Twenty-One: a baseline for multilingual multimedia retrieval

Franciska de Jong
University of Twente
Department of Computer Science /CTIT
P.O. Box 217, 7500 AE Enschede, The Netherlands.
E-mail: fdejong@cs.utwente.nl

ABSTRACT

In this paper we will give a short overview of the ideas underpinning the demonstrator developed within the EU-funded project Twenty-One; this system provides for the disclosure of information in a heterogeneous document environment that includes documents of different types and languages. As part of the off-line document processing that has been integrated in the system noun phrases are extracted to build a phrase-based index. They are the starting point for the generation of both a fuzzy phrase index and a translation step that is needed for the realisation of cross-language retrieval functionality.

Keywords: language technology, multimedia information retrieval, cross-language information retrieval

1 INTRODUCTION

In many environments, such as the World Wide Web, full text retrieval tools appear to be attractive for the searching and accessing of unstructured information content. Twenty-One intends to contribute to the need for more powerful approaches to content disclosure in a number of ways. The project aims to develop a demonstrator system supporting the disclosure of information in a heterogeneous document environment that includes documents of different types and languages.

The technology that has been developed within the project and integrated with useful background components was evaluated within two tasks of the international IR evaluation conference TREC-7. Both in the main task and in the cross-language task, the Twenty-One system performed at the level of today's world leading experimental IR systems. Cf. [8].

The project has resulted in the first on-line text retrieval system in Europe supporting cross-language retrieval (accessible since 1996). It has set a baseline for a series of other EU-funded projects, and has led already to some spin-off applications, such as the retrieval engine supporting the web-site of the Dutch national Millenium Platform[1].

Section will 2 will describe the user perspective, and the system design, section 3 analyses the role of natural language in disclosure of multimedia collections, section 4 addresses the relationship with some other projects, and in section 5 an overview of the current functionality of the demonstrator will be given.

2 THE TWENTY-ONE PROJECT

The full name of the project is 'Twenty-One: development of a multimedia dissemination and transaction tool', and the main objective of Twenty-One is to develop domain-independent technology to improve the dissemination level of digitised and non-digitised multimedia information, and to make it more readily and cheaply accessible to a larger group of people. The system can be inspected through the project homepage: `http://twentyone.tpd.tno.nl/` Cf. also [2].

2.1 USER-ORIENTATION

The project focus is on the information need in the field of ecology and sustainable development. The project's user group consists o five environmental organisations involved in the publishing of information in this field. Because of the generic characteristics of the distinct software modules they can also be applied outside the domain of environmental information. Information will be disseminated either via Internet or via a periodically distributed CD-ROM (suited for rapid access to static document bases). The Twenty-One information transaction model, also called the Galilei model, forms an important prerequisite for employing the technology developed within the project. This information transaction model triggers different environmental organisations to exchange information.

The nature of the information to be handled by the system varies enormously, both in format, source,

[1] `Cf. http://www.mp2000.nl/`

and content. A lot of material for which a publication need exists can be characterised as 'grey literature', for which few bibliographic details and no electronic source are available. Considerable emphasis in Twenty-One is put on the development of preprocessing modules for the digital conversion of paper documents. Another example of a domain or application specific focus is the indexing of documents on web sites maintained by organisations outside the Twenty-One user group via a web crawler.

2.2 SYSTEM DESIGN

Though the primary focus may vary with the kind of users for which the technology is put to use, it can in principle deal with information objects in various different media: paper, word processor texts, pictures, video, and audio. Also information in database-format, either from local or from remote sources, falls within its scope. The language elements in the documents to be disclosed are the basis for the automatic generation of a text based index that enables the kind of functionality commonly known as full text retrieval. This provides users access to information not via a controlled set of search terms, but via any word in the document. It also allows users not only to look for entire documents, but also for information within the documents. This functionality is particularly suited for large collections of *unstructured* data. Two crucial sets of software can be distinguished:

- Software to disclose multimedia information
- Software to retrieve multimedia information (with state-of-the-art browsing applications) from remote or local servers, or from a local CD-ROM.

The core of the retrieval software is based upon proprietary software from TNO-TPD and consists of a search kernel supporting several query modes and interface languages. In Figure 1 it is depicted how the various document types are submitted to a three-stage off-line disclosure process.

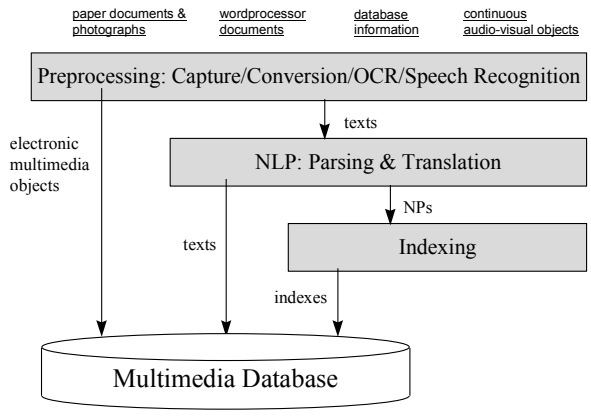

Figure 1 Multimedia disclosure (off-line)

2.2.1 Disclosure

First there is a preprocessing step in which all objects are converted to a format suitable for storage into a database. This includes the isolation of language material, recognition of characters (in the case of paper input) and formatting (SGML/HTML) of language elements. The information objects to be indexed may be stored in multiple representations. For example, a paper document is stored both in HTML-format and as bitmap. In a second step the language elements are submitted to a series of natural language processing modules. This stage includes morphological analysis and part-of-speech tagging[2], parsing (noun phrase extraction)[3] and translation. The parser output consists of a version of the original document in which the noun phrases (NPs) -which are considered to be potential index terms- have been marked. The role of the translation modules is to facilitate cross-language retrieval. This aspect of the functionality is described in detail in [3] and [4]. The third stage is the building of an index on the basis of the output generated by the parser. The results of all stages are stored into one database. A series of modules can be distinguished for each of the disclosure stages. Together the modules can be viewed as a multimedia indexing and retrieval workbench. Each new application may focus on a specific type of document, and can make a different selection out of the modules from the workbench accordingly. Because of this modular design, the system appeared to be a very useful baseline for a series of applications, covering tools for disclosing both static and continuous data types, such as video and audio.

2.2.2 Retrieval

Retrieval relies on language as the medium for indexing and querying and it also exploits language as a means to filter and narrow down in several steps the space of potentially relevant target objects. One of the obvious advantages of this stepwise process is that the downloading of condense data objects such as images and pictures can be postponed until there is confirmed evidence that there is a match with the actual information need.

Searching the index on the basis of a query will support the retrieval of the stored textual representations and (fragments of) the objects linked to the index terms. The automatically acquired text based index is the link between the disclosure and retrieval modules. The index is, unlike most ordinary

[2] The modules for morphological analysis and POS-tagging make use of the Xelda-toolkit developed by Xerox parsing

[3] The parsing modules make use of the NLP-toolkit from TNO-TPD (the Netherlands). Grammars have been developed by DFKI (Germany), University of Twente (the Netherlands) and Xerox (Grenoble, France)

retrieval systems, not limited to an index based on single words or lemmata. In fact it is a combination of several indexes, among which a fuzzy phrase based index, a lemmatized vector space index and a bibliographic index. Using a phrase based index, users are allowed to query the system by using not only simple keywords, but also complete phrases, such as: 'effects of acid rain on forests in the Netherlands'. The matching between query text and index can be done via a one-run fuzzy match that ranks documents on the basis of similarity and number of matching phrases. The incorporation of a vector space index allows a user to improve the initial retrieval results by feeding the most relevant pages back into the retrieval system to get similar documents returned. This mixed approach taken has been proven to yield a considerable improvement in retrieval performance. Recall profits from the morphological analysis (compound splitting) and fuzzy matching, step-wise retrieval with user interaction and relevance feedback improves precision. Cf. also [5, 6, 7].

3 MULTIMEDIA & NATURAL LANGUAGE

Though the focus in Twenty-One is on the disclosure of paper documents, from a more programmatical perspective it is meant to set a framework for a wider range of formats. Ideally indexing and retrieval of multimedia objects should be based on technologies for the automated processing and analysis of information content, for example, image feature extraction. However, though bit-wise and pattern-based recognition of sound, pictures, still images, film sequences etc., already is or may soon become feasible, the state-of-the-art in the relevant technological domains allows only very limited automatic interpretation of the objects involved. (For an overview of problems and approaches, see for example the introductory chapter to [1]). Progress in the development of applications not restricted to specific domains is not to be expected in the short run. More advanced multimedia retrieval could only be achieved if long term research efforts are put in the improvement of content analysis. And even then it will be questionable what the appropriate medium for representing and querying such content could be and whether the human language will not remain *the* access and search medium after all. In any case, as indicated above, some of the needs for multimedia information access can already be solved by applying human language technologies in combination with state-of-the art retrieval technology. Evidently, automated indexing of textual objects is supported by a relatively matured technology and fortunately, natural language is often part of the various media. The disclosure of objects that are not purely or primarily textual can therefore benefit from the advances in indexing based on natural language processing. And as projects such as Pop-Eye and Olive show, Twenty-One has offered a very useful baseline for the proof of concept.

4 RELATION WITH OTHER PROJECTS

In two other EU-projects a similar approach towards indexing, retrieval and translation is applied, but there the focus is on the disclosure of video material and the preprocessing modules needed to capture the language elements: subtitle capturing (Pop-Eye[4]) and speech recognition (Olive[5]). In DRUID, a project carried within the Telematics Institute (Netherlands), the results from Twenty-One will be the starting point for the development of information filtering techniques and for the application of speech recognition in disclosing digital archives in the Netherlands.

5 THE DEMONSTRATOR

The retrieval functionality of the Twenty-One demonstrator can best be described in three steps, that each will be illustrated with a screen shot of a specific part of the interface. We will distinguish:

- querying
- browsing and selection
- intermediate presentation
- presentation of original

We will ignore here the possibility to query the document base with bibliographic keys, but only discuss the so-called 'Normal Query' mode.

As Twenty-One discloses documents and supports querying in four languages (English, German, French, Dutch), the user can select the query language of his preference in the left hand bar of the interface. In Figure 2 the selected query language is German and the query is '*Kompostierung von Haushaltsabfall*'. In the result screen to the right of

[4] Full project name: "Pop-Eye: a multilingual continuous video disclosing tool based on subtitle indexing and partial translation". Pop-Eye is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE1-4234). Duration: 1997-1998. For further information, cf. the project homepage `http://pop-eye.tros.com/`. Cf. also the contribution by Wim van Bruxvoort in this volume.

[5] Full project name: "Olive: a multilingual indexing tool for broadcast material based on speech recognition". Olive is a EU-funded project within the Telematics Application Programme, sector Language Engineering (LE4-8364). Duration: Spring 1998- Spring 2000. For further information, cf. the project homepage `http://twentyone.tpd.tno.nl/~olive/`.

the query bar a table is presenting the documents that contain this query phrase or a phrase that according to some similarity measure is related to it. The table gives the relevant document, its source (which can be either the Twenty-One database containing electronic versions of paper documents, or documents from remote sites that have been marked as relevant to the application domain), the page in the document containing the matching phrase, the matching phrase itself, and an icon indication the original language of the document.

By default the results are ordered on document relevance (Doc-Score). Optionally a user may ask for ordering on the basis of phrase ranking (Phrase Score).

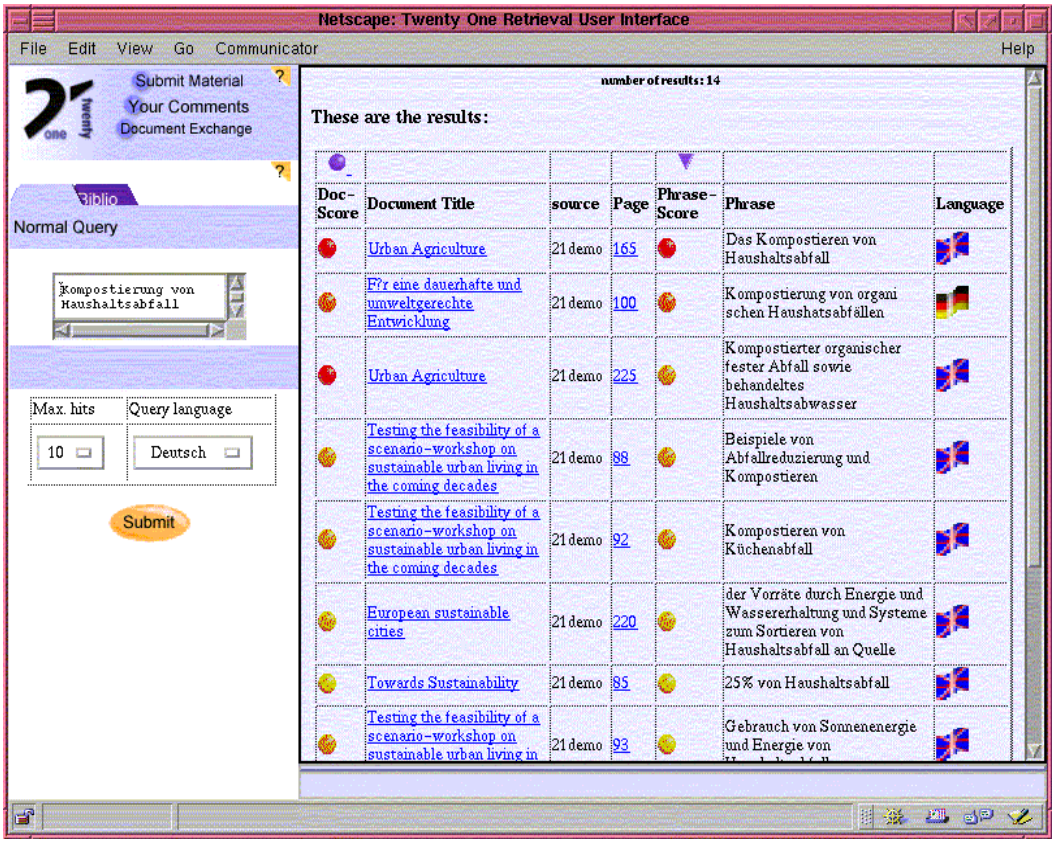

Figure 2, Query screen

The results in Figure 2 illustrate that the system offers the user contextual information on the basis of which he can select the phrase that is most likely to match with his information need.

After clicking the Page cell number for the selected document/page, the screen in Figure 3 pops up. It gives the HTML-version of the document at the point where the selected phrase occurs. This HTML-version can be either in the language in which the document was originally submitted to the disclosure modules, or it can be the result of translation. Figure 3 shows the HTML-version of an originally English document. The original text has been translated off-line into German with LOGOS (commercial MT software) and the translation has been stored in the database, together with the source text. The link with the German query could be established because the German translation has been indexed off-line, in the same way as source language documents.

Via the buttons at the bottom of the screen, the user can ask for bibliographic information, for other pages from the same document, for other documents with similar content (via button 'search similar'), for a version in another language (if available), and (via the button in the lower left corner) the user can also ask for a presentation of the document in its original lay-out. See Figure 4.

The latter option allows the user to view the bitmap of the original page, which is of course always in the original language. This part of the functionality is especially useful in cases where the relevant page contains tables or figures than can not be captured properly by the OCR-module. The initial presentation in HTML-format prevents the unnecessary downloading of irrelevant massive objects. It is one of the aspects of the Twenty-One concept that makes the approach suitable for application in multimedia retrieval.

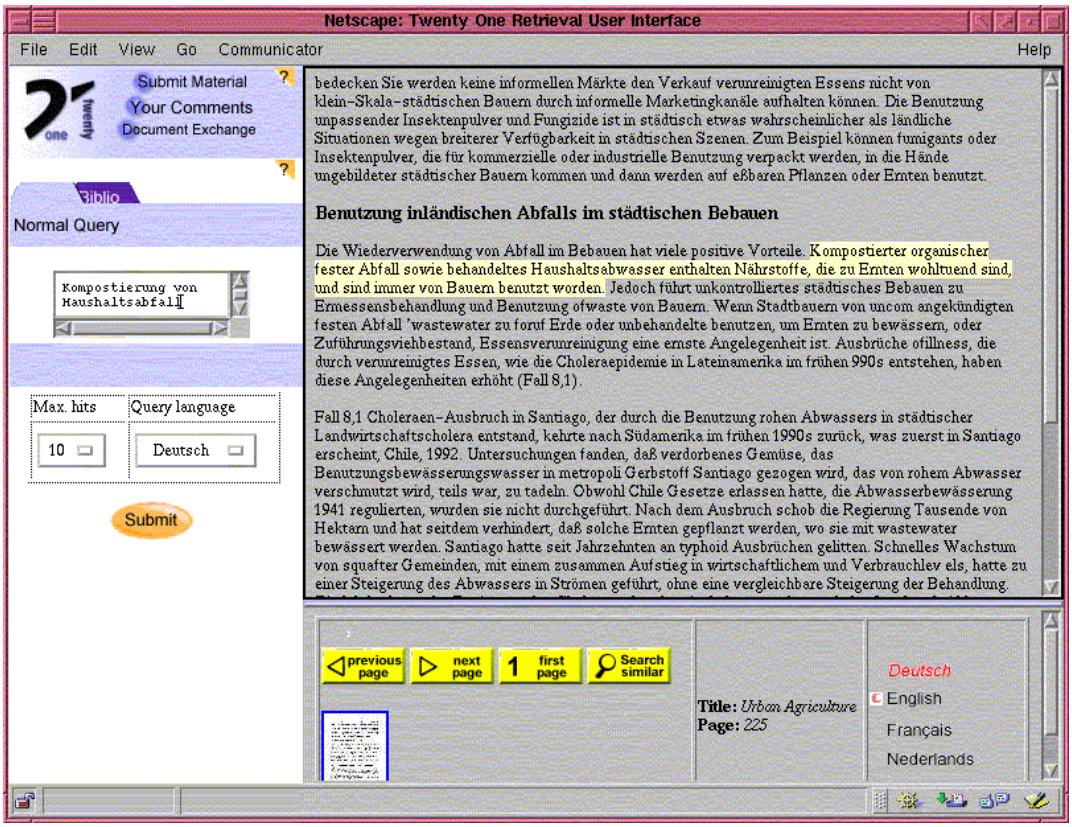

Figure 3, HTML-presentation

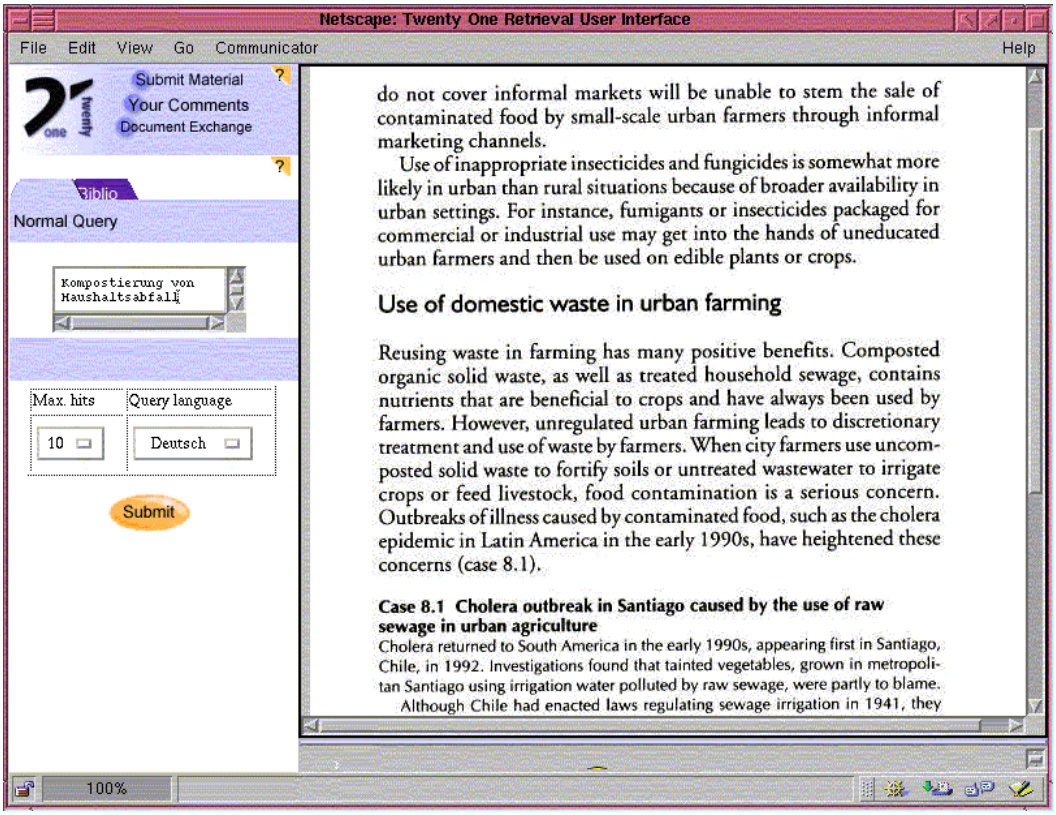

Figure 4, Bitmap presentation

REFERENCES

[1] M. Maybury (ed.), "Intelligent Multimedia Information Retrieval", MIT Press, Cambridge, 1997.

[2] W.G. ter Stal. J-H Beijert, G. de Bruin, J. van Gent, F.M.G. de Jong, W. Kraaij, K. Netter and G. Smart, "Twenty-One: Cross-language disclosure and retrieval of multimedia documents on sustainable development", *Journal of Computer Networks and ISDN Systems* Vol. 30, Elsevier, pp. 1237-1248, 1998.

[3] D. Hiemstra, F.M.G. de Jong and W. Kraaij, "A Domain Specific Lexicon Acquisition Tool for Cross-Language Information Retrieval", *Proceedings of RIAO'97 Montreal*, L. Devroye and C. Chrisment (eds.), pp. 217-232, 1997.

[4] W. Kraaij and D. Hiemstra, "Cross Language Retrieval with the Twenty-One system", *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. In: Ellen Voorhees and Donna Harman (eds.), Proceedings of the sixth Text Retrieval Conference TREC-6, NIST, Special Publication 500-240, pages 753-761, 1998.

[5] W. Kraaij and R. Pohlmann, "Viewing Stemming as Recall Enhancement", *Proceedings of the 19th ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR96), H.P. Frei, D. Harman, P. Schauble and R. Wilkinson eds., Zürich, pp. 40-48, 1996.

[6] R. Pohlmann and W. Kraaij, "The Effect of Syntactic Phrase Indexing on Retrieval Performance for Dutch Texts", *Proceedings of RIAO'97*, L. Devroye and C. Chrisment (eds.), pp. 176-187, 1997.

[7] W. Kraaij and R. Pohlmann, Comparing the Effect of Syntactic vs. Statistical Phrase Indexing Strategies for Dutch. In: Christos Nicolaou and Constantine Stephanidis, editors, Proceedings of the second European Conference on Research and Advanced Technology for Digital Libraries: ECDL'98, Springer-Verlag, pages 605-614, 1998.

[8] Ellen Voorhees and Donna Harman (eds.), Proceedings of the sixth Text Retrieval Conference TREC-7, NIST, Special Publication, to appear.