Procesamiento del Lenguaje Natural, nº39 (2007), pp. 287-288 recibido 05-06-2007; aceptado 22-06-2007

NowOnWeb: a NewsIR System*

Javier Parapar
IRLab, Computer Science Dept.
University of A Coruña, Spain
Fac. Informática, Campus de Elviña
15071, A Coruña, SPAIN
javierparapar@udc.es

Álvaro Barreiro
IRLab, Computer Science Dept.
University of A Coruña, Spain
Fac. Informática, Campus de Elviña
15071, A Coruña, SPAIN
barreiro@udc.es

Resumen: Hoy en día existen miles de sitios web de noticias. Los modos tradicionales para acceder a este inmenso repositorio de información no son adecuados. En este contexto presentamos *NowOnWeb*, un sistema de recuperación de noticias que obtiene los artículos de la red y permite buscar y navegar entre los mismos.
Palabras clave: Sistemas de noticias, extracción de información, detección de redundancia, generación de resúmenes.

Abstract: Nowadays there are thousands of news sites available on-line. Traditional methods to access this huge news repository are overwhelmed. In this paper we present *NowOnWeb*, a news retrieval system that crawls the articles from the internet publishers and provides news searching and browsing.
Keywords: News system, information extraction, redundancy detection, text summarization.

1. Introduction

The huge amount of news information available on-line requires the use of Information Retrieval (IR) techniques to avoid overwhelming the users. The main objectives of these IR methods are: reduce the time spend in reading the articles, avert the redundancy and provide topic search capability. Given this context we present *NowOnWeb*[1], a NewsIR system that deals with the on-line news sources to provide an effective and efficient way to show news articles to the user through a comfortable and friendly interface. It is based on our previous research and solutions in the IR field and serves as a research platform to test and asses the new solutions, algorithms and improvements developed in the area.

2. System Overview

NowOnWeb was designed as a Model-View-Controller web-application following a component-based architecture. The main system components are: a crawler and an indexer to maintain an incremental index with a temporal window, a news recognition and extraction module that allows dynamic source adding, a news grouping component that uses a redundancy detection approach, and an article summariser based on relevant sentences extraction.

Figura 1: A snapshot of the application appearance.

Our application offers the user: news searching among all the indexed publishers, query suggestion, query spelling correction, redundancy detection and filtering, query biased summary generation, multiple format outputs like PDF or syndication services, and

* Acknowledgements:This work was cofunded by the "Secretaría de Estado de Universidades e Investigación" and FEDER funds (MEC TIN2005-08521-C02-02) and "Xunta de Galicia"(PGIDIT06 PXIC10501PN).

[1] An operative version with international news is available in http://nowonweb.dc.fi.udc.es

personalisation options such as source selection. All these characteristics aim to facilitate the use of the system, for this reason the results are showed in a friendly and natural way (see Figure 1). In this sense technologies like AJAX were applied in order to improve the user experience and the system possibilities.

3. Research Issues

Three are the main research topics involved in the development of *NowOnWeb*: news recognition and extraction, redundancy detection and summary generation

3.1. News Recognition and Extraction

The problem here is to extract from an heterogenous set of pages, most of them without articles, the news articles present. So first we have to filter the pages without interesting content, and second from those with an article inside, extract the fields (title, body, date and image if present) among many not desired content.

We developed a news recognition and extraction technique based on domain specific heuristics over the articles structure that resulted in an efficient and effective algorithm.

3.2. Redundancy Detection

The objective of this point is to filter the redundant articles in order to avoid the overload of the user. To get this we developed and algorithm based on traditional techniques of the information filtering field (Zhang, Callan, and Minka, 2002).

Generally speaking our method takes as input a ranking of documents sorted in base of their relevance with the user query. The algorithm dynamically assigns a redundancy score to each document respect to the already created redundancy sets. If that score is over a threshold with one of the sets, the document will be included in that set, other way it will constitute a new redundancy group.

3.3. Summary Generation

The system offers the user summaries about the relevant articles respect to the query. These summaries are dynamically generated in retrieval time, they are query-biased.

To get this we used a technique based on the extraction of relevant sentences. Each sentence is scored (Allan, Wade, and Bolivar, 2003) with respect to its relevance with the query. The sentences with higher score are chosen to get a summary of the desired size and they are resorted to maintain the original article relative position.

4. Conclusions and Future Work

NowOnWeb resulted in a NewsIR system that satisfies the user needs of information, allowing them to be up-to-date without time waste.

We got an original solution different from the existing ones in the academic (Columbia NewsBlaster (McKeown et al., 2002), Michigan NewsInEssence(Radev et al., 2005)) and commercial (Google News, Yahoo News or MSN Newsbot) fields.

As further work we will approach architectural system improvements, efficient query logging storage and mining, and evaluation of our news extraction algorithm.

References

Allan, James, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 314–321, New York, NY, USA. ACM Press.

McKeown, Kathleen R., Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*.

Radev, Dragomir, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. 2005. Newsinessence: summarizing online news topics. *Commun. ACM*, 48(10):95–98.

Zhang, Yi, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88, New York, NY, USA. ACM Press.