Text Mining and Multimedia Search in a Large Content Repository

Gerhard Paaß, Stefan Eickeler, and Stefan Wrobel

Fraunhofer Institute Intelligent Analysis and Information Systems
St. Augustin, Germany

Abstract. Methods of acquiring, seeking and processing knowledge are a strategically vital issue in the context of globalized competition. One of the main subjects currently being researched is the development of semantic technologies that are capable of recognizing and classifying the content and meaning of information (words, pictures or sounds). In the context of the joint project CONTENTUS we show how different text mining techniques in a workflow are able to extract useful semantic information from text. In a comprehensive multimedia search engine these annotations together with text, metadata, and semantic clues extracted from multimedia documents (speech, music, video) may be used to give more focused access to information.

1 Introduction

Information and communication technologies (ICT) now contribute more to value creation in Germany than the classic technologies of automotive and mechanical engineering [Tec09]. Methods of acquiring, seeking and processing knowledge are a strategically vital issue in the context of globalised competition. The German THESEUS[1] research program [The09] aims at faster and more effective online knowledge processing in future. One of the main subjects being researched is the development of semantic technologies that are capable of recognizing and classifying the content and meaning of information (words, pictures or sounds). These technologies allow 'smart' computer programs to recognize and replicate the context in which data has been stored. In addition, computers will be trained to draw logical conclusions from content by applying rules and classification principles, and subsequently recognize and construct links between various items of information from diverse sources.

Within THESEUS researchers from the public and industry sectors collaborate to develop and design innovative basic technologies and technical standards. Developed prototypes of the new technologies are tested in six application scenarios. The purpose of the tests is to find short-term ways of converting the new technologies into innovative tools, commercially-viable services and potentially profitable business models for the World Wide Web and other internet-based networks.

[1] funded by the German Federal Ministry of Economy and Technology (BMWi)

At the current time, 30 research institutions, universities, and companies have joined the THESEUS program. Among them are the Fraunhofer Society, the German National Library, empolis, SAP, and Siemens. The program has a duration of five years and will last until 2012.

Fraunhofer IAIS is engaged in the Core Technology Cluster to develop advanced technologies and two application scenarios: CONTENTUS and ORDO. CONTENTUS [Con09] is devoted to the preservation of cultural heritage and the safeguarding of cultural diversity in Europe. It aims at digitizing text and multimedia collections, annotate them semantically and make them freely accessible to a wide audience online. ORDO [Ord09] is intended to research and develop semantic technology, to create software tools that will enable users to organize their entire store of digital information. This personalized linking allows unstructured and structured data to be organized in a uniform manner, making efficient, individual knowledge management possible, especially in the economic framework.

In this paper we describe in the framework of CONTENTUS, how text mining and semantic technologies can be orchestrated to allow problem-oriented access to text and multimedia. In the next section we describe the CONTENTUS use case in more detail. After outlining the target multimedia collection we characterize the annotation workflow for text documents where techniques developed in the Core Technology Cluster are applied. The following section gives the details for the comprehensive multimedia search platform, which integrates text, speech, video and music. We close with a summary and outlook.

2 CONTENTUS

The German National Library (DNB) - as the central archival library and national bibliographic center for the Federal Republic of Germany - collects, permanently archives, comprehensively documents and records bibliographically without gap all German and German-language publications from 1913 on and makes them available to the public. DNB does not only archive printed content but also covers multimedia content. Its subsidiary Deutsches Musikarchiv (DMA) collects every audio recording published since the 1950ies and has an extensive collection of even older material and recordings. The raw media in this vast collection – however well it may be maintained – is subject to deterioration and in danger of becoming unusable. Paper erodes, colors fade, even modern media like CDs are prone to a slow decay due to chemical breakup of the information layers. Hence there is an urgent need to conserve the contents.

A common problem of these collections is the limited possibility of searching and finding the desired media as the available metadata typically contains only the most basic attributes like author/artist and title, but no references to the content itself, or to related media. A solution to this dilemma is the digitization of the content to index it and make it accessible by the computer.

Fig. 1. Recordings and Photos from the MIZ Collection

2.1 The MIZ Collection

To demonstrate the workflow of digitization, restoration, semantic indexing and the provision of the content to users a specific collection of the DMA was selected, which urgently requires restoration. This is the archive of the former "Musikinformationszentrum des Verbandes der Komponisten und Musikwissenschaftler der DDR" (MIZ), the center of the association of composers and musicologists of the GDR, which contains material from 1945 to 1990 [MIZ09]. It contains different types of media:

- About 9200 audio tapes with music recordings, 200 audio tapes with speech recordings, and 6800 disc records (cf. figure 1).
- 850 pieces of sheet music.
- About 2500 books and brochures about GDR music together with a collection of 19500 programs of performances (cf. figure 2).
- 143000 newspaper stories about GDR music.
- 5000 still images.
- 2200 file cards of members of the association of composers and musicologists of the GDR.

For most items there is an associated file card archive. As the photos show the material is partly in a bad shape and urgently requires restoration. This collection contains an enormous treasure of information on the musical trends in the GDR and the relation of the musicians among each other, to the GDR officials, to the communists in the Socialist Unity Party of Germany (SED) as well as to the secret police (Stasi). A better public access to the material over the Internet and the semantic linking between different documents is highly desirable for research and the interested general audience.

Fig. 2. Printed Media from the MIZ Collection

2.2 Digitization and Quality Control

The processing of printed material is done in the following steps:

- **Digitization** of content. This involves the scanning of newspaper articles, books and file cards, as well as the digitization of audio and video recordings.
- **Restoration**. For scanned print material stains, tilted book scans, and lens distortion have to be detected and to be removed automatically.
- **Transcription** by optical character recognition (OCR). Pages of printed material are processed by an OCR module to transform the image into a textual representation.
- **Document structure analysis**. The layout of pages together with transcribed contents from books, journals and newspapers are analyzed with respect to structure to extract articles, headings and titles, sections, images, graphical elements, etc..
- **Metadata extraction**, e.g. authors, titles and publication dates. For printed material this can in part be done automatically using machine learning approaches described below. For other media like audio recordings, images and videos this is usually done manually by supplying the required information.

These steps form a workflow which is typical for the digitization of legacy media. At each step a quality control has to be done and – if necessary – previous steps have to be repeated. Analogous digitization steps are performs in CONTENTUS for audio, speech, image and video content.

Fraunhofer IAIS has extensive experience with the digitization of archives. The process of scanning, OCR, layout detection and article segmentation has, among others, been implemented für the Neue Zürcher Zeitung (NZZ) [EBH05]. It turned out that most steps may be done automatically, even for an archive of 225 years requiring a throughput of 12000 pages every day.

3 Semantic Annotation

Current web search engines are excellent in retrieving documents containing specific strings, but they have large problems if queries are posed like "Geliebte

of Wolf Biermann", i.e. "girl friend of Wolf Biermann". Although it is well known that Wolf Biermann had a long relation to the famous actress Eva-Maria Hagen, this is very difficult to retrieve by a web search. To achieve this we need to identify all women who had a liaison with the person "Wolf Biermann" mentioned in text documents. Semantic annotation aims at identifying specific types of concepts or entities, e.g. persons, men, women, locations, professions, in a document. and subsequently identify relations between these objects. An example would be the relation has_liaison(woman, man).

One approach to add semantics to documents is proposed by the Semantic Web. It assumes richly annotated and explicitly structured web pages based upon ontologies and thesauri. This should enable intelligent query processing and semantic reasoning. A large problem with this approach is that only a tiny fraction of documents has been manually annotated in this way. In addition the diversity and uncertainty of terminologies will make precise assignment of concepts nearly hopeless. We therefore think that statistical methods are required to extract vague concepts, entities and relations from text, e.g. has_liaison(woman, man). These can be combined with knowledge available as ontologies, thesauri, etc. to perform more or less precise reasoning and inference.

In CONTENTUS semantic annotation has been implemented as a workflow leading from low-level to high-level annotations:

- Tokenization, sentence detection, lemmatization and part-of-speech tagging.
- Recognition of named entities, e.g. persons, locations, organizations, etc. This is done by classifiers and Markov probability models and has reach a sufficient level of maturity [SM03]. In addition we extract terminology terms, i.e. multiword phrases.
- Detection of co-references of named entities or personal pronouns to other named entities within a document. Here we use Markov random field models with promising results.
- The detection of higher level ontology concepts for words and phrases. The aim is to assign each content word to its semantic category. Again Markov models yield good results for the WordNet concept hierarchy [PR09].
- The disambiguation of detected name phrases. We use kernel techniques to assign name phrases to the corresponding Wikipedia articles [PMP09], which can be considered as an ontology. This will also be done for other Wikipedia concepts.

By establishing links to Wikipedia we get access to structured knowledge, which already in part has been converted into an RDF database called DBpedia [ABK+07]. Among others it currently contains 213,000 persons, 328,000 places, 57,000 music albums, and 20,000 companies. DBpedia covers many domains, and is automatically updated as Wikipedia changes, and it is truly multilingual. In the CONTENTUS project we will use further structured resources like the catalog of the German National Library.

The extraction of relations between concepts or named entities in text is a difficult task. We currently use graph kernel methods [HR08] to include information on words, phrases there annotations as well as the syntactic structure

represented by parse trees. While for many relations the precision may be quite high the recall currently is a problem because of the diversity in formulation semantic connections.

4 Semantic Multimedia Retrieval

In CONTENTUS semantic retrieval is applied to the following content:

- Text of documents with metadata like title, author, source, etc.
- Images with captions and metadata
- Speech recordings including transcribed text and metadata, e.g. speakers.
- Music recordings with metadata like title, artists, genre.
- Video with transcription text and metadata.

Search is mainly based on text. For multimedia metadata text and transcriptions may be enhanced by the annotations described above. Additional features which are employed are phonetic syllables for the search in speech [KLdJ+08] or musical or image similarity.

Usually a user does not only want to search metadata of documents like title, author, and publication date, but also likes to retrieve facts from inside the document. In addition to customary full-text search the user may explicitly search for named entities like persons, organizations, locations and dates. These annotations are stored in the search engine index in the context of other words and annotations. As it is possible to search in different media types the ranking algorithms simultaneously takes into account all media types and in addition included different types of annotations (cf. [CPVP08]).

As the user often is not aware which entities and concepts are relevant in a context we use a faceted search paradigm [SVH07]. After an initial search the most frequent entities, ontology terms, keywords, etc. are extracted from the search results and presented to the user side by side with retrieved snippets from the ranked documents. In addition categories extracted by document classifiers and clusters created on the fly from the search result are presented with other facets. By adding this information we may to specialize or relax the search.

The content itself is represented in specific "players", e.g. for newspaper pages, video clips or recorded speech. Query terms are highlighted to emphasize relevant positions in the text or multimedia content. By linking the extracted content to resources like Wikipedia we are able to offer the user specific explanations and details for specific concepts and articles, which is available by clicking a term.

There are also plans for experimental development of a new, dual-purpose services platform that will allow the community-based semantic annotation of text and multimedia content.

5 Summary and Outlook

Within the project CONTENTUS we described an automated process chain for the presentation of multimedia knowledge, ranging from the initial digitization

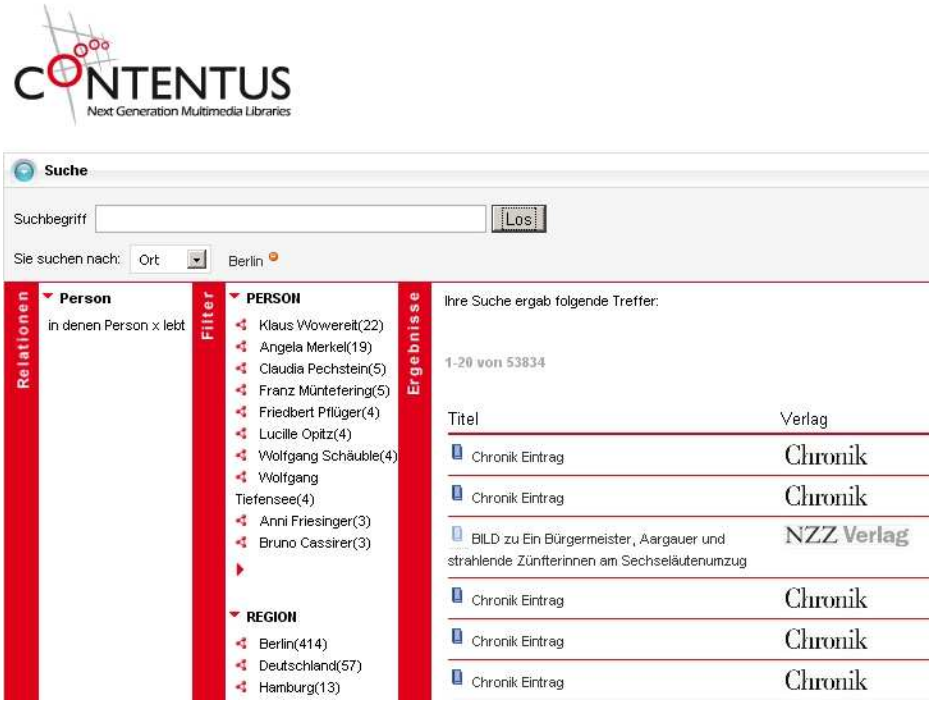

Fig. 3. A first GUI for the CONTENTUS Prototype shown at Frankfurt Book Fair 2008. Search terms are entered in the search field (here: Berlin) and automatically disambiguated (here: location). The ranked search results are shown in the third column, containing text documents, images, etc.. Important related entities and concepts from the search results are listed in the second column.

and automatic quality optimization of multimedia sources, to automated annotation of contents and an comprehensive retrieval engine. A first version of the retrieval engine was demonstrated at the Frankfurt Book Fair (cf. figure 3). Such a portal will provide a much better way to retrieve relevant information from cultural heritage resources like the MIZ collection.

One of the aims of CONTENTUS is to find short-term ways of converting the new technologies into innovative tools, commercially-viable services and potentially profitable business models. The CONTENTUS platform has the potential to be used for more comprehensive document collections than the MIZ collection. We currently use parts of the system to build a revamped portal and search engine for Fraunhofer Gesellschaft with its associated 57 institutes. Within Fraunhofer IAIS these techniques are used for many public and commercially sponsored projects ranging form newspaper archive digitizations to marketing applications and the monitoring of web forums and opinion mining.

6 Acknowledgement

The work presented here was funded by the German Federal Ministry of Economy and Technology (BMWi) under the THESEUS project.

References

[ABK+07] Sren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC 2007*, pages 722–735, 2007.

[Con09] Contentus - safeguarding cultural heritage. http://theseus-programm.de/scenarios/en/contentus.html, Retrieved on Feb. 22, 2009.

[CPVP08] Joan Codina, Emanuele Pianta, Stefanos Vrochidis, and Symeon Papadopoulos. Integration of semantic, metadata and image search engines with a text search engine for patent retrieval. In *Proc. SemSearch 2008*, 2008.

[EBH05] Stefan Eickeler, Lars Bröcker, and Ruth Haener. NZZ: 225 Jahre Old economy vernetzt - Realisierung des digitalen Archivs der Neuen Zürcher Zeitung. In *GI Jahrestagung*, pages 73–77, 2005.

[HR08] Tamas Horvath and Jan Ramon. Efficient frequent connected subgraph mining in graphs of bounded treewidth. In *Proc. ECML/PKDD*, 2008.

[KLdJ+08] Joachim Köhler, Martha Larson, Franciska de Jong, Wessel Kraaij, and Roeland Ordelman. Spoken content retrieval: Searching spontaneous conversational speech. *ACM SIGIR Forum*, 42:66–75, 2008.

[MIZ09] Sammlung MIZ. http://www.d-nb.de/sammlungen/sondersammlungen/miz.htm, Retrieved on Feb. 22, 2009.

[Ord09] Ordo - organising digital information. http://theseus-programm.de/scenarios/en/ordo.html, Retrieved on Feb. 22, 2009.

[PMP09] Anja Pilz, Lukas Molzberger, and Gerhard Paa. Entity resolution by kernel methods. In *Proc. Sabre TMS*, 2009.

[PR09] Gerhard Paaß and Frank Reichartz. Exploiting semantic constraints for estimating supersenses with crfs. In *Proc. SDM 2009*, 2009.

[SM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[SVH07] Osma Suominen, Kim Viljanen, and Eero Hyvnen. User-centric faceted search for semantic portals. In *Proc. ESWC 2007*, 2007.

[Tec09] Federal Ministry of Economics and Technology. Id2010. Germany: building the information society. http://www.bmwi.de/English/Navigation/Technology-policy/the-information-society,did=79428.html, Retrieved on Feb.22, 2009.

[The09] Theseus. http://theseus-programm.de/theseus-basic-technologies.html. 2009.