PROBABILISTIC RETRIEVAL BASED ON DOCUMENT REPRESENTATIONS

Wolfgang Macherey, Jörg Viechtbauer, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{w.macherey,viechtbauer,ney}@informatik.rwth-aachen.de

ABSTRACT

Accessing information in multimedia databases encompasses a wide range of applications in which spoken document retrieval (SDR) plays an important role. In the past, research increasingly focused on the development of heuristic and probabilistic retrieval metrics that are suitable for retrieving spoken documents. So far, many heuristic retrieval metrics, e.g. theabilistic retrieval metrics that are suitable for retrieving spoken increasing focused on the development of heuristic and prob-

2. BASELINE RETRIEVAL METRIC

The SMART-2 metric is an enhanced version of the SMART metric and was published in [1] the first time. Due to its good performance on text and SDR tasks, we utilize SMART-2 as baseline metric. In this section, we give a brief introduction to the SMART-2 metric in order to introduce the terminology used in this paper. Let \( D := \{d_1, \ldots, d_K\} \) be a set of \( K \) documents and \( Q := \{q_1, \ldots, q_L\} \) denote a set of queries. Then, documents and queries are given as sequences of index terms \( t \in T \):

\[
d_k = \{t_{k1}, \ldots, t_{ks}\} \quad d_k \in D
\]

\[
q_i = \{t_{i1}, \ldots, t_{is}\} \quad q_i \in Q
\]

The term frequency, i.e. the number of occurrences of an index term \( t \) in a document \( d_k \) is denoted by:

\[
n(t, d_k) := \sum_{i=1}^{s_k} \delta(t, t_{ki})
\]

According to [2], each index term \( t \) of a document \( d \) is associated with a weight \( g(t, d) \) that depends on the ratio of the logarithm of the term frequency \( n(t, d) \) to the logarithm of the average term frequency \( m(d) \)

\[
g(t, d) := \begin{cases} 
1 + \log n(t, d) & \text{if } t \in d \\
1 + \log m(d) & \text{if } t \notin d 
\end{cases}
\]

with

\[
\log 0 := 0 \quad \text{and} \quad m(d) := \frac{\sum_{t \in T} n(t, d)}{\sum_{t \in T \delta(t, d)>0} 1}
\]

The logarithms in Eq. (4) prevent documents with high term frequencies from dominating those with low term frequencies. In order to obtain the final term weights, \( g(t, d) \) is divided by a linear combination between a pivot element \( c \) and the number of singletons \( n_1(d) \) in document \( d \):

\[
\omega(t, d) := \frac{g(t, d)}{(1 - \lambda) \cdot c + \lambda \cdot n_1(d)}
\]

with \( \lambda = 0.2 \) and

\[
c := \frac{1}{K} \sum_{k=1}^{K} n_1(d_k) \quad \text{and} \quad n_1(d) := \sum_{t \in T \delta(t, d)=1} 1
\]

Unlike document terms, query terms are weighted with the inverse document frequency idf(t)

\[
\omega(t, q) = \left[1 + \log n(t, q)\right] \cdot \text{idf}(t)
\]
Here, \( idf(t) \) is defined by
\[
idf(t) = \log \left( \frac{K}{n(t)} \right)
\]
(8)

The SMART-2 retrieval function is defined as the product over the document and query specific index term weights:
\[
f(q, d) = \sum_{t \in T} \omega(t, q) \cdot \omega(t, d)
\]
(9)

Note that due to the floor operation in Eq. (8) a term weight will be zero if it occurs in more than half of the documents.

3. A NEW STATISTICAL APPROACH TO SPOKEN DOCUMENT RETRIEVAL

Even though many probabilistic retrieval metrics (e.g. [3], [4]) are able to outperform basic retrieval metrics as for example the term-frequency/inverse-document-frequency (tf-idf) metric, they usually do not achieve the effectiveness of advanced heuristic retrieval metrics such as SMART-2 or OKAPI [5]. In particular for SDR tasks, probabilistic metrics often turned out to be less robust towards recognition errors than their heuristic counterparts. To compensate for this shortcoming, we propose a new statistical approach to information retrieval that is based on document similarities [6].

3.1. Probabilistic Retrieval Using Document Representations

A fundamental difficulty in statistical approaches to information retrieval is the fact that typically a rare term is well suited to filter out a document. On the other hand, a reliable estimation of distribution parameters requires that the underlying events, i.e. index terms are observed as frequently as possible. Therefore, it is necessary to properly smooth the distributions. In our case, document specific term probabilities \( p(t | d) \) are smoothed with term probabilities of documents that are similar to \( d \). The similarity measure is based on document representations which in the simplest case are document specific histograms of the index terms. The starting point of our approach is the joint probability \( p(q, d) \) of a query \( q \) and a document \( d \):
\[
p(q, d) = \prod_{i=1}^{\#d} p(q_i, d | q_i^{-1})
\]
(10)
\[
= \prod_{i=1}^{\#d} p(q_i, d)
\]
(11)

The conditional probabilities \( p(q_i, d | q_i^{-1}) \) in Eq. (10) are assumed to be independent of the predecessor terms \( q_i^{-1} \). Document representations are now introduced via a hidden variable \( r \):
\[
p(q, d) = \prod_{i=1}^{\#d} \sum_{r \in R} p(q_i, d, r)
\]
(12)
\[
= \prod_{i=1}^{\#d} \sum_{r \in R} p(q_i | d) \cdot p(d | r) \cdot p(r)
\]
(13)
\[
= \prod_{i=1}^{\#d} \sum_{r \in R} p(q_i | d) \cdot \prod_{j=1}^{\#d} p(d_j | r, d_j^{-1}) \cdot p(r)
\]
(14)
\[
= \prod_{i=1}^{\#d} \sum_{r \in R} p(q_i | r) \cdot \prod_{j=1}^{\#d} p(d_j | r) \cdot p(r)
\]
(15)

Here, two model assumptions have been made: first the conditional probabilities \( p(q_i | d, r) \) are assumed to be independent of \( d \) (cf. Eq.(13)) and secondly, \( p(d_j | r, d_j^{-1}) \) shall not depend on the predecessor terms \( d_j^{-1} \) (cf. Eq.(15)). Finally, it remains to specify models for the document representations \( r \in R \) as well as the distributions \( p_q(t | r) \), \( p_d(t | r) \), and \( p(r) \). Since we want to distinguish between the event that a query term \( t \) is predicted by a representation \( r \) and the event that the term to be predicted is part of a document, \( p_q(t | r) \) and \( p_d(t | r) \) are modeled differently. In our approach we identify the set of document representations \( R \) with the histograms over the index terms of the document collection \( D \):
\[
n_q(t) \equiv n(t, d) \quad n_r(\cdot) \equiv |d|
\]
(16)
\[
n(t) \equiv \sum_{d \in D} n(t, d) \quad n(\cdot) \equiv \sum_{d \in D} |d|
\]
(17)

Thus, we can define the following interpolations:
\[
p_q(t | r) := (1 - \alpha) \cdot \frac{n_q(t)}{n_r(\cdot)} + \alpha \cdot \frac{n(t)}{n(\cdot)}
\]
(18)
\[
p_d(t | r) := (1 - \beta) \cdot \frac{n_d(t)}{n_r(\cdot)} + \beta \cdot \frac{n(t)}{n(\cdot)}
\]
(19)

Since we do not make any assumptions about the a-priori relevance of a document representation, we set up a uniform distribution for \( p(r) \). Note that Eq. (19) is an interpolation between the relative counts \( n_q(t) / n_r(\cdot) \) and \( n(t) / n(\cdot) \). Instead of interpolating between the relative frequencies as in Eq. (19), we can also interpolate between the absolute frequencies:
\[
p_d(t | r) := \frac{(1 - \beta) \cdot n_q(t) + \beta \cdot n(t)}{1 - \beta \cdot n_r(\cdot) + \beta \cdot n(\cdot)}
\]
(20)

Both interpolation variants will be considered in the following section.

4. TASKS AND EXPERIMENTAL RESULTS

Experiments were performed on the TREC-7 and the TREC-8 SDR task. The TREC-7 task comprises 2866 spoken documents and 23 test queries. The TREC-8 task comprises 21745 spoken documents and 27 test queries. Table 1 summarizes some corpus statistics. All speech recognition outputs were produced using the RWTH large vocabulary continuous speech recognizer (LVCSR) (cf. [7]) for the TREC-7 corpus and the Byblos “Rough ‘N Ready” [8] and Dragon LVCSR system [9], respectively, for the TREC-8 SDR corpus. Due to the small number of test queries for both retrieval tasks, we made use of a leaving-one-out (L-1-O) approach [10, p. 220] in order to estimate the interpolation parameters \( \alpha \) and \( \beta \). Additionally, we carried out a cheating experiment by adjusting the parameters \( \alpha \) and \( \beta \) to maximize the MAP on the complete set of test queries. This yields an optimistically upper bound of

<table>
<thead>
<tr>
<th>Task</th>
<th>TREC-7</th>
<th>TREC-8</th>
</tr>
</thead>
<tbody>
<tr>
<td># documents</td>
<td>2866</td>
<td>21745</td>
</tr>
<tr>
<td># queries</td>
<td>23</td>
<td>27</td>
</tr>
<tr>
<td>avg. doc. length</td>
<td>267.4</td>
<td>265.5</td>
</tr>
<tr>
<td>rel.</td>
<td>348</td>
<td>16.79</td>
</tr>
<tr>
<td>irr.</td>
<td>2518</td>
<td>20066</td>
</tr>
<tr>
<td>rel.</td>
<td>580.1</td>
<td>283.9</td>
</tr>
<tr>
<td>irr.</td>
<td>265.5</td>
<td>169.4</td>
</tr>
</tbody>
</table>
the possible retrieval effectiveness. All experiments conducted are based on the document representations according to Eq. (16) and Eq. (17), i.e. each document is smoothed with all other documents in the database.

In a first experiment, the interpolation parameter \( \alpha \) was estimated. Fig. 1 shows the MAP as a function of the interpolation parameter \( \alpha \) with fixed \( \beta \) on the reference transcriptions of the TREC-7 corpus. Using the L-1-0 estimation scheme, the best value for \( \alpha \) was found to be 0.742 which has to be compared with a globally optimal value of 0.875, i.e. the cheating experiment without L-1-O. The interpolation parameter \( \beta \) was adjusted in a similar way. Using the interpolation scheme according to Eq. (19), the retrieval effectiveness on both tasks is maximum for values of \( \beta \) that are very close to 1. This effect is caused by singletons, i.e. index terms that occur once only in the whole document collection. Since the magnitude of the ratio of both denominators in Eq. (19) is approximately

\[
\frac{n_\alpha(\cdot)}{n(\cdot)} \approx \frac{1}{D}
\]

the optimal value for \( \beta \) should be found in the range of \( 1 - 1/D \), assuming that singletons are the most important features in order to filter out a relevant document. In fact, using \( \beta = 1 - 1/D \) exactly meets the optimal value of 0.99965 on the TREC-7 corpus and 0.99995 on the TREC-8 retrieval task.

However, since the interpolation according to Eq. (19) runs the risk of becoming numerically unstable (especially for very large document collections), we investigated an alternative smoothing scheme that interpolates between absolute counts instead of relative counts (cf. Eq. (20)). Fig. 2 depicts the MAP as a function of the interpolation parameter \( \beta \) for both interpolation methods on the reference transcriptions of the TREC-7 SDR task. Since the interpolation scheme according to Eq. (20) proved to be numerically stable and achieved slightly better results, it was used for all further experiments. Table 2 shows the obtained retrieval effectiveness for the new probabilistic approach on the TREC-7 SDR task. Using L-1-O, the retrieval performance of the new proposed method lies within the magnitude of the SMART-2 metric, i.e. we obtained a MAP of 45.8% on manually transcribed data, which must be compared with 46.6% using the SMART-2 retrieval metric. Using automatically generated transcriptions we achieved a MAP of 40.4% which is quite close to the performance of the SMART-2 metric. Fig. 3 shows the recall-precision graphs for both SMART-2 and the new probabilistic approach.

Table 2. Comparison of retrieval effectiveness measured in terms of mean average precision (MAP) on the TREC-7 spoken document retrieval task for the SMART-2 metric and the new probabilistic approach PROB. Interpolation was performed according to Eq. (20).

<table>
<thead>
<tr>
<th>TREC-7 metric</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>MAP [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>SMART-2</td>
<td>0.851</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>PROB</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>speech</td>
<td>SMART-2</td>
<td>0.742</td>
<td>0.270</td>
</tr>
<tr>
<td>(Byblos)</td>
<td>PROB</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>speech</td>
<td>SMART-2</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>(Dragon)</td>
<td>PROB</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>text</td>
<td>SMART-2</td>
<td>0.950</td>
<td>0.650</td>
</tr>
<tr>
<td></td>
<td>PROB</td>
<td>0.947</td>
<td>0.646</td>
</tr>
<tr>
<td>speech</td>
<td>SMART-2</td>
<td>0.801</td>
<td>0.287</td>
</tr>
<tr>
<td>(Byblos)</td>
<td>PROB</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>speech</td>
<td>SMART-2</td>
<td>0.875</td>
<td>0.300</td>
</tr>
<tr>
<td>(Dragon)</td>
<td>PROB</td>
<td>0.875</td>
<td>0.300</td>
</tr>
</tbody>
</table>
The same applies to the results obtained on the TREC-8 SDR task. Here, the new probabilistic approach even outperformed the SMART-2 retrieval metric. Thus, we obtained a MAP of 51.3% on the manually transcribed data in comparison with 49.6% for the SMART-2 metric. This improvement over SMART-2 is also obtained on recognized transcriptions even though the improvement is smaller. Thus, we achieved a MAP of 44.4% on the automatically generated transcriptions produced with the Byblos speech recognizer, which is an improvement of 3% relative compared to the SMART-2 metric, and 44.1% MAP using the Dragon speech recognition outputs, which is an improvement of 5% relative. Fig. 4 shows the recall-precision graphs for SMART-2 and the probabilistic approach.

5. CONCLUSION

In this paper, we presented a new probabilistic approach to spoken document retrieval that is based on interpolations between a document specific term histogram and a global term histogram that is pooled over all documents. To this purpose, the set of documents was mapped onto a set of document representations. These document representations were identified with document specific histograms and can be interpreted as a kind of nearest neighbor concept. Two smoothing schemes were discussed and investigated. Experiments performed on the TREC-7 and the TREC-8 spoken document retrieval task showed comparable or even better results for the new probabilistic approach than an enhanced version of the SMART-2 retrieval metric. In addition, the new probabilistic approach turned out to be robust towards recognition errors.

6. REFERENCES


