CHeM: A System for the Automatic Analysis of e-mails in the Restoration and Conservation Domain

Luciana Bordoni

ENEA, Funzione Centrale Studi
Via Anguillarese,301, 00060 Roma
E-mail: bordoni@casaccia.enea.it

Leonardo Pasqualini

University of Roma Tre
Via della Vasca Navale 79, 00146 Roma
E-mail: l_pasqualini@hotmail.com

Filippo Sciarrone

Open Informatica srl
Via dei Castelli Romani 12/a, 00040 Pomezia (RM), Italy
E-mail: f.sciarrone@openinformatica.org

Abstract

In this paper, we present the CHeM system, Cultural Heritage e-mail Manager, a support system for the analysis of e-mails of the *Restoration and Conservation* newsgroup, hosted by the Yahoo portal from December 2000 to January 2003. The complexity of the domain as well as the specificity of the e-mails, prompted us to build the first system prototype based on a client-side architecture, to help less expert users in classifying information contained in e-mails. The system goal is therefore to provide an instrument capable of classifying the received messages, downloaded onto the users' desktops, into standard categories, based on their content, using the well-known techniques of Data Mining and Information Retrieval. The categories thus obtained are then used to label the messages in order to provide valuable information on the domain and therefore support specific information retrieval and produce new user groups by an automatic generation of mailing lists. The methodology presented and the first test results are encouraging with a view to porting the system in other similar domains.

1. Introduction

The development of advanced data analysis tools, such as those offered by Data Mining is today one of the most studied fields in informatics. This research area is highly interdisciplinary, involving aspects of data modelling, storage and retrieval, as well as aspects of statistical analysis and artificial intelligence. In particular, the branches of artificial intelligence have contributed enormously to the development of the sector, providing solutions for information representation, automatic knowledge extraction and interfacing with inexperienced users (Berry & Linoff, 1997; Berry & Linoff, 2002). Therefore, there is clearly an enormous amount of solutions developed by software houses that support decision-making nowadays and that make use of Data Mining and Information Retrieval techniques, so many it would be impossible to list them all. Precisely for this reason our system, in such a context, is not so much distinguished by its uniqueness or innovative techniques, than by the field it was contemplated and developed for. In fact, of all the instruments available on the market today, there is not one that was developed specifically for the domain of *Restoration and Conservation of Artistic Heritage*, although many can be adapted to various application contexts. In our opinion, a dedicated system produces better results for such a specific field.

The domain under consideration comprises the messages of the *Restoration and Conservation* newsgroup, on the Yahoo! portal, from December 2000 to January 2003 (http://it.groups.yahoo.com/group/restauro). Sergio Tinè, an architect, created the group with the aim of promoting cooperation and exchanges of experience and know-how between operators in the field. The messages, produced by members of the newsgroup, were simultaneously published in a forum on the Yahoo! site and received by all members in the form of e-mails. We addressed the problem to provide users with an instrument capable of categorizing the messages downloaded by the e-mail server onto their desktops into standard categories, on the basis of their content, so as to facilitate a first analysis. In fact, users currently base the analysis of e-mails on only their subject field that distinguishes the e-mail and enables the identification of messages on the same subject. However, it is evident that, too often, this is insufficiently discriminating and, in most cases, lacks detail and therefore to use it as the only element to sum up the body of messages is not plausible in order to make a structured analysis of the domain.

Everyone knows that it is often impossible, just by reading the subject of an e-mail, to recognise the exact subject of discussion and that one is obliged to open irrelevant e-mails to make sure that no important information is lost. This project aims to provide a solution to this problem, using the clusters identified by the system, appropriately indexed, so as to provide a detailed label for each message that sums up their content. The categories are univocally identified by keywords that are characteristic of the domain under examination.

This article is composed of the following sections. In Section 2 the architecture of the system and its constituent blocks are illustrated. Section 3 briefly illustrates the technological platform used. Lastly, the results of the first test are given in Section 4.

2. The System

Figure 1 illustrates the general architecture of the system. It consists of four separate blocks, and each one interacts with the user through a suitable graphic interface: *mail manager*, *vector creator*, *cluster creator and cluster viewer*.

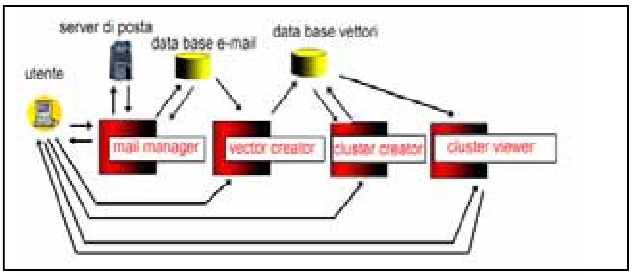


Fig.1: General Architecture of the System

2.1. Mail Manager

This module's task is to manage e-mail downloads from the user's electronic mail pop3 server and to store them in a table of the relational database embedded in the system. Track of the key information in every message is kept: address, sender ID, subject and text.

2.2. Vector Creator

The task of this module is to produce a dictionary of terms that are meaningful to the domain and to represent documents in a vectorial form, which simplifies their manipulation and comparison. The classic Vector Space Model, used by Information Retrieval to represent documents, was applied to achieve this, according to which, each document is described by a set of representative terms called *features*, which are simply terms in messages whose semantics help remember the subject of the message (Baesa-Yates & Ribeiro-Neto, 1999).

It can clearly be observed that, with regard to message content, not all the terms have the same representational capacity, therefore when creating a representation model, it is important to select only the most significant ones, useful for indexing texts (indexing phase). These will go to make up the *template*, that is, the dictionary of keywords of the domain. In order to reduce its size, a *stemming* algorithm (Porter, 1997) is used to reduce each term to its morphological root (stem). This template will therefore only be made up of the stems of the features identified in the previous step. Although the stemming operation considerably reduces the size of the initial dictionary, this is not always sufficient to have a reasonably sized template, which means that further filtering of the set of features is required.

This is performed by selecting the most significant terms in the template on the basis of an evaluation function which can either be *Information Gain* or *Document Frequency* (Yang & Pedersen, 1997; Yang & Liu, 1999).

In order to obtain a good template, as a product of the indexing and stemming phase, knowledge of the domain was essential and enabled non-generic results to be produced, that nevertheless were well adapted to the particular environment considered.

2.3. Cluster Creator

This module is the most important function block of the system; it has the task of categorizing the messages vectorized in the previous step, by means of clustering algorithms. Several clustering techniques can be used in the system: *K-means* (Likas, Vlassis & Verbeek, 2003), *DBSCAN* (Ester et al., 1996) and *Hierarchical Clustering* (Boudaillier & Hebrail, 1997; Dash et al., 2003; Serna, 1996). The user can select one or two or all three to study the different results and select the best one. These three techniques were opted for because they are representative of the three main categories of clustering algorithms: partitioning algorithms, density-based algorithms and hierarchical algorithms, and because in their categories, they are the ones capable of producing sufficiently adequate results also for users that are not familiar with the domain in question (Michaud, 1997).

2.4. Cluster Viewer

The task of this module is to display the results obtained from the clustering process. In order to do so, during the design stage, the aim was to provide the user with user-friendly tools. We opted to display the most representative parameters, such as the internal and external statistical standard deviations, and display clusters in their totality (showing the number of messages contained, the text, their representative features and a significance index for each one) and to provide a user-friendly graphic tool. For this purpose, the *Multidimensional Scaling* technique was implemented, using a metric method (Borg & Groenen, 1997). The enormous advantage of graphic displays is that the result presented is clear, essential and immediate and is comprehensible even to users inexperienced in statistics and data mining techniques. At the same time, it was also necessary to provide a yardstick for the various clusters obtained, to enable the user to select the best one and make a possible evaluation of them all.

3. Technological Platform

The entire system was developed using Java technology in a Microsoft operating environment; Microsoft Access was used to manage the Databases. This choice was motivated by the substantial potentials of the Microsoft package, its considerable efficiency that is its distinguishing mark, and the technological environment where the first version of the system was installed.

Lastly, the *Enterprise Miner* tool of the *SAS Institute* (http://www.sas.com*)* was used to assess the quality and accuracy of the findings of the tests for statistical analysis.

4. Test Results and Conclusions

The aim of the experiment was to test the system on a representative sample of the domain in order to evaluate performance in terms of its clustering capacity in order to help less expert users to study this particular domain and to find other similar users. It was also an opportunity to compare the implemented clustering algorithms to evaluate the support given by the system to help users classify documents, comparing the output obtained by manual classification.

The following test variables were used to evaluate the performance of the different clustering algorithms:

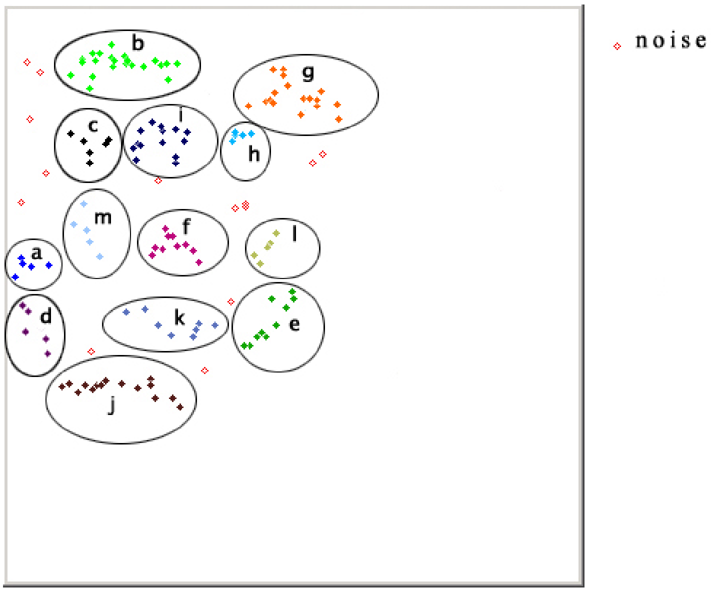
1. Execution time
2. Number of clusters identified
3. Validity of the clusters, in terms of significance, as determined by an expert of the domain
4. Number of elements contained in every cluster

This test was carried out on a sample of 150 e-mail messages. So that these would be representative of the population, the messages of the domain were divided into separate categories and the way they distributed themselves among these was observed in order to replicate the same proportions in the sample.

The categories, identified with the help of an expert of the domain, were as follows:

1. *Documentation*, containing messages on document requests and sharing, such as texts, photographs and articles
2. *Events*, containing messages regarding shows, meetings, conferences, exhibitions and fairs, related to the world of restoration
3. *Training*, messages containing information on schools, training courses, Master degrees, etc.
4. *Newsgroup*, containing messages discussing the mode of operation of the forum and associated problems
5. *Legislature*, messages containing information on texts of law, ministerial decrees, and government regulations
6. *Technical problems*, containing messages discussing technical problems concerning restoration and conservation

This first phase of analysis and categorization of the domain was followed by a sampling phase, during which the sample set of messages was randomly collected, as suggested in literature (Devore, 1995). The analysis carried out to make up the sample was followed by the extraction of terms constituting the domain's dictionary, the elimination of stop words and, lastly, the replacement of each word with a corresponding stem. The results of the Porter algorithm were used as a basis for the stemmer; then, with the help of an expert of the domain, they were adapted to the peculiarities of the domain analyzed. The features, at the end of this process, amounted to 438 and were reduced to 142, using a domain frequency technique. In this way, 145 of the 150 initial messages were able to be represented, the remaining 5 being extremely short and irrelevant for analysis purposes. The clusters performed by the DBSCAN algorithm are shown in Figure 2. The first findings show that the partitioning obtained using DBSCAN is very specific and easy to label, useful for developing a database. However, the clusters obtained with the other two techniques, k-means and hierarchic, were just slightly more generic. It is also true that if, on the one hand, they lost in specificity since they united certain clusters, on the other, they identified a cluster that DBSCAN had missed.



DBSCAN

a) MER method for calculating the state of deterioration
b) Events
c) Information on degree theses
d) Planning and design support software (CAD and others)
e) Consolidating lapidary material
f) Microclimate analysis
g) Regulations
h) Entry examinations
i) Training
j) Mode of operation and problems connected to Yahoo! groups and mailers
k) Work on vaulted roof structures
l) Work on supporting structures
m) Dissipaters and dampers

Fig. 2: DBSCAN clustering results

With regard to comparisons with manual classification, made when the samples were selected, it can easily be observed that it remained very generic and was characterised by poor discriminating capacity, even though carried out by an expert of the domain, without taking into account the considerable expenditure of energy, in terms of time and resources, which this involved. These findings show that, analyzing the three algorithms in the specific, DBSCAN was found, not only to be faster but also simpler to use, thanks to the values the system proposed for the two main tuning parameters. Hierarchical clustering is also very simple to use, as it does not require any parameters from the user when logging-on, but the execution time is very long. Lastly, with regard to the noise identification, it should be noted that DBSCAN was the only algorithm capable, not only of identifying, but also of isolating it from the other elements of the domain. The clusters were labelled by the user after examination of the representative features associated to them. For example, the features of the cluster extracted by DBSCAN and labelled, *work on vaulted roof structures,* were as follows: estradoss 100% capp 100% intradoss 98% volt(a) 100% caten 100% pietr 34% lapid 21%.

Finally, the system automatically generates a mailing list of all the users belonging to each cluster. From these initial findings, the utility of the system presented appears clear: users are facilitated in the study of the newsgroup's e-mail domain thanks to the automatic clustering the system offers. A graphic and user friendly interface to the system permits each cluster to be verified and studied, jointly with the associated user group. In the future, we plan to develop a simpler desktop architecture and carry out a wider-ranging test using other clustering methods and machine learning methods in general.

References

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press Addison Wesley Publishing Company.

Berry, Michael J. A. & Linoff, Gordon S. (1997). Data Mining Techniques: For Marketing, Sales and Customer Support. Wiley Publishing, Inc

Berry, Michael J. A. & Linoff, Gordon S. (2002). Mining the Web: Transforming Customer Data into Customer Value. Wiley Publishing, Inc.

Borg, I. & Groenen, P. (1997). Modern Multidimensional Scaling: Theory and Applications. Springer Series in Statistics. Springer & Verlag.

Boudaillier, E. & Hebrail, G. (1997). Interactive Interpretation of Hierarchical Clustering. In Principles of Data Mining and Knowledge Discovery: Proceedings of the First European Symposium, Pkdd '97, Trondheim, Norway, June 24-27, 288 – 298.

Dash, M., Liu, H., Scheuermann, P. & Lee Tan, K. (2003). Fast Hierarchical Clustering and its Validation. Data and Knowledge Engineering, 44(1).

Devore, J. L. (1995). Probability and Statistics for Engineering and the Sciences. Brooks/Cole Publishing Company, Monterey, California, Fourth Edition.

Ester, M., Kriegel, H. P., Sander J. & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, KDD-96.

Likas, A., Vlassis, N. & Verbeek, J. J. (2003). The Global k-means Clustering Algorithm. Pattern Recognition, 36(2), 451 – 461, February 2003.

Michaud, P. (1997). Clustering Techniques. Future Generation Computer System, 13, 135 - 147.

Porter, M.F. (1997). An Algorithm for Suffix Stripping. In K. Sparck Jones and P.Willet, editors, Readings in Information Retrieval, 313-316, Morgan Kaufmann Publishers, Inc.

Serna, A. (1996). Implementation of Hierarchical Clustering Methods. Journal of Computational Physics, 129, 30-40.

Yang, Y. & Liu, X. (1999). A Re-examination of Text Categorization Methods. In Proceedings of the 22nd Annual International SIGIR, SIGIR-99.

Yang Y.& Pedersen J. P. (1997). A comparative Study on Features Selection in Text Categorization. In Jr. D. H. Fisher, editor, The Fourteenth International Conference on Machine Learning, 412 – 420. Morgan Kaufmann.