Categorizing photographs for user-adapted searching in a news agency e-commerce application[1]

José María Gómez Hidalgo, Raúl Murciano Quejido, Alberto Díaz Esteban, Manuel de Buenaga Rodríguez, and Enrique Puertas Sanz

Departamento de Inteligencia Artificial
Universidad Europea – CEES
28670 Villaviciosa de Odón, Madrid, Spain
{jmgomez, murciano, alberto, buenaga, epuertas}
@dinar.esi.uem.es

Abstract. In this work, we present a system for categorizing photographs based on the text of their captions. The system has been developed as a part of the system CODI, an e-commerce application for an Spanish news agency. The categorization system makes able to the user the personalization of their information interests, improving search possibilities in the CODI application. Our approach for photograph categorization is based on linear text classifiers and Web mining programs, specially selected due to their suitability for industrial applications. The evaluation of our categorization system has shown that it meets the efficiency and effectiveness requirements of the e-commerce application.

1 Introduction

In recent years, innumerable providers are making their presence in Internet a main strategic goal. Their success depends on the ability of users to find the desired items among thousands or documents, images or commercial products. The providers must help customers to find these products, which are frequently supplemented with text descriptions that help customers to find interesting items in web sites with search functions.

A common feature in many of the information access providers that supply these services is that they are very generic, making it difficult for users to specify their information needs. As a consequence, users spend lots of time seeking for information relevant to their particular interests, because these services do not take into account their goals, experience or knowledge. Moreover, an additional problem is that user's interests change over time as a direct result of interaction with information [4]. These factors have resulted in the appearance of various personalized information services, although most of them are implemented using very simple methods. Most of

[1] The research described in this paper has been partially supported by the Spanish Ministry of Industry as part of the Iniciativa ATYCA program.

them provide personalization based on content, but the levels of satisfaction obtained with the methods employed are low, and the field is still on need of further exploration.

We have taken part in the development of an e-commerce application named CODI ("Sistema de Comercialización de Objetos Digitales en Internet", Digital Objects Commercialization System in the Internet) in which personalization for products search plays a key role. The application, developed for the leading news and photograph Spanish agency Agencia EFE[2], is designed for the commercialization of the historic image archive of the agency. The consulting and development company Grupo GESFOR[3] and our university, have built the e-commerce system. The project includes conventional search functions that allow the customers to find the pictures they are interested in purchasing, and a set of functionalities that allow a degree of personalization in the search operations. As photographs come with text captions, we make use of textual content analysis techniques [19] to achieve a elaborate user model that enrich search possibilities.

Text categorization, the assignment of subject labels to text items, is one of the most prominent text analysis and access tasks nowadays [21]. We have implemented a text categorization module that automatically assigns Yahoo! Spain subject labels to pictures based on their captions. Among other information, the users of the system can specify some Yahoo! Spain categories to define their interests, which are later taken into account in the advanced picture search functions. The text categorization module is based on linear classifiers [8, 7, 14] and a program that mines Yahoo! Spain web pages.

This work is organized as follows. In the next section, we describe the main functionalities of the system, specially those related to user-adapted searching. In the Section 3, we show the system architecture and how we have implemented its user-adapted search capabilities. In the following section, the fundamentals of the text categorization model are described, including the details of the Yahoo! Spain mining module and the linear classifiers. Afterwards, we describe the approach we have followed to evaluate the performance of the categorization module, and we discuss the results of our experiments. Finally, our conclusions are presented, along with future research directions.

2 The CODI Project

The first participant of the project is a software development company, which developed a photograph commercialization system through Internet, CODI, for the Spanish news agency Agencia EFE. The role of our team in the project was the design and implementation of advanced user-adapted search funcionalities, improving the abilities of personalization of the system. The CODI project was funded by the Spanish

[2] http://www.efe.es
[3] http://www.gesfor.es

Ministry of Industry as part of the Iniciativa ATYCA program to improve the transfer of technology between universities and private companies.

Fig. 1. Conventional search fields and results of a search in the CODI system.

The system implements all functionalities for the commercialization of the historic archive of the news agency. An important set of functions are those related to the search of pictures in the database. When a user has registered in the system, he or she can locate images according to keywords or standard domain-dependent information, like author, date, and other domain specific categories. The search fields and the results of a search are shown in Figure 1. We have enhanced these abilities letting the user define his or her interests, and adapting the search results to them.

When the user logs in the system, he or she can define or modify a user profile in terms of two kinds of information (as shown in Figure 2):

- A set of preferred subject categories obtained from the two first levels of Yahoo! Spain. Internet users are familiar to web directories like Yahoo!, and they understand the meaning and utilization of subject categories.
- A set of saved searches. A saved search has a label and a set of keywords that the user has find useful other times he or she has used the system.

When a user performs a search in the system, he or she can choose the option for advanced search, shown in Figure 3. This option lets the user specify a set of search criteria that include:

- A category and/or subcategory from Yahoo! Spain.

– The option of using personal categories in the search.
– The option of using a saved search.

Fig. 2. Interface for profile edition.

When the user selects all options, a logical “and” is computed across the search results.

The utilization of categories extracted from Yahoo! Spain leads to the application of a text categorization function when the news agency employees update the database with new photographs. These images came with a caption that permits the automatic classification of them according to the categories in Yahoo! Spain. So, this task is performed by the system when the employees have entered some pictures in the system and they select the option.

3 System Implementation and Architecture

The advanced search capabilities developed by our team have been structured a four modules represented in UML in the Figure 4.

The Yahoo! Spain Mining module is a set of Java and SQL programs that extract information about the categories from the Yahoo! Spain web site and introduce it in the database, making it available for the categorization module. Basically, a Java program identifies the information about indexed web pages in the two first levels of Yahoo! Spain directory, extracting a set of words specially relevant for each category. These words are entered in the database by an SQL program.

The Text Categorization module is a set of Java and SQL programs that assign Yahoo! Spain category labels to the pictures according to their captions and the information extracted by the mining module. Text categorization is usually performed by a classifier, a program that is built by Machine Learning algorithms [21].

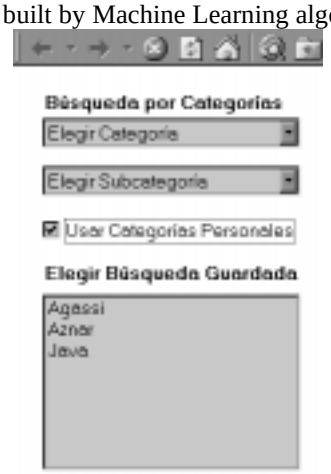

Fig. 3. Advanced search interface.

We have selected the linear algorithm Rocchio because it fits the efficiency and effectiveness requirements of the system. The description and evaluation of our approach to text categorization of image captions is the main focus of this work.

We have finally built the User Model and the Search modules as a set of Java, JavaScript and SQL programs that allow the user de definition and modification of his or her interests, and to search in the images database according to his or her profile.

4 Caption Text Processing and Categorization

Text categorization – the assignment of subject labels to text items [21] – of photograph captions plays a key role in the CODI project. The user profile is defined in terms of a set of categories extracted from Yahoo! Spain, and images must be classified according to them through an automatic process. An automatic text categorization system has been built to perform this task. The system learns a linear classifier using the information extracted from Yahoo! Spain and then it automatically classifies photographs according to their captions.

4.1 Categorization of Photographs Using Captions

Search and categorization of photographs is the center of interest of an important research in the latest years (see, e. g., [3, 15, 16, 18, 22]). Using the information in the

pictures (colors, shapes, etc.) themselves leads to systems that perform much worse than using the text captions of the images or the text surrounding them [18]. For instance, the image categorization system AdEater, designed to detect banners, makes use only of few picture features (height, width) but takes advantage of the text surrounding the image (hyperlink text, etc) to perform categorization [10].

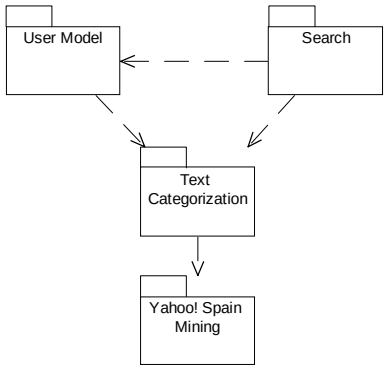

Fig. 4. Advanced search subsystem module design.

Pictures from the news agency come with a medium length caption that make the categorization process possible. The captions are processed to extract separate words, converted to lower case, which are considered the terms. Each caption is represented as a term weight vector, according to the Vector Space Model [19]. For a caption, we set the weight of a term as 1 if the term occurs in the caption, and 0 if not.

We have taken the categories in the first level of Yahoo! Spain, that we also represent as a term weight vector. These vectors are built using the information extracted from Yahoo! Spain, and the Rocchio linear classifier [17]. For assigning a category to a picture, the vectors for the category c_j and for the caption of the image p_k are compared using the following similarity formula:

$$sim\left(c_j, p_k\right) = \sum_{i=1}^{N} c_{ij} \cdot p_{ik}$$

Being c_{ij} the weight of the ith term for the vector of category c_j, p_{ik} the weight of the ith term for the vector of image caption p_k, and N the number of different terms. This formula does not take into account the length of picture captions and categories [19], because all captions have got the same number of words (approximately), and also all the categories.

For the practical prototype, a image is assigned to a category if the similarity among them is over a threshold determined empirically. In the prototype, the threshold has been set to five, in order to obtain around two categories per picture. However, for the experiments described below, image p_k is assigned to the most similar category to it, making the evaluation easier.

4.2 Data acquisition by Mining Yahoo!

For constructing the vectors that represent the categories, we have extracted information from Yahoo! Spain. One of the requirements of the categorization system was that it had to work autonomously, and so, no prelabeled pictures were available for learning. Then, information from Yahoo! Spain was taken as training data for the learning process.

We have designed a Java program that mines the Yahoo! Spain web site to get the information necessary for learning the classifier or classification program. The mining program downloads the web pages reachable from the two first levels of the Yahoo! Spain web site, restricted to the Yahoo! domain. For each of the categories, the pages downloaded are taken as training data.

It is important to note that a common assumption made when applying a learning approach is that items to learn from and items to be classified are of the same type. The approach we describe here violates this assumption, but it is forced by the requirement stated above. Also, this approach has been followed by Attardi *et al.* [2] when classifying web pages, where the documents to be classified are artificially constructed using the text surrounding the hyperlinks to the pages to be categorized.

4.3 Machine Learning Linear Classifiers

For training a text categorization system, a number of Machine Learning approaches have been tested, including decision tree and rule-based learners [1, 5, 9, 13], probabilistic classifiers like Naive Bayes [9, 11, 12], neural networks [6, 20], instance-based classifiers like kNN [9, 25], etc. See [21] for other approaches.

An important subclass of learning approaches are those which learn linear classifiers, like Rocchio, Widrow-Hoff, or Winnow algorithms [6, 7, 8, 14]. These approaches examine training instances a finite number of times, and construct a prototype instance (a term weight vector) for each category, which is latter compared to the instances to be classified. Linear classifiers show interesting properties that make them ideal for industrial applications [21]:

- Linear classifiers are very efficient. Both the learning and the classification steps are linear on the number of terms, documents and categories, being far more efficient than most of the other learning approaches.
- Linear classifiers are simple to interpret. When the prototype vector for a category shows a high weight for a term, this term can be considered a good predictor for the category.
- Linear classifiers show good performance. Some of the algorithms are nearly top performing on standard test collections.
- Linear classifiers can take advantage of an standard IR system. The representation of categories is similar to documents in a text collection. The categories vectors can be indexed by an IR system, and documents to classifiy can be sent as queries to the system. The classification system assign he most relevant category(es) to the document.

For our prototype, we have selected the linear classifier Rocchio [17], as we discuss in the next subsection.

4.4 Implementation of the Rocchio Algorithm

The Rocchio algorithm is one of the most popular learning approaches tested for text categorization [7, 8, 11, 20]. This algorithm was originally designed for the relevance feedback process in Information Retrieval systems [17]. It was first adapted to text categorization by Hull in [8]. We have applied a restricted version of this algorithm in the prototype of the system CODI. Let be the c_{ij} the weight of the ith term in the vector for category c_j, and d_{im} the weight of the ith term in the vector for the Yahoo! web page d_m. The Rocchio formula is:

$$c_{ij} = \beta \frac{\sum_{k \in R_j} d_{im}}{\left|R_j\right|} - \gamma \frac{\sum_{k \in \overline{R_j}} d_{im}}{\left|\overline{R_j}\right|}$$

Where R_j is the set indexes of prelabeled documents in the jth category. The parameters β and γ control the impact of the positive and negative training. We have implemented a simplified version of this algorithm. Because of the characteristics of our training data, we have set $\gamma = 0$, stripped out the denominator of the first term, and normalized the resulting weights to 1.

4.5 Efficiency Issues

A main requirement of the advanced search subsystem that we have implemented was efficiency. On a Sun Ultra Enterprise 450, with 1 Gb. of RAM and two 250 Hz. processors, the training process spends less than an hour, and the categorization of a picture spends around a second. The news agency usually enters around two hundred pictures each day, and so, the categorization process meets the efficiency requirements. The training process must be run once and offline, and so, also meets the requirements.

5 Experiments and Results

We have developed a set of experiments to validate the feasibility of our approach. We have constructed a test collection from real data in the application, and we have tested our approach according to standard performance metrics for text categorization. The results support the election of the algorithm, which performs efficiently and with enough efficacy for the prototype requirements.

5.1 Experiments Setup

Two important issues when performing experiments on text categorization systems are the definition of the test collection and the evaluation metrics. For the construction of our test collection, we have collected a set of 121 photograph captions in Spanish. Each caption has approximately 39 words.

The image captions have been manually classified according to the categories in the first level of Yahoo! Spain. One category has been assigned to each caption. Among the fourteen categories in the first level of Yahoo! Spain, only nine have got at least one image assigned. The set of categories and their distribution in our test collection are shown in Table 1.

Table 1. Test collection categories and statistics.

Category number and name	Equivalent Yahoo! category	Frequency (%)
1 Arte y cultura	Arts & Humanities	12 (9,9%)
2 Ciencia y tecnología	Science	6 (4,9%)
4 Deportes y ocio	Recreation & Sports	38 (31,4%)
5 Economía y negocios	Business & Economy	2 (1,6%)
7 Espectáculos y diversión	Entertainment	7 (5,7%)
11 Política y gobierno	Government	35 (28,9%)
12 Salud	Health	1 (0,8%)
13 Sociedad	Society & Culture	17 (14,0%)
14 Zonas geográficas	Regional	2 (1,6%)

As shown in the table, the categories are asymmetrically distributed. This fact motivates the election of the evaluation metrics. The most common evaluation metrics for text categorization are recall, precision and F_1 [24].

Table 2. Contingency table for a category c.

	Docs. assigned to c	Docs. not assigned to c
Docs. in c	A	C
Docs. not in c	B	D

Given a category *c* and the contingency table shown in Table 2, the recall is defined as A/A+C, and the precision as A/A+B. The recall shows the ability of the system to detect if a document pertains to a category, and the precision represents the hits over the decisions taken by the system. The measure F_1 balances the tradeoff between recall and precision. These measures can be calculated by micro or macroaveraging. For microaveraging, the contingency tables for all categories are summed up and just one value is computed. This way, most populated categories have got more importance. When macroaveraging is performed, the metrics are calculated for each category and then are averaged. This approach gives equal importance to all categories. Both approaches show a complete picture of the system effectiveness [24].

5.2 Results and Interpretation

We show the results of the evaluation of our system in the Table 3.

Table 3. Results of the experiments.

Category	Recall	Precision	F_1
1	1,00	0,86	0,92
2	0,66	0,50	0,57
4	0,36	1,00	0,53
5, 7, 11, 12, 13, 14	0,00	0,00	0,00
Macroaveraging	0,22	0,26	0,22
Microaveraging	0,25	0,72	0,37

The analysis of these data let us conclude:

- The overall performance of the system is acceptable for the image categorization module requirements. Most of the errors made by the categorization module do not affect the performance of the advanced search functions. The microaveraged effectiveness achieved is nice according to the amount and kind of information used for the categorization process (format and style of training documents do not match documents to be classified).
- While not bad, effectiveness should be improved, specially for highly populated categories (4, 11 and 13). Classification errors in these categories strongly influence overall performance.
- any style files, templates, and special fonts you may have used,

The main reason for these results is that our approach, based on the extraction of training information from Yahoo! Spain, violates the assumption that training and test documents are similar in style and format.

6 Conclusions and future work

We have presented an image categorization system developed in the framework of an e-commerce application for a news agency. The system is based on the utilization of the captions of the pictures, linear text classifiers and web mining programs. The utilization of this categorization module allows the improvement of search capabilities of the commercialization system. Specifically, the categorization system makes able to the user the personalization of their information interests. The categorization module meets the efficiency and effectiveness requirements of the search functions. In particular, the categorization process spends around a second per image, and effectiveness metrics applied in the evaluation show acceptable values.

Our future work will address the improvement of the effectiveness of the categorization module. For this task, we plan to develop a better mining module that will extract more valuable information from Yahoo! Spain web site. Also, we plan to

make use of lexical databases like EuroWordNet [23]. Lexical databases are repositories of information about the lexical items of one or several languages. The utilization of EuroWordNet will allow to improve the lexical information extracted from Yahoo! Spain.

References

1. Apté, C., Damerau, F.J. and Weiss, S.M. (1994) Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), pp. 233-251.
2. Attardi, G., and Gullí, A. and Sebastiani, F. (1999) Automatic Web Page Categorization by Link and Context Analysis. In Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence, pp. 105-119.
3. Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horovitz, B., Humphrey, R., Jain, R.C., and Shu, C. (1996) The VIRAGE Image Search Engine: An Open Framework for Image Management. In Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases IV, IS&T/SPIE.
4. Belkin, N. J. (1997) User Modeling in Information Retrieval. In Proceedings of the Sixth International Conference on User Modeling, UM97, Chia Laguna, Sardinia, Italy.
5. Cohen, W.W. and Singer, Y. (1996) Context-sensitive learning methods for text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 307-315, ACM Press, New York, US.
6. Dagan, I., Karov, Y., and Roth, D. (1997) Mistake-driven learning in text categorization. In Proceedings of EMNLP-97, 2nd Conference on Empirical Methods in Natural Language Processing, pp. 55-63, Association for Computational Linguistics, Morristown US.
7. de Buenaga, M., Gómez, J.M. and Díaz B. (1997) Using WordNet to Complement Training Information in Text Categorization. In Proceedings of the Second International Conference on Recent Advances in Natural Language Processing (RANLP).
8. Hull, D.A. (1994) Improving text retrieval for the routing problem using latent semantic indexing. In Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, pp. 282-289, Springer Verlag, Heidelberg, DE.
9. Joachims, T. (1998) Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, Lecture Notes in Computer Science, Number 1398, pp. 137-142, Springer Verlag, Heidelberg, DE.
10. Kushmerick, N. (1999) Learning to remove Internet advertisements. In Proceedings of AGENTS-99.
11. Larkey, L.S. and Croft, W.B. (1996) Combining classifiers in text categorization. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 289-297, ACM Press, New York, US.
12. Lewis, D.D. (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In Proceedings of ECML-98, 10th European Conference on Machine Learning, Lecture Notes in Computer Science, Number 1398, pp. 4-15, Springer Verlag, Heidelberg, DE.
13. Lewis, D.D. and Ringuette, M. (1994) A comparison of two learning algorithms for text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93.

14. Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R. (1996) Training algorithms for linear text classifiers. In Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, pp. 298-306, ACM Press, New York, US.
15. Niblack, W., Barner, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., Yanker, P., Faloutsos, C. and Taubin, G. (1993) The QBIC Project:Quering Images by Content Using Color, Texture, and Shape. In Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases, SPIE.
16. Pentland, A., Picard, R.W. and Sclaroff, S. (1994) Photobook: Tools for Content-Based Manipulation of Image Databases. In Proceedings of the Symposium on Electronic Imagic: Science and Technology-Storage and Retrieval for Image and Video Databases II, SPIE.
17. Rocchio, J.J. Jr. (1971) Relevance feedback in information retrieval. In G. Salton, editor, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall.
18. Sable, C.L. and Hatzivassiloglou, V. (1999) Text-Based Approaches for the Categorization of Images. In Proceedings of the European Conference of Digital Libraries (ECDL), published as Research and Advanced Technology for Digital Libraries, lecture Notes in Computer Science 1696, Springer.
19. Salton, G. (1989) Automatic Text Processing: the transformation, analysis and retrieval of information by computer. Addison Wesley.
20. Schütze, H., Hull, D.A. and Pedersen, J.O. (1995) A comparison of classifiers and document representations for the routing problem. In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, pp. 229-237, ACM Press, New York, US.
21. Sebastiani, F. (1999) A Tutorial on Automated Text Categorisation. In Proceedings of the First Argentinean Symposium on Artificial Intelligence (ASAI-99).
22. Smith, J.R. and Chang S.-F. (1997) Visually Searching the Web for Content. IEEE Multimedia, 4(3):12-20, July-September.
23. Vossen, P. (1997) EuroWordNet: a multilingual database for information retrieval. In Proceedings of the DELOS workshop on Cross-language Information Retrieval, March 5-7, Zurich.
24. Yang, Y. (1999) An evaluation of statistical approaches to text categorization. Information Retrieval, Vol. 1, Number 1-2, pp. 69-90.
25. Yang, Y. and Pedersen, J.O. (1997) A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning, pp. 412-420, Morgan Kaufmann Publishers, San Francisco, US.