The second stage in text-mining process is the **text classification** or text clustering. Text classification, namely text categorization, dates back to 1960s, but it became a major subfield in the early 1990s. It is defined as assigning predefined categories to text documents, where documents can be news stories, technical reports, web pages, etc. [4]

However, text clustering is known as an important and automatic technique for unsupervised document organization into clusters (i.e. the documents sharing the same topic are grouped together), and fast information retrieval or filtering. [5]

The text clustering is different from text classification because there is no training stage by using labeled documents, and the number of clusters is unknown prior to the clustering.

There are various methods of classification that can be categorized into statistical and machine learning method.

Numerous statistical methods have been introduced, such as Hidden Markov models [6], regression models [7], discriminant analysis [8], etc.

Machine learning methods represent a general inductive process automatically builds a classifier by learning, from a set of pre-classified documents, the characteristics of the categories [9]. A wide range of learning methods has been applied to this purpose, such as k-Nearest Neighbor [10], Naive Bayes [11], Support Vector Machines [12], Voting [3], centroid classifier [9], etc.

2.2. E-mail processing

Electronic mail can be viewed as a special type of document as it is primarily text along with some identifying information unique to it (e.g., from, to, subject, cc, attachments and so on). [13] In the past few years, with the advent of text-mining, the examination of e-mail started to get an increased attention of a growing number of researchers.

As already mentioned in the onset of this section, there has been a vast literature on text-mining. Furthermore, there has been many works on classification and clustering e-mails that have been applied to e-mail processing in order to reduce information overload.

In the aim to scout what had been done previously by others, existing research to automatically classify incoming e-mails can be broadly categorized into: rule based classification [14]; machine learning method such as Decision Trees learning [15], Support Vector Machines [16], Naïve Bayes [17]; ontology classification e-mails [18].

The original motivation of this work is based on the fact that, although various techniques have been applied to e-mail categorization, it still confronted some challenging issues. The most significant issues are the high costs of classification errors due to the unformed content of e-mails; earlier works was focused especially for the commercial field and they prevent on the level of e-mails classification and filtering; and a few researches establish semi automatic answers which require a human intervention for the choice of the recommended reply template or the checking.

For example, Weng & al. [19] classifies customer e-mails using multiple concepts, suggests relevant reply templates to quickly and accurately answer customer e-mails, and then customer service staff still has to choice and send the correct reply template to the customer. Besides, the process of sending the reply e-mail still manual.

Moreover, the major problem which arises in the majority of research of classifications of e-mail is

that they do not process e-mails which comprise more than one function of classification.

To address those issues, we propose in this study a new solution of automatic processing and response of e-mails in e-learning field in order to decrease the number of e-mails processed by human tutors and accelerate the process of reply e-mails.

3. Proposed system

Within a virtual learning environment, the proposed system EQASTO (E-mails Question Answering System using Text-mining and Ontological techniques) aims to improve e-learning by means of integrating an e-mail processing system based on intelligent agents. The overall flow of the e-mail processing architecture is shown in Figure 1.

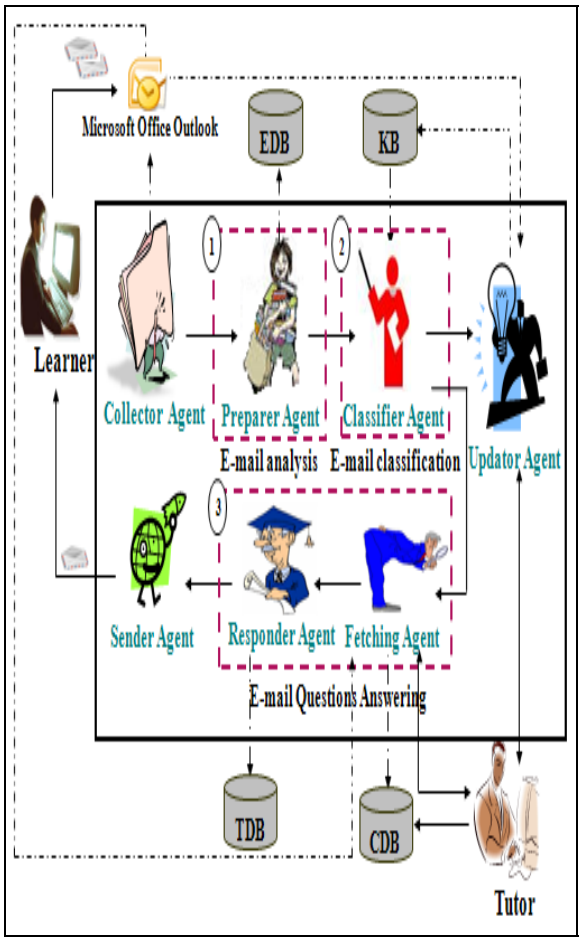

Figure 1. EQASTO's architecture

Legend :
EDB : E-mails Data Base.
KB : Knowledge Base.
TDB : Template Data Base.
CDB : Courses Data Base.

As they mentioned in Figure 1, three major layers are outlined: e-mails analysis, e-mails classification and E-mails questions answering.

A brief description of each processing layer is given below.

3.1. E-mails analysis

The first step in our e-mail processing is to carries out an analysis of the learner's e-mails. The purpose of this step is to get a structured representation that will be used to cluster e-mails accordingly to their semantics. For that, we propose to use the text-mining techniques as a strategy for parsing learner's e-mails. Within a Multi-Agents System (MAS), the structure of e-mails analysis is depicted in Figure 2.

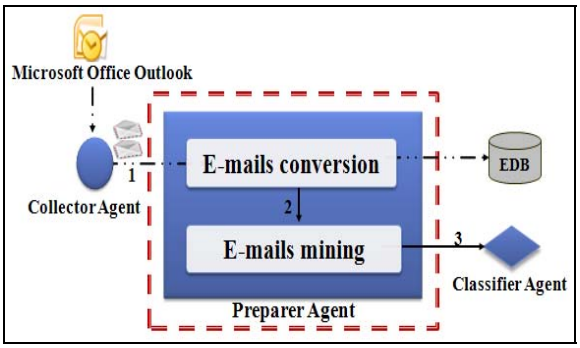

Figure 2. E-mails analysis

The e-mails analysis structure uses a software agent called ''Preparer Agent" that receive (1) e-mails from the "Collector Agent'' and transfer (3) the result of her process to the "Classifier Agent''.

As its name indicates, the ''Collector Agent'' makes only the collection and the multiple transmission of received e-mails without any processing to the ''Preparer Agent''. This last one made up a first processing (2) of the received e-mails in order to facilitate the step of classification and extraction of information. This first module of e-mail analysis consists on two tasks:

3.1.1. E-mails conversion. E-mails are unstructured by nature. So, the ''Preparer Agent'' converts every e-mail into a structured representation. In this study, we choose to represent the e-mail's HTML format in a textual file that contains mainly two parts: the first one contains information of addressing (such as the Sender, Recipients and Subject) and the second part contains the body of e-mail.

In the e-mail conversion task, the ''Preparer Agent'' focus only on the first part that will be parsed and tokenized through the text-mining techniques to get information about: Sender (From), Recipient (To, CC, Bcc) and Subject.

Attachments are considered part of the body and are processed in future work.

3.1.2. E-mails mining. The second part of each textual file is parsed now using the text-mining process with some adjusts in order to extract the relevant e-mail body features. Indeed, the text-mining is applied to textual data. And since our treatment is carried out on e-mail, a special form of textual data, we named this task the process of E-mail mining. The sequence of execution of this process is attested in Figure 3.

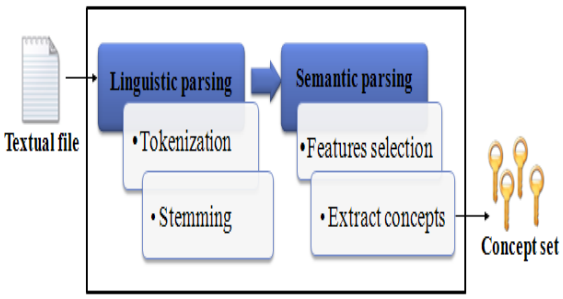

Figure 3. E-mails mining

Firstly, the ''Preparer Agent'' apply the linguistic parsing, that consists to tokenize the second part of textual file into words. A standard stop-word list is used to remove stop words, such as whitespace or special characters. Then, remaining words are stemmed to their root in order to facilitate the next parsing.

Second, in order to extract key words that identify correctly the e-mail, a semantic analysis is applied to choose the most important words of e-mails which would be preserved for the classification process. For that, the "Preparer Agent" picks the most significant words of e-mails using an e-learning taxonomy. Then, the "Preparer Agent" extracts the concept connected to each retained word. In fact, these concepts maintain the semantic significance of original terms in order to be used in the inference of classification rules. Next, an automatic transposition of textual questions into structured form must be run in the goal to be exploited automatically in the process of e-mails questions answering. This transposition must be entirely explicit, i.e. to follow a structured model and to be expressed in a clear and definite vocabulary (for example the question ''who is'' seeks a person, the question "Why" waits a raison). Thus, we propose to use ontology techniques which are the receptacle of this treatment.

Finally, each e-mail is represented by a set of concepts that will be useful for the next process called e-mails classification.

3.2. E-mails classification

E-mail classification process can be stated as follows. Given a training set of labeled e-mails $E_{train}=\{(e_1, c_1) (e_2, c_2) \dots (e_n, c_n)\}$, where e_i is an e-mail from an e-mail set E and c_i is the label chosen from a predefined set of categories C. This process attempt to infer a classifier that can correctly classify a test set e-mails E_{test}.

In order to classify e-mail by similar topics, we must first determine the set of categories, and then automatically classify e-mails into the appropriate one. For that, an approach oriented software agents for a semantic classification that links e-mails to ontological concepts is proposed.

Figure 4 summarizes the e-mail classification process used in predicting the category of e-mails.

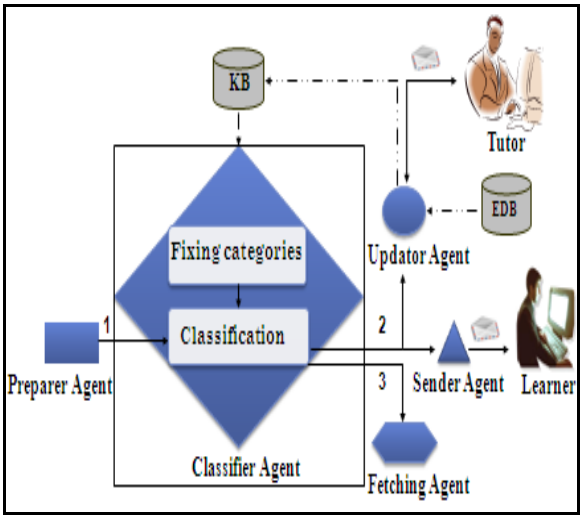

Figure 4. E-mails classification

The framework of this second stage consists on two tasks: fixing categories, and classification.

3.2.1. Fixing categories. In order to identify the various significances and functions of communication via e-mails between learners and their tutors, we decided to invest a similar study of Thao Lê & al. [20] which 1478 learner's e-mails was analyzed. This study acknowledges that topic's e-mails were divided into 10 categories, which can be grouped in three principal clusters called:

- **Cluster 1:** procedural function.
- **Cluster 2:** social function.
- **Cluster 3:** cognitive function.

Figure 5 proves clearly that learners' e-mail are concentrated on procedural function (64%), and less on social subjects (34%) and seldom on the cognitive function (2%).

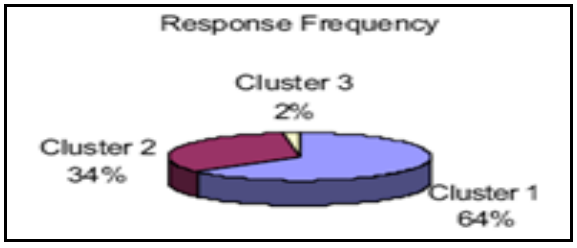

Figure 5. Percentage of three clusters [20]

The categories are sorted according to an order of prevalence. This order helps us, in the case of a manual processing, to identify the urgent e-mails of those which aren't.
Table 1 shows the 10 most frequent categories which are ranked in the following enumeration:

Table 1. E-mails clusters

Procedural function	Social function	Cognitive function
Requesting.	Thinking.	Discussing.
Confirming.	Referring.	
Clarifying.	Complimenting.	
Complaining.	Greeting.	
	Sharing.	

We notice nowadays that learners use synchronous methods of communication such as Facebook and Skype for the cognitive function. Thus, we admit that this function become out of date. Hence, our system will focus only procedural and social e-mail's clusters.
Basing on these two functions, we develop taxonomy of e-learning e-mail's questions. This taxonomy represent a set of rules of classification describing ontology associated with learner's e-mails, and will be stored in the Knowledge Base (KB). The drawback in the study of Thao Lê & al. [20] is it doesn't process e-mails which comprise more than one function. In this paper, we use category priority to solve this kind of classification problem, i.e. we classify this type of e-mails in the most dominant function and we mention that they comprise other functions in order to process them in a specific way of the other e-mails.

3.2.2. Classification process. The semantic classification of e-mail is carried out by several classification agents. Each one is associated to particular cluster of e-learning's ontology. Using the KB, each classifier agent applies the rules of classification in order to find the relation between the set of retained concepts of each e-mail with the category which is entrusted to him. The combination of the discovered relations gives the set of categories that will be ranked as it is shown in the previous section. Then, the result of this process will be transmitted (3) to the ''Fetching Agent''.
When no relation is found, the "Classifier Agent" sends (2) the e-mail's information to the "Updator Agent", and a notification of waiting to the "Sender Agent". This last one sends an acknowledgement of delivery to the learner to reassure him that his e-mail was taken into consideration.

The "Updator Agent" fetches the original e-mail, stored in EDB, and sends it to the tutor to be processed.
After receiving a special file from the tutor, the "Updator Agent" can make two actions. It updates the set of rules of an existing category fixed by the tutor, or adds a label of new category to the KB to indicate that a new category appeared and then create their rules in order to be employed in future classifications.

3.3. E-mail Question Answering

The E-mail Question Answering (EQA) process represents the task of extracting the right answer from a large collection of documents where the answer to a natural language e-mail's question lies. In this study, we develop an EQA system able to answer e-mail's questions according to the e-learning questions taxonomy.

This EQA system is focused on procedural and social categories. Its main components are summarized in two steps: fetch answer, and formulate answer. Figure 6 graphically shows the execution sequence of these components which are related to each other and executed by specifics agents.

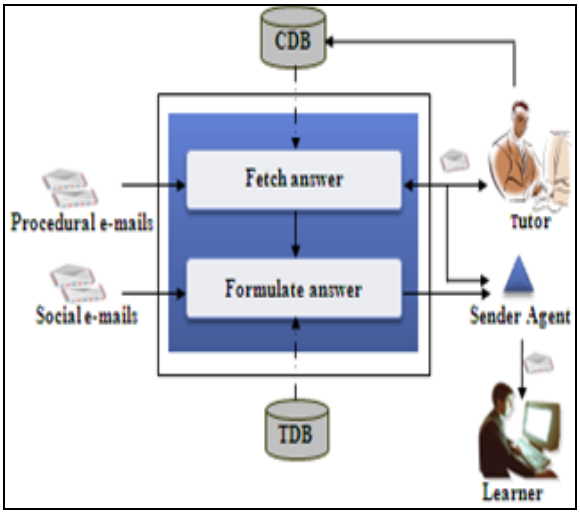

Figure 6. Architecture of the EQA process

3.3.1. Fetch answer. The answer of the e-mail's question is fetch, by the "Fetching Agent", in our Courses Data Base (CDB) using the e-learning ontology. This step is only reserved for the procedural function's e-mails. To do it, two steps are necessary:

- Documents retrieval: The obtained information from the analysis process is used by this step to perform a selection of relevant documents from our CDB.

- Relevant passages selection: with the purpose of identifying and picking out relevant text passages that are susceptible of containing the search answer, the "Fetching Agent" performs an ontological analysis of the relevant documents set using the types of questions detected in the phase of e-mails analysis. The selected fragments of documents will be sending automatically to the "Responder Agent".
  If no document is found, the "Sender Agent" sends an acknowledgement of delivery to the learner. Simultaneous, the "Fetching Agent" sends the original e-mail towards the e-mails box of the tutor for manual processing. The tutor is so invited to answer manually e-mails by order of importance of the categories according to two alternatives. Either he will select the fragment containing the response if there exists in one of the courses stored in our DBC. Or, in the event of absence of answer, he updates the DBC by adding another course which must be indexed and finally he locates the fragment containing the answer. In the two alternatives, the tutor transmits the response to the ‘’Fetching Agent’’. This later will treat the received information, and thereafter transmits the answer towards another agent in charge of the procedure of formulation the final answer.

3.3.2. Formulate answer. The set of alternative text fragments obtained from the previous process will be processed by the "Responder Agent" with the purpose of choosing the search answer. Then, this agent chooses an appropriate response template from a several standard templates, stored in our TDB, on which it can apply to formulate the final answer.
Finally, the answer of the e-mail’s question will be sending by a "Sender Agent" to the learner.

4. Conclusion

E-Mail has proven to be one of the main communication means between learners and tutors within a virtual campus. It is due to its simplicity of use, popularity, and speed of delivery. But the exponential increase in the volume of e-mails can make the manual processing of e-mails tedious and time consuming for tutors. Consequently, learners receive response of their e-mails after many days and perhaps no answers will be sent. In order to assist the learners in learning effectively, it is necessary to provide an appropriate learning environment.

This work describes the architecture of an e-mail processing system using text-mining and ontology techniques in e-learning framework in order to decrease the number of e-mails processed by human tutors and accelerate the process of e-mail’s reply.

The proposed system called EQASTO (E-mails Question Answering System using Text-mining and Ontological techniques) incorporates several dimensions such as: ***e-mail analysis*** carrying out a transformation of the original e-mail into a structured representation, ***e-mail classification*** ensuring a semantic classification of e-mails and finally ***e-mail question answering*** allowing to seek, formulate and send automatically answers to learners.

Our future work is to generalize this system to other languages such as English and Arabic. Also, we will implement an intelligent tutorial system in order to remove human tutor by a virtual one.

5. Acknowledgements

I am particularly grateful to Mr. Mahmoud NEJI for his thoughtful and creative comments, his encouragement, his guidance and support from the initial to the final level of my work.
Also, I offer my regards and thanks to all who encouraged me to write this paper.

6. References

[1] Feldman and Sanger. (2007) The Text Mining Handbook-Advanced Approaches in Analyzing Unstructured Data, USA: New York.

[2] Kristof Coussement and Dirk Van den Poel. (2008) ‘Integrating the voice of customers through call center emails into a decision support system for churn prediction’, in Information & Management, vol. 45, pp. 164–174.

[3] Md Rafiqul Islam and Wanlei Zhou. (2007) ‘Email Categorization Using Multi-Stage Classification Technique’, in Eighth International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 51-58.

[4] Wen Zhang, Taketoshi Yoshida and Xijin Tang. (2008) Text classification based on multi-word with support vector machine, Knowledge-Based Systems.

[5] Shi Zhong. (2005) ‘Efficient streaming text clustering’, in Neural Networks 18, pp. 790–798.

[6] Khmelev and Tweedie. (2001) ‘Using Markov chains for identification of writers’, in Literary and Linguistic Computing, vol. 16, n°3, pp. 299–307.

[7] Yang and Chute. (1984) ‘An example-based mapping method for text categorization and retrieval’, in ACM Transactions on Information Systems, vol. 12 n°3, pp 252–277.

[8] Can and Patton. (2004) ‘Change of writing style with time’, in Computers and Humanities, vol. 38, n°1, pp. 61–82.

[9] Songbo Tan. (2006) 'An effective refinement strategy for KNN text classifier', in Expert Systems with Applications 30, pp. 290–298.

[10] Kucukyilmaz, Cambazoglu, Aykanat and Can. (2006) 'Chat mining for gender prediction', in Proceedings of the fourth Biennial conference on advances in information sciences, Turkey, pp. 274–284.

[11] Ding-An Chiang, Huan-Chao Keh and Hui-Hua Huang, Derming Chyr. (2008) 'The Chinese text categorization system with association rule and category priority', in Expert Systems with Applications, vol. 35, pp. 102–110.

[12] Shiqun Yin, Yuhui Qiu and Jike Ge. (2007) 'Research and Realization of Text Mining Algorithm on Web', in International Conference on Computational Intelligence and Security Workshops, pp. 413- 416.

[13] Manu Aery and Sharma Chakravarthy. (2005) 'eMailSift: Email classification based on structure and content', in Proceedings of the Fifth IEEE International Conference on Data Mining.

[14] Crawford, Kay and McCreath. (2001) 'Automatic induction of rules for e-mail classification', in Proceedings of the Sixth Australasian Document Computing Symposium, Coffs Harbour, Australia.

[15] Yun-Qing Xia, Jian-Xin Wang, Fang Zheng and Yi Liu. (2007) 'A binarization approach to email categorization using binary decision tree', in Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp. 3459-3464.

[16] Qing Yang and Fang-Min Li. (2005) 'Support Vector Machine for customized email filtering based on improving Latent Semantic Indexing', in Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, pp. 3787–3791.

[17] Haiyi Zhang and Di Li. (2007) "Naïve Bayes Text Classifier", in IEEE International Conference on Granular Computing, pp. 705–711.

[18] Taghva, Borsack, Coombs, Condit, Lumos and Nartker. (2003) 'Ontology-based classification of email', in ITCC, Las Vegas, NV, pp. 194–198.

[19] Weng Sung-Shun and Liu Chih-Kai. (2004) 'Using text classification and multiple concepts to answer e-mails', in Expert Systems with Applications, vol. 26 n° 4, pp. 529–543.

[20] Thao Lê and Quynh Lê. (2002) 'The Nature of Learners' Email Communication', in Proceedings of the International Conference on Computers in Education.

