Bildähnlichkeitssuche in der LostArt Meta-Suchmaschine

Visual Retrieval for Searching in a LostArt Metasearch Engine System

E. Schallehn I. Schmitt N. Schulz
Otto-von-Guericke-Universität Magdeburg
Institut für Technische und Betriebliche Informationssysteme
Universitätsplatz 2, 39106 Magdeburg
Tel.: 0391 / 67 – 18665, Fax.: 0391 / 67 – 12020
Email: schallehn|schmitt|nschulz@iti.cs.uni-magdeburg.de
Internet: http://wwwiti.cs.uni-magdeburg.de/

Zusammenfassung:
Informationen über in der Zeit von 1933 bis 1945 verloren gegangene Kulturgüter wurden weltweit durch eine Reihe von Institutionen gesammelt und öffentlich zugänglich gemacht. Ungenaue und unvollständige Informationen, die von einer großen Anzahl von Informationssystemen angeboten werden, machen erweiterte Anfragekonzepte notwendig, um einen effizienten Zugriff zu gewährleisten. Im LostArt-Projekt untersuchten wir den Einsatz von Multimediatechniken zur Verbesserung der Anfrageergebnisse, sowie einen integrierten Zugriff auf viele Anbieter relevanter Informationen. Beide Konzepte erwiesen sich als sehr nützlich für eine umfassendere Suche. An dieser Stelle präsentieren wir unseren Ansatz der Kombination beider Techniken.

Abstract:
Information on cultural assets lost during the period from 1933 to 1945 have been gathered and made publically available by a number of institutions worldwide. Due to vague and incomplete information available from a great number of sources advanced query facilities are necessary to provide efficient access. Within the LostArt project we investigated visual retrieval to improve query results as well as mediated access to information systems providing relevant information. Both concepts proved to be very useful for a more comprehensive search. In this paper we describe an approach to combine both techniques.

1. Introduction

During the period from 1933 to 1945 many private persons, museums, libraries etc. lost their cultural assets as a result of persecution by the Nazi regime and World War II. The documentation and publication of these losses has been a concern of various institutions world-wide ever since, aiming at returning the objects to their rightful owners. Besides losses of art objects, there are many cultural objects without exact information about their provenance. Altogether, documented objects are ranging from assets disseized from the Jewish community as well as `war trophies' taken from museums all over Europe, to collections of found objects where the ownership or provenance in the mentioned time period could not be clarified conclusively. The Internet opens new opportunities to publish information about these cultural assets, and a number of projects have been established since the mid '90s to make corresponding databases and collections available to a broader audience. Such projects aim at finding links between loss and found reports and at supporting the search for lost and found art objects.

One of these projects is the Lost Art Internet Database (www.lostart.de), that facilitates the registration of and the search for cultural assets. This information system is a German-wide network of loss and found reports from more than three hundred museums, archives, libraries and private persons. Textual descriptions of the cultural assets enriched by images are stored in the database. The system lostart.de provides support for various search alternatives. A keyword search as well as browsing through the objects via structured lists are based on the textual descriptions.

In the early stages of the project two major problems in this domain became obvious, namely

1. vague and incomplete information and
2. a great number of sources of related or possibly overlapping information.

The poor data quality mainly results from the fact that in the respective time period data acquisition was very often not possible or considered less important, while at the same time and in the long time span afterwards knowledge about the assets got lost. Later on, the acquisition and management of these information was passed on to various institutions worldwide, based for instance on local responsibilities, e.g. location of findings or losses. As the history of a single object often includes many steps, some of them, including the origin or current location, possibly unknown, a person searching for this object will have to contact a number or even all of the institutions.

To address these problems advanced concepts for querying the available information are required. On the one hand, the LostArt system was extended to support similarity search based on visual information of paintings, drawings, and etchings. Furthermore, querying structured textual descriptions from various sources through a metasearch system is made possible by a cooperative effort with other institutions. To provide the comprehensive search facilities required by users in this domain, the approaches have to be combined. The combination of both is a topic of ongoing research, and in this paper we present the concepts applied for local visual retrieval and global mediated access, as well as our approach of combining them.

2. Visual Retrieval and Mediated Data Access

2.1. Visual Retrieval

Without visual retrieval, a user needs for a successful search a textual description of an art object in mind, e.g. the artist name and/or title of the object, as search criteria. Often, due to several reasons, such search criteria are not correct and therefore do not match database objects. In such scenarios the use of visual information, e.g. images taken from paintings, can help to identify art objects. Based on a given query image the system searches for the most similar images in the database. Alternatively, the user can visually describe the content of a searched painting and the systems uses that description as a search criterion. In order to support visual retrieval, for each image in the database and the query image features, such as color, texture, and shape, are extracted by the system automatically [1]. These features describe the image content.

In our prototype for visual retrieval, implemented in the lostart.de system, the user chooses a query image and specifies relevance weights for certain features, as shown in Figure 2.1 (a). The query image can be either an image uploaded from the user's hard disk or an image already stored in the database. Then, the system searches for the most similar images in the database. This is done by applying a similarity function to determine the similarity between the query image and the images stored in the database. The degree of similarity between two images is expressed by a numerical value and calculated using a scoring function considering the relevance weights given by the user [2]. The result list containing the most similar database images according to the query image is then presented to the user, as shown in Figure 2.1 (b). Usually the result list is ranked in decreasing order based on the images' degree of similarity.

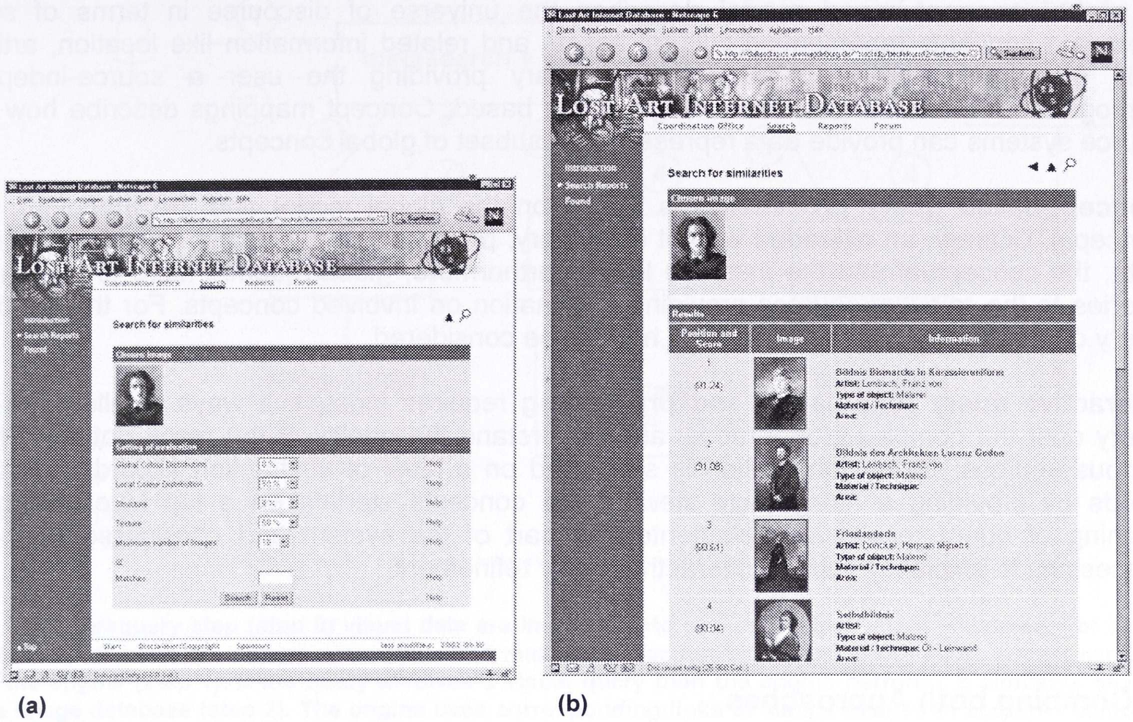

Figure 2.1: Screenshots of the visual retrieval facility in lostart.de. Screenshot (a) shows the query image at the top and the parameter panel for weight adjustment. The retrieved most similar images are displayed in screenshot (b). The result list includes the degree of similarity, thumbnails as well as short descriptions of the similar images.

At this point the user can interactively refine the query by marking images as relevant or non-relevant and initiate a relevance-feedback cycle. Based on the user's judgment the query is reformulated, e.g. weights are adjusted. The systems then retrieves all images that are similar to the refined query [3].

2.2. The LostArt Metasearch

Many different countries and institutions provide publicly available information systems to allow a search for lost or found art objects. Unfortunately, the data collections of such different systems have been developed independently from each other. As result, data about art objects is often stored heterogeneously in more than one system. For completeness, a search has to be performed on every single system. Such a scattered search, however, appears to be impractical to users. A promising approach is to establish a centralized point of access for searching on lost cultural assets. The idea is to start a search at one access point, to propagate the query automatically to several systems, and then to present the collected search results to the user.

In a pilot project at the University of Magdeburg a prototype for such a metasearch engine was developed, providing access to publicly available databases from the Czech Republic, the Netherlands and Germany. The implemented system is based on a common mediator architecture [4] using XML for data transfer, RDFS for representing a global conceptual model of the application domain and XML Web-Services to link to a source system or a wrapper to a system. This way it is open for a non-cooperative as well as a cooperative proceeding on the integration, the latter granting a better stability and accuracy of the solution [5]. More detailed, the following issues are addressed.

A global concept-based model describes the universe of discourse in terms of semantic metadata concepts representing cultural assets and related information like location, artists etc. This model is used as a shared vocabulary providing the user a source-independent, homogeneous view, on which the integration is based. Concept mappings describe how various source systems can provide data representing a subset of global concepts.

Concept-based query processing is based on the global model and the mappings to local concepts. CQuery, an extended variant of XQuery, provides the necessary operations working on both, the conceptual and the instance level. Furthermore, global queries have to be mapped to queries to the relevant sources providing information on involved concepts. For this purpose the query capabilities of the source system have to be considered.

Interactive query formulation and processing requires innovative ways to allow the user to query over the complex global model and understand the quality of the result gathered from the various sources. Query formulation is supported on a level of abstraction according to the user needs by providing a navigatable view of the concepts starting on a top level. Furthermore, caching of query results is implemented as part of the system and considered during query processing to efficiently support interactive query refinement.

3. Combing both Approaches

Our approach is a trade-off between two contradictory goals. On one hand, the institutions behind the existing search systems want to keep autonomy of their systems including their data. Therefore, they would not import their data into one centralized search system. On the other hand, many different autonomous systems create redundancy, heterogeneity and inconsistencies and burdens the user with these problems. Therefore, the goal would be to integrate all the data into one centralized system.

As trade-off, we provide a virtual integration described above. In this way, the burden of querying many different heterogeneous systems is moved from the user to a metasearch system while respecting autonomy aspects. However, there is still the problem of redundancies and inconsistencies among the data. A prerequisite to overcome these problems is the ability to detect duplicate entries. Therefore, we propose to integrate a visual search into the metasearch system. A centralized image database stores the images of all lost or found paintings, drawings and etchings. Obviously, this data must be copied from the different search systems. These images are equipped with information about their origins, i.e. which search system they come from. Following these links further information can be fetched from the different search systems. As benefit for giving the images to the metasearch system the institutions profit from an additional access point to their data and systems, and from information from the metasearch system about found duplicate entries and inconsistencies.

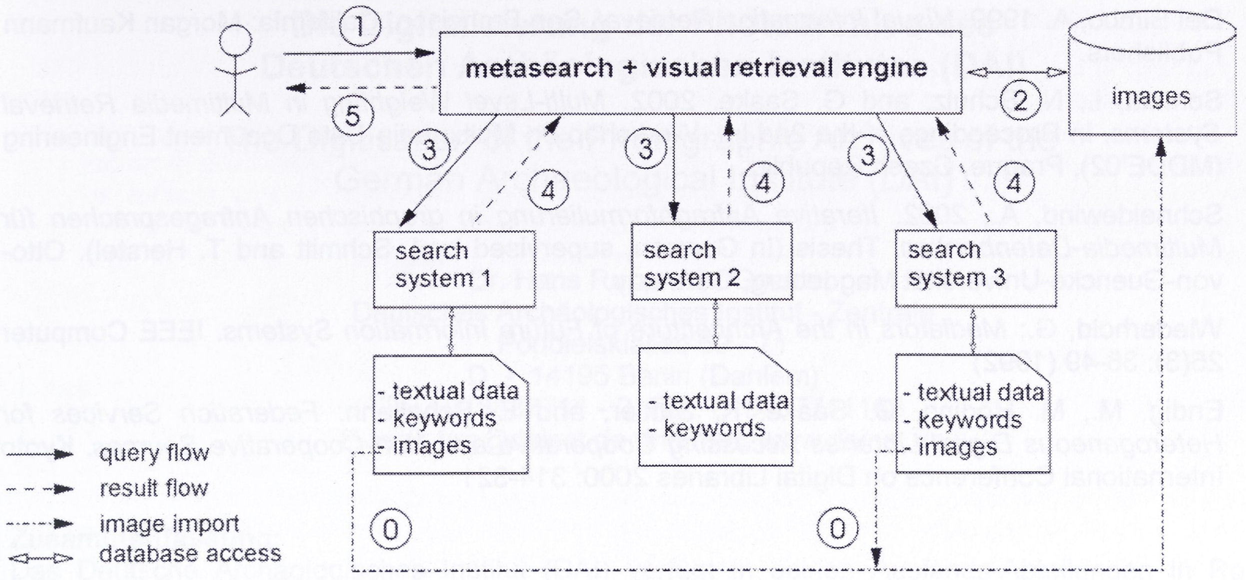

Figure 3.1: In a prequery step (step 0) visual data are imported into one centralized image database. For every image a link to the corresponding source system is maintained for further textual information. A user sends a query to the engine (step 1). If the query involves a visual query then the engine performs a similarity search using the image database (step 2). The engine uses corresponding links of similar images to acquire additional textual information from the sources. Furthermore, textual or keyword queries are propagated to the respective systems (step 3) and their query results are gathered and analyzed (step 4). As the last step, the engine returns the overall result to the user.

The advantages of combining a metasearch system with visual retrieval can be summarized as follows:

1. Improved facility to detect duplicates: Duplication detection on textual data often suffers from misspellings, synonyms, and missing or wrong data. Computing visual similarities between images however improves the chance to detect duplicates.
2. The metasearch engine supports both traditional queries and visual queries and gives, henceforth, the user more flexibility to express queries. A visual query is performed on all images.

4. Conclusions

In this work we presented a new approach to access search systems storing and managing information on cultural assets. Our approach combines a metasearch engine together with a visual retrieval engine. It offers a comfortable access while respecting autonomy aspects of the underlying, independently developed search systems.

At the current stage the basic components of the metasearch system are implemented. Furthermore, a visual retrieval engine is already running. The next steps are to link them together and to connect them to the existing search systems. We hope, that the different institutions are willing to provide us with necessary access information and visual data in order to run first tests.

Due to space restrictions, the presented idea is very roughly described. There are many further issues to be discussed. For example, which concrete system architecture should be applied and how information about duplicate entries can be managed and reported to institutions.

1. Del Bimbo, A. 1999. *Visual Information Retrieval.* San Francisco, California: Morgan Kaufmann Publishers.

2. Schmitt, I., N. Schulz, and G. Saake. 2002. *Multi-Level Weighting in Multimedia Retrieval Systems.* In Proceedings of the 2nd Int. Workshop on Multimedia Data Document Engineering (MDDE'02), Prague, Czech Republic.

3. Schneidewind, A., 2002. *Iterative Anfrageformulierung in graphischen Anfragesprachen für Multimedia-Datenbanken.* Thesis (In German, supervised by I. Schmitt and T. Herstel). Otto-von-Guericke-Universität Magdeburg, Germany.

4. Wiederhold, G.: *Mediators in the Architecture of Future Information Systems*. IEEE Computer 25(3): 38-49 (1992)

5. Endig, M., M. Höding, G. Saake, K. Sattler, and E. Schallehn*: Federation Services for Heterogeneous Digital Libraries Accessing Cooperative and Non-Cooperative Sources.* Kyoto International Conference on Digital Libraries 2000: 314-321