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.

3.1 Objectives

The specific goal of the project is to experiment and implement strategies for promotion and dissemination of regional heritage through a effective communication style. The project aims offering a set of selected documents and cultural routes by which to increase the understanding of historic and landscaped resources, exploiting the potentialities of Web 2.0. The aim is to join the maintenance and preservation activities and enhancement and promotion of the cultural heritage initiatives with an increasing opening toward the general user.

3.2 Expected results

The most important results of the project are the creation of proposals for use of the cultural and landscaped heritage as well as the creation of a prototype for the dissemination of domain information. The contents extracted will be articulated in cultural and tourist itineraries, thematic areas (such as historical periods) and types (castles, churches, etc.), according to a specific analytical work and research.

4. ILC experiences

The tools provided by ILC belong to next important projects: "Linguistic Miner" and "Text Power", both based on modules, methods and resources developed by our research team.

4.1 Linguistic Miner (LM)

LM was created for the automatic extraction and acquisition of linguistic knowledge from large collections of text material in Italian to develop:

- procedures for the automatic downloading (like automatic spiders and parsers) and for the analysis of large bulk of Web texts;
- statistical and linguistic analysis tools.

4.2 Text Power (TP)

TP is the natural evolution of LM and aims at identifying implicit semantic knowledge in the documents which is expressed through the text annotation and classification. Named entities identification and terminology (“mono / polirematica”) represent a fundamental part of text enrichment. Both the named entities and the terminology were extracted by means of the tool of NER (Name Entity Recognition) developed in TP, named PiNER. The relevant elements are typically proper names, names of institutions, locations, and other similar terms. The key feature of this system is the ability to develop strategies for the classification and recognition of terms and named entities, very important prerequisites for a more effective text analysis.

4.2.1. DBT Faccette

Our experience in the treatment of large amount of data has not only allowed the refinement of extraction tools for semantically relevant information, but also the creation of terminological resources toolkit.

Identification of all semantic information has contributed to the creation of a system of textual analysis. It is clear that this process also produced a knowledge network that can be used for functions of clustering and intelligent browsing: this constitues the richness of the instrument and the service offered by the system.

The more a text is "enriched" with annotations, the better it can be processed by tools for analysis, categorization, browsing and Information Retrieval. The system here described is not limited to the identification and classification of entities; it also identifies the particular relations between the entities involved. In an open domain like "cultural heritage" the information can hardly be classified just via hierarchical criteria. An approach based on criteria of "semantic similarity" is more useful as it allows to link different types of information which may belong to different domains, but which satisfy a particular information need nonetheless.

5. Strategies for Cultural Heritage texts acquisition

The text materials about cultural heritage of Basilicata were collected in two steps. We built a starting text corpus (hereafter C0) which was exploited to generate new linguistic resources and to enrich the ones available from LM and TP projects. Then on the basis of these resources, we enlarged C0 to create the reference corpus.

5.1 First acquisition strategy

The APT Basilicata provided us with a list of Basilicata websites. The items of the list more related to historical, artistic and landscaped regional heritage were selected and browsed by using automatic spiders and parsers for the creation of a cultural heritage text corpus. Then, all textual materials acquired, were indexed and after a tagging phase by using PiTagger[2] we could identify all lemmas and relative POS in each document. PiTagger associates each word to the related lemma by using the morphological component of the Italian language PiMorfo[3]. Then it solves the ambiguities by following a statistical approach on the basis of a training corpus statistically analyzed and summarized. Later on the multiword were extracted from C0, by exploiting pattern matching techniques. We refer the multiword also as "Facets[4]” in the article. Typically, in the Italian syntactic construction, the most productive linguistic patterns are N-preposition-N and ADj-N/N-ADj. Statistical algorithms analyze the distributions frequency of each pattern identified. On the basis of results we extracted a set of semantically relevant terms and concepts for cultural heritage domain. As a matter of fact, analyzing

[2] PiTagger is an important component for text lemmatization and tagging and constitutes a software module of PiSystem: integrated system for processing of textual and lexical materials.

[3] PiMorfo: system for morphological analysis of the italian language.

[4] We use the term "facets" to refer to such classes, extending the traditional notion of this term as in the literature.

the collected texts material by means of linguistic tools (morphological engine and tagger) is fundamental for a productive application of the statistical functions of extraction. Semantically relevant terms extracted for cultural heritage domain are terms e.g.:

belonging to a domain terminology
- "campanile a vela"
- "castello angioino"
- "scavi archeologici"

concerning historically events:
- "invasione dei normanni"
- "fine del neolitico"

concerning proper names, surnames, geographic locations, institutions, etc.:
- "Carlo Levi"
- "Lago del Pertusillo"
- "Museo archeologico nazionale Domenico Ridola"

We annotated all relevant terms in the texts which were again indexed to regenerate C0 corpus.

5.1.1. Text corpus to linguistic resources

C0 was exploited to generate new linguistic resources and integrate the ones already existent. Three weighed domain lexicons were built:
- Cultural heritage of Basilicata;
- I Normanni in Basilicata
- Sassi di Matera

Starting from a small set of relevant pivot terms[5], each lexicon is obtained by means of mutual information criteria. Statistical algorithms analyze and weigh the frequency of the cooccurrency of each word with the pivot terms. The domain lexicons can be used to evaluate the relevance of a document for that domain. In this case it is most important to establish a minimal threshold.

5.2 Second acquisition strategy

The next step involved the browsing of the item links excepted in the previous phase of work, such links were mainly official websites of municipalities and provinces. The textual materials collected were filtered by using the Basilicata cultural heritage weighed lexicon. The specific tools that, create the domain topic, were used to rank the documents and evaluate their relevance with the cultural heritage domain. The extracted documents were joined in C0 corpus to build the reference (text) corpus. The reference corpus is constituted of almost 2 million and a half words.

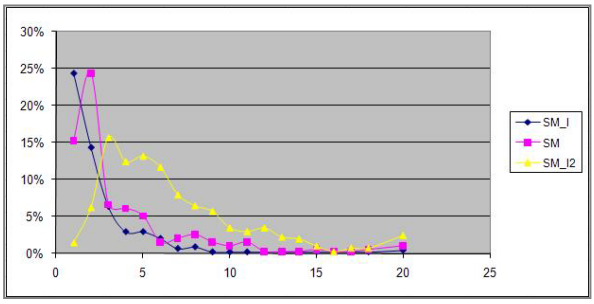
6. Topics

The term "topic" denotes a field of interest chosen according to the requirements of the project. Any topic is made with linguistic-statistical procedures and is aimed at creating a domain lexicon weighed, can be also used for text classification. In fact, the procedure allows to identify automatically the specific text domain based on the terminological lexicon, and to measure the relevance (ranking) of an article to the selected topic. Once the lexicon has been created, the system recognizes the relevant documents available and provides the user with a list of documents sorted by relevance to the topic chosen. Basically, the system "decides" if a document deals with the topic. Various phases of testing have shown a high degree of success of the procedure, but the creation of each topic remains a delicate task and requires a drafting process *a posteriori*. Often, in the short documents, the system fails to properly weigh the relevant terms in relation to the other words of the text. This requires careful evaluation of cases and also progressive phases of tools tuning.

6.1 Comparison between two methods

In this section we are going to show a comparison between the results of two specific topics about "Sassi di Matera". The topics "Sassi di Matera" (hereafter SM) e "Sassi di Matera in Internet"(hereafter SM_I) were built by means of two different strategies and using specific weighed lexicon "Sassi di Matera" as semantic filter:

1) SM generated from the reference corpus;
2) SM_I obtained by using tools crawler like (TP-crawler[6]) that, starting from a query consisting of a key words sequence, retrieves the documents proposed from on-line main search engines. The condition set in the query is not sufficient by itself to ensure the relevance of the documents recovered for the chosen topic. It is fundamental to prepare semantic filter to extracting a coherent set of web pages, without loss of significant documents.



Graph 1: distribution of documents depending on the occurrences of the word families in the three topics

The relevant documents in both corpora compared by means of DBT-Query System, which enables advanced searches on the texts. A set of words combined using logical operators (*word families*), were searched in SM and SM_I to evaluate their frequency in each document. We searched for six word families in the corpora setting the word distance parameter at two:

- Sassi (and) Matera;
- Rioni (and) Sassi
- Città (and) Sassi

[5] We chose some terms about "Sassi di Matera" among a set of relevant terms suggested by domain experts

[6] TP-crawler is a specific crawler system developed in TP project

- Sasso(and) Caveoso;
- Sasso (and) Barisano;
- Chiesa/e (and) Rupestre/i

The results show in SM a relevant documents percentage of 13% higher than in SM_I.

In the light of previous results we tried to improve TP-crawler. We refined parsing and topic generation modules, particularly the ranking documents procedure; finally we acted on acceptability threshold.

The topic “Sassi di Matera in Internet 2” (hereafter SM_I2) was generated on the basis of the new TP-crawler and a percentage of relevant documents of 20% higher than in SM was obtained.

As shown in the graph, the largest number of documents in the last topic contains from 3 to 10 occurrences of the searched families of word. Otherwise the documents of the first two topics are focused on low number of occurrences.

7. Conclusion

We would like to give below a summary of the problems encountered in the creation of different corpora but also of two topics "Sassi di Matera" e "sassi di Matera in internet".

7.1 Updating resources

The updating of language resources is an open question. The manual generation of resources is expensive and it requires the intervention of human experts. The manual approach is also prone to errors of omission. Otherwise, an approach based on the automatic updating of language resources is not efficient because the terminology is constantly evolving and cannot be entrusted to the exclusive use of automatic application. Especially dealing with tourism, the web is constantly offering new ideas: a new park, an archaeological site discovered, etc... We were look for a diplomatic solution: we try to find a compromise.

Currently the project is not finished yet and the various applications have to be integrated in a single system yet. In any case, to have always current content, you need updated resources, both text material and language resources. Specifically, we have planned periodic phases of updating frequency, which is however determined by mutual agreement with the other project partners.

7.2 Assessment results

The model of linguistic resources acquisition which we have developed can be applied to other domains or sub domains. For example, we used a model derived from the model described in this article to generate linguistic resources related to the issues of food safety and restoration of cultural heritage. One of the critical phase of our approach is the determination of the threshold in relation to a weighed domain lexicon. In fact we have to perform several testing phases to identify the optimal value, value that may be difficult to be establish about generic domains. The system can propose a threshold, but it is only a guideline.

We should use TP-crawler mainly if we have already domain linguistic resources available. In this case the TP-crawler allows us to collect large amount of domain information, including news, and updates the domain corpora already built. It may also be useful in creating sub domains o derived domains: as “Restauro” in cultural heritage domain.

The web integrated system of project is not on line available yet, then we not had the possibility to evaluate the usability of the navigation system "DBT Faccette" by the end-user. This phase of evaluation will be made in later stages of the project. In order to test the search system, we allowed the browsing in the all text corpora as well as the visualization of all repositories of relevant terms. We realized an demo website available at the URL: http://serverdbt.ilc.cnr.it/DBTFaccette.

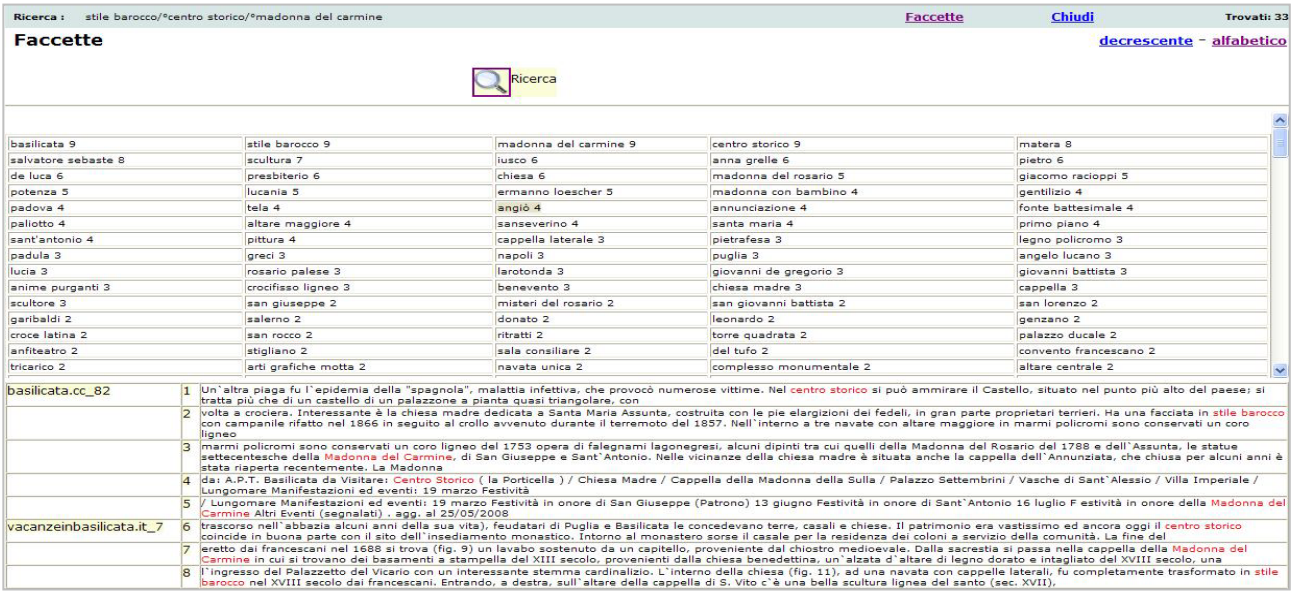


Figure 1: DBT Faccette search example: facets and related contexts

8. Future work

If we consider which countries are more attractive as a tourist destination, we see Italy is among the first places in the world. The factors that contribute to its popularity are mainly art, culture and lifestyle. Other sectors are losing effectiveness: history and gastronomy. Since the web is by far the most important channel for the collection of tourist information, to develop of tools to promote our cultural heritage become more and more important. In this context, is very important developing of new methodologies and tools to overcome traditional categorization systems and their rigidity. On the contrary a set of horizontal open and adaptive classes ("facets"), it can guide the end user in refining his/her search. The aim is to structure a knowledge system of domain-specific information, which could automatically suggest possible cultural routes for tourist purposes. Such knowledge systems can also provide valuable support for applications e.g. for mobile devices, for the realisation of. geo-referenced tourist guides.

9. References

Picchi, E. (1994). Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian. In Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), *Proceedings of Euralex '94*, Amsterdam, The Netherlands.

Picchi, E., Ceccotti, M. L., Cucurullo, S., Sassi, M., Sassolini, E. (2004). Linguistic Miner: an Italian Linguistic Knowledge System. In *Proceedings of LREC 2004*, Volume V, ELRA, Paris, France, pp. 1811--1814.

Picchi, E., Cucurullo, S., Sassolini, E., Bertagna, F. (2008). Mining the News with Semantic Press. In *Proceedings of LangTech 2008*, Roma, Italy, pp.141--144.

Picchi, E., Sassolini, E.. (2009). "Text Power": tools for Cultural Heritage. In *Proceedings of in 4th International Congress on "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin"*, IMC (CNR) Rome, Italy, pp. 277--278.

Granieri, G., Perri, G. (2009) Linguaggi digitali per il turismo. Edizioni Apogeo, November 2006.

Avancini, H., Lavelli, A., Sebastiani, F., Zanoli, R. (2006). Automatic Expansion of Domain-Specific Lexicons by Term Categorization. In *Proceedings* of ACM Transactions on Speech and Language Processing, Vol 3 N.1, New York, NY USA, pp 1--30.