Automatic based-on-ontology classification of learners’ e-mails

Summary: The Information and Communication Technologies (ICT) emerged all the fields and became part and partial of the foundations of the modern society. It is for this reason that new forms of teaching appeared to substitute the traditional teaching for the so-called distance learning or E-learning which eliminated the inadequacies of the former and especially those temporal and spatial constraints. To integrate at best these new technologies in the field of teaching, numerous tools are indispensable to guarantee it among which we are going to focus on the exchange of e-mails between the tutor and their learners. The objective then of this article is to manage this big volume of e-mails by the display of a system of automatic classification of e-mails to ease the task of the tutor.

Key words: E-learning; automatic classification; e-mail; ontology.

I. INTRODUCTION

Because of the development of distance learning or (e-learning), many new technologies are installed for the best management and assurance of favorable means of communication to facilitate the exchange and the distribution of pedagogical resources via Internet.

Given this development, the e-mail became the main mechanism of communication between the tutor and his learners to assure the exchange of ideas, the request of information and of the clarification or instructions.

Of this fact the volume of e-mails that the tutor receives is getting more and more increased to influence the performance and the quality of the education because the tutor cannot control all the e-mails what leads to neglect some of them.

Therefore, the implementation of an automatized system for the management of the e-mails becomes indispensable, necessary and primordial.

This system has to contain essentially three modules:

- E-mails’ analysis.
- E-mails’ automatic classification.
- Prompt feedback of e-mails.

In this paper the automatic classification of e-mails is what interests us the most in that when the system affects a very specific category for every e-mail it facilitates well the task of the automatic answer.

In our team, one colleague proposed architecture of a system of classification based on a multi-agents system which allows classifying the e-mails in themes similar to what is presented in the figure below.

![Figure 1: Module of automatic classification of e-mails](image)

II. RELATED WORK

A. E-learning

1) Definition

The word E-learning is made up of “E” which is an abbreviation of electronics and “Learning”.

By evoking the E-learning, the various actors of the training often hint at the electronic learning that is at the provision of contents of training on the Internet or Intranet.

The E-learning is also defined as « a mode of learning based on the use of the new technologies, allowing the access to on-line training through Internet, Intranet or the other electronic media, to develop the skills, while making the process of learning independent of time and place. (Quebec’s office of the French language).
Thus we can say that the E-learning indeed has well acted on the elimination of the barriers of distance and the feeling of isolation, the easiness of exchange and the possibility of pooling of the resources and the complementarity.

The E-learning allows:

- to reduce the constraints of time and place,
- to individualize the paths of training,
- to favor the autonomy of the learner;
- to develop on-line pedagogical relationships with the tutor or the other learners.

2) Principle of evolution of E-learning: of the face-to-face interaction in E-learning:

Cegos which is a company of professional training particularly in the E-learning, drew up statistics which concern the distribution of the various types of training in France more in 2001 than in 2005 as illustrated in the following figure 3.

![Figure 3: Distribution of the various types of training in France from 2001 till 2005 (Source: Cegos on 2005)](image)

This distribution accentuates the deployment which the E-learning has known for four years in comparison with the other disciplines, this increase can be explained by the fact that this type of learning answers better the common and current needs of the learners.

This migration from face-to-face learning to the E-learning brings us to draw up a comparative study between them.

In fact, the E-learning possesses the big asset of the entire elimination of the barriers of time and space because the information is always available. Thus it is the learner who is in the center of concentration and not the tutor. Therefore, the learner is incited to be a transmitter of information and to participate significantly in the education and to not be limited to be just a receiver of information.

This advantage of the E-learning can never deny the big weight occupied by the education in face-to-face learning which still persists thanks to the simplicity of the collaborative work [3].

3) The modes of communication in E-learning:

The E-learning paradigm is based on three major methods of communication each has advantages and drawbacks. This communication can be made in a synchronous, asynchronous or mixed learning between the learners or between the learners and their tutors.

a) The synchronous communication:

In a synchronous communication, an exchange, a sharing of information and the communication between the learners and their tutors is made in real time.

We can note some elements used to assure the synchronous communication:

- Chat: tool of immediate communication, via the remote educational platform or with another system of Chat, as for example MSN Messenger.
- Board: still known for interactive digital board or interactive pedagogical board which indicates a board connected with a computer to liven up the group by realizing interactive educational sequences.
- Videoconference: characterizes a gathering of persons who are not in the same place during the conference but can see and hear the participants thanks to the video.
- Skype: software which takes back the same principle as the instant messaging with the same features, but by adding the feature of the videoconference.

b) The asynchronous communication:

In the asynchronous learning, the exchange with the other learners or with the tutor is recorded via modes of communication that do not require simultaneous connection.

We can note same examples:

- Forum: allows exchanging messages or questions with the tutor and the learners. This communication is profitable because messages are visible and clear.
- E-mail: is the tool of bilateral communication that offers the opportunity to communicate with one or several persons to assure the transfer of documents and the request of assistance or instructions in a remote education under forms of private messages.

c) The mixed communication

This method of learning combines the elements of the electronic learning and the traditional learning in traditional classes.

This modality of learning requires certain kinds of preparation, in that, before benefiting from a traditional education in class, the learners prepare for materials resorting to the technologies of information. The time of interaction is thus decreased by this
preparation and the training in class concentrates on the thorough exchanges.

Here is the figure which compares between the three already mentioned modes of learning:

**Figure 4: comparison of the 3 modes of learning**

<table>
<thead>
<tr>
<th>Mode</th>
<th>Distance</th>
<th>Time</th>
<th>Tutor</th>
</tr>
</thead>
<tbody>
<tr>
<td>The synchronous communication</td>
<td>Different place</td>
<td>Real</td>
<td>Present</td>
</tr>
<tr>
<td>The asynchronous communication</td>
<td>Different place</td>
<td>Delayed</td>
<td>Available in delayed time</td>
</tr>
<tr>
<td>The mixed communication</td>
<td>Different place</td>
<td>Real/Delayed</td>
<td>Present/Available in delayed time</td>
</tr>
</tbody>
</table>

B. The previous works of e-mails’ classification

The majority of the researches are focused on the processing of textual data because of the increasing number of textual documents put on the Internet which were led to surmount in the human intervention which is demanding especially in terms of time.

Among the existing textual data via Internet we find the electronic e-mails which are considered as the most popular means of communication, it is for this that we look for working out automatic methods which allow to deal, to analyze, to classify and to answer automatically these e-mails.

The following figure summarizes a general system of automatic classification of e-mails.

**Figure 5: General overview of a system of the e-mails’ management**

1) Method of classifiers’ combination: this method is based on a combination of several classifiers to gather their data with the aim of having a more relevant classification.

This method represents an innovation with regard to the previous works of classification of e-mails.

The general idea of this method is generated from the use of a set of classifiers instead of a unique classifier which can lead to an improved result [4] of the fact that it proposes a multi-staged technique of classification while using algorithms of learning’s with an analyzer which increases the precision of the classification.

The figure 6 illustrates the approach of this method:

**Figure 6: Method of the classifiers’ combination [6]**

This method aims essentially at detecting the unwanted e-mails (Spam) among those which are received in an e-mail box through techniques of filtering by using algorithms of classification to learn to predict those which are unwanted of a total corpus of e-mails.

The results obtained are of three categories:

- **TP (TRUE POSITIVE):** common exits of classification containing e-mails considered justifiable and not containing threats.
- **TN (True Negative):** common exits of non classification containing e-mails considered unwanted (Spam).
- **GL (Grey List):** not common exits.

At the end to succeed in this reliable classification it is necessary to respect and to follow two essential rules: the first one is that of the good choice of the classifiers which must be strictly independent and the second is that of the good choice of the function of combination.

This method as we have just mentioned is an innovation with regard to those previous ones but it does not imply that this method contains some inadequacies among which we can note the slowness in term of execution, the rate of errors which is very high, the task of combination of classifiers is a little bit difficult because e-mails are not the same and do not belong generally to the same subject what makes the choice of the methods of classifications the most suited a soft task.

2) Classification of e-mails by an approach based on the linguistic tools: this approach is proposed by O.Nouali, A.Regnier and P.Blache [7] and is based on the use of a set of tools which focus on the content and the structure of the e-mail rather than the lexical indicators.

The approach is allocated at first to the e-mail categorization among a predefined set of categories. Then, we enchain by the stage of classification which consists of grouping the e-mails which are similar in alike classes, the classes in this stage are not known in advance. Finally this approach ends with the use
of the linguistic modules reduced to what we associate linguistic properties [7].

This approach is made up of four main modules:
- A module of preprocessing
- A module of linguistic analysis
- A module of classification and filtering

An overview of this approach is described in the figure below:

![Figure 7: Architecture globale du système de classification des e-mails par une approche basée sur les modèles linguistiques [7]](image)

In this architecture the module of classification and filtering interests us most. The latter as its name indicates is composed of two stages: the classification and the filtering.

The classification is a task which comes just after the linguistic analysis, it consists of affecting a predetermined set of classes, the category the most suited to the e-mail received by the system. This task consists of prefiltering [7].

Once the task of the classification is achieved, the process of filtering starts. This task consists of undertaking actions of filtering according to the ambitions of the user. The modeling of the interests of the users is an important task for a system of filtering [7].

Though this approach contains several advantages, it also contains some inadequacies of the fact that in the practice the user himself cannot always judge the typology of his message.

3) The classification of content-based e-mails: this approach, known as "eMailSift", takes as an entry a corpus of e-mails that passes through a set of stages classified in similar classes.

This approach rests upon the content and the structure of the e-mail as a method of automatic classification.

The global flow of this system is described in the following figure:

![Figure 9: Global Architecture of « eMailSift » system [5]](image)

In this figure we notice the use of six modules:

a) The file of e-mail or initial corpus of e-mails (Email Folder):

This module is considered as a stage of preprocessing to eliminate the empty words (stop word), the spaces, the punctuation marks …

Then, the words whose frequencies is between 60 and 80 % of the sum of all the frequencies are tidied up and held to be used as an entry for the following stage.

b) Graphic representation (Graph Generator):

This module consists of transforming the electronic messages in a graphic format known as "graph mining".

This representation has an impact on the classification because the generation of a graphic representation of a text is different from that of an e-mail.

c) Extraction of under-structures (Graph Mining Technique):

This module resorts to the technique of graph mining to extract under-structures.

d) Reduced Representation (Pruning):

The exit generated by the techniques of graph mining consists of a large number of under-structures. The objective of this stage is to identify those which are necessary to reduce calculation times that are in proportion to the number and to the size of the generated graphs.

Those which are retained must be frequent in the initial corpus of e-mail.

e) Ranking of under-structures (Ranking):

The under-structures retained are tidied up according to their frequencies of occurrence, the size of the e-mail and of the under-structure.

The rank associated in every under-structure is used to classify the received e-mails.
f) The classification:

The classified under-structures are compared with every entering e-mail to determine whether one of them appears in the graphic representation. If there is a correspondence then the e-mail will be classified in the corresponding file, otherwise, the rank associated with under-structures is used to decide which file will be classified.

III. Proposed Solution

A. General description of the proposed solution

In this section we present our entitled approach “Automatic Classification of e-mails on Ontological basis”.

Our approach consists of classifying automatically the e-mails sent by the learners to a prospect of distance learning; it then allows attributing precise categories for every e-mail to ease the task of the automatic answer.

Our approach requires an e-mail supposedly analyzed, thus the work starts by a continuation of analyzed words establishing a particular e-mail, the latter passes through a continuum of stages to obtain a well classified e-mail.

A global view of our approach is presented in the following figure:

According to the global view of the module of the automatic classification of e-mails, every e-mail reaching our system is registered in a Data Base of Analyzed E-mails (DBAE) followed by the stage of the analysis of e-mails. This base stores all the XML documents of the analyzed e-mails. Every document contains the analyzed words by every e-mail which is going to serve in the phase of the automatic classification.

The result of this module is the classified e-mails which will afterward be stored in a Datum Base of Classified E-mails (BDCE).

At this stage, we judge that the module of automatic classification of e-mails will be effective only after making a semantic classification of e-mails. To come to terms with this objective, our intervention will particularly question not only the search for an existing ontology but instead the creation of an appropriate ontology for the domain of the distance learning which will be very useful in our module of automatic classification for ontological base.

B. Detailed description of the proposed solution

1) Conception of the ontology

Our solution consists first of all of building ontology. To do it, this construction has to be preceded by a conception which draws the detailed chain of the construction described in the figure below.

To create our ontology, it is necessary to follow miscellaneous stages to end in a threshold of desired reliability and certainty. To achieve this goal, the generation of the ontology begins with:

- Stage 1: in this first stage we—as designers of ontology are going to end in the collection of the documents that we need for the creation of the ontology; it is a question in fact of defining the field of study (which is in our case that of the domain of distance learning); by crossing every corpus.

Legend:

DBAE: Data Base of Analyzed E-mails
DBCE: Data Base of Classified E-mails

Figure 10: Global overview of the module of automatic classification of based-on-ontology e-mails
Stage 2: it is the main stage in the creation of our ontology. This stage can be decomposed into three sub-stages:

- Terms’ extraction: in this stage we are going to be interested essentially in fixing the primitives of the domain through extracting the candidate terms which will be used later in the creation of the ontology.
- Search for the synonyms: this phase consists of looking for synonyms for every term.
- Extraction of the semantic relations: at this stage we fix the relations connecting the terms. In our ontology we are going to be satisfied by two types of relations which are:
  - The relations of equivalence: which manage the relations of synonymy between the terms.
  - The hierarchical relations: they are decomposed into two sub-relations given:
    - The son-of relation.
    - The father-of relation.

Stage 3: this stage consists of the representation of the ontology in a well defined language to be treated. In our case we used the XML language (stretchable Markup Language).

Stage 4: through the previous stages, we acquired a well organized and structured vocabulary which can be interpretable by a machine.

2) Detailed architecture of the module of automatic classification of e-mails

In what follows, we are going to describe the various under modules of our module of automatic classification on ontological base which can be divided into two main shutters:

- Module of extraction of concepts.
- Module of automatic classification

![Figure 12: Detailed architecture of the module of automatic classification of based-on-ontology e-mails](image1)

a) Module of extraction of concepts

The ontology supplies the concepts, the synonyms and the relations between them. The idea here is to benefit not only from assets of use of ontology but also to make appeal to multi-system agents.

Agent’s paradigm is described according to Russell and Norwig [9] as being a software entity capable of perceiving its environment thanks to very specific sensors; and which allows to answer the objectives of conception [11].

A Multi-Agent System (MAS) is described as being a set of agents communicating between them by means of messages [1]; and which are created to offer a very precise service and which are in cooperation and in collaboration between them. Seen the flexibility of this modeling, our choice of use of the MAS is concretely explained.

For this, we led a module of automatic classification of the e-mails of the learners which contains a set of agents among whom each holds a very specific role and which will be explained in what follows.

An overview of the module of extraction of concepts is presented in the figure below:

![Figure 13: Sub-module of concepts’ extraction](image2)

1) Process of ontological research

In our approach, we create an ontology which concerns the domain of distance learning to improve the process of research for information and for extraction of key concepts to solve the problems of processing and facilitate the extraction of the
information which will be communicated between the intelligent agents.

To do it, the ontological agent receives the list of the analyzed words and extracts the concepts in relation by making a search in the ontology. If the concept exists already then it will be transferred to the classification agent, otherwise we pass through a stage of enrichment of ontology to make its update.

The process of ontological search is realized by going through the ontology block by block; these blocks present all the possible combinations of the sentence.

If we find the significant combination of our sentence received by the process of analysis, the latter will be transmitted in the process of automatic classification. Otherwise we pass to the module of enrichment of the ontology.

② Process of ontological enrichment

Seen the fast evolution of the information over time, the ontology becomes sometimes incapable of including everything, hence the necessity of updating our ontology with new terms of the domain. Our module of automatic classification needs ontology much richer to have better results. The process of enrichment of ontology is specially led to complete the missing and to limit the human intervention as far as possible.

The Figure 14 presents a detailed description of sub module of ontological enrichment:

The sub-module of enrichment of ontology requires the presence of a global ontology which includes quite the concepts and the terms which are not existing in the initially created ontology. For this, we are also going to create this global ontology which is more specific, richer and more voluminous than the initial ontology.

In this stage if a concept is not found in the ontology then the agent researcher will be in charge of going through the global ontology to find all synonyms for the concept in question. If the agent researcher finds it, he crosses quite synonyms to the agent actualisor concept which updates the initially created ontology. And if the agent researcher does not find a synonym, then it indicates an “indefinite class” and it is the role of the tutor to classify manually the e-mail then it updates the base of datum of classified e-mails and updates in its turn the initial ontology.

b) Module of automatic classification

In our work we are based on a classification of the e-mails of the learners which presents a part of the work of Thao and al. 2002 [10]. Out of 1478 e-mails which were collected for the analysis of data, the results showed that these e-mails concern different subjects which can be grouped according to three main functions:

- Procedural Function.
- Social Function.
- Cognitive Function.

This analysis shows well the ascendancy of the procedural function as the main objective of the communication of the students with their tutors via e-mail is mainly designed for the request of assistance or the clarification of a given subject.

These functions are afterward distributed according to the study led by Thao and al. 2002 [10] in sub-function:

① Procedural Function
  - Confirmation: confirm a given point.
  - Request: ask for something from the tutor.
  - Clarification: ask for a clarification.

② Social Function
  - Greeting: greet the tutor.
- Reference: discussion about another person.
- Thanking: thank the tutor.
- Complain: complain about a certain procedure.
- Share: distribute personal experiences.

Cognitive Function
- Discussion: discuss a given topic.

Given that we focus here on e-mails for the communication between the learners and their tutor, we are going to be satisfied by the first two functions (procedural and social) and we are going to neglect the cognitive function that is mainly used for the discussion and that there are other means more appropriate than the use of e-mails assuring the discussion as the Chat, the Facebook, ...

The figure below schematizes the general speed of it or the classes that an e-mail with certain modifications can set as we consider more suited to our approach:

Figure 16: The set of classes that can take an e-mail

So, according to the key concepts obtained by the module of extraction of concepts the classification agent will be able detect the class of every targeted e-mail by executing a well precise algorithm described in the Figure 17.

CONCLUSION

With the aim of decreasing the burden of the tutor who is sometimes incapable of managing the gigantic volume of the received e-mails, our approach comes to solve this problem while benefiting from the use of ontology in a perspective of multi-agent system.

Our future works do not consist only of using an ontology which concentrates on a very specific domain but also of a global ontology gathering more than one domain (medical, agriculture) which will be divided into local ontologies.

ACKNOWLEDGMENT

Special thanks to Mr. Mahmoud NEJI and Mrs. Awatef Aloui for your thoughtful and creative, your encouragement, your advice and support in the first final of my work.

Also, I offer my regards and thanks to all who encouraged me to write this paper.

REFERENCES