

# WELKIN: automatic generation of adaptive hypermedia sites with NLP techniques\*

Enrique Alfonseca, Diana Pérez,, and Pilar Rodríguez

Computer Science Department, Universidad Autonoma de Madrid,  
Carretera de Colmenar Viejo, km. 14,5,  
28043 Madrid, Spain

{Enrique.Alfonseca, Diana.Perez, Pilar.Rodriguez}@ii.uam.es  
<http://www.ii.uam.es/~ealfon>

**Abstract.** The demonstration shows the system WELKIN, a multilingual system that analyses one or several source texts with a cascade of linguistic-processing modules, including syntactic analyses, text identification and text classification techniques, in order to fill in a database of information about it. That information is later used to generate on-the-fly adaptive on-line information sites according to some user profiles.

## 1 Introduction

WELKIN<sup>1</sup> is an on-going project which started three years ago, whose aim is to build automatically adaptive on-line hypermedia sites from electronic texts using Natural Language Processing techniques [1, 2]. It differs from previous approaches in that the knowledge base is generated automatically from source texts. Figure 1 shows the architecture of the system, which is divided into two steps: an off-line processing in which the texts are analysed and several internal databases are created, and an on-line processing step, in which the web pages are generated, when users access the system, in an adaptive way.

## 2 Off-line processing step

During the **off-line** processing, the domain-specific texts provided by the user are analysed with some modules, which can be configured by the user. Currently, the following modules are available:

- Linguistic processing tools: tokenisation, sentence splitting, stemming, chunking, and a shallow dependency parser.
- Term extraction: the unknown domain-specific terms that have a high frequency of appearance in the documents are automatically considered relevant terms and collected.

---

\* This work has been sponsored by CICYT, project number TIC2001-0685-C02-01.

<sup>1</sup> <http://agamenon.ii.uam.es/welkin/>

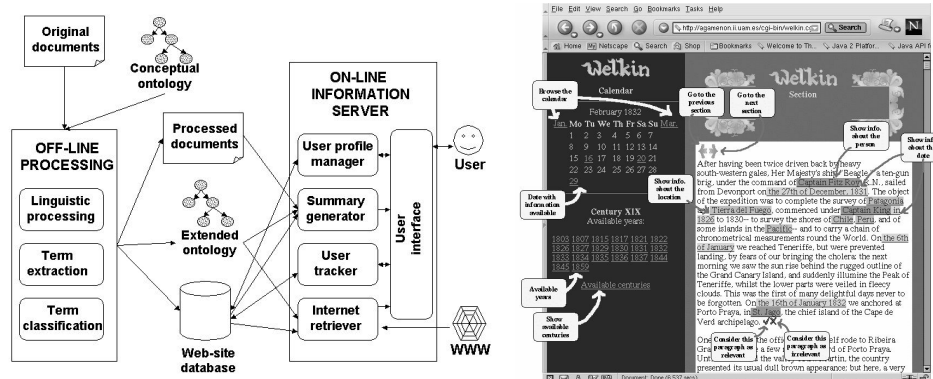


Fig. 1. *Left*: High-level architecture of WELKIN. *Right*: Screenshot.

- Special terminology identification: there are modules which are specialist for identifying dates and scientific names inside the texts.
- Finally, a module that classifies the unknown terms, automatically, inside a semantic network, so some meaning can be inferred from their position.

The cited modules are currently available for English and Spanish, so texts in both languages can be processed. The output of this step, apart from annotating the texts with the output of the modules, is also to build a database with all the information that could be extracted. This step is fully automatised, so the whole web site can be generated just by executing a script that indicates which modules to use.

### 3 On-line processing step

After a brief description and show of the off-line step, the demonstration will centre on the on-line step, which involves the ways in which users can interact with the system. This includes the creation of their initial profiles, and the manners in which WELKIN adapts the contents according to them. The possibilities for adaptation include:

- An automatic summarisation of the generated pages, depending on the user's reading speed and available time.
- A selection of the contents inside each page, according to a profile of interests.
- An additional tool to guide a search on Internet for terms from the web site.

### References

1. Alfonseca, E.: An Approach for Automatic Generation of Adaptive Hypermedia using Knowledge Discovery, Text Summarisation and other Natural Language Processing Techniques. Ph.D. thesis, Universidad Autónoma de Madrid (2003)
2. Alfonseca, E., Rodríguez, P.: Modelling users' interests and needs for an adaptive on-line information system. Volume 2702 of Lecture Notes in A.I. (2003) 76–80