MultiMatch: Multiple Access to Cultural Heritage

Giuseppe Amato, Franca Debole, Carol Peters, Pasquale Savino

Institute of Information Science and Technologies - CNR
Pisa, Italy
{franca.debole}@isti.cnr.it

Abstract. Our shared cultural heritage (CH) is an essential part of our European identity, transcending cultural and language barriers. The aim of the MultiMatch project is to enable users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of large and complex cultural heritage information. MultiMatch aims at heterogeneous digital object retrieval and presentation.

1 Introduction

Online Cultural Heritage (CH) content is being produced in many countries by organisations such as national libraries, museums, galleries and audiovisual archives. Additionally, there are increasing amounts of CH relevant content available more generally on the World Wide Web. While some of this material concerns national or regional content only of local interest, much material relates to items involving multiple nations and languages, for example concerning events in Europe or Asia. In order to gain a full understanding of such events, including details contained in different collections and exploring different cultural perspectives requires effective multilingual search technologies. The EU FP6 MultiMatch project is concerned with information access for multimedia and multilingual content for a range of European languages. MultiMatch tries to offer "complex object retrieval" through a combination of focused crawling, and semantic enrichment that exploits the vast amounts of metadata available in the cultural heritage domain. The MultiMatch search engine was developed with specialised search facilities for multilingual access to cultural heritage material in diverse media. The aim was to present the user with the "detailed picture" of complex CH objects to a very large part of the people interested on CH. The overall goal of the project was to build a fully operational system prototype, designed and refined according to the requirements of numerous and different user classes. Users can search information using their preferred language, searching for all types of digital objects, accessing only the sites that contain information potentially relevant to their request, retrieving mainly relevant items, and viewing the

query results in an organized structured fashion. Standard and ontology-based descriptions of content are used, thus providing an interoperable semantic framework for intelligent multimedia object delivery. Metadata automatically mapped from the original CH material and enriched with supplementary information was ingested onto this framework. At the moment the system has been demonstrated for the main languages of the cultural heritage institutions in the consortium: Dutch, Italian, Spanish, English and also German and Polish, but it is extendible to other languages.

2 Motivation

Europe's vast collections of unique and exciting cultural content are an important asset of our society. On the web, cultural heritage (CH) content is everywhere, in traditional environments such as libraries, museums, galleries and audiovisual archives, but also reviews in popular magazines and newspapers, in multiple languages and multiple media. CH objects on the web are no longer isolated objects, but situated, richly connected entities, equipped with very heterogeneous metadata, and with information from a broad spectrum of sources, some with authoritative views and some with highly personal views. The aim of the MultiMatch project is enabling users to explore and interact with online internet-accessible CH content, across media types and language boundaries, in ways that do justice to the multitude of existing perspectives. This has been achieved through the development of a search engine targeted for the access, organisation and personalized presentation of cultural heritage information. The main source of information stored in the MultiMatch prototype system is composed of cultural heritage objects obtained through crawling and indexing of material obtained from cultural heritage sites, web encyclopedias (e.g. Wikipedia), digital libraries of specific cultural heritage organizations, OAI compliant digital resources, and RSS feeds from cultural web sites. The cultural heritage search and navigation facilities envisaged by MultiMatch cater for these information needs by presenting users with a composite picture of complex CH objects. For instance, in reply to a user's request for information on Van Gogh, the MultiMatch engine can present information on Van Gogh from multiple museums around Europe, in multiple languages; it could complement this with pointers to Van Gogh's contemporaries, with links to exhibitions on Van Gogh, to reviews of these exhibitions, to blog entries by visitors to these exhibitions, and to background information taken from online resources or dedicated sites. The MultiMatch search engine has been developed with specialised search facilities for multilingual access to cultural heritage material in diverse media.

3 The System

The concepts underlying the system are shown in Figure 1. On the left-hand side of the figure, we show users querying the system in different languages for a range of information on the Dutch artist Vincent van Gogh, including critical

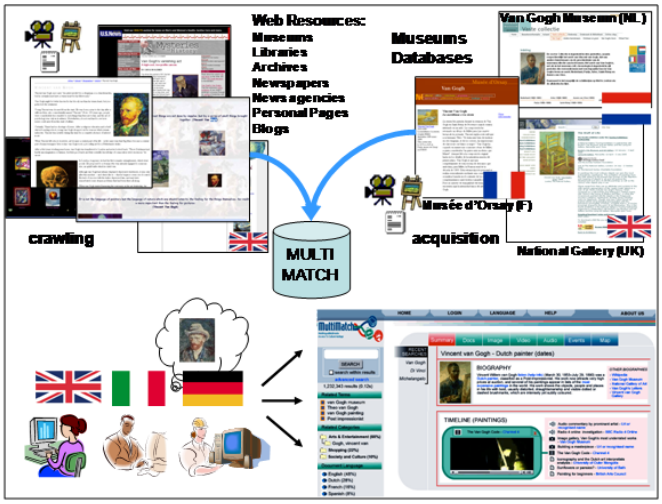

Fig. 1. Overview of the MultiMatch Integrated System.

analysis, biographies, details of exhibitions. The system displays the information retrieved in an integrated fashion, and in a format determined by the particular user profile. On the right-hand side, we show possible sources of this information and the ways in which it can be acquired. The project developed a system prototype that supports six languages: Dutch, English, Italian, German, Polish, and Spanish, and extendible to others. The MultiMatch search engine is able to:

- identify relevant material via an in-depth crawling of selected CH institutions, accepting and processing any semantic web encoding of the information retrieved;
- crawl the Internet to identify websites with CH information, locating relevant texts, images and videos, regardless of the source and target languages used to write the query and/or describe the results;
- automatically classify the results on the basis of a document's content, its metadata, its context, and on the occurrence of relevant CH concepts;
- automatically extract relevant information which will then be used to create cross-links between related material, such as the biography of an artist, exhibitions of his/her work, critical analysis, etc.;
- organise and further analyse the material crawled to serve focused queries generated from information needs formulated by the user;
- interact with the user to obtain a more specific definition of initial information requirements;
- the search results are organized in an integrated, user-friendly manner, allowing users to access and exploit the information retrieved regardless of language barriers.

The achievement of these objectives implied a significant research effort in the general field of multilingual/multimedia information access and retrieval, and in particular in the following areas:

Focussed search engine In its simplest form, a vertical search engine, i.e., one that caters for domain-specific searches, simply filters a subset of the web believed to be relevant to a topic. In a more useful form, a vertical search engine is able to extract information from web pages, allowing for more sophisticated query interfaces and presentation of results adapted to the task. MultiMatch aimed at taking a significant leap forward from today's vertical search engines, by offering "complex object retrieval" through a combination of focused crawling and semantic enrichment that exploits the vast amounts of metadata available in the cultural heritage domain, presenting both certified and non-certified information together (while clearly distinguishing one from the other). The MultiMatch project fits into the category of advanced, domain-specific search engines, with some salient features: i) it is a search engine that combines automatic classification and extraction techniques with semantic web compliant encoding standards; ii) it considers complex user profiles and search scenarios; iii) it is able to search across language boundaries and across different media.

Multilingual/multimedia indexing Instead of returning documents in isolation, MultiMatch provides complex search results that put documents of various media types into context. For the indexing-end of MultiMatch, complex object retrieval generates special challenges. First, documents of various media types (text, audio, image, video, or mixed-content) and accompanying metadata are indexed. Existing generic standards such as MPEG-7 cater for such data models by incorporating multimedia content and metadata in a single semi-structured document. The indexing strategies used also recognise and cater for multilingual content. Particular Cultural Heritage relevant Internet domains or subdomains were spidered using a state-of-the-art crawler [1] [2] and, in parallel, where supported by the CH institution, the engine interfaces with information sources using open standards [3]. CH information was also gathered from the Web at large, employing existing focused crawling techniques specifically targeting cultural heritage information [4],[5].

Information extraction and classification MultiMatch allows users to interpret the wealth of CH information by presenting objects not as isolated individual items, but as situated, richly connected entities. A range of classifications, as well as various links to reviews, experience reports, and general background knowledge, are provided. Documents are classified on the basis of diverse dimensions, such as topical, geographical, and temporal. MultiMatch uses large scale information extraction from documents to identify entities and their relations in large Web corpora [6] [7].

Multilingual/multimedia information retrieval For many years information retrieval research concentrated primarily on English language text documents.

However, recent years have seen a significant increase in research activity extension to information retrieval techniques for multimedia and multilingual document collections. Unfortunately, so far, there has been little transfer of research advances to real world applications. MultiMatch aimed at bridging this gap. Multimedia data can be classified according to its constituent media streams: audio, visual and textual. Research in audio retrieval has largely been concentrated in Spoken Document Retrieval (SDR), where the key challenge is accurate automatic content recognition. Research in Visual Information Retrieval (VIR) for images and video data streams has similarly been underway for over 10 years. Problems of VIR relate to both recognition of visual content and the definition of visual content for IR. MultiLingual Information Retrieval (MLIR) has also become an established area of research in recent years. MLIR focuses on the problem of using a request in one language to retrieve documents from a collection in multiple different languages. MultiMatch developed components for both document and query translation and procedures for matching one against the other. Much effort was dedicated to the building of domain-specific multilingual resources catering for the terminology adopted in the CH domain [8].

User-centred interaction Although there has been huge progress, content-based information retrieval (e.g. video and image retrieval by visual content) still faces significant barriers when attempting to create truly effective and comprehensive retrieval with respect to the user's needs. Users look mainly for concepts (e.g. individuals, facts, places) and far less for features (e.g. mountains, sunset, clouds). A "semantic gap" exists: human beings intrinsically interpret images depending on a subjective viewpoint while computers remain at the most objective and elementary level. To bridge the semantic gap, human intervention is still needed to add high-level features (i.e. metadata) [9]. However, recent advances in the areas of information retrieval and information extraction make it possible to automatically associate concepts to objects when text is available. The MultiMatch user interface integrates automatic techniques for low level feature extraction and automatic concept classification. Structures for browsing are created, allowing users to explore content or search results following multiple facets. A key research problem for MultiMatch was enabling the user to adequately formulate their query using the language of their choice and specify both low-level and high-level multimedia feature [10].

4 Data and Evaluation

The MultiMatch prototype included a wealth of documents derived from multiple sources, annotated by using different formats and schemas, and composed of different types of data, and different languages. During the project lifetime we selected, analysed, harmonized and mapped partners collections and external cultural heritage data. The following Figure (2) describes some of the data ingested on the MM engine from a quantitative (and global) point of view. The prototype system was evaluated in order to measure its quality with respect to

Origin	Sources	Totals (PT1+PT2)	Languages estimations (%)
Dynamic resources collected from feeds (text-based feeds and feed items)	WWW (UvA)	> 25,000	Not specified
Constrained Focused Crawl (Web pages)	WWW (UvA)	>130,000	ITA (~3,5%), SPA (~10,6%), GER (~3,8%), ENG (23%), POL(~11,3%), NDL(~16%) Wikipedia (all languages) (~32%)
Metadata - Web	BVCM	>2,600	SPA (100%)
Metadata - Text	BVCM National Library Austrian National Library OAI (Alinari), MICHAEL, Alinari, BandG	>280,000	SPA (<0,05%), ENG(>70%), GER (<0,2) ITA(<0,03%) POL(<13%) NED(<0,09%)
Still images	Alinari, AISA (Alinari)	> 12,000	ITA (~83%), SPA (~41%), GER (~41%), ENG (100%), POL(~41%), NDL(~41%)
Video	BandG, Alinari, TecheRAI (Alinari)	>570	NDL (>90%) ENG (<10%)
Audio podcasts	WWW (UvA)	>100 h	ITA (~14%), SPA (~8,7%), GER (~31%), ENG (22%), NDL(~24%)

Fig. 2. Cumulative contents collected (PT1+PT2) and language distribution. Brackets, in the language column, indicate the percentage of the content available in the specific language with respect to the total data volume available

performance and usability. In particular, we conducted a laboratory-based and a user-centred evaluation based on field trials. The evaluation of the different components developed in the project showed that the technical solutions adopted provided an advance w.r.t. existing systems. Furthermore, in order to get some basic idea of how fast, and how many concurrent users the search service can support, some simple load testing, was performed. The search service and the content cache proxy component were deployed in a Tomcat instance on a desktop test machine (an Intel Core Duo @ 2.66GHz with 2Gb of RAM). The tests, conducted for the most expensive search tasks (image and video search), showed that the prototype is able to efficiently process up to 40 users even with the limited computer resources used, while for more than 100 concurrent searches approximately 10% of requests time out. The Field Trials (FT) were conducted in

three different application settings, simulating the day-by-day use of the system. The selected application areas were:

- Education (20 users), for the use of MultiMatch in the Humanities
- Tourism (9 users), finding cultural events related to locations
- Cultural Heritage (26 users), by evaluating the use of MultiMatch in three professional environments.

Each Field trial (FT) attendee performed a set of tasks, (i.e. sequence of operations) chosen according to the User Group to which the attendee belonged.

Finally, as part of the evaluation, users of WIND Libero portal (with 28 million registered users) have been invited to use the MultiMatch Search engine. Users had to fill a questionnaire and were interviewed. They provided comments regarding both performance problems and also comments regarding the new functionality offered and the potential of the system. Most users felt that MultiMatch compares very well with other tools they are currently using and in many cases offers solutions that are not actually available. Users found the interface intuitive and, although participants had some problems with the use of the language tools, they did see a lot of potential for them and appreciated the functionalities offered. Improving the way multilingual search is supported by the system will be a priority in future research.

5 Conclusion

Multimatch manages a lot of various file types with highly different internal structure and complex semantics, it also generates various types of relations between the resources and it deals with several languages and user needs. The project was completed on October 31st, 2008 and all planned objectives were achieved. The MultiMatch project involved 11 partners: Istituto di Scienza e Tecnologie dell'Informazione - Consiglio Nazionale delle Ricerche (ISTI-CNR), University of Sheffield, Dublin City University, University of Amsterdam, University of Geneva, Universidad Nacional de Educacion a Distancia, Fratelli Alinari Istituto Edizioni Artistiche SpA, Netherland Institute for Sound and Vision, Biblioteca Virtual Miguel de Cervantes, OCLC PICA, WIND Telecomunicazioni SpA. The project website is active (http://www.multimatch.org) and an online demo of the MultiMatch prototype can be accessed by registered users (free registration available).

Acknowledgement
Work partially supported by European Community under the Information Society Technologies (IST) programme of the 6th FP for RTD - project MultiMATCH contract IST- 033104. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

References

1. Heritrix: Heritrix.internet archive's web crawler project (2005) http://crawler.archive.org/.
2. Nutch: Nutch. open source web-search software (2005) http://lucene.apache.org/nutch/.
3. OAI: Open archives initiative (2005) http://www.openarchives.org/.
4. Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery (1999)
5. Chakrabarti, S.: Mining theWeb: Analysis of Hypertext and Semi Structured Data. Morgan Kaufmann (2002)
6. Bruno, E., Moenne-Loccoz, N., Maillet, S.M.: Interactive video retrieval based on multimodal dissimilarity representation. In: 1st Workshop on Machine Learning Techniques for Processing Multimedia Content. (2005)
7. Clough, P.: Extracting metadata for spatially-aware information retrieval on the internet. In: In GIR-05 Workshop at CIKM 2005. (2005)
8. J.-Y. Nie, M. M.Simard, P.S., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: In 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1999)
9. Wactlar, H., Christel, M., Gong, Y., Hauptmann, A.: Lessons learned from the creation and development of a terabyte digital video library (1999)
10. Daniela, P., Stephen, L., Micheline, B., Mark, S.: Which user interaction for cross-language information retrieval? design issues and reflections. J. Am. Soc. Inf. Sci. Technol. **57**(5) (2006) 709–722