Retro-conversion of Library Catalogues and Multilingual Information Retrieval

Peter Schäuble, Páraic Sheridan
Swiss Federal Institute of Technology (ETH)
CH-8092 Zürich, Switzerland

Abstract

We present an overview of two areas of ongoing research at ETH, directly related to digital libraries in Europe. We are presently co-operating with the Zentralbibliothek Zürich in the digitisation of their 2.2 million card catalogue. We have developed new techniques to allow effective retrieval on the full texts of the scanned cards, even though OCR is only achieving a word recognition rate of 67%. These retrieval techniques have been integrated into our information retrieval system, SPIDER, and have been shown to increase retrieval performance by 35% in experiments on a sample of digitised cards. We are also actively working in the area of multilingual information retrieval, allowing users to query a system in one language and retrieve documents in other languages. We have adapted proven information retrieval techniques (the use of corpus-based similarity thesauri for query expansion) to the multilingual problem, and we have recently demonstrated their effectiveness, again using the SPIDER retrieval system. In experiments over a collection of more than 90,000 Italian documents we have shown that the SPIDER system can retrieve Italian documents in response to German queries with *better* effectiveness than a baseline system retrieving Italian documents in response to Italian queries.

Digitisation of Library Catalogues.

ETH Zürich has worked in cooperation with the Zentralbibliothek Zürich on a project to digitise the complete central catalogue of the library and to make it available through an electronic text retrieval system. The Alphabetischer Zentralkatalog (AZK) consists of about 2.2 million paper

cards referring to 1.7 million documents. The AZK is an important catalogue because 1.58 million of the libary's documents (93%) are only accessible through this catalogue. The AZK includes entries in six main European languages plus other languages like Russian, Arabic, Turkish and Latin. This aspect of the catalogue presents difficulties for Optical Character Recognition (OCR), since these languages contain many accented characters and also make it impossible to do automatic correction of OCR errors, since the automatic recognition of language for lexicon lookup is infeasible. This is further complicated by the large number of proper nouns and abbreviations in the collection. A typical index card of the AZK contains, on the average, 23 words. An example index card from the catalogue is shown below. In a pilot study of the problem of digitising the catalogue we focused on a sample of 8,500 index cards. The index cards were scanned at 300dpi and processed through a commercial OCR system (OmniPage Professional) to produce OCR text. The OCR process resulted in a word accuracy of 67%, which means that one of every three words in the catalogue sample was incorrectly recognised.

A central objective of our study was to establish the extent to which it was feasible to deliver effective search and retrieval facilities over the digitised catalogue with minimal manual intervention in the digitisation process. We therefore proposed a probabilistic approach to retrieval, taking into account the error probabilities associated with the OCR process.

{ INCORPORER "Word.Picture.6" *
fusionformat **}**

Husin. Social stratification in Kar,-
pong Bagans a study of class, status,
con'ict and
rlobility in a raral Eilay community, X +
170 S.,
Eart., Tab.
(Monographs of the S-P.G .w7 sia-n
BraSnch tof ths
Royal Asiatic Society. 1).
Singapore (1D64),

Z TA 375C : 1

TNG 755 / Klassen (Siz.).: Einz. C3'J. /
Sosialer
Wandel / Johore: 50Z ol,kl 5+h

Digitised Index Card **OCR Version**

Based on a comparison of the OCR output against a sample of 650 cards that were manually corrected, we established probabilities associated with

Search Result	
Direct Hit	78%
1 Card Away	92%
2 Cards Away	98%

typical OCR errors. For example the probability that an *i* is mis-recognised as an *l* is much greater than the probability that an *e* is recognised as an *m*. Our estimation also included probabilities for characters being completely omitted or randomly inserted by the OCR process. Given these probabilities, we can take OCR errors into account and compute the probability that an OCR string represents an occurrence of a term being sought by a user. We can compute expectation values for the frequencies of query features within individual index cards and across the whole catalogue and then use these expectation values in a standard information retrieval model. We have implemented this probabilistic retrieval model in our retrieval system SPIDER.

The present physical catalogue is searched by relying on the alphabetic order of cards. A library user will select a drawer with the appropriate letter and browse backwards and forwards through the cards based on the alphabetic order. In order to duplicate this search facility in the digitised version, we needed to recognise automatically the key word of each index card for the alphabetic order. This relies on the assumption that the catalogue is scanned in its proper order. We used certain rules about the structure of the cards to estimate the key token from the OCR output of each scanned card and then used a maximum liklihood technique to identify the proper alphabetic order of cards in the digitised catalogue. Note that a probabilistic approach is necessary here because of the corruption of alphabetic keys by the OCR process.

We evaluated the effectiveness of searching alphabetically in our digitised catalogue experimentally over a sample of 3,500 index cards. This involved specifying an author name query and searching through the index generated automatically by our maximum liklihood techniques for the desired key. The results show that in 98% of searches the user must flip no more than two cards to land at the position sought for. Our experimental results further demonstrate that we can *always* position a user within 10 cards of the position sought.

An additional feature that our full-text indexing provides that is *not* possible in the current physical catalogue, is the ability to search for words from the title of the document, or in fact from anywhere in the index card. We evaluated this feature of our system with a sample of 64 queries over

the collection of 8,500 sample catalogue cards. We take as a baseline for performance on this task the scenario where searching is performed on the OCR output without the benefit of our probabilisitc model. This baseline results in retrieval performance equivalent to the word recognition rate of the OCR device; 67% in our case. Using the SPIDER retrieval system with the probabilistic model of OCR errors, we can achieve a retrieval rate of 90.62%, and improvement of 35%. The experiment consisted of searching for single-word queries that occured in only one card of the collection. We therefore based our retrieval performance on the percentage of queries for which the proper index card was returned as the highest ranked card of the list of cards returned by our system.

Search Results	
Word-Based	67.00%
Probabilistic	90.62%

In summary, the retroconversion to digital form of the 2.2 million Zentralkatolog of the Zentralbibliothek Zürich is now being undertaken fully automatically (though with manual quality control). The cost of this process represents a saving of 90% of the cost of manually entering the contents of the catalogue. ETH Zürich has developed new probabilistic retrieval tecniques which provide effective keyword or full-text retrieval over the catalogue, despite OCR accuracy of only 67% word recognition. Catalogue browsing is supported with a 98% success rate of positioning the user within 2 cards of the desired author in response to an author keyword search. Further, content based retrieval (e.g. using title words) can be provided in the digitised catalogue, something which is not possible in its current form.

Multilingual Information Retrieval.

Given the multilingual environment in Switzerland, where there are three official languages in everyday use, there is a great deal of interest in multilingual information systems. We define multilingual information retrieval as the process whereby a user can specify a search request to a system in one language and retrieve relevant information in *other* languages. We observe that many users may have sufficient language comprehension to understand, at least at a high level, a text in a given language even though they may not be able to adequately express an information need in that language. In this case, it is more natural for the user to specify the information need in her own native language.

ETH Zürich is at the leading edge of research into multilingual retrieval. The SPIDER retrieval system was initially developed at ETH to work with English texts. This system has now been extended to provide retrieval over French, German and Italian documents and we have recently conducted experiments into multilingual retrieval, submitting German queries and retrieving Italian documents. Our approach to multilingual retrieval is based on the use of *similarity thesauri* constructed across a collection of documents. A *similarity thesaurus* represents, for a given query, those terms in the document collection that are most similar to the query. The use of similarity thesauri has been demonstrated effectively for performing query expansion over a large single-language collection of documents. We have adapted the ideas of using similarity thesauri for query expansion by applying the techniques over collections of comparable documents in multiple languages in order to derive *multilingual similarity thesarui.* By performing query expansion through a multilingual similarity thesaurus and then filtering the most similar terms, we can retrieve a list of terms from the document collection in language *y* which are most similar to the query submitted in language *x*. For example, in our recent experiments we used a German-Italian similarity thesaurus to determine for a German query the most similar Italian terms in the document collection and then used these terms as a query to the SPIDER information retrieval system to retrieve Italian documents. It is important to note that a multilingual similarity thesaurus is not a translation table in the sense of a bilingual dictionary. It does however provide a *translation effect* by, for example, returning Italian terms that are used in mainly the same contexts as submitted German terms. An example of this effect is given below.

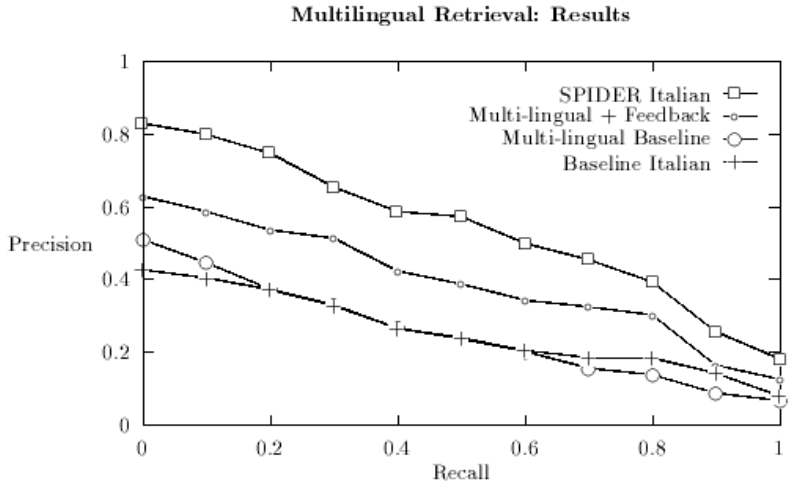
Pub Loughinisland Nordirland getötet

nord loughinisland pub ulster cattolico protestare irlanda irlandese gento terror terrorista provincia strago uvf accadere conservatore diano uccidere famoso bar lontano rivendicare iro trono

√	1.	GB: Ulster, lacrime e terrore rappresaglie dopo strage
√	2.	Gb: ulster, sparatoria in un pub, vittime (3)
√	3.	GB: Ulster, strage in un pub: sei morti

This example illustrates a German query submitted to the system, then the top 25 most similar Italian terms returned by the multilingual similarity thesaurus, and finally the top 3 Italian documents returned by the SPIDER

retrieval system, which are all relevant to the query. The above example is just one out of a test suite of 65 queries that we used in an evaluation of our system. The evaluation of the 65 queries was carried out over a test collection of 93,000 documents of news stories from the Schweizerische Depeschen Agentur (SDA), a Swiss news agency. As a baseline for comparison we used the SPIDER system in its barest form, without the use of our word normalisation techniques for Italian, to evaluate Italian queries against the Italian documents. We then evaluated our multilingual retrieval system by running equivalent German queries through the similarity thesaurus to retrieve Italian documents (including the use of Italian word normalisation). The comparison of German queries versus Italian queries is represented by the lower two curves in the figure below.



Having observed that users may be able to read a document in Italian even though they can not express a query in Italian, we also performed an experiment where the user examined some of the highest ranked documents and made judgements as to their relevance. SPIDER used this relevance information to refine the search and provide a new list of relevant documents. The result of using this relevance feedback technique is represented as the middle curve of the above graph. The relevance feedback loop provides a 60% improvement in average precision over the fully automatic approach to multilingual retrieval and comes within 25% of the best performance we have achieved on Italian retrieval using the full SPIDER system with Italian word reduction.

Our work in multilingual retrieval is continuing. We are currently concentrating on improving our thesaurus construction techniques, improving our German lexical analysis and introducing French lexical analysis into the SPIDER system.