A Method for Transferring Retrieval Scores Between Collections with Non-Overlapping Vocabularies

Fernando Diaz*
Yahoo! Inc.
1000 Rue de la Gauchetiere, Suite 2400
Montreal, QC
diazf@yahoo-inc.com

ABSTRACT

We present a method for projecting retrieval scores across two corpora with a shared, parallel corpus.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Retrieval models

General Terms: Algorithms

Keywords: regularization, cross-lingual retrieval

1. INTRODUCTION

In many retrieval scenarios, the collection of retrievable documents consists of several, disjoint sub-collections. This is generally referred to as distributed information retrieval or federated search. We focus on the situation where each sub-collection uses a unique vocabulary. For example, we might have a sub-collection of text documents written in english, a sub-collection of text documents written in french, and another sub-collection of images.

Given a query, we often are able to score documents in one sub-collection but not in others. When our sub-collections consist of documents written in different languages, this is known as cross-lingual retrieval. In cross-lingual information retrieval, a user is interested in documents written in a foreign or *target* language and provides a query in her native or *source* language. Traditional approaches to this problem usually perform some sort of query translation from the source to the target language [2]. When sub-collections consist of documents in different media, this is known as cross-media retrieval.

We focus on transferring the scores from one sub-collection to another sub-collection. We accomplish this by scoring source parallel documents and using these scores as the basis for regression in the target collection. Like other methods, we only require a parallel corpus. However, we do not translate the query and hence do not require a second retrieval.

2. TRANSFERRING SCORES BETWEEN COLLECTIONS

Formally, we have a target collection of n_t documents with a vocabulary size of m_t. Some relatively small number, n_s, of the target documents have been translated into the source language with a vocabulary size of m_s. Sets of translated collections are common in the machine translation community and are referred to as parallel corpora. More generally, we only require some representation of the target documents in the source vocabulary. For example, we may have captions associated with images. We will further assume that, given a query in the source language, we have some method for scoring the source language documents.

Transferring scores between collections is a process of scoring the source parallel documents and then assuming that the n_s parallel target documents should have the same score. If the user were interested in retrieving the parallel target documents, the retrieval process could terminate at this stage. However, the user is more often interested in those target documents which do not have source translations. We will score these non-parallel target documents by using the score information from the parallel documents.

Assume that the translated documents are all indexed identically from 0 to n_s for both corpora. Additionally, we will assume that we have an $n_t \times n_t$ affinity matrix, $\mathbf{W}_t$, for the target collection of documents. An affinity matrix contains the similarity information between all pairs of documents in the target collection. For text documents, similarity measures have been studied in the context of topic link detection as well as clustering for many languages. More generally, we can use any kernel defined on documents of the target collection. We process $\mathbf{W}_t$ by keeping only the k-nearest neighbors of each document and then making this matrix symmetric. Further let $\mathbf{D}$ be the $n_t \times n_t$ normalizing matrix such that $D_{ii} = \sum_j W_{ij}$.

Let the $n_t \times 1$ vector $\mathbf{f}_t$ contain the scores transferred to the target corpus documents. We would like to search over the space of all $n_t \times 1$ vectors to find a vector for which similar documents—as represented by $\mathbf{W}_t$—have very similar scores—as measured by some score similarity function—subject to the constraint that the projected scores for the first n_s elements are similar to the original retrieval scores. We represent the *dissimilarity* of scores of related documents using the function $\mathcal{S}(\mathbf{f}_t)$; we represent the *dissimilarity* of scores of the first n_s documents with the original retrieval scores using the function $\mathcal{E}(\mathbf{f}_t, \mathbf{y}_s)$ where $\mathbf{y}_s$ is the $n_s \times 1$ vector of source collection scores. We linearly combine these into a composite function,

$$\mathcal{Q}(\mathbf{f}_t, \mathbf{y}_s) = \mathcal{S}(\mathbf{f}_t) + \mu\mathcal{E}(\mathbf{f}_t, \mathbf{y}_s) \quad (1)$$

where μ is a scalar parameter combining both objectives.

*Work conducted at the Center for Intelligent Information Retrieval.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

The constraints are defined as,

$$\mathcal{S}(\mathbf{f}_t) = \mathbf{f}_t^\mathsf{T} \Delta_t \mathbf{f}_t \qquad \mathcal{E}(\mathbf{f}_t, \mathbf{y}_s) = \|\mathbf{f}_t^\mathsf{T} - \mathbf{y}_t\|_2^2$$

where $\mathbf{y}_t = [\mathbf{y}_s^\mathsf{T} \mathbf{0}^\mathsf{T}]^\mathsf{T}$ is a vector of projected scores and $\Delta = \mathbf{I} - \mathbf{D}_t^{-1/2} \mathbf{W}_t \mathbf{D}_t^{-1/2}$ is known as the combinatorial Laplacian. The combinatorial Laplacian as a measure of score similarity has been previously used successfully in the situation of document re-ranking [1]. The closed form solution for computing $\mathbf{f}^*$ is,

$$\mathbf{f}_t^* = (1-\alpha)(\alpha \Delta_t + (1-\alpha)\mathbf{I})^{-1} \mathbf{y}_t \tag{2}$$

where $\alpha = \frac{1}{1+\mu}$.

3. CROSS-LINGUAL RELEVANCE MODELS

Let θ^t refer to a language model over the target vocabulary; similarly, θ^s models the source language. If we have a query in the source language, we score each source parallel document, d, according to the query likelihood, $P(Q|\theta_d^s)$. The *cross-lingual relevance model* is estimated as [2],

$$P(w|\theta_R^t) = \sum_d \frac{P(Q|\theta_d^s)}{\mathcal{Z}} P(w|\theta_d^t) \tag{3}$$

where $\mathcal{Z}$ is a normalizing constant. With the cross-lingual relevance model, we are applying the score for a source document to the parallel target document, allowing us to build a relevance model in the target language using source document scores as the interpolation weights. This solves our problem of not having a query in the target language. We then score a document by its cross-entropy with $P(w|\theta_R^t)$.

Interestingly, we can show that scoring by cross-lingual relevance models is very similar to our method of transferring scores. The proof follows from combining the relevance model estimation and cross-entropy scoring and rearranging summations. The resulting scores, $\mathbf{f}_t$, can be are related to the original target corpus scores according to,

$$\mathbf{f}_t = \frac{1}{\|\mathbf{y}_t\|_1} \mathbf{A}_t \mathbf{y}_t \tag{4}$$

where $\mathbf{y}_t$ is composed of $P(Q|\theta_d^s)$ scores and $\mathbf{A}_t$ is an $n_t \times n_t$ affinity matrix based on inter-document cross-entropy between the target documents. This is a single step of an iterative version of Equation 2.

4. EXPERIMENTS

We compared the performance of our method to cross-lingual relevance models (CLRM) using a cross-lingual retrieval task involving a source query written in English and a target collection written in Mandarin [3]; machine translated TDT5 documents were used as a parallel corpus [2]. We indexed and retrieved english parallel documents using the open source Indri retrieval system. We indexed Mandarin documents using Indri, treating each character as a word. We used character tf.idf vectors and cosine similarity for computing $\mathbf{W}_t$. We perform 10-fold cross-validation to tune free retrieval parameters. We used the paired t-test to measure statistical significance.

We present results in Table 1. Perhaps due to the theoretical similarity of the approaches, there is no statistical difference between our method and CLRM. However, our method tends to perform better at low-recall areas. Therefore, in Table 2, we evaluate each algorithm by the precision for the top k documents.

	CLRM	transfer	p
0.00	0.5694	0.6238	0.3879
0.10	0.3737	0.4456	0.0604
0.20	0.3194	0.3535	0.3383
0.30	0.2789	0.2943	0.5576
0.40	0.2424	0.2502	0.7381
0.50	0.2049	0.2010	0.8509
0.60	0.1673	0.1520	0.3691
0.70	0.1301	*0.0989*	0.0201
0.80	0.0916	*0.0536*	0.0017
0.90	0.0361	*0.0154*	0.0045
map	0.2027	0.2064	0.8897

Table 1: Cross-lingual relevance models compared to transferring scores.

P@K	CLRM	transfer	p
5	0.3556	0.4185	0.2566
10	0.3167	**0.4037**	0.0418
15	0.3123	0.3617	0.1962
20	0.3102	0.3389	0.3958
30	0.3006	0.3228	0.4781

Table 2: Cross-lingual relevance models compared to transferring scores.

These preliminary results indicate an interesting direction for information retrieval research for several reasons. First, our perspective is general and can be applied to cross-lingual retrieval, distributed information retrieval, and cross-media retrieval. Second, our method generalizes a previous, high-quality retrieval method and allows us to study its components—computing source collection scores and interpolating target collection scores—independently. Third, unlike previous query-translation approaches, we do not need to expand the query into a potentially large query, allowing us to optimize retrieval engines for short queries instead of having to handle short and long queries. Finally, in addition to links between the two corpora, we only require a kernel to be defined on the target corpus. This means that, in cases where a query translation is ill-specified or target corpus retrieval is poor but we understand a kernel, we can still apply our algorithm.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the author and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] F. Diaz. Regularizing query-based retrieval scores. *Information Retrieval*, 10(6):531–562, December 2007.

[2] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR 2002*, pages 175–182, 2002.

[3] A. F. Smeaton. Spanish and chinese document retrieval in TREC-5. In *TREC*, 1996.