Automatic vs. Manual Categorisation of Documents in Spanish

Carlos G. Figuerola, Angel Francisco Zazo Rodríguez, José Luis Alonso Berrocal

Universidad de Salamanca, Facultad de Documentación

e-mail: [figue|afzazo|berrocal]@gugu.usal.es

Abstract: Automatic categorisation can be understood as a learning process during which a programme recognises the characteristics that distinguish each category or class from others, i.e. those characteristics which the documents should have in order to belong to that category. As yet few experiments have been carried out with documents in Spanish. Here we show the possibilities of elaborating pattern vectors that include the characteristics of different classes or categories of documents, using techniques based on those applied to the expansion of queries by relevance; likewise, the results of applying these techniques to a collection of documents in Spanish are given. The same collection of documents was classified manually and the results of both procedures were compared.

1. Introduction

The automatic classification of documents has been widely studied by diverse researchers. Its usefulness is based on the subsequent possibility of adequate retrieval, assuming that those texts dealing with the same subject are classified together or in nearby sections. Various techniques have been proposed for some time now. Thus, Fairthorne [1] and Hayes [2] separately suggested the possibility of using classification systems as a way to increase efficacy in information retrieval. Although Salton himself [3] believed that grouping documents was of interest, he felt that it made their retrieval less effective.

A good part of these techniques are based on the use of measurements of similarity (or

disparity, depending on the point of view) between two documents. A description of the most important ones of both types can be found in [4]. By using a system that associates documents, complex classifying schemes can be made automatically.

Occasionally, however, they are based on a set of classes or categories designed *a priori*, such that any new documents going into the system must fit into one of these classes or categories. Several mechanisms have been proposed for achieving this kind of categorisation automatically and many of them have been tested experimentally; a recent review of these aspects can be found in [5]. Manual categorisation carried out by trained specialists does not give perfect results, but can be considered as a valid point of reference when estimating the viability of automatic procedures. It thus seemed of interest to compare the effectiveness of some of these systems with the effectiveness of manual categorisation.

2. Automatic Categorisation

To carry out automatic categorisation one must have some kind of mechanism that will permit the construction or development of patterns or models representative of the different classes in question and some type of mechanism that will measure or estimate the similarity or disparity between the document to be classified and each of the patterns of each of the categories. For the first case a vector model was used, and for the second the so-called Rocchio algorithm.

2.1. The Vector Model

The vector model, defined by Salton quite a few years ago, [3] is widely used in IR operations and can also be used to explain the automatic categorisation process. Accordingly, a document can be considered as a vector

$$D=(c_1, c_2, c_3 \ldots c_j)$$

that is, as a set of characteristics, up to a total of j, and in which c_1 is a numerical value that expresses to what degree document D possesses characteristic *1*, c_2 the same for characteristic *2*, and so on.

The concept "characteristic" usually refers to the occurrence of certain words in the document, although other factors may be taken into consideration. In the simplest case, binary values can be applied exclusively, so that if word *1* appears in document D the value of c_1 would be *1* and in the opposite case, *0*. Since a word may appear naturally more than once in the same document, and since, furthermore, some words can be considered as more significant than others, the numerical value of each one of the components of the vector is the result of rather more sophisticated calculations which take more factors into account than the simple occurrence or not of a term.

Diverse systems have been proposed for calculating this numerical value, i.e. the weight of each term considered for each document. In general, the inverse frequency (IDF) is taken into account for this and combined in some way with the frequency of the term in the document [6]. Salton and Buckley [7] experimented with more than 200 allocation or weight calculation systems, hence there are plenty to choose from.

In classic document retrieval operations, the query made can also be represented by a vector with an equal number of elements. The value of each of these elements would express the degree to which each of the terms represents the information needs of the person making the query.

Thus, the solving of the query consists of a process of establishing the degree of similarity between the query vector and each of the vectors of each of the documents. For a particular query, then, each document gives a particular degree of similarity; those with the highest degree of similarity will better fit the needs expressed in the query, i.e. they will be more relevant with respect to this query.

The simplest way to calculate this similarity is to find the product of both vectors, query and document. Usually a normalisation of the results is desirable in order to avoid distortions caused by different sized documents. Various methods have also been proposed for calculating similarity; a chart

with the most important ones can be found in [8].

2.2. The Construction of Class Patterns

Based on these ideas, and returning to the field of categorisation, we can now attempt to establish a vector for each of the possible classes or categories which reflects the characteristics of each of these classes. For classification or categorisation operations the basic mechanism consists of measuring the similarity of the vector of each document with each of the pattern vectors that contain the characteristics of the classes or categories. Obviously the class pattern vector that offers greatest similarity to the vector of the document will be the one that most reliably indicates the class or category to which the document belongs or should be allocated.

However, the question is how to build pattern vectors representative of each category or class. Once again we can borrow our ideas from document retrieval. Many systems apply a feedback mechanism, through which after a first query and its corresponding results, the documents indicated by the user as most relevant are used to automatically reformulate the query by extracting the most relevant terms from these documents, adding them to the original query and recalculating the weights of the terms.

Thus, if we have a collection of documents classified manually, and allocated to a particular class, it is possible to apply these feedback mechanisms in order to build a pattern vector representative of that class. The new documents to be classified can be contrasted with this pattern vector, and their similarity calculated. Based on this degree of similarity, it will or will not be allocated to that class.

Diverse systems are used in feedback processes to build a new query vector [9]. These systems can be applied to categorisation in order to construct pattern vectors for each class or category.

One of the most used is Rocchio's algorithm [10], which, in its standard form, has the following formula

$$Q_1 = Q_0 + \beta \sum_{i=1}^{n_r} \frac{R_i}{n_r} - \gamma \sum_{i=1}^{n_{nr}} \frac{NR_i}{n_{nr}}$$

where

Q_0 is the vector of the original query

R_i is the vector of relevant document *i*

NR_i is the vector of non-relevant document *i*

n_r is the number of relevant documents

n_{nr} is the number of non-relevant documents

β and γ are constants which make it possible to adjust the impact of relevant and non-relevant documents.

There are other algorithms that can be used, some of which can be seen in Harman's study [9]. A review of several algorithms directly applied to categorisation can be found in [11].

2.3. The Automatic Categorisation Experiment

We carried out an experiment of automatic categorisation of texts using two collections of press news items taken from the Spanish newspaper 'EL MUNDO' both for the training and for the actual categorisation. These collections were chosen because this newspaper has issued a complete CD ROM version every six months of everything published daily in the newspaper since1994; this allowed us easy access to a large number of texts or documents.

Furthermore, each news item has already been classified, since the CD indicates in which section of the newspaper it was originally published. This simplifies the training operations, as it

eliminates the need for manual categorisation, and also makes verification of the results of the experiment easier.

The news items are obviously in Spanish, giving an added interest to the experiment. Indeed, practical research in IR on documents in Spanish is scarce [12], although in recent years studies on texts in this language have begun to appear. Particularly notable is the inclusion of documents in Spanish among the collections that served as a basis for some of the TREC experiments [13, 14, 15], as well as in the CLEF conferences [16].

2.3.1. The Training Collection

As a training collection we used 2.741 news items published in January 1994. Each item had an average of 3.603 characters, and they were notably uniform in size. It should be pointed out that we worked exclusively with news items and rejected materials such as opinion columns, editorials, etc.

The news items, moreover, correspond to different newspaper sections. The number of items or documents used in each section was fairly similar, although not exactly the same. It should be taken into account that we were seeking to encompass a compact time range (as regards the dates of the news items), since the nature of each section can vary considerably over time when dealing with a daily newspaper. Table 1 shows the number of news items and the sections from which they were taken, both for the training phase and for the system tests.

We started from the basis that each of the sections forms a class or category. And, although the thematic areas are differentiated, it should be noted that some of them could overlap: for example, Stock Exchange and Economy or Campus (Education) and Culture.

The only pre-processing operation carried out was the conversion of all the letters into capitals, and the elimination of accents. Although accents are an important element in Spanish, to the extent that they can in themselves define completely different words, in fact there is an increasing tendency to

avoid them, to use them incorrectly or, at least, to use them carelessly. This means that, from the point of view of processing strings of characters, they are an element of distortion.

Furthermore, we did not use any stemming system, which would have allowed us to work with normalized terms. Indeed, stemming depends greatly on the particular language [17] and Spanish is a particularly rich and complex language from a morphological point of view. Experimental studies have demonstrated the failure of systems used with English when they are applied to Spanish [18]. Notwithstanding, this does not seem to have been a major difficulty, probably owing to the syntactic simplicity and the limited morphological variety which seem to characterise newspaper texts [19, 20].

2.3.2. Training

In the training process, pattern vectors were built for each of the classes, and to do this, a system of weights was used, calculated based on Salton's proposals [8], such that the weight of term *t* in document *d* is obtained using the formula:

$$\frac{(ftd \cdot \log(N/n))}{\left(\sum_{i=1}^{nt} ftd_i \cdot \log(N^2/nd_i)\right)^{1/2}}$$

where

ftd is the frequency of term *t* in document *d*

nt is the number of terms in document *d*

N is the number of documents in the whole collection

n is the number of documents in which the term *t* appears

nd_i is the number of documents in which the term *i* appears

Pattern vectors were subsequently constructed for each of the nine classes in question based on Rocchio's algorithm, and the weight of each term was obtained according to the formula described above, taking into account that, in each case, there was no initial query, and thus Q_0 was set at *0*.

Furthermore, and taking into account the problems derived from working in negative values, the entry vectors were modified so that those weights with a negative value were also set at *0*, while the constants *?* and *?*, following the recommendations of Buckley et al. [21], were set at 16 and 4, respectively.

2.3.3. Categorisation Test

To try out the system, a collection of 4.250 news items from the same newspaper and approximately the same dates was used. The general characteristics of these documents were similar to those used in the training. All the news items, moreover, pertained to one of the sections or classes used in the training phase.

To estimate the degree of similarity between the documents to be classified and the patterns of each of the classes, the cosine coefficient was used, as it is widely applied to Information Retrieval operations [6, 8]:

$$SIM(P_x, D_y) ? \frac{\sum_{i?1}^{n} p_{xi} d_{yi}}{\sqrt{\sum_{i?1}^{n} p_{xi}^2 ? \sum_{i?1}^{n} d_{yi}^2}}$$

where

P_x is the pattern vector of class *x*

D_y is the vector of document y

p_{xi} is element i of P_x

d_{yi} is element i of D_y

n is the number of elements or terms in the vectors

By contrasting the documents to be classified and the patterns of the classes considered, nine coefficients of similarity were obtained in each document, one for each class contemplated. In a manual work situation, these coefficients could be presented to the user in decreasing order, and then they could manually determine the most suitable class or classes.

When working totally automatically, however, it is necessary to define a threshold in these coefficients so that those of the classes situated above the threshold would indicate in which categories the document to be classified could be located. This threshold should be established experimentally, with a view to optimising the results[11, 22].

However, the use of thresholds presupposes that a document can be allocated to more than one class. In some real situations, and owing to external conditioners or restrictions, it may be necessary to choose a single class instead of several. In fact, the documents used in our experiment actually belong to only one section of the newspaper. With this restriction, the class chosen should be the one where the similarity coefficient is highest.

2.4. Evaluation

Traditionally, the effectiveness of IR operations is calculated by the classic Precision and Recall measurements [8]. This practice has also been followed in works of categorisation, although in some cases the results were presented in terms of percentages of successes and failures For our part, and with a view to possible comparisons, we preferred to use Precision and Recall, calculating them according to

the following formulae [11]:

$$R = \frac{a}{a + c} \qquad P = \frac{a}{a + b}$$

where

R is Recall

P is Precision

a is the number of documents belonging to a class and allocated to that class

b is the number of documents not belonging to a class but assigned to that class

c is the number of documents belonging to a class but not assigned to that class

Naturally, one must evaluate the results for each of the classes. Some measurements have also been proposed to unify Precision and Recall in one result. One of these is the measurement F_β [4]:

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

β is a parameter which makes it possible to adjust the relative influence of both components, Precision and Recall. $\beta = 1$ gives equal weight to both components of the measurement. F_β has often been used in categorisation studies [23].

Table 2 gives the results obtained for each class, taking into account that each document to be classified can only be allocated to a single class, obviously the one with the greatest similarity.

For some classes the results are frankly good (for example, SPORT, with an F_1 of 0.91), but

even the average results can be considered interesting.

3. Manual Categorisation

Manual categorisation is not an infallible operation and the results are far from perfect. The problems of inconsistency in operations of manual indexing of documents are well known [24, 25], since different indexers consider different descriptors applicable to the same documents. Presumably this type of problem also occurs in the manual allocation of documents to different classes or categories. Consequently, a certain level of error or disparity is to be expected between the classes assigned manually and those that the documents or news items to be classified really have.

For the manual categorisation, 34 students from the last year of the Licentiate Degree Course in Information Science were asked to collaborate. They were all familiar with the newspaper 'El Mundo'. The same collection of 4 240 news items or documents used in the automatic categorisation was used, and each of the participants was given 125 documents, each of which had to be allocated to a single category.

Table 3 gives the results measured as described above. These results are good, as was to be expected, but there were major disparities with the original allocation of the news items to the newspaper sections. Only five of the categories were above 0.8 of F_1, and two of them were around 0.6.

Comparison with the results obtained automatically confirmed the expected superiority of manual categorisation. However, the difference (except in one of the categories) is small, and in another of the categories automatic categorisation is even slightly better than manual categorisation. Broadly speaking, the results of the automatic study come quite close to, although they do not quite reach, those of manual categorisation.

4. Conclusions

We have shown the possibilities of elaborating pattern vectors that include the characteristics of

different classes or documents using techniques based on those applied in the expanding of queries by relevance. At the same time, a description is given of an experiment consisting of the application of these techniques to classify a collection of news items from the Spanish press. The results obtained were on the whole promising, and frankly good for some of the categories.

The same collection of documents was classified manually, with better results, but which did not differ greatly from those obtained automatically. This seems to indicate the feasibility of using automatic mechanisms in categorisation.

CLASSES	Training	Test
STOCK MARKET	287	467
CAMPUS	292	472
CULTURE	301	464
SPORT	315	484
ECONOMY	301	473
INTERNATIONAL	293	451
MOTOR	324	481
NATIONAL	338	482
SOCIETY	290	476
TOTAL docs.	2741	4250

Table 1. Number of documents for training and test, by sections

Class	Prec.	Recall	F_1
STOCK MARKET	0.29	1	0.44
CAMPUS	0.28	1	0.43
CULTURE	0.89	0.83	0.86
SPORT	0.93	0.88	0.91
ECONOMY	0.73	0.57	0.64
INTERNATIONAL	0.88	0.73	0.8
MOTOR	0.8	1	0.89
NATIONAL	0.83	0.75	0.8
SOCIETY	0.86	0.48	0.62

Table 2. Results of automatic categorisation with single class allocation

Class	Prec.	Recall	F_1
STOCK MARKET	0.69	0.5	0.58
CAMPUS	1	0.73	0.84
CULTURE	0.88	0.9	0.89
SPORT	0.99	0.97	0.98
ECONOMY	0.63	0.7	0.67
INTERNATIONAL	0.73	0.99	0.84
MOTOR	0.96	0.93	0.94
NATIONAL	0.69	0.75	0.72
SOCIETY	0.91	0.75	0.82

Table 3. Results of manual categorisation

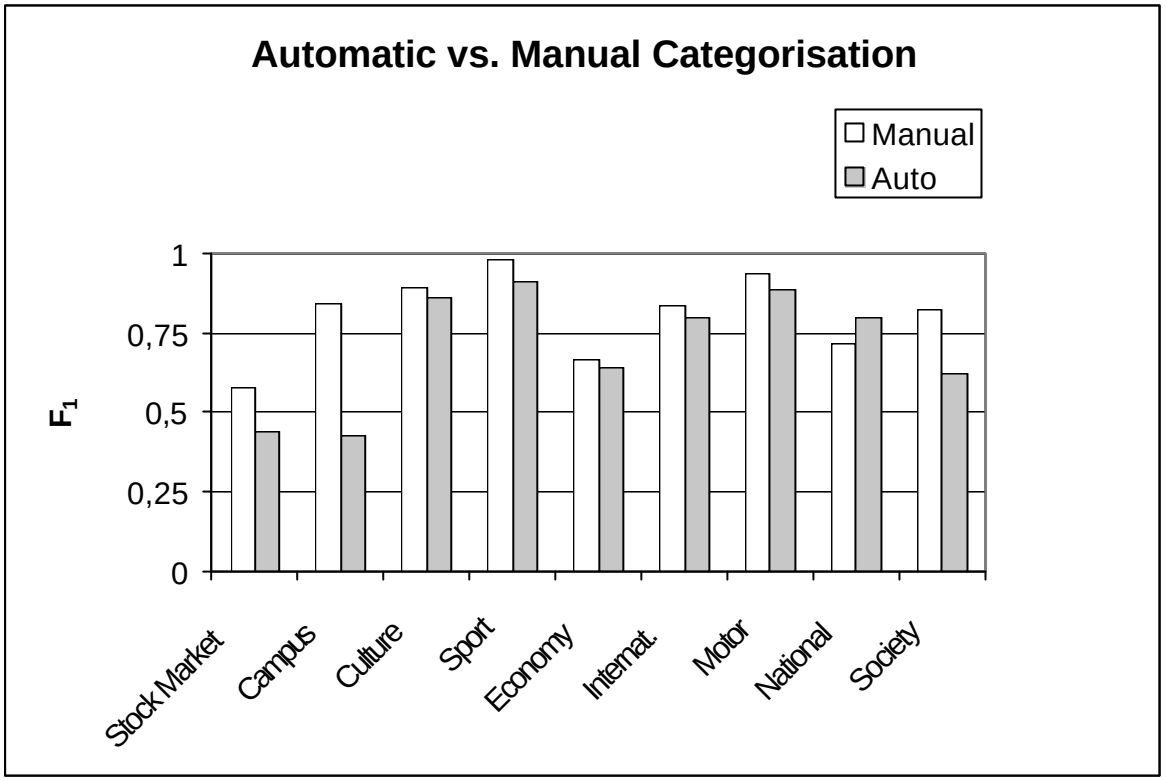

Figure 1. Results (F_1) with single class allocation

References

1. Fairthorne, R. A. *The mathematics of the classification: Towards Information Retrieval.* London: Butterwoths, 1961 .
2. Hayes, R. M. Mathematical models in information retrieval. In: Garvin, P. L., ed. *Natural Language and the Computers.* New York: McGraw-Hill, 1963, 268-309.
3. Salton, G. *Automatic Information Organization and Retrieval.* New York: McGraw-Hill, 1968.
4. Rijsbergen, K. Van *Information Retrieval.* London: Butterworths, 1979.
5. Yang, Y. and Liu, X. A re-examination of text categorization models, *ACM SIGIR 99*, 1999, 42-49.
6. Harman, D. Ranking Algorithms. In: Frakes, W.B. and Baeza-Yates, R.eds. *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs (NJ): Prentice-Hall, 1992, 363-392.
7. Salton, G. and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24(5), 1988, 513-523.
8. Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval.* New York: McGraw-Hill, 1983.
9. Harman, D. Relevance feedback and other query modification techniques. In: Frakes, W.B. and Baeza-Yates, R.eds. *Information Retrieval: Data Structures and Algorithms.* Englewood Cliffs (NJ): Prentice-Hall, 1992, 241-263.
10. Rocchio, J. J. Relevance Feedback in Information Retrieval. In Salton, G. ed. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Engelwood Cliffs (NJ):

Prentice-Hall, 1971, 313-323.

11. Lewis, D.D., Shapire, R.E., Callan, J.P. and Papka, R. Training Algorithms for Linear Text Classifiers, *ACM SIGIR 96*, 1996, 298-306.

12. Figuerola, C. G. La investigación sobre Recuperación de la Información en español. In Gonzalo García, E. and García Yebra, V. eds. *Documentación, Terminología y Traducción*, Madrid: Síntesis, 2000, 73-82.

13. Harman, D. ed. *NIST Special Publication 500-225: Overview of The Third Text Retrieval Conference (TREC-3)*, Gaithersburg, 1995

14. Harman, D. K. ed. *NIST Special Publication 500-236: The Fourth Text Retrieval Conference (TREC-4)*, Gaithersburg, 1996

15. Voohees, E. and Harman, D. eds. *NIST Special Publication, 500-238: The Fifth Text Retrieval Conference (TREC-5)*, Gaithersburg, 1997

16. Peters, C. ed. *First Results of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign*, Lisbon, 2000

17. Paice, C.D. Method for Evaluation of Stemming Algorithms Based on Error Counting, *JASIS*, 47(8), 1996, 632-649

18. Gómez Díaz, R. *La Recuperación de Información en español: evaluación del efecto de sus peculiaridades lingüísticas*. Master's Thesis, Salamanca: Universidad de Salamanca, 1998.

19. EFE, Agencia *Manual de español urgente*, Madrid: Cátedra, 1991, 18-22 and 36-60.

20. Elena García, P. La traducción de textos informativos (noticias). In Elena Garcia, P. ed. *Curso práctico de traducción general alemán – español*. Salamanca: Ediciones Universidad de Salamanca, 1994, 11-67

21. Buckley, C., Salton, G. and Allan, J. The effect of adding relevance information in a relevance feedback environment, *ACM SIGIR 94*, 1994, 292-300.

22. Cohen, W.W. and Singer, Y. Context-sensitive learning methods for text categorization, *ACM SIGIR 96*, 1996, 307-315.

23. Lewis, D.D. and Gale, W. A sequential algorithm for training texts classifiers, *ACM SIGIR 94*, 1994, 3-12.

24. Hooper, R. S. *Indexer consistency tests-origin, measurements, results and utilization.* Bethesda: IBM Corp., 1965

25. Stubbs, E. A., Mangiaterra, N.E and Martínez, A. M. Internal quality audit of indexing: a new application of interindexer consistency, *Cataloguing & Classification Quaterly*, 28(4), 2000, 53-70.