Hermes: Intelligent Multilingual News Filtering based on Language Engineering for Advanced User Profiling

Ignacio Giráldez, Enrique Puertas, José María Gómez
Dpto. Inteligencia Artificial, Universidad Europea de Madrid
{giraldez, epuertas, jmgomez}@dinar.esi.uem.es

Raúl Murciano
Dpto. Programación e Ingeniería del Software, Universidad Europea de Madrid
murciano@dpris.esi.uem.es

Inmaculada Chacón
Dpto. Periodismo Especializado, Universidad Europea de Madrid
inmaculada.chacon@fcp.cin.uem.es

Abstract

In this paper, we describe Hermes, a multilingual news filtering service that sends its users a customized e-newspaper by email through the use of several text classification techniques, including categorization, summarization and relevance feedback. Hermes is based on a user model far richer than most current newspapers personalization services. The prototype has been evaluated by final users reporting a high degree of satisfaction.

Keywords: Crosslingual Text Filtering, Crosslingual Information Retrieval, User Modelling, Text Categorization

1. INTRODUCTION

Nowadays many newspapers offer digital access to their contents. Moreover users can subscribe to newspapers' services and receive daily news by e-mail. Unfortunately, most of them are simple transcriptions of their printed version. More advanced systems include user-profiling options which allow users to define what kind of information they want to receive.

There are two main approaches to define user interests about content. First, category-based systems list some categories –usually newspaper sections– and users pick up what categories are considered interesting. After user profile definition, the service sends all news stories contained at selected categories at a daily basis. The second approach uses term-based descriptions to define user interests, so that users give interest-related keywords –and sometimes their interest degree–. Before sending daily messages to each user, the system selects which news items contain stored user-keywords, orders selected items by relevance, and includes the most relevant items in the final message to be sent by e-mail. Both approaches can be integrated, letting users define personalized categories by keywords (that is, each user-defined category is represented by a keyword list), in a way similar to Yahoo!'s stored searches. In this way,

services are required to automatically categorize each news item for each user-defined category to build each user final message.

Through last years our efforts were oriented to offer personalized information access by integrating all previously defined user-profiling systems in a monolingual setting, such as Mercurio [Díaz 00a]. However, there are many circumstances which favour multilingual information systems –specially at the actual European Union context, where information flow involves interactions with documents in several languages. This kind of requirements promote automatic translation and multilingual services –for instance, EU official organisms make use of translation systems like EC Systran, multilingual information access like CELEX, and multilingual systems for analysis content on the Internet as POESIA [Gómez 02]. In this paper we present Hermes[1], a multilingual news filtering system which allows users to receive personalized messages containing most interesting news extracted from digital versions of several European newspapers, using several languages.

We present an overview of Hermes functionalities in Section 2, before getting into more detail in Section 3, which describes: (i) user model characteristics, (ii) machine translation algorithms used for filtering, (iii) user feedback and its influence in user model evolution, and (iv) news summarization process guided by user preferences. Hermes evaluation results are included in Section 4. Section 5 presents related work, leading to our conclusions.

2. SYSTEM DESCRIPTION

Hermes has been designed for providing personalized news according to user interests, sending by e-mail a message with a set of news (title and a summary). For generating each message, both user and news information are retrieved, formally represented and properly processed to obtain the final resulting news representation for each user. As we describe below, news selection is based on the Vector Space Model (VSM), applying a simplification of Rocchio algorithm that was previously applied with satisfactory results [Gervás 99].

[1] *Hermes has been partly funded by the Spanish Ministerio de Ciencia y Tecnología.*

The first thing required for the system to work is the user information. In the sign-in process, each user fills in a form which collects preferences about news, interests and delivering settings (e.g. days to receive messages). Hermes collects each user information and stores it internally twice, a copy in English and another one in Spanish. Of course, users can also access again to this form to modify their preferences.

Then, the system acquires news information. Everyday Hermes connects to e-newspapers, one in Spanish and another one in English, and gets textual content from each news item. These are processed to obtain their summaries and an internal representation according to the VSM: each news item is mapped to a vector which ponderates each term relevance. Each item is also categorized to be sent to users who selected resulting category. That categorization process involves a generally accepted categories system (Yahoo! and Yahoo! Spain first level categories), which was processed to extract a VSM representation for each category.

Once user prefences and news items are equally represented by term weight vectors, Hermes is ready to obtain the relevance between a news story *i* belonging to a newspaper section *k* and a user model *j,* in Spanish or English*,* using the following formula:

$$s_{ij} = \alpha_j S_{kj} + \beta_{kj} \sum_{h=1}^{n} C_{hj} \cdot sim_c(d_i, c_h) + \gamma_j \cdot sim_k(d_i, l_j)$$

Where:

- α_j is the significance of newspaper sections for user *j*,
- S_{kj} is the interest of section *k* for user *j*,
- β_{kj} is the significance of section *k* for user *j*,
- *n* is the number of categories,
- C_{hj} is the interest of category *h* for user *j*,
- d_i is the vector of weights for the news story *i*,
- c_h is the vector of weights for the category *h*,
- sim_c is the dot product,
- γ_j is the interest of keywords for user *j*, being $\alpha_j + \beta_j + \gamma_j = 1$,
- l_j is the vector of keywords for user *j*, and
- sim_k is the cosine formula of the Vector Space Model.

A ranking of the news items is obtained according to their relevance for a given user. Top items in the ranking are selected for delivery to the user

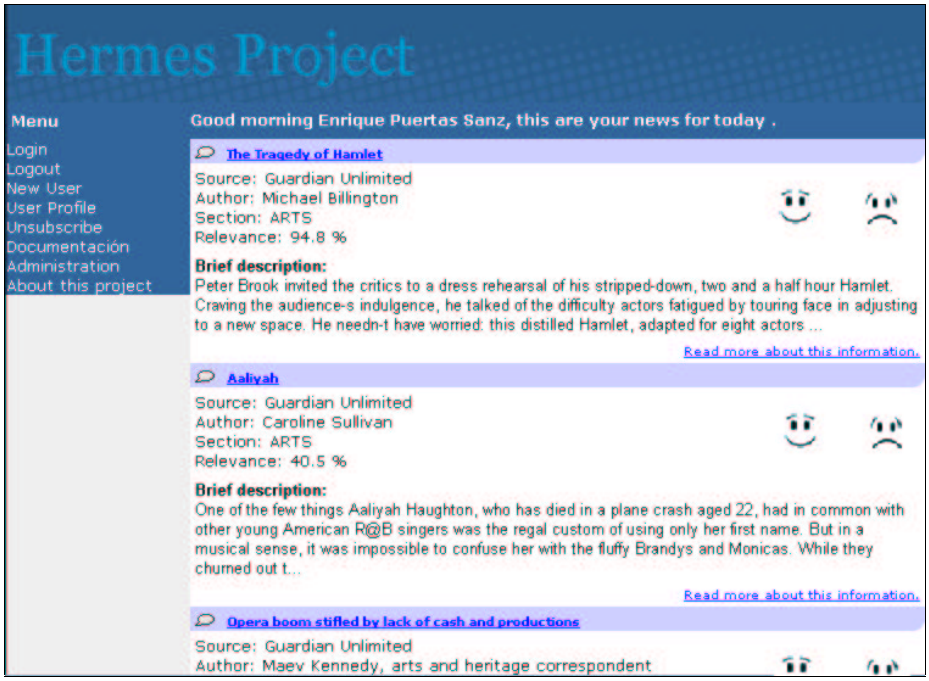

Figure 1. An example of automatically generated message produced by Hermes.

according to the upper bound on number of items per message specified in their profile. We applied Text Categorization using category-pivoted categorization ([Lewis 96, Ureña 98, Larkey 96]) with the categories against the news to obtain a ranking of the different news ordered by relevance for each category. We applied also Information Retrieval [Salton 89] with all the keywords against the news to obtain a list of relevant documents for the user. Also all news items are processed to check if they belong to one of the sections selected in the user model.

When all the documents have been sorted according to the different sources of relevance, the resulting orderings are integrated by using the level of interest that the user assigned to each of the different reference systems. This implies that users looking for the same information but having chosen different methods to specify their interest may get different results. For the relevance values provided to the user to be easy to interpret, they are normalised over the number of selection methods involved in obtaining them. In this way, the system can quote a final relevance value in the range 0-100% to every user regardless of the number of selection methods that he chose. Selected news summaries for each user are included in his e-message.

There is a feedback easy-to-use interface in each item, which allows user to specify his interest degree on that item.

An example of a message generated by Hermes is shown in Figure 1.

3. LANGUAGE ENGINEERING FOR ADVANCED USER PROFILING

3.1 The User Model

User model stores information about three different types of information:

- Management information: that includes name, login, password and email of the user.
- Messages information, including the days the user wants to receive the messages, a maximum and minimum of news per mail, a way to indicate that no more message should be delivered (e.g. for holydays), preferred language for the model and from which

sources the user want to obtain the news.

- Interests, with information about which sections of the newspaper and which categories the user is interested in. Also the terms chosen by the user are stored in the model. These terms are words, written in the language defined in the model, that the user considers interesting.

For each user, two profiles are stored: one in Spanish and another one in English. Once the user has filled the forms with his interests (in his language), the model is automatically translated to the other language. Storing two copies of the user profile is more efficient than translating all the news. The English profile is used with the news retrieved from the English e-newspaper. Analogue for the Spanish profile.

3.2 Machine Translation for Cross-Lingual Filtering

First we would like to remark that, on one hand, the addition of bilingual capabilities to the filtering task significantly enriches the quality of the output. On the other hand, the filtering task does not require a translation quality as good as the one you would need for other translation applications, such as legal document translation or literary translation. These tasks are particularly sensitive to translation quality, since the output text is required to be precise (most remarkably in the first application) as well as compliant with the underlying writing style (the way it happens in the second application). In these applications both parameters will greatly influence the quality of the output text, and consequently the satisfaction of the end user, and the suitability of the machine translation approach.

Nonetheless, the situation is quite different when translation is carried out in the context of a filtering application, since the parameters used for quality assessment in other machine translation applications are not suitable in the context of filtering. For instance, the violation of number agreement is highly relevant for assessing the translation quality of a legal document, but it is harmless within the context of a filtering application. Consider a document containing the sentence “the journalists sent their reports early in the morning”. This document will be very likely classified in the same group even if the prior sentence was erroneously translated as “the journalists sent his reports early in the morning”, which violates the number agreement and alters the meaning of the sentence. The robustness of the filtering application makes it immune to certain flaws of the translation tool. As long as the relevant keywords are correctly translated and the context is clearly determined, the translation quality will be sufficient to meet the demands of the filtering task.

With this considerations in mind, we decided to use the PROTRANS automatic translator, developed by one of the members of the Hermes development team for a headline translation application. PROTRANS was written in Prolog and its software architecture contains the following modules:

- **Interface**: supports command line user interface as well as file processing. Contains an input checker for handling erroneous spellings and unknown words.
- **Grammar**: contains approximate English and Spanish grammars using the Prolog Definite Clause Grammars representation formalism.
- **Lexicon**: contains 6000 entries with relevant information attached (such as meaning, gender, number, and others).
- **Lextool**: a tool for management of the lexicon that makes its editing unnecessary.

Using PROTRANS, the relevant terms in the lexicon have been collected in both languages and later edited into SQL commands for their inclusion in a database for fast access.

3.3 Model Evolution through Relevance Feedback

Information interests can not be assumed static in an information filtering application [Belkin 97]. Hermes has the ability to make the user models evolve according to user’s information interest changes along time. In order to change the model, some explicit feedback from the users is required. Users are allowed to state which news items provided in the message are of interest for them. The system adds keywords extracted from relevant news stories summaries to the model, using the techniques described in [Nakashima 97]. In short, the weight of terms is obtained by using information from the terms themselves, and a speed rate that defines which is the importance of terms in relation to other

terms that were already present in the user model (a kind of learning rate). Terms are slightly down-weighted each day to emulate a kind of forget function.

3.4 Automatic News Items Summarization

One of the key values of Hermes is the automatic summarization or abstracting facility. When a user receives a message from the system, it includes not only the title but a user-adapted summary of each news item that has been selected for him by the system. This summary allows the user to take a quick decision of the relevance of the news item to his information interests. The summary provided also explains, to some extent, which are the reasons why the system has selected the news item for them, increasing the confidence of the users in the decisions taken by Hermes.

According to the taxonomy presented in [Hahn 00], automatic abstracts generated by Hermes are:

- With a scope restricted to one document, i.e., a single news item (instead of generating a summary of all selected news items as a whole).
- Indicative, i.e., with the purpose of helping the user to make a relevance judgment (instead of being informative, replacing the reading of the full text, or critical, which incorporate opinion statements on content).
- User adapted, i.e., taking into account information about the user (information) interests (instead of generic, addressing a broad community with no specific needs).

Many current automatic abstracting systems are based on knowledge rich techniques, aiming at a nearly full understanding of the information content of the text. We have instead followed a knowledge poor approach, by using statistical techniques that basically consist of selecting best candidate sentences from the news item text to be abstracted, and appending those sentences to build the summary.

Most statistical based automatic summarization systems follow the approach of using a set of heuristics for computing a combined score for each sentence in the text, and selecting the top scoring sentences to build the summary. The heuristics include the position of the sentence in the document (initial sentences are more likely to give general ideas, and thus are better candidates for being in the summary), the occurrence of good keywords on them (the best sentences are those that discuss the central topics in the document), among others. We have also considered a customization heuristic that gives a higher score to those sentences that deal with the topics selected by the user in his profile. In short, the score for a sentence is computed as a weighted average of the scores of the sentence according to the three heuristics discussed here. The personalization heuristic score is computed by using a formula similar to that used for selecting the news items for a user with respect to the information need stated by them in their profile. More details about the summarization system can be found in [Acero 01].

4. USER-ORIENTED EVALUATION

4.1 Evaluation system

In order to evaluate the system, we have to consider two aspects: evaluation of the performance achieved by the system, analyzing measurable parameters, and an evaluation that considers user global satisfaction with the system (usability) [Chacón 00a, Chacón 00b, Díaz 00a, García 00, Pastor 99].

The evaluation process involved a group of 23 users taken from the research team, a group of students (Computer Science and Journalism) and some users that are neither related with computers nor journalism. Evaluation takes into account quantitative and qualitative aspects of the system, made by the selected users. Qualitative analysis have been made using data taken from objective parameters of the system, and a final valuation that reflects user satisfaction with the system and its results. Then, a report is done by evaluators showing positive and negative aspects of the system.

In order to make the quantitative analysis, we designed a test with different groups of questions about the interface, categories and sections, summaries, and the bilingual feature, which allows us to evaluate the system. Once we processed the

answers, we used the results to support the qualitative analysis of the system, reducing or reinforcing the valuation given by the evaluators when describing system. Studying obtained results, taken from both types of analysis, allows us to select the positive features and to identify its deficiencies.

4.2 Evaluation results

The overall satisfaction of users is very high. They show a high degree of satisfaction with the interface; the existence of a tutorial and an introduction manual to the system is very helpful, as well as the method for helping the user when entering the profile data. However this last issue may be improved.

Profile configuration has also been graded very well. Users are highly satisfied by the feedback system and with the possibility of choosing the days of the week in which they receive the news. Also, high satisfaction has been reported regarding the way the categories, sections and terms can be selected when building the profile. Abstracts are considered very good, specially regarding the way they are shown and the information provided in them.

Another well appreciated aspect is the possibility of choosing the language. However, users think that the translation of the selected terms on the user profile could be improved. Another feature suitable for improvement is the correspondence between received news in both languages according to the terms selected in the profile.

The overall evaluation of retrieved documents analyzed is very high, especially regarding the quality of the contents and the relevance to the interests of the user. Evaluators reported that contents of the final documents satisfy their information needs. They were also satisfied with the way the received news offered new knowledge about other documents related to them. As a final observation, evaluators consider the system as a very interesting tool.

We also performed an empirical evaluation of the quality of the summaries, that aims to test if they show the most important information present in the original news items. We have defined three profiles and manually found the most relevant news items for each of them. We have compared the ability of the system to select right news items according to those profiles by using the full text, and by using only the automatically generated summary. The results of this evaluation, are promising and support our claim that it is better to present user adapted summaries instead of generic ones (see [Acero 01] for more details).

5. RELATED WORK

In this section, we present a number of systems that aim at the same goals as Hermes, or propose similar methods for different problems.

On the recent years, there has been an increasing interest on personalization of information systems. For instance, web portals as Yahoo! offer free personalization services like My Yahoo! to promote traffic across their web pages. This service allows users to access a set of news delivered to Yahoo! by third-part content generation companies, as news agencies. My Yahoo! users can select a number of information sources from which they wish to see the headlines when accessing the service. The information sources come from a number of countries, and news are provided in several languages. The users of My Yahoo! can select some several country-based information services among those available for a small set of thematic language-independent categories. For instance, a user can select the sources “España: Finanzas (Reuters)”, “España: Finanzas (Europa Press)” and “US: Business from AP” from the broad topic “Economía y negocios” (Business and finance) when accessing My Yahoo! in Spain. News headlines selected from these information sources are shown to the user via the My Yahoo! page, and changed as a daily basis.

While the multilingual nature of these kind of services is out of doubt, it is clear that they are not cross-lingual, in the sense that the user does not specify a monolingual profile and is sent news items in several languages. For instance, a multilingual retrieval system would allow to retrieve documents in several languages, but it may retrieve only documents written in the query language. Instead, a Cross-Lingual Information Retrieval (CLIR) system would allow to retrieve documents in languages different from the query language [Gilarranz97].

CLIR is the focus of very active research nowadays. However, Information Retrieval and Filtering are quite different tasks. In Information Filtering, the system sifts through a stream of incoming information to find documents relevant to a set of

user needs represented by profiles. Unlike the traditional search query, user profiles are persistent, and tend to reflect a long term information need [Robertson 00]. Despite of the differences between Retrieval and Filtering, many techniques are shared.

CLIR approaches can be roughly classified in two groups: both that make use of machine translation and those that aim at building language independent index of the documents. Machine translation programs can be used for translating the query to the target documents languages, and/or to translate the documents to a selected language. The machine translation based approach is nowadays dominant in CLIR (see e.g. the systems presented at the Cross-Language Evaluation Forum (CLEF) workshops [Peters 01]). On the other side, resources like EuroWordNet have been developed with CLIR in mind, and can be used as the basis for a language independent index of documents [Gilarranz 97]. Systems representing this approach include ITEM and Hermes-UNED[2] [Verdejo 00].

According to the classification of CLIR approaches, our approach to Cross-Lingual Information Filtering follows the prevalent tendency of using a form of machine translation to translate queries, which happen to be profiles in the Filtering context. This cross-lingual feature is also the most remarkable difference between the system described in this work, and those which it evolves from, including Mercurio [Díaz 00b] and CODI [Gómez 01]. From these earlier systems, we have also improved relevance feedback functionalities and the implementation technology, among other features.

6. CONCLUSIONS

After evaluating Hermes we can conclude that the main goal has been reached. Hermes is one of the first cross-lingual personalized e-news systems, and has proven a valid prototype to demonstrate and evaluate how several text analysis and natural language processing techniques can be combined to improve communication across different language environ-ments. Of course there are several improvements that can be made in the system.
Hermes has demonstrated to be a valid cross-lingual system, but it may be improved by supporting another languages, and extracting news from more sources. Another interesting improvement is to add semantic-based elements to Hermes language processing techniques. Adding semantic information should improve its results, obtaining several desirable benefits (e.g. clarifying ambiguity situations).

[2] *We identify the Hermes project leaded by the Universidad Nacional de Educación a Distancia (UNED) at Madrid, Spain, as Hermes-UNED, to avoid confusion with our own system.*

REFERENCES

[Acero 01] Acero, I., Alcojor, M., Díaz Esteban, A., Gómez Hidalgo, J.M., Maña López, M. Generación automática de resúmenes personalizados. *Procesamiento del Lenguaje Natural*, 27, 2001.

[Belkin 97] Belkin, N. J. (1997) User Modeling in Information Retrieval. In *Proceedings of the Sixth International Conference on User Modeling, UM97*, Chia Laguna, Sardinia, Italy.

[Chacón 00a] Chacón I., García A., Díaz A. Y Gervás, P. "Sistemas de información en Internet: estudio de un caso", *Investigación Bibliotecológica*, julio-diciembre 2000, vol. 14, nº 29, p. 114-129.

[Chacón 00b] Chacón I, García, A. y Guede, A. Sistemas de almacenamiento y recuperación de imágenes fotográficas en Internet, *Documentación de Ciencias de la Información*, 2000, nº 23, p. 109-122.

[Díaz 00a] Díaz, A., Gervás, P. and García A. (2000). Evaluating a User-Model Based Personalisation Architecture for Digital News Services. *Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries*, Lectures Notes in Computer Science, Springer Verlag, Lisbon, pp. 259-268.

[Díaz 00b] Díaz, A., Gervás, P., Gómez Hidalgo, J.M., García, A., de Buenaga, M., Chacón, I., San Miguel, B., Murciano, R., Puertas, E., Alcojor, M., Acero, I., Proyecto Mercurio: un servicio personalizado de noticias basado en técnicas de clasificación de texto y modelado de usuario. *XVI Congreso de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)*, Vigo, Spain, September 2000.

[García 00] García, A., Chacón, I., Díaz, A. and Gervás, P. (2000). Nuevos sistemas de información: tendencias y evaluación, *Cuadernos de*

Documentación Multimedia, nº9, http://www.ucm.es/info/multidoc/multidoc/revista/num9/prensa/jime-chacon.htm

[Gervás 99] Gervás, P., San Miguel, B., Díaz, A. and García, A. (1999) "Mercurio: un servidor personalizado de noticias basado en modelos de usuario obtenidos a través de la WWW", *III Congreso de Investigadores Audiovisuales (Los medios del tercer milenio)*, 10-12 noviembre 1999, Facultad de Ciencias de la Información, Universidad Complutense de Madrid, Madrid.

[Gilarranz 97] Gilarranz, J, J. Gonzalo y M.F. Verdejo. (1997) Language-Independent Text Retrieval using the EuroWordNet Multilingual Semantic Database. *Proceedings of the Second Workshop on Multilinguality in the Software Industry: the AI contribution, Workshop 21 of the 15th International Joint Conference on Artificial Intelligence*, IJCAI'97.

[Gómez 01] Gómez Hidalgo, J.M., Murciano Quejido, R., Díaz Esteban, A., de Buenaga Rodríguez, M., Puertas Sanz, E. Categorizing photographs for user-adapted searching in a news agency e-commerce application. *Proceedings of the 1st International Workshop on New Developments in Digital Libraries (NDDL-2001), International Conference on Enterprise Information Systems (ICEIS 2001)*, Setúbal, Portugal, 7-10 July, 2001.

[Gómez 02] José María Gómez Hidalgo, Enrique Puertas Sanz, Manuel de Buenaga Rodríguez, Francisco Carrero García. Text filtering at POESIA: a new Internet content filtering tool for educational environments. *Procesamiento del Lenguaje Natural*, no. 29 (2002), pp. 291-292.

[Hahn 00] Hahn, Udo, and Mani, Inderjeet. The challenges of automatic summarization In *Computer*, 33 (11), 2000, pp. 29-36.

[Nakashima 97] Takuo Nakashime, Ryozo Nakamura. Information filtering for the Newspaper. *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. August 20-22, 1997, Victoria, B.C., Canada.

[Pastor 99] Pastor, J. A. and Asensi, V. (1999). Un modelo para la Evaluación de Interfaces en Sistemas de Recuperación de Información, *IV Congreso Isko-España Eoconsid'99*, Granada (Spain), 1999.

[Peters 01] Carol Peters. Results of the CLEF 2001 Cross-Language System Evaluation Campaign. *Working Notes for the CLEF 2001 Workshop*, 3 September, Darmstadt, Germany, 2001.

[Robertson 00] S. Robertson and D.A. Hull. The TREC-9 Filtering Track Final Report. In *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC 9)*, 2000.

[Salton 89] Salton, G. (1989). Automatic Text Processing: the transformation, analysis and retrieval of information by computer. Addison Wesley. 1989

[Verdejo 00] Verdejo, M.F., Gonzalo, J., Peñas, A., López, F. and Fernández, D. (2000) Evaluating wordnets in Cross-Language Text Retrieval: the ITEM multilingual search engine. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC'2000),* Athens.