EVALUATING DIFFERENT INFORMATION RETRIEVAL ALGORITHMS ON REAL-WORLD DATA

Manfred Weber
Interactive Systems Labs
Am Fasanengarten 5
76129 Karlsruhe, Germany

Thomas Kemp
SONY International (Europe) GmbH
Stuttgart Advanced Technology Center
Hedelfinger Str. 63
70327 Stuttgart, Germany

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

More and more data is produced in the form of videos, which are opaque to textual queries. To allow searching in video data collections, two problems have to be solved: The automatic generation of a searchable index, and the effective search in the automatically produced and therefore imperfect index. The ISL View4You system is a prototype of a video indexing and retrieval system which both generates the index and provides a search engine to access it. An end to end evaluation was carried out using real-world data and queries from naive subjects. From the results it can be concluded, errors of the overall system are not due to the index generation, but are introduced by the information retrieval engine (the search). Therefore, the focus of this paper is a comparison of two different search algorithms, LSI (latent semantic indexing) and Okapi (a flavor of the traditional classic vector model approach). The evaluation is carried out on the automatically produced index on a relatively small database, which allows for full manual relevance judgement.

1. INTRODUCTION

More and more data is produced in the form of videos, which are opaque to textual queries. To allow searching in video data collections, a searchable index of the data must be created. This sort of processing is generally referred to as *metadata generation*, and can be done by several methods, including image analysis and speech recognition of the audio stream. If the metadata is generated automatically, a variety of errors will usually be introduced due to the imperfections in the automatic generation process. Such errors might include segmentation errors (a story or scene cut boundary has been missed, or accidentally inserted), classification errors (a commercial has been classified as relevant, or relevant material has been classified e.g. as music), and speech recognition transcription errors. Clearly, the information retrieval component of a video library system has to be robust against errors in the underlying metadata.

Previous analysis [2] however has shown, that the majority of errors (not found or wrongly found items) in a real-world scenario is not due to errors in the metadata but due to the information retrieval component of the system: even with perfect, manually constructed metadata, the system output is frequently not satisfying. This result has also been confirmed by other researchers [4].

Following these results, this paper focuses on the information retrieval component of the ISL View4You video indexing and retrieval system.

The paper is organized as follows. First, the ISL View4You video indexing and retrieval system is described. In the following section, two different information retrieval algorithms are introduced. The main part of the paper describes the results of constrastive retrieval experiments using both perfect (hand-made) and automatically generated metadata.

2. THE ISL VIEW4YOU SYSTEM

The ISL View4You system [2] is a fully operational prototype of a video indexing system. German newscasts recorded from public television are used as input. The newscasts are first segmented into topic stories [1]. The audio track of each story is then transcribed using a large vocabulary continuous speech recognition system. User queries can be given in natural language, like e.g. 'Please tell me everything you have about the visit of President Herzog in Japan'. Using the transcriptions, the systems information retrieval component searches through the database and returns the most relevant parts of the video data in a sorted list. The user can select one item from the list, which is then displayed on the screen. Figure 1 shows a screenshot of the system.

Figure 1. User interface

In the last few years, several systems [5] [4] have been introduced which are similar in scope.

3. END TO END EVALUATION

As the View4You system is a fully operational prototype of a video indexing and retrieval system, it is very interesting to run an end to end evaluation to spot the weaknesses of the parts that make up the system. Such an evaluation can serve as a starting point to govern the direction of future research.

For the evaluation, a database was automatically created by the View4You system by processing a set of TV broadcasts. Both steps of the database creation process - segmentation, and speech recognition - were carried out without human intervention.

A set of 10 queries in natural language was defined as follows (the original questions are in German):

1. Are there reports about Jerusalem?
2. Will Helmut Kohl run for chancellor again?
3. I want to know the winning numbers of the Lottery!
4. Is there anything about Benyamin Netanyahu?
5. I am interested in anything that recently happened in Africa!
6. What is the situation in Albany?
7. What are the results of the National Soccer League?
8. Are there any reports about refugees?
9. I'd like to see reports about President Herzogs visit to Japan!
10. Is there anything new in the Mykonos trial?

To generate the reference, the database was segmented manually in two steps. In the first step, the newscasts were segmented into topic stories, where a segment boundary was inserted only if the (semantical) *topic* of the segment changed. In the second step, the so defined segments were further segmented with respect to their acoustic background. A segment boundary was introduced whenever the acoustic background changed significantly, e.g. from anchor speaker to field speech, or from conference noise to battlefield noise.

Each of the segments was then manually judged relevant or irrelevant with respect to each of the test queries. This constitutes a major difference to the (much larger) TREC spoken document retrieval track [3], where the list of relevant segments is obtained by a refinement of the joint search result from different engines - an approach which masks errors that are common to all search engines.

During the evaluation, each of the queries was presented to the system and the segments that were returned were compared with the list of the segments that were judged relevant for this query. However, the interpretation of the results is not trivial. Most information retrieval algorithms assign a *relevance* score to each of the segments, and return a list of segments sorted by this relevance score. This list can sometimes comprise the whole database. The evaluation result therefore depends on the number of resulting segments which are taken into account during the evaluation. This number, however, cannot be chosen a priori, but depends on the - generally unknown - number of 'true hits' in the database.

Due to this problem, there is no generally accepted single way to represent the results of an information retrieval evaluation. In the text retrieval literature (see, e.g., [8]), the results are usually presented in one or more of the following ways:

- a plot of Precision (PRC) over Recall (RCL)
- average Precision (AveP), and
- R-Precision.

Average precision is defined as the average of the eleven values of PRC at Recall 0, 0.1, 0.2, ..., 1.0. If a given Recall value cannot be accurately achieved, the corresponding precision is determined by interpolation. There is some freedom of choice for the PRC at Recall zero, which can be defined in different ways. In our evaluation, we chose the PRC value computed at the first found item.
In our parameterization of the Okapi algorithm, many segments are assigned a relevance score of zero. The maximum possible value RCL_{max}, therefore, is obtained if all segments that have a nonzero relevance score are included into the evaluation. When computing average precision, the values for PRC are set to zero for all RCL values higher than RCL_{max}.
R-Precision is defined as the PRC value which is obtained when the number of segments evaluated is set to the number of 'true hits' in the database (with respect to the current query). At this operating point, PRC equals RCL.

3.1. Evaluation data

Three different segmentation strategies were chosen for the end to end evaluations: manual segmentation, a slow, high-performance segmentation engine and a fast segmenter. The test set contains 65 newscasts (roughly 16.5 hours of speech) dating between April 03, 1997 to January 16, 1999. The stop word list contained the 595 most frequent German words. A stemming algorithm was applied both to the query and the document in the database. Table 1 summarizes the structure of the test database.

Segmentation	documents	terms
manual	3165	13830
automatic (good)	2122	13895
automatic (fast)	3207	13567

Table 1. Structure of the test data

On average, 27.7 documents per query were judged relevant.

3.2. Results

The result of an end to end evaluation of the View4You system is summarized in table 2. To evaluate the influence of the speech recognition errors, both manually generated transcriptions and speech recognizer output were used to generate the metadata.

Segmentation	WER	IR algorithm	R-prec.
manual (topic)	0%	manual	1.0
manual (acoustic)	0%	Okapi	0.45
manual (acoustic)	22.7%	Okapi	0.43
automatic	22.7%	Okapi	0.39

Table 2. End to end evaluation results (R-precision)

From these results it can be concluded, that 90% of the total performance loss (as compared to perfect performance) is due to the information retrieval engine and the segmentation strategy (acoustically homogeneous segments rather than topic stories). From earlier experiments [2] it

is known, that the influence of the segmentation strategy is significantly smaller than the influence of the information retrieval component. Hence it can be concluded that the information retrieval component is responsible for the majority of the total errors in the View4You system.

4. THE INFORMATION RETRIEVAL ENGINE

4.1. The classic vector model approach

We chose the Okapi similarity measure [11] for our experiments. This measure has been evaluated thoroughly in the context of NIST's TREC information retrieval contests [8], and has been found to be very powerful. The Okapi measure can be parameterized to meet the special requirements of a given task. We use a parameterization that has been found to be very good for short queries [12]:

$$d(q,d) = \sum_{t \in Q \wedge t \in d} \left(\frac{f_{d,t}}{f_{d,t} + \frac{\sqrt{f_d}}{E(\sqrt{f_d})}} \right) \log\left(\frac{N - f_t}{f_t} \right) \quad (1)$$

$$= \text{Okapi}(k_1 = 1, k_2 = 0, k_3 = 0, b = 1, r = 0, R = 0) \quad (2)$$

where $E(.)$ denotes the expected value, N is the number of documents in the collection, f_t is the number of documents containing term t, $f_{d,t}$ is the frequency of term t in document d, and f_d is the number of terms in document d, which is an approximation to the document length. A *term* in this context is the same as a word, however, the 500 most frequent words ('I', 'other' and the like) are excluded. Morphological stemming is applied to both the query and the database records. The database engine computes the distance between a query and each article in the database and returns the articles sorted in decreasing order of similarity to the query.

4.2. Latent semantic indexing (LSI)

A more recent approach which makes implicit use of co-occurrence statistics between words is latent semantic indexing (LSI). Here, a term-by-document matrix is constructed from the database. Each term is weighted in the following way:

$$d(\text{termi}, \text{documentj}) = \log(1 + tf_{ij})(1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log(ndocs)} \quad (3)$$

$$p_{ij} = \frac{tf_{ij}}{\sum_j tf_{ij}} \quad (4)$$

tf_{ij} is the frequency of term i in document j, and $ndocs$ is the number of documents in the collection.

The resulting term-by-document matrix is decomposed using singular value decomposition (SVD). By discarding all but N of the singular values, an approximation of the original term-by-document matrix can be achieved, reducing its rank to N (in our experiments, we used both 100 and 200 for N). This dimensionality reduction results in a (intentional) loss in precision: the crisp distinction between different documents is weakened, and documents sharing several terms are mapped to points close to each other in the reduced space. The search itself is performed in the reduced dimensional space. Latent semantic indexing is described in detail in [13]. In the experiments in this paper, the original software as provided by [10] was used.

5. EXPERIMENTS

In a contrastive experiment, the same set of 10 user queries was evaluated on the same 65 news shows, where the metadata was computed fully automatically by the system. For the information retrieval component, both the Okapi algorithm in the parameterization (1) and LSI with N=100 were used. Several versions of metadata were evaluated. First, the segmentation of the videos was done with three different segmentation strategies (manual segmentation, a slow but high performance hybrid segmentation algorithm and a fast model-based segmentation algorithm [1]).

The result of the evaluation is summarized below.

Segmentation	SR error rate	LSI	Okapi
manual	0%	0.58	0.45
manual	22.7%	0.56	0.43
hybrid (good)	22.7%	0.49	0.39
model-based (fast)	22.7%	0.47	0.30

Table 3. End to end evaluation (R-precision)

Segmentation	SR error rate	LSI	Okapi
manual	0%	0.60	0.46
manual	22.7%	0.59	0.43
hybrid (good)	22.7%	0.51	0.36
model-based (fast)	22.7%	0.49	0.28

Table 4. End to end evaluation (average precision)

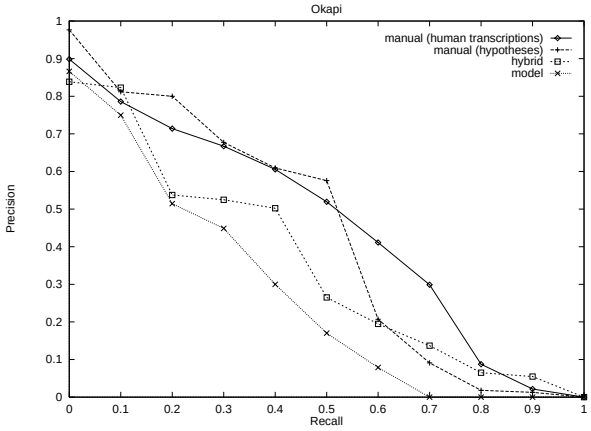

Figure 2. Precision over Recall for Okapi

The results show that latent semantic indexing outperforms the classic vector model approach on this dataset. The influence of the speech recognition errors on the final result is low for both information retrieval algorithms. The tf-idf based IR algorithm suffers less from a moderate deterioration of the segmentation quality than LSI, but shows a higher performance loss if the segmentation quality drops further.

5.1. Database expansion

LSI makes implicit use of co-occurrence statistics, i.e. whether a given keyword appears frequently together with another. Since the size of the test database is limited and contains speech recognition and segmentation errors, the co-occurrence statistics that can be derived from the test database's metadata is suboptimal. Therefore, we added additional data to the database which could be used to

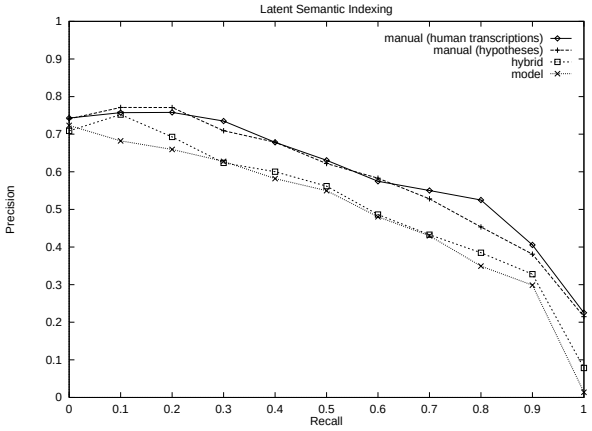

Figure 3. Precision over Recall for LSI

compute the co-occurrence statistics. The additional material was taken from radio broadcast transcripts, which were taken from the same 65 days as the 65 test newscasts. No selection of the transcripts took place, i.e. all transcripts from the 65 days were added to the database. In total, 7832 radio broadcast transcripts were added, roughly 4 times the total length of the original testset. The term by document matrix was then computed on the union of the test metadata and the radio transcripts. The computation of the SVD took about 2 hours on a 300 MHz Sparc Ultra 2 machine. When retrieving documents, hits that referred to the radio transcripts were discarded. Using this sort of database expansion, another end to end evaluation was carried out. The results are shown in table 5. Both R-precision and average precision improve by 8% relative.

expansion	Segmentation	WER	R-prec.	AveP
no	automatic (good)	22.7%	0.49	0.51
yes	automatic (good)	22.7%	0.53	0.55

Table 5. Effect of database expansion

5.2. LSI search dimensionality

For LSI, the dimensionality of the reduced term by document matrix can be varied. A higher dimensionality means a more precise reflection of the original term by document structure of the database in the search space, which corresponds to less generalization and generally a more keyword-like behaviour. To evaluate the effect of the dimensionality of the reduced search space, two end to end evaluations were run using two different dimensionalities, both with database expansion (see 5.1). The results are shown in table 6.

Dimensionality	AveP	R-PRC
100	0.55	0.53
200	0.58	0.51

Table 6. Effect of dimensionality in LSI

Although average precision increases when the dimensionality is increased from 100 to 200, R-precision drops. It can therefore be concluded that 100 is a reasonable value for the dimensionality on the data set used in this evaluation.

6. CONCLUSIONS

Two standard information retrieval approaches, latent semantic indexing (LSI) and a vector model based approach (Okapi) have been evaluated on automatically generated metadata from a video indexing machine. On this kind of data and the relatively small size of the database, LSI consistently outperformed Okapi both in terms of R-precision and average Precision. LSI also showed less performance loss if the quality of the underlying metadata was degraded. By adding easily obtainable context information to the system, the end to end performance using LSI could be increased by another 10% relative to a maximum value of 0.53 R-precision.

7. ACKNOWLEDGEMENTS

The authors wish to thank all members of the Interactive Systems Labs for useful discussions and active support.

REFERENCES

[1] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, *Strategies for automatic segmentation of audio data*, in Proc. ICASSP 2000, Vol 3, pp 1423 ff, Istanbul, Turkey, June 5-9, 2000

[2] T. Kemp, M. Weber, A. Waibel, *End to end evaluation of the ISL View4You broadcast news transcription system*, in Proc. RIAO-2000, Paris, France, April 12-14, 2000

[3] J. Garofolo, C. Auzanne, E. Voorhees, *The TREC Spoken Document Retrieval Track: A Success Story*, in Proc. RIAO 2000, Paris, France, April 12-14, 2000

[4] S.E. Johnson, P. Jourlin, K. Sparck-Jones, P.C. Woodland, *Audio Indexing and Retrieval of Complete News Shows*, in Proc. RIAO 2000, Paris, France, April 12-14, 2000

[5] H. Wactlar, A. Hauptmann, M. Witbrock, *Informedia: news-on-demand experiments in speech recognition*, Proc. of the ARPA SLT workshop, 1996.

[6] P.C.Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, S. Young, *Experiments in broadcast news transcription*, Proc. ICASSP 1998, pp. 909 ff, Seattle, Washington, May 1998

[7] J. L. Gauvain, L. Lamel, G. Adda, *The LIMSI 1997 Hub-4E Transcription System*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[8] http://trec.nist.gov/

[9] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, R. Gadde, *The development of SRI's 1997 broadcast news transcription system*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[10] http://www.cs.utk.edu/ lsi/

[11] M.M. Beaulieu, M. Gatford, X. Huang, S.E. Robertson, S. Walker, P. Williams, *Okapi at TREC-5*, Proc. of the 5th Text Retrieval Conference, NIST, Gaithersburg, MD, Januar 1997

[12] R. Wilkinson, J. Zobel, R. Sacks-Davis, *Similarity Measures for Short Queries*, Proc. of the 4th Text Retrieval Conference, NIST, Gaithersburg, MD, November 1995

[13] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, *Indexing by latent semantic analysis*, Journal of the Society for Information Science, 1990 **41**(6), pp 391-407