A METHODOLOGY FOR INTELLIGENT E-MAIL MANAGEMENT

Francisco P. Romero,
Soluziona Software Factory, R+D Center Soluziona-UCLM, Ronda de Toledo, s/n, 13071 Ciudad Real, SPAIN

Jose A. Olivas
Dep. Of Computer Science, University of Castilla La Mancha,Paseo de la Universidad 4, 13071 Ciudad Real, SPAIN

Pablo Garcés
Dep. Of Computer Science. University of Alicante,Carretera San Vicente del Raspeig, s/n, 03080 Alicante, SPAIN

Keywords: e-mail, soft-computing, fuzzy logic, automatic classification, clustering.

Abstract: We present, in the context of the intelligent Information Retrieval, a soft-computing based methodology that enables the efficient e-mail management. We use fuzzy logic technologies and a data mining process for automatic classification of large amounts of e-mails in a folder organization. It is also presented a process to deal with the incoming messages to keep the achieved structure. The aim is to make possible an optimum exploitation of the information contained in these messages. Therefore, we apply Fuzzy Deformable Prototypes for the knowledge representation. The effectiveness of the method has been proved by applying these techniques in an IR system. The documents considered are composed by a set of e-mail messages produced by some distribution lists with different subjects and languages.

1 INTRODUCTION

Every day e-mail becomes more important as a method for global communication. The Wordtalk Corporation (Harrys, 2002), estimates that 60 million professionals use e-mail. According to a report from technology analysts IDC, on an average day in 2000, 9.7 billion e-mail messages were sent worldwide. By the year 2005 that number will grow to 35 billion.

Along with the increase in messages size and traffic, comes an increase in users' expectations. People send e-mail to communicate critical information across continents and time zones: they expect their information to get through. E-mail glitches are high profile events, generating negative press for the organization, frustrated users, loss of productivity, and most importantly loss of business.

Every stored message contains information that, depending on the circumstances, may become relevant. To make the recovery of the messages received in the previous weeks, months or years easier, the mail management programs allow the organization of them in specific folders well-defined by the user.

The most common thing for a user is to have from ten to a hundred folders hierarchically organized. Moving a message to its proper folder involves a considerable effort and a waste of time. Some commercial programs allow the definition of simple rules to help in this task, but these rules only cover a very small percentage of the range of messages received, generating lots of erroneous classifications, irrelevant for the user. The task of managing the information inherent in the messages of a list with a great amount of daily entries is quite complex.

This study is concerned with Web mining. This term is used to describe three different types of data mining, namely *content mining*, *usage mining* and

P. Romero F., A. Olivas J. and Garcés P. (2005).
A METHODOLOGY FOR INTELLIGENT E-MAIL MANAGEMENT.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 11-16
DOI: 10.5220/0002533000110016

structure mining. The mining of textual data is a common web mining task, often for the purposes of information retrieval. This type of mining is becoming increasingly necessary as finding information on the Web is almost impossible without automated assistance.

In this context, the most of the related work are based on Naive-Bayes classifiers applied to the documental categorization (Turenne, 2003). Also it is necessary to mention the classic works of electronic mail processing (Sahami, 1998), (Kiritchenko, 2002), or the application of supervised learning to this problem (Joachims, 1998).

In this study, a methodology based on different *soft-computing* techniques (Nikravesh, 2001) to manage great amounts of e-mails is proposed. The aim is to solve a specific problem in the control of the e-mail system: The management of distribution lists. A distribution list is nothing but a list of users who are associated in order to exchange information related to a specific subject by using the e-mail.

The main targets to reach are the following:

- Hierarchical (and fuzzy) organization of a large amount of messages received (1.500 messages approx.) based on the concepts they involve.
- Automatic sorting of the incoming mail on the list from their contents and without the intervention of the user.
- Navigational or conceptual searching inside the messages of the distribution list.

To test the effectiveness of the techniques and make the necessary adjustments, diverse methods have been used (outliers, similarity degree…). Studies about distribution list concerning both technological domains and general and historical matters have been made. The aim is to show the different problems present in e-mail management tasks and the general applicability of the proposed method.

The rest of this paper is organized as follows. In Section 2, we present a methodology for fuzzy hierarchical e-mail classification and we introduce and explain a method to deal with the incoming messages to keep the achieved structure. We explain the experimental results in Section 3 and finally, we conclude this work in Section 4.

2 A METHODOLOGY OF E-MAIL MANAGEMENT

The quantity of received messages in the distribution list is frequently very large. This means that the organization to be build should be the most efficient possible to make the later exploitation optimum. The characteristics of the later results will be the following:

1. Hierarchical organization of the messages in folders following the criteria based both on the "conceptual" content of the message and its structured fields. The conceptual representation will allow us to achieve a concept based search.
2. Definition of the folders through relevant terms/concepts and membership functions.
3. Possibility for a message to get stored in more than one group, depending on its contents.
4. Every message should have a degree of affinity (or membership) with each of the groups into which it is sorted. There should exist the opportunity to arrange the messages from every group according to this degree of membership.
5. Basic "matching" mechanism between documents and folders. It will be used to store an incoming message in one or more folders depending on the affinity with the different ones.

The construction process is based on the following stages: linguistic pre-process, conceptual representation, message clustering, post-process and results (Figure 1).

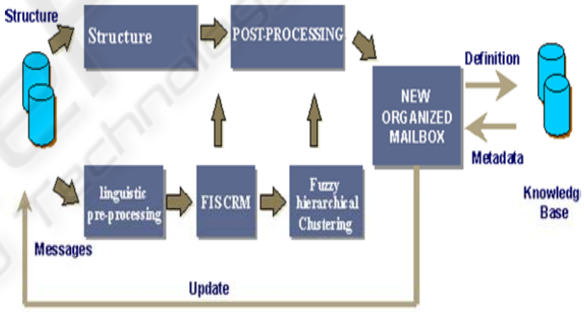

Figure 1: Building the new structure

2.1 Linguistic pre-process

The Linguistic Preprocess consists of the following tree steps: *previous transformation*, *lexical analysis*, *stoplist word* removal and *stemming*, and candidate keyword detection.

- *The Previous Transformation* has been performed in the separation of the structured fields (subject, sender, etc.) of the contents of the message. Also we are grouping ("*starred*" in *gmail* tool) the messages according to conversations, meaning groups of messages related to the same subject (ignoring Re:, Rv: and other particles) and uniformly arranged in sequences.
- *The Lexical Analysis* is the process of converting an input stream of characters into a stream of words or tokens (Frakes, 1992).

Lexical analysis tasks are: management of numbers, punctuation, singulars, special words and proper nouns. This stage produces candidate terms that are further checked and retained if they are not in a *stoplist*.

- *Stoplist Word Removal*: *Stoplist* is a list of words that are most frequent in a text corpus and are not discriminative of a message contents, such as prepositions, pronouns and conjunctions.
- *Stemming*: *Stemming* is the process of suffix removal to generate word stems. Several different methods for automatic *stemming* are described in (Frakes, 92). One of them, Porter stemming algorithm, is the most common and it was used in the system.
- *Ranking Candidate KeyWords*: a candidate keyword had to appear in at least three documents and in no more than the 33% of all documents. Only the candidate keywords are useful for the following stages.

2.2 Conceptual Representation

To be able to work with the messages in an abstract way, securing a logical representation of them is essential. The vector space model and its extensions (Pasi, 2002) have been used traditionally. FIS-CRM (Olivas, 2003) can be considered as an extension of this traditional model, in charge of the representation, within the vector attached to the document, of the concepts inherent in the words displayed.

Therefore FIS-CRM, is based on two main points:

a) If a word appears in a document, its synonyms that represent the same concept underlie it.
b) If a word appears in a document, the words that represent a more general concept underlie it.

The fundamental basis of FIS-CRM is to "share" the occurrences of a contained word among the fuzzy synonyms that represent the same concept, and to "give" a fuzzy weight to the words that represent a more general concept that the contained one. To obtain this aim, documents must be first represented by their base weight vectors (based on the occurrences of the contained words) and afterwards, a weight readjustment process is made to obtain a new vector (based on concept occurrences). In this way, a word may have a fuzzy weight in the new vector even if it is not contained in it, as long as the referenced concept underlies the document.

To carry out the readjustment, the synonymy and generality fuzzy interrelations has to be taken into account, respectively obtained from a fuzzy dictionary of synonyms (Fernandez-Lanza, 2001) and an ontological (Kiryakov, 1999) one. The process to be used in the conceptual representation of already pre-processed e-mail messages on a distribution list consists of the following steps:

1) Indexation of all the terms obtained in the pre-process.
2) Building synonymy and ontology matrices by storing synonymy and generality degrees from each pair of words in the index.
3) Representation of the messages using the classic vector space model.
4) Readjustment of the vector weights using the FIS-CRM formulae group.
   a) The vector readjustment made using the synonymy interrelation is hindered by the fact that there are lots of polysemic words (words with several meanings).
   b) The vector readjustment made using the generality interrelation is linear and proportional to the generality degree between term A and term B.
5) Generation of the similarity matrix which will store the degree of similarity from every pair of messages in the collection. The matrix will be the input to the later clustering process.
6) Storage of the essential information as meta-data to allow the management of later incoming messages.

2.3 Messages Clustering

Using the clustering process we will achieve the splitting up of the collection of messages in a reduced number of groups made up of messages with enough conceptual similarity. Each group will contain one or more relevant terms which will make it different from the rest.

In this work, a hierarchical fuzzy clustering approach is presented. The clustering procedure is implemented by two connected and adapted algorithm. It uses a fuzzy hierarchical clustering algorithm to determine an initial clustering which is then refined using the SISC (King-Ip, 2001) clustering algorithm used in FISS meta-searcher structure.

This algorithm is characterized by creating an initial number (automatically calculated) of centroid clusters, followed by an iterative process that includes each document in the clusters whose average similarity is upper than the threshold of similarity (automatically calculated, but user specified if wanted). The algorithm also considers merging clusters and removing documents for clusters when their average similarity decreases under the threshold. In order to get a hierarchical structure, big clusters and the bag cluster (formed by the less similar documents) are reprocessed with the same method.

The resulting organization is hierarchical, so, from a large mailbox we will obtain a tree folders organization. It will also be a fuzzy organization in which the messages will be located in more than one group with different degrees of connection to each one.

2.4 Post-Process and Results

The results obtained in the clustering process should be completed to give way to a new organization of folders which carries out the required characteristics. This process is divided into different tasks.

- *Securing the complete definition of each folder:* identification, most relevant terms and hierarchical relations.
- *Processing the results to achieve the complete definition of the folder* using Fuzzy Deformable Prototypes (Olivas, 2000) which will make both analysis of the incoming messages and structure updating easier.

The update of the structure is accomplished by order of the user because disorganization of the mailbox or changes in his preference criteria. Periodically, in a batch clustering process, the structure can be re-built reapplying the clustering process and reusing the previous organization.

2.5 Administration of Incoming Messages

The task of classifying each incoming message automatically and correctly is complex so it is essential that the analysis and sorting operations are carried out in a process which is clear to the user, who only has to be aware of the secured results, without delays or loss of effectiveness.

The process to deal with each of the incoming messages is the following:

1. Linguistic pre-process of the message: Elimination of *stop words* and *stop zones*, and *stemming*. Determination of its being within an open conversation.
2. Construction of conceptual representation using FIS-CRM techniques and matrices calculated in the previous process.
3. Comparison between the characteristics of the message and the characteristics of each folder. If the message were within a conversation, the comparison would be made with a subgroup of folders and not with all of them. To calculate the connection to each folder, one should use inference with Fuzzy Deformable Prototypes (Olivas, 2000).
4. Storage of the message in those folders in which has been reached a positive relation. Updating of the model if the amount of incoming messages or the state of the mailbox (folders which are too overloaded) require it.

3 EXPERIMENTAL RESULTS

3.1 Linguistic Pre-process

In this stage, the reduction achieved can be observed through the steps already explained: conversations grouping, elimination of superfluous elements and reduction to significant lexemes through stemming.

These results will allow a later management of the relevant aspects of the messages (Figure 2 and Table 1).

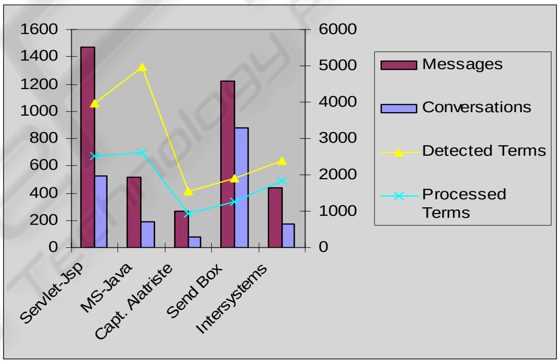

Figure 2: Pre-processing Results.

Table 1: Preprocessing results.

Collections	Messages	Convers.	Detected terms	Processed terms
Servlet-Jsp	1468	521	3958	2746
MS-Java	516	192	4960	2909
Alatriste	270	81	1545	1087
Send Box	1218	881	1890	1367
Intersystems	436	176	2376	1541

3.2 Conceptual Representation

The use of FIS-CRM as a conceptual representation method generates an increase in the degree of similarity between messages (see table 2). We used two metrics:

- Mean similarity of each element with the rest of the set.

- Number of Outliers: An outlier is an element that has no similes (or a low number), and it is difficult to conveniently group it.

In this study, the synonyms dictionary presented in (Fernandez-Lanza, 2001) and put into practice in FISS (Olivas, 2003) has been used in collections of messages in Spanish. At the same time, the management of distribution lists whose subject matter was the JAVA technology has also supported itself on ontologies generated in a semi-automatic way.

Table 2: Similarity Differences.

Collections	Similarity without FISS.	Similarity with FISS	Outliers without FISS.	Outlier FISS
Servlet-Jsp	14,53	38,89	489	242
MS-Java	12,82	21,87	181	48
Alatriste	14,06	20,02	133	23
Send Box	10,05	17,36	529	311
Intersystems	8,77	13,67	145	72

3.3 Clustering

The number of obtained groups and their average depth is given in the following table (table 3).

All groups cover a particular but not disjunctive subset from that of the messages in the distribution list domain. At the same time, we can observe that the number of associations remains within several limits in which the user can keep his message collections controlled.

Table 3: Clustering results.

Collections	Root Groups	Total Groups	Unclassified
Servlet-Jsp	12	35	110
MS-Java	6	11	50
Alatriste	6	14	12
Send Box	24	51	87
Intersystems	10	18	28

3.4 Post-process and results

The obtained message folders become easier to use and more significant than the folders to be built through user rules. Each folder has a complete definition of its characteristics which will allow the processing of incoming messages.

3.5 Administration of incoming messages

In the following table, the management of a particular case (in Spanish) is shown (table 4):

Table 4: Management of a particular message.

1. INCOMING MESSAGE	*Error Message in Tomcat***:** hola amigos de la lista: alguno sabe la manera de personalizar las páginas de mensajes de error de status HTTP en tomcat 4.1.12 en particular me interesan los mensajes correspondientes a acceso denegado((403) recurso inexistente (404)	
2. PREPROCESSSING	Messages 12 tok. Subject(3 tol.) Keys (4)	stemming: 12 Sin./ Ont: 20
3. CONCEPTUAL REPRESENTATION: FIS-CRM		
4. MATCHING	Relevant Terms:	Tomcat, error, access, resource, status, HTTP
	Comparison between the characteristics of the message and the characteristics of each folder.	
5. RESULTS	Apache Tomcat	80%
	Errors	60%

To test the real utility of the obtained structure, it has been used the rest of the collections of messages (over the 33% of the messages used in the training process).

To evaluate the automatic classification process, the "Number of correct classifications (NC)" has been used as a discriminator metric. It separates the number of correct classifications at level 1 (NC1), at level 2 (NC2) or at end groups level (NCH). For all of them it is used the number of well classified messages divided into the total of messages processed. The obtained results for each one of the collections can be observed in Table 5.

Table 5: Results of the load test.

Collections	Messages	NC1	NC2	NC3
Servlet-Jsp	387	81,8%	78,8%	62,73%
MS-Java	163	75,4%	66,7%	54,12%
Capt. Alatriste	113	85,6%	79,8%	64,56%
Send Box	320	83,2%	66,7%	63,19%
Intersystems	143	77,4%	57,6%	43.29%

The performance of the management of incoming messages process is acceptable if it happens within the period of time from the moment the message comes into the server to that of its discharge by the user. The obtained classification generates practically no changes in the folders structure obtained in the initial process; this means that the set of initial messages was relevant and the obtained organization close to the optimum one.

4 CONCLUSION AND FUTURE WORK

In this exposition, a working method to solve the administration problems of overloaded e-mailboxes has been presented. To do so, linguistic pre-processing, advanced conceptual representation using FIS-CRM and soft clustering algorithms specifically modified for this task have been used. At the end of the process we have a hierarchically structured e-mailbox that allows the highest degree of exploitation with a minimum effort.

The validity of the method has been tested through experimentation on e-mail messages from mail distribution lists of different kinds and languages.

Compared with other similar systems, the proposed methodology provides a richer hierarchical structure due to the use of fuzzy logic and fuzzy interrelations in its construction. Concerning the performance, the use of an improved iterative clustering algorithm and the FIS-CRM based conceptual representation processes make the performance closer to responses based on classic algorithms such as *fuzzy c-means* or *k-means*.

For its practical use, a user-friendly Web tool, that allows users the administration of their messages in an efficient way, has been built. The building process of this tool is based on top technologies like Java and XML. Its flexibility and portability makes it useful in any environment.

Nevertheless, there are several points in which the process needs improvement, so further studies are essential. Some main points are the following:

- Linguistic functions for the management of specific characteristics of the messages and spelling correction of terms.
- Improvements in identification of the language used in the messages and inclusion of Multilanguage dictionaries.
- Improvements in conceptual representation through the exploitation of context and other factors.
- More efficient clustering algorithms using improvements and/or adaptations of traditional algorithms such as fuzzy c-means and Kohonen Maps.
- Widening of the direct, conceptual and navigational search with the possibility of a phonetic search.

REFERENCES

Fernandez-Lanza, S., 2001. A contribution to the automatic processing of the synonymy using Prolog, PhD Thesis, University of Santiago de Compostela.

Frakes, W. B., & Baeza-Yates, R., 1992. Information Retrieval: Data Structures & Algorithms. Prentice Hall. Englewood Clifss, N. J.

Harrys, D. & Heather C. 2002. Wordtalk releases first Internet e-mail corporate usage report; concludes e-mail abuse at epidemic levels. http://www.wordtalk.com.

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of the European Conference on Machine Learning, Germany.

King-ip L., Ravikumar, K., 2001. A similarity-based soft clustering algorithm for documents. Proc. of the Seventh Int. Conf. on Database Sys. for Advanced Applications.

Kiritchenko, S. & Matwin, S. 2002. Email Classification with Co-training, Proceedings of the CASCON'02 (IBM Centre for Advanced Studies Conference), Toronto.

Kiryakov A. K. & Simov K. I., 1999. Ontologically supported semantic matching, Proceedings of NODALIDA'99: Nordic Conference on Computational Linguistics.

Nikravesh, M. & Azvine, B. (eds.), 2001. Proceedings of the 2001 BISC International Workshop on Fuzzy Logic and the Internet, University of California-Berkeley.

Olivas, J. A.; Garcés, P.; Romero, F. P.,2003. An application of the FIS-CRM model to the FISS metasearcher: Using fuzzy synonymy and fuzzy generality for representing concepts in documents. Int. Jour. of Approx. Reasoning (Soft Computing in Recognition and Search).

Olivas, J. A., 2000. Contribution to the experimental study of the prediction based on Fuzzy Deformable Categories, PhD Thesis, University of Castilla-La Mancha, Spain.

Pasi, G., 2002. Flexible information retrieval: some research trends. Mathware & Soft Computing.

Sahami, M. et al. 1998. A Bayesian Approach to Filtering Junk E-Mail, in Proceedings of the AAAI Symposium.

Turene, Nicolas, 2003. Learning Semantic Classes for improving Email Classification, Proceeding of the Text Link Conference.