NowOnWeb: News Search and Summarization

Javier Parapar, José M. Casanova, and Álvaro Barreiro

IRLab, Department of Computer Science , University of A Coruña,
Campus de Elviña s/n, 15071, A Coruña, Spain
{javierparapar,jcasanova,barreiro}@udc.es
http://www.dc.fi.udc.es/irlab

1 Introduction

In the last years the number of news sites available on Internet has experienced an amazing growing; this produced the overwhelm of the manual techniques to cover all this information. In this situation the use of Information Retrieval (IR) strategies has been a successful solution.

In this context we present a system to manage the news articles from different online sources and capable to provide search ability, redundancy detection and summarization. We called it *NowOnWeb*[1] and is based on our previous research and solutions in the IR field.

Within this paper our aim is to introduce the general architecture of our news system and also to go deep with the three main research subjects for which we provided useful solutions.

2 System Architecture

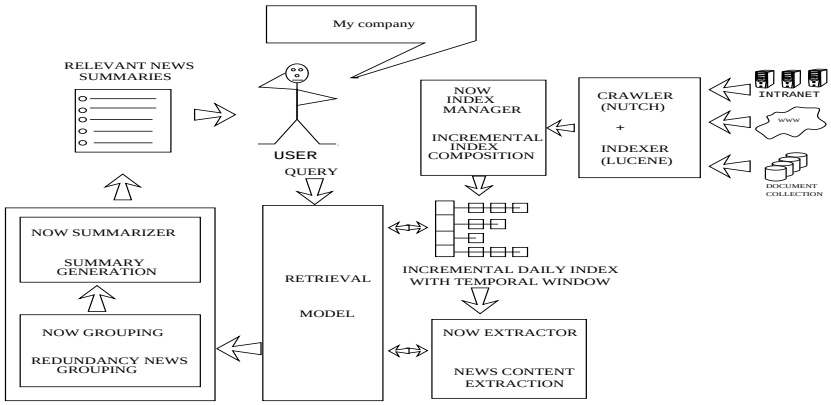

Fig. 1. System architecture overview

[1] A version with Spanish and international news is operative in http://irlab.dc.fi.udc.es

Figure 1 is a general overview of our system architecture. It shows the main components of the application and how they interact.

3 Research Issues

NowOnWeb shows innovative solutions to the three main problems of an IR news system, i.e., news recognition and extraction, redundancy detection and summary generation. Next we will introduce the issues and solutions.

News Recognition and Extraction: There are computationally expensive previous works [1, 2] providing a generic method to automated data extraction. We developed a computationally effective alternative based on a few heuristics about the structure of the news body, title and images, with linear complexity on the document length

Redundancy Detection: In this field our aim was to avoid to the user the overloading of read repeated news. As solution we created a new algorithm based upon the calculation of redundancy using the cosine distance[3] over a previous ranked collection of relevant news represented as vectors.

Summary Generation: For this problem we extract the relevant sentences [4] of an article with respect to the user query. The selected sentences are joined to form a coherent summary within the 30% of the original size[5].

Acknowledgements. The work reported here was cofunded by the "Secretaría de Estado de Universidades e Investigación" and FEDER funds under the project MEC TIN2005-08521-C02-02 and "Xunta de Galicia" under project PGIDIT06PXIC10501PN. The authors also want to acknowledge the additional support of the "Galician Network ot NLP&IR" (2006/23).

References

1. D. C. Reis and P. B. Golgher and A. S. Silva and A. F. Laender: Automatic web news extraction using tree edit distance. WWW '04: Proceedings of the 13th international conference on World Wide Web (2004) 502–511
2. Valter Crescenzi and Giansalvatore Mecca: Automatic information extraction from large websites Journal ACM Vol 51 **5** (2004) 731–779
3. Yi Zhang and Jamie Callan and Thomas Minka: Novelty and redundancy detection in adaptive filtering SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (2002) 81–88
4. James Allan and Courtney Wade and Alvaro Bolivar: Retrieval and novelty detection at the sentence level SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (2003) 314–321
5. Eduard Hovy: Text Summarization. In R. Mitkov Ed. The Oxford Handbook of Computational Linguistics, chapter 32 (2005) 583-598