The TREVI Project

Personalized Information Filtering, Linking, and Delivery for the News Domain

Reginald Ferber and Costas Tzeras, GMD - IPSI, Dolivostr. 15,
D-63293 Darmstadt, Germany, {ferber,tzeras}@darmstadt.gmd.de,
http://www.darmstadt.gmd.de/~ferber http://www.darmstadt.gmd.de/~tzeras

The goal of the TREVI project (Text Retrieval and Enrichment for Vital Information) is to offer a solution to the problem of "information overflow", i.e. the problem experienced by companies and individuals in extracting useful information from distributed textual information sources. These information sources are available through public distribution channels such as the Internet and the World Wide Web, or through proprietary networks. At the same time, more and more archival or encyclopedic data collections are becoming available in electronic format, providing background knowledge to particular business domains.

The TREVI approach aims to filter information from streams of incoming news (information sources) based on individual user profiles. Furthermore, TREVI aims to enhance the filtered information by enrichment with background data sources in accordance with user profiles. The filtered and linked information will be presented in a coherent and comprehensible way to end-users (document publication).

TREVI is an ESPRIT joint project (ESPRIT Programme 23311) of GMD-IPSI with

- Economisch Instituut Tilburg (EIT), Netherlands;
- FEND Association, Spain;
- ITACA s.r.l., Italy;
- Lyras Shipping LTD, United Kingdom;
- REUTERS LTD, United Kingdom;
- SARENET SA, Spain;
- Vrije Universiteit Brussel (VUB), Brussels.

It will run from January 1997 to June 1999.

GMD-IPSI's work on TREVI is divided into two parts: (1) Text Enrichment and (2) Document Publication. GMD-IPSI also assists the project partners in specifying a representation formalism for user profiles.

TREVI will be applied to four test environments:

- The Italian Health online service (ARAKNE) that provides news, research results, and information documents for the medical domain from various sites. As background material there will be archives of this service.
- The ECO PRENSA service that provides Spanish abstracts of newspaper articles from the economics domain. As background material there will be databases with information from the stockmarket.
- An experimental subset of Reuters news service. As background material there will be a set of historical information and selected news articles, stockmarket and company information.
- The distribution of business circulars within the Lyras shipping company. These circulars include news, guidelines, and business informations that have to be directed to the appropriate persons within the company.

The incoming information streams will be heterogeneous. They will rank from unstructured texts to news that are structured by different fields like author, city of origin, subject etc. The background material will also be heterogeneous. It will include unstructured text documents, weekly structured material and databases as highly structured information.

The main tools for filtering are a lexicon system and the user profiles. The lexicon system is a kind of enriched ontology based on WordNet that allows to specify concepts. WordNet is enriched with specific information from the domain and with linguistic and terminology information for parsing and tagging.

The user profiles contain various types of information: Content information is specified by concepts from the lexicon. Some metadata specify formal information like sources to be used, cost limits, time and geographical restrictions as far as they can be identified in structured sources. A second part of the user profiles is information concerning strategies of enrichment. This can be the time at which the enrichment shall be made, different search strategies, and formal restrictions or properties for background material like source, length, price, age... The last package of information concerns the selection and configuration of modules used for the specific user. This selection depends on the information sources, the availability of specified lexical information and the retrieval methods to be used. It will affect the speed and the costs of linking. Probably there will be a fixed selection of configurations for the most likely scenarios.

The user profiles will be created either by information experts or brokers for their clients, or by expert clients themselves. Such experts will be able to change their profiles temporarily or permanently. There will be also some predefined profiles for casual users.

The publication and user interaction component of TREVI will supply three different modes to be selected depending on the user habits and the network and hardware situation. In e-mail mode the filtered news and the respective background information will be send to the user in fixed time intervals or upon a arrival. In the two other modes - the HTML and APALO mode - the filtered news will be collected and shown when the user logs in. In this case users can select if they want to see all items that arrived since they logged in last or only those from a given time period. The APALO mode is named after a layout system developed by GMD IPSI that puts strong emphasis on a structured and content sensitive presentation of text and images. Both the HTML and the APALO mode will provide personal archives, to store and retrieve documents. Retrieved documents can be used to select similar information from the news stream or the background material. Advanced users can select retrieval strategies. In addition they can use relevance feedback and a profile editor to change their user profiles.