Cultural Heritage: knowledge extraction from web documents

Eva Sassolini, Alessandra Cinini
Consiglio Nazionale delle Ricerche – Istituto di Linguistica Computaizonale "Antonio Zampolli"
Via Giuseppe Moruzzi n° 1, 56124 Pisa, Italy
{eva.sassolini|alessandra.cinini}@ilc.cnr.it

Abstract

This article presents the use of NLP techniques (text mining, text analysis) to develop specific tools that allow to create linguistic resources related to the cultural heritage domain.
The aim of our approach is to create tools for the building of an online "knowledge network", automatically extracted from text materials concerning this domain. A particular methodology was experimented by dividing the automatic acquisition of texts, and consequently, the creation of reference corpus in two phases. In the first phase, on-line documents have been extracted from lists of links provided by human experts. All documents extracted from the web by means of automatic spider have been stored in a repository of text materials. On the basis of these documents, automatic parsers create the reference corpus for the cultural heritage domain. Relevant information and semantic concepts are then extracted from this corpus. In a second phase, all these semantically relevant elements (such as proper names, names of institutions, names of places, and other relevant terms) have been used as basis for a new search strategy of text materials from heterogeneous sources. In this case also specialized crawlers (TP-crawler) have been used to work on a bulk of text materials available on line.

1. Introduction

The diffusion of the internet and the information technologies are creating continuous information flows. There is a widespread awareness of the added value and of the role that the web has in the dissemination, exploitation and promotion of the Italian cultural heritage. Moreover an open philosophy causes problems of authoritativeness in the production of contents because it is characterized by a strong interaction among users thus creating a distance between knowledge and communication. The diffusion of the spread of the network is followed by significant changes of communication paradigm. Nowadays the competition among contents decreases, even among from sources published in potential competition with them. In network logic, all nodes are interdependent and represent a single large hypertext. The proliferation of paths boosts circulating of ideas and can bring out most interesting contents. Consequently, in our experience, the only use of crawling tools is not sufficient; you must first build a knowledge base of specific domain to build the acquisition strategies.

2. ICT and Cultural Heritage

In many European countries initiatives aimed at developing knowledge and enhancement of digital cultural heritage have been undertaken. Among these, "Minerva" and "Michael" have been coordinated by the Italian Ministry for Cultural Heritage. Minerva has developed a platform of guidelines and recommendations, which are shared by European member states, for the digitization of cultural heritage and its network access.
Since from October 2006, the Minerva project has been enlarged to MINERVA EC, which is a Thematic Network in the area of cultural, scientific information and scholarly content.

The Michael project[1] (Multilingual Inventory of Cultural Heritage in Europe), will establish an international on-line service that will allow users to search, browse and examine multiple national cultural portals from a single access point.

2.1 Communicative models

Some basic rules make on-line communication models more effective. These models should pay special attention, defining assets, which cannot be ignored:

- Relation generating: communication goes through people;
- Potential users: the catching is essential to arose the users' curiosity;
- Innovation: innovation is a value and a content at the same time and it is an important repeater on traditional media;
- Talk: the network is compared to a "Big Conversation", in that communication is bidirectional.

In summary, the aim is to develop and translate a popular approach that is focused on the user.

3. "The on-line dissemination of the historical artistic and landscaped, regional heritage" project

The project was born within the framework of a collaboration between the Pisa ILC-CNR and the APT Basilicata (i.e. Agenzia di Promozione Territoriale della regione Basilicata) to experiment and implement strategies for the promotion and dissemination of regional heritage.
The ILC contribution consisted of defining a linguistic analysis model of texts and of acquiring domain linguistic resources. Semantic information and terminology acquired have been later used for text categorization.

[1] Michael was called "Michaelplus" since 2006.