Filtering Documents with Subspaces

B. Piwowarski, I. Frommholz, Y. Moshfeghi, M. Lalmas, and C.J. van Rijsbergen

University of Glasgow, Department of Computing Science,
Glasgow G12 8QQ, UK

Abstract We propose an approach to build a subspace representation for documents. This more powerful representation is a first step towards the development of a quantum-based model for Information Retrieval (IR). To validate our methodology, we apply it to the adaptive document filtering task.

1 INTRODUCTION

We explore an alternative representation of documents where each document is not represented as a vector but as a subspace. This novel way of representing documents is more powerful than the standard one-dimensional (vector) representation. Subspaces are a core component of the generalisation of the probabilistic framework brought by quantum physics [1], which enables to combine both geometry and probabilities.

Sophisticated document representations have already been explored. Melucci proposed to use subspaces to describe the locus of relevant documents for capturing context in IR; however documents are still represented as vectors [2]. Zuccon et al. [3] showed that the cluster hypothesis still holds when representing documents as subspaces. In our work, we propose a different approach to build such subspaces, where we suppose that a document can be represented as a set of information needs (IN), each being represented as a vector. We also show how to build a user *profile* from relevance feedback that can be used to compute the probability of a document to be relevant.

Knowing how to represent documents is the first step towards a working IR system, and here we focus on how to build such a representation and leave out (among others) the problem of the query (or topic) representation. This makes information filtering a suitable task to investigate our proposed subspaces since it does not necessitate to represent a profile from a set of keywords like in an ad-hoc task. We evaluate our approach on the adaptive document filtering task of TREC-11 [4].

2 Document Filtering with Subspaces

In the adaptive filtering task [4], for each topic, three relevant documents from the training set are given to build a profile representation. Then, documents

are filtered one by one in a specified order, and each time the system decides whether to retrieve the incoming document or not. Only when the document is retrieved by the system, its associated relevance assessment can be used to update the profile representation before the system evaluates the relevance of the next incoming document. This process simulates a user interactive relevance feedback, since the user can only judge a document if it is retrieved.

2.1 Building the Document Subspace

Our main hypothesis is that a document can be represented as the subspace S_d spanned by a set of vectors, where each vector corresponds to an IN covered by the document. In practice, we assume that we can decompose a document into text excerpts that are associated with one or more INs. For a document d, we denote $\mathcal{U}_d$ the set of such vectors.

There are various possibilities to define the excerpts and how to map an excerpt to a vector, ranging from extracting sentences, paragraphs to using the full document as the single excerpt. As a first approximation, we chose to use sentences as excerpts (simple heuristics were applied to detect sentence boundaries[1]), and to transform them into vectors in the standard term space after stop word elimination and stemming. The weighting scheme used to construct vectors was either tf or tf-idf (see Section 3).

To compute the subspace S_d from the set of vectors of $\mathcal{U}_d$ (which are then spanning the subspace), an eigenvalue decomposition is used. The eigenvectors associated with the set of non-null eigenvalues of the matrix $\sum_{u \in \mathcal{U}_d} uu^\top$ define a basis of the subspace spanned by the vectors from $\mathcal{U}_d$. As the vectors from $\mathcal{U}_d$ are extracted from a corpus, we are not interested in all the eigenvectors but only in those that are associated with high eigenvalues λ_i, since low eigenvalues might be associated with noise. We used a simple strategy to select the rank of the eigenvalue decomposition, where we only select the eigenvectors with eigenvalues superior to the mean of the eigenvalues.

2.2 Profile Updating and Matching

The representation of the filtering profile is closely related to the above described document representation. We rely on the quantum probability framework to compute the probability of a document matching this profile.

The profile is updated whenever a document is retrieved. At each step, we can construct two sets Ψ^+ and Ψ^- that correspond to the set of all the INs of the retrieved documents that are relevant (resp. non relevant). From the set Ψ^- we build a *negative* subspace N (as described in the previous section) and assume that vectors lying in this subspace correspond to non-relevant INs. This process is the underlying motivation of using a subspace for the negative sub-profile. We denote $N^\perp$ the subspace orthogonal to this negative subspace.

[1] We use `http://www.andy-roberts.net/software/jTokeniser/index.html`

To determine if a document d, represented as a projector D on the subspace S_d (constructed as described in section 2.1), is retrieved or rejected with respect to the profile, we first project each (unit) vector $\psi_i \in \Psi^+$ of the positive profile onto the subspace $N^\perp$, in order to remove its non-relevant part. The result is a vector ψ'_i . We then suppose that a relevant document should "contain" as much as possible of these vectors ψ'_i. It is possible to give a probabilistic definition of this containment, by letting the probability that the document contains the IN ψ'_i be $\Pr(D|\psi'_i) = {\psi'_i}^\top D\psi'_i$ which has a value between 0 and 1, since D is a projector and ψ'_i has a norm less than 1.

As we have no preference about which of the vectors ψ'_i should be contained, we assume that each of the vectors is picked with a uniform probability, so that the probability of the document being relevant is given by $\Pr(D) \propto \sum_i \Pr(D|\psi'_i) = \mathrm{tr}(\rho D)$ where ρ equals $\sum_i \psi'_i {\psi'_i}^\top$ and tr is the trace operator. We can compute the actual probability by dividing $\mathrm{tr}(\rho D)$ by $\mathrm{tr}(\rho)$, which is a normalisation constant. If the value $\Pr(D)$ is over a given threshold, we retrieve the document; otherwise, we reject it. For simplicity, we only use a fixed threshold in the experiments, whereas a better approach would be to use a threshold that depends on the topic and the current state of the profile.

3 Experiments

We experimented with the adaptive filtering task of TREC-11 [4] and followed the task guidelines. Note that we ignored documents for which there was no relevance judgement. One important issue is to set a threshold so that a document whose score (as determined by the profile) is above the threshold is retrieved. As we wanted to focus on showing how the subspace approach performs compared to a baseline, we used a fixed threshold. We tried several values for this threshold, and selected the best performing runs. Comparing to approaches reported in [4], we have the unfair advantage of reporting the best performing settings but at the same time are penalised by the fact that our threshold is constant.

Subspace	Neg	$\overline{\text{Neg}}$	**Rocchio**
TF	0.44^a	0.30^b	0.35^a
TF-IDF	0.41^a	0.31^b	0.44^b

Table 1. Mean F-0.5 measure for the TREC-11 adaptive filtering task, for the Subspace and the Rocchio-based approach. The corresponding threshold values are (a) 0.05 and (b) 0.10.

We report results using one of the official metrics, the mean of F-0.5 metric (harmonic mean biased towards precision) which is less sensitive to the threshold policy. As a simple baseline, we report results obtained using the Rocchio-based approach for user profiling [5], and use a constant threshold (for a fair comparison with our approach) and a cosine similarity measure between a profile and a document (since it allows to experiment with the tf and the tf-idf weighting schemes). For the subspace approach, we experimented with the following parameters: (1) Using a negative subspace as described above (Neg) or

not ($\overline{\text{Neg}}$, where we do not project ψ_i onto $N^{\perp}$) (2) using a tf-idf or tf weighting scheme to construct the ψ_i vector. Note that as for [5], idf values were estimated using an external collection (in our case Wikipedia) and updated with statistics from filtered documents. Eventually, for all models, we did an exhaustive search using a 0.05 step for the threshold. Values reported in Table 1 should be regarded as the maximum achievable with a fixed threshold; due to the small scale of the experiment, we do not report here statistical significance.

Our best runs are able to compete with those reported in [4], although it should be noted that we selected our best performing run (but also the baseline) *a posteriori*. We can then outline two facts from the results. First, using negative subspace was beneficial both for tf and tf-idf schemes: Using orthogonality to define non relevance is thus meaningful. Second, our subspace approach is competitive with a Rocchio-based baseline without relying on idf values.

4 Conclusion

There is the view that using subspaces instead of one dimension space is essential for sophisticated IR tasks like e.g. interactive IR [6]. In this paper, we showed through document filtering experiments that both the subspace representation of documents and the way we construct it lead to positive results. To exploit this representation, we also showed how to construct a user *profile* as a weighted set of vectors (and not a single vector as in e.g. Rocchio). This profile was constructed from documents, and our future work is to show how to construct and update this profile through sophisticated user interaction (query formulation, clicks, etc.), thus further exploiting the proposed subspace document representation.

Acknowledgements This research was supported by an Engineering and Physical Sciences Research Council grant (Grant Number EP/F015984/2). M. Lalmas is currently funded by Microsoft Research/Royal Academy of Engineering.

References

1. van Rijsbergen, C.J.: The Geometry of Information Retrieval. Cambridge University Press, New York, NY, USA (2004)
2. Melucci, M.: A basis for information retrieval in context. ACM TOIS **26**(3) (2008)
3. Zuccon, G., Azzopardi, L., van Rijsbergen, C.J.: Semantic spaces: Measuring the distance between different subspaces. In: Third QI Symposium. (2009)
4. Robertson, S., Soboroff, I.: The TREC 2002 filtering track report. In NIST, ed.: TREC-11. (2001)
5. Zhang, Y., Callan, J.: Yfilter at TREC-9. NIST special publication (2001) 135–140
6. Piwowarski, B., Lalmas, M.: A Quantum-based Model for Interactive Information Retrieval (extended version). ArXiv e-prints (0906.4026) (September 2009)