Modular System Design for Multimedial Information Handling

Botond Pakucs[1] **Björn Gambäck**[1,2] **Preben Hansen**[1]

botte@sics.se gamback@sics.se preben@sics.se

[1] Information and Language Engineering
Swedish Institute of Computer Science
Box 1263, S-164 29 Kista, Sweden

[2] Computational Linguistics
University of Helsinki
P.O. Box 4, SF-00014 Helsinki, Finland

Abstract

Often, information retrieval from various other media is analogous to text-based retrieval; however, accessing documents in e.g. audio or video formats causes some extra problems, in particular with respect to document segmentation, choice of indexing features, and robustness. We review these difficulties, together with some previous attempts to overcome them, and then describe a very flexible, modular IR system which has been designed with a specific eye towards these issues.

1 Introduction

When interacting with information, people use a variety of ways to decide whether to interact with some particular information resource, or some particular information object (or set of objects). The topic of the objects, how those topics are treated, the relationships of the objects to the work task and to the user's goals, knowledge, status, and preferences are different relevance criteria affecting the design of a flexible, multimodal information accessing system (Marchionini, 1995). At the end of the day, there is a need to support users in a wide range of different tasks.

Until recently, most of the supporting tools have been focused on texts; not only has most of the material processed been in text format, but even when other media has been considered in a system, text has been the primary concern. Consequently, most multimedial retrieval systems are modifications of existing text-based IR systems, disregarding the particular problems caused by the new media types. However, the amount of material in other formats increase all the time, increasing the need for tools that handle this information both from a system-oriented and a user-oriented perspective. An important issue for digital libraries is how to represent these objects and collections of objects to best support retrieval. Another issue is to let users search in several modalities. An example would be when a person wants to search a news database: the database contains both news in text format and audio sequences of spoken material. These are stored and indexed in different ways, but the user wants to search both archives at the same time.

In Section 2 we review the difficulties caused by trying to access multimedial documents, in particular audio documents, and some previous attempts to overcome them. We are building a flexible toolkit specifically designed to allow for different approaches to addressing these difficulties and for handling data in different types of media. The toolkit is still under development, however, Section 3 discusses the underlying design philosophy.

2 Problematic Issues

An obvious difference between information retrieval from written documents and from, for example, audio or video documents, is that the written ones actually are physical documents, while the boundaries of the other ones, in contrast, have to decided on inside large files possibly containing quite diverse information. We will start out by looking at the segmentation question and also at some other common problems in non-text-based retrieval, in particular with respect to choice of indexing features and robustness.

2.1 Document Segmentation

Many research projects, such as the ones in the TREC SDR Track (Garofolo *et al.*, 1997), have left document segmentation aside and instead manually tagged and segmented the documents. A straight-forward, automatic approach is to break down, e.g., audio documents into equally long time slices (windows) and treat the windows as individual documents (Smeaton *et al.*, 1998).

The linguistically most correct solution would be to automatically achieve meaningful segmentation of the audio documents through analysis of the prosodic information, as in (Falk, 1997). However, not much work has been done on prosody in speech recognition. So far, the major success story was in the German spoken dialogue machine translation system Verbmobil (Bub *et al.*, 1997).

2.2 Indexing Features

In text-based information retrieval methods, the most obvious and frequently used indexing features are the words. This approach is also the dominating one in spoken document retrieval systems. However, in spite of rapidly improving speech recognition techniques, automatically produced transcriptions contain a multitude of errors and operate with limited dictionaries. For small domains, it is possible to select *a priori* fixed sets of keywords for recognition and indexing. This is, however, not feasible if large and diverse domains are covered, in which case a large-vocabulary word-based speech recognizer can be used together with language understanding system (applied as a filter), as in the CMU Informedia project (Hauptmann, 1995). An additional problem is that out-of-vocabulary (OOV) words tend to be proper names or technical terms which generally have low frequencies in document collections. Thus, it is difficult to train speech recognizers for them, even though these word types are very common in user queries. This has led to the investigation of subword unit approaches.

The drawback with the use of subword indexing features is that elaborated information retrieval and linguistic techniques, such as stop word elimination and morphological analysis, are not suitable. This is specially cumbersome when the spoken information is in a language with more serious morphological properties than English. However, Ng and Zue (1997) showed that with appropriate subword units, it is possible to achieve performance comparable to that of word-based retrieval methods, if the underlying phonetic units are recognized correctly. Both syllable- and phoneme-based indexing features have been tried.

Syllablic units as indexing features was proposed by Glavitsch and Schäuble (1992). The syllablic units can be generated with a rule-based system (Glavitsch *et al.*, 1994). A problem with this is that the syllablic units are not based on acoustic data and do not take the characteristics of the recognition system into consideration.

n-grams, i.e., overlapping, fixed-length, phonetic sequences, are the most straight-forward phoneme-based subword units. With large enough length-units n, cross-word features can be captured. As indicated by, e.g., Ng and Zue (1997), trigrams are normally long enough to capture information sufficient for indexing, while not being so long that they corrupt retrieval performance.

m-grams, non-overlapping, variable-length phonetic sequences with some maximum length m, can be discovered automatically by applying iterative unsupervised learning algorithms. Wechsler (1995) used an efficient algorithm to find variable length phone sequences for indexing. After frequency analysis, low frequency sequences were discarded, while high frequency sequences were extended. In this way an indexing vocabulary covering most parts of the phonetic transcription was obtained.

bclass, broad phonetic class sequences, are derived via unsupervised hierarchical clustering of the original phones. Acoustically similar phones are grouped into the same class, so that some of the typical speech recognition errors can be avoided. Ng and Zue (1997) claimed that even after collapsing the number of phones from 41 down to 20 bclasses, enough information is preserved to perform resonable retrieval: Given perfect recognition, they got a precision of 0.82.

2.3 Robustness

The robustness issues facing multimedial retrieval systems are similar to when text retrieval systems encounter optical character recognition errors in scanned text documents. Thus, even given an ideal speech recognition component, low quality audio data will still effect the speech recognition accuracy and thus the retrieval performance. Accordingly, tolerance of the retrieval models to different disturbing factors such as noisy speech, conversational speech, telephone bandwidth speech, and multiple speakers is important. Quite a few solutions have been proposed to cope with speech recognition errors caused by low quality audio data, however, in the area of robust methods for spoken document retrieval, there is still much left to be done.

In word-based systems, OOV-words can cause significantly weaker retrieval results. THISL (Abberley *et al.*, 1998) applies query-time word-spotting based on posterior probability estimates derived from the recurrent network acoustic model, in order to detect index terms not in the recognizer's vocabulary. Jones *et al.* (1996) used a continuous-speech large vocabulary recognition system in combination with the phone lattice based word-spotting method of James (1995). They showed that the two methods are complementary and work best in combination.

In phoneme-based solutions, one possibility is to expand the query with errorful variants of the original terms, in order to improve the chance of matching wrongly recognized terms. A similar method is to expand the spoken document representation by including high scoring recognition alternatives to increase the retrieval precision. Ng (1998) tried to combine the various types of information captured by the different subword unit representations. He obtained a marginal improvement in retrieval performance by using n-best recognition hypotheses and linear combination of the individual retrieval scores obtained with different subword unit methods.

Crestani and Sanderson (1997) aimed at taking advantage of some particular features of the automatic transcriptions. They used the *Abbot* (Robinson *et al.*, 1996) speaker independent continuous speech recognition system which associates a measure of uncertainty to each word it recognizes. These measures were used in a probabilistic term frequency weighting scheme for improving retrieval efficiency. The results were encouraging, even though the performance was very bad on some particular queries.

2.4 Retrieval result presentation

Presenting multimedial retrieval results to the users introduces some new problems into the IR field. Determining whether a retrieved text document contains an answer to a query is quickly done by simple browsing. However, browsing information objects such as TV or radio news broadcasts can be rather cumbersome due to the sequential nature of the video/audio files.

Having decided on what piece of information to retrieve, we are still far from having decided on *how* to present the information. Most interfaces to text-based retrieval systems simply show the user a list of documents, sorted according to some kind of relevance measure, which often is incomprehensible to the user. A notable exception is the SICS–Telia prototype Easify (Bretan *et al.*, 1998) where search results are presented as matrixes, rather than lists. Easify applies a machine learning method to create collection specific stylistics for genre prediction together with rapid topical clustering in order to represent documents as members of topically and stylistically homogenous clusters.

Present multimedial retrieval systems have no, or fairly simple, graphical user interfaces. VoiceGraph (Slaughter *et al.*, 1998) is a graphical interface developed for audio and multimedia retrieval. The interface was designed to facilitate rapid browsing of spoken documents. Another, web-based, graphical interface was developed for the Taiscéalaí radio broadcast retrieval system (Smeaton *et al.*, 1998). In Taiscéalaí, the retrieved documents are presented as time series of window scores. The users are allowed to select parts of a broadcast. The selections are delivered from the archive and transformed to a RealAudio file which is played to the user.

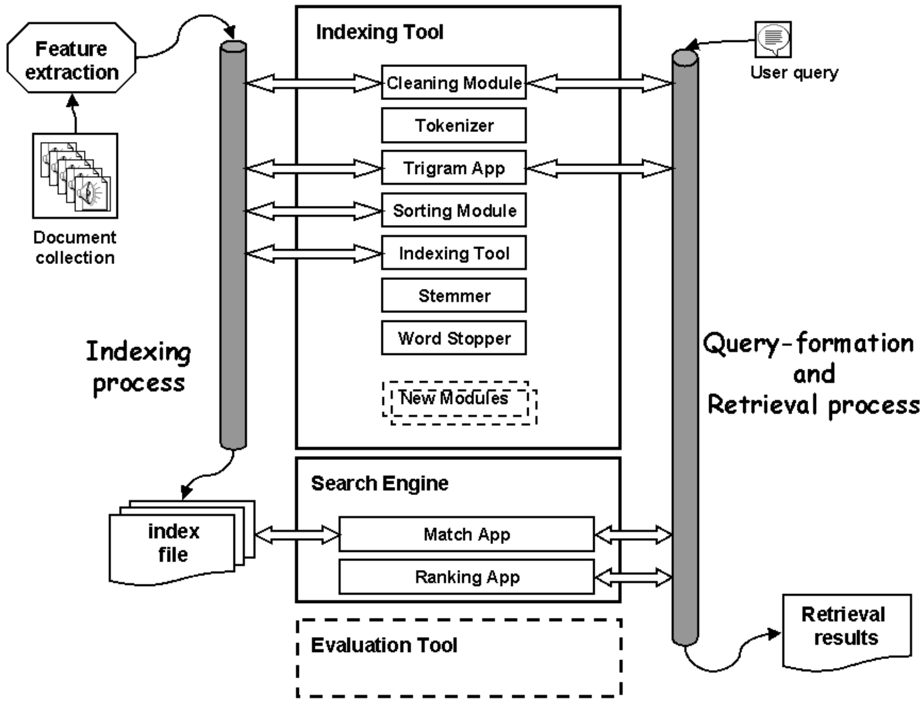

Figure 1: The design of SIREN

3 System Design

Most multimedial retrieval systems are modifications of existing text-based IR systems, disregarding the particular problems caused by the new media types. Instead of using existing platforms, we decided to design a new IR system, in order to manage the problems presented in Section 2. The requirements for a retrieval toolkit can be formulated as:

- A *modular* design is desirable: it should be easy and fast to make small changes to specific parts of the toolkit without reimplementing the whole system.
- The system must be *extensible* and *open*, since we want to be able to experiment with new algorithms and new media types. Thus, adding new modules has to be fairly straight-forward.
- The toolkit has to run be *dynamically* and *flexibly*, allowing for quick adaptations to new requirements at any time by connecting/disconnecting modules.
- Another desirable quality is *transparency*: we might want to examine the data flow under certain conditions.
- Due to the distributed nature of digital libraries, *interoperability* within a DL architecture is necessary. To achieve a high level of access and sharing, we need a general framework for handling information across various domains such as different communities and different types of information objects.
- *Portability* is not necessarily a major requirement, but in itself a desirable quality of software systems.

We chose the Java environment in order to address the portability issue.[1] Java's object-oriented design makes it easy to reuse modules in the query-formation process. When implementing the search engine, we used the built-in `Thread` class features, allowing us the option of adapting the system to the use of parallel search threads for searching multiple index files. This is essential when several media types are searched simultaneously.

[1] Certainly, the choice of programming language is crucial for system performance, and Java is not well-known for it's speed qualities. However, it is still under development, promising better speed properties. Also, modular system design coupled with the Java environment makes reimplementation of vital components in more time-effective languages tractable.

The two major components of our IR toolkit are the indexing tool and the search engine as shown in Figure 1. A third component, an evaluation tool, is planned. The IR solutions and algorithms are based on well-founded results: We use a common inverted-file structure for storing indexing and posting information. The index-term weighting scheme is the *tf∗idf* formula. The search engine implements the vector-space model and the cosine metric matching function. Thus, the novelty of the approach lies not in the algorithms used, but in the way these algorithms are implemented in the system, and in the system design which allows for fast reimplementation and integration of new modules.

The main tools are built up of several small, independent modules, each of them taking care of one well-defined atomic task. The modules are implemented as independent Java classes as shown in Figure 2.

As indicated in Figure 1, the modules are connected together in a pipelined architecture, using the Java built-ins `PipedReader` and `PipedWriter` for input and output, respectively. In this way, pipe syncronization is taken care off automatically. The connections between the modules are dynamic, so that modules can be invoked "on the fly". Another advantage of the piped architecture is that the need of using and saving intermediate files is eliminated. The system is *open*, thus, the pipes are open in both ends and adding new modules with specific preprocessing task such as, for example, prosodic analysis is possible, as is also connecting the pipe directly to a speech recognizer. Appending new modules for relevance feedback or retrieval evaluation is also feasible. Integrating new modules is fairly straight-forward: only new method calls have to be declared.

The modules are communicating through a common token stream, a solution slightly inspired by University of Glasgow's SIRE system (Sanderson and Crestani, 1998). The token stream consists of simple ASCII text. The modules are operating on the stream by modifying the tokens, by adding new tokens, or by removing some of them. Each token holds one term and a number of additional attributes connected to it. However,

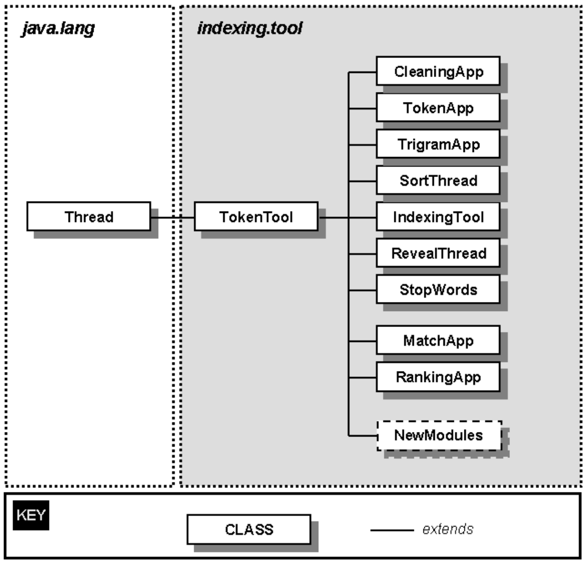

Figure 2: The SIREN class structure

the underlying data representation is hidden from the modules in a object-oriented manner. The modules extend a specially implemented `TokenTool` class which provides the primitives necessary to manipulate the tokens. Thus, the modules are separated from the data representation and rather use general IR concepts, such as index-term, weight, frequencies, etc. This solution allows a more general implementation of the IR modules and operations, so that data manipulation is independent of the addressed media type.

4 Conclusions / Future Work

We have described SIREN, a flexible, modular information retrieval system designed to allow for different approaches to addressing the particular problems which are encountered when attempting to access non-text documents. The system is currently under a performance evaluation and will soon be used as a platform for IR studies in several ongoing projects. The toolkit is already functional and useful for text and spoken document retrieval. New modules with specific tasks will soon be added, such as stylistic analysis and name recognition.

References

Abberley, D., Renals, S., and Cook, G. 1998. Retrieval of Broadcast News Documents with the THISL System. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, Washington. IEEE.

Bretan, I., Dewe, J., Hallberg, A., Karlgren, J., and Wolkert, N. 1998. Web-Specific Genre Visualization. *Proc. 3rd World Conf. on the WWW and Internet*, Orlando, FL. AACE.

Bub, T., Wahlster, W., and Waibel, A. 1997. Verbmobil: The Combination of Deep and Shallow Processing for Spontaneous Speech Translation. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 71–74, München, Germany. IEEE.

Crestani, F. and Sanderson, M. 1997. Retrieval of Spoken Documents: First Experiences. Technical report, Dept. of Computing Science, University of Glasgow, Scotland.

Falk, J. 1997. Pauses in Synthesized Speech: Automatic Prediction of Silent Intervals in Swedish. MA Thesis, Dept. of Linguistics, Göteborg University, Sweden.

Garofolo, J. S., Voorhees, E. M., Stanford, V., and Jones, K. S. 1997. TREC–6 1997 Spoken Document Retrieval Track Overview and Results. *Proc. 6th Text Retrieval Conf.*, pp. 83–91, Gaithersburg, MD. NIST.

Glavitsch, U. and Schäuble, P. 1992. A System for Retrieving Speech Documents. *Proc. 15th Int. Conf. on Research and Development in Information Retrieval*, pp. 168–176, København, Denmark. ACM SIGIR.

Glavitsch, U., Schäuble, P., and Wechsler, M. 1994. Metadata for Integrating Speech Documents in a Text Retrieval System. *SIGMOD Record*, **23**(4):57–63.

Hauptmann, A. G. 1995. Speech Recognition in the Informedia^{TM} Digital Library: Uses and Limitations. *Proc. 7th Int. Conf. on Tools with AI*, Washington, DC. IEEE.

James, D. A. 1995. *The Application of Classical Information Retrieval Techniques to Spoken Documents.* PhD Thesis, Engineering Dept., University of Cambridge, England.

Jones, G. J. F., Foote, J. T., Sparck Jones, K., and Young, S. J. 1996. Retrieving Spoken Documents by Combining Multiple Index Sources. *Proc. 19th Int. Conf. on Research and Development in Information Retrieval*, pp. 30–38, Zürich, Switzerland. ACM SIGIR.

Marchionini, G. 1995. *Information Seeking in Electronic Environments.* Cambridge University Press, Cambridge, England.

Ng, K. 1998. Towards Robust Methods for Spoken Document Retrieval. *Proc. 5th Int. Conf. on Spoken Language Processing*, Sydney, Australia.

Ng, K. and Zue, V. 1997. Subword Unit Representations for Spoken Document Retrieval. *Proc. 5th European Conf. on Speech Communication and Technology*, pp. 1607–1610, Rhodes, Greece.

Robinson, T., Hochberg, M., and Renals, S. 1996. The Use of Recurrent Networks in Continuous Speech Recognition. In Lee, C.-H. and Soong, F. K., editors, *Advanced Topics in Automatic Speech and Speaker Recognition*, Ch. 7. Kluwer, Dordrecht, Holland.

Sanderson, M. and Crestani, F. 1998. Mixing and Merging for Spoken Document Retrieval. *Proc. 2nd European Conf. on Research and Advanced Technology for Digital Libraries*, pp. 397–407, Heraklion, Greece.

Slaughter, L., Oard, D., Warnick, V., Harding, J., and Wilkerson, G. 1998. A Graphical Interface for Speech-Based Retrieval. *Proc. 3rd Digital Library Conf.*, Philadelphia, PA. ACM.

Smeaton, A., Morony, M., Quinn, G., and Scaife, R. 1998. Taiscéalaí: Information Retrieval from an Archive of Spoken Radio News. *Proc. 2nd European Conf. on Research and Advanced Technology for Digital Libraries*, pp. 397–407, Heraklion, Greece.

Wechsler, M. 1995. Eine neue Indexierungsmethode für Information Retrieval auf Audiodokumente. *HIM95 Konferenz: Hypertext–Information Retrieval–Multimedia*, pp. 117–128. (in German).