Probabilistic Hyperspace Analogue to Language

Leif Azzopardi[*]
University of Paisley
School of Computing
azzo-ci0@paisley.ac.uk

Mark Girolami
University of Glasgow
Dept. of Computing Science
girolami@dcs.gla.ac.uk

Malcolm Crowe
University of Paisley
School of Computing
crow-ci0@paisley.ac.uk

ABSTRACT

Song and Bruza [6] introduce a framework for Information Retrieval(IR) based on Gardenfor's three tiered cognitive model; Conceptual Spaces[4]. They instantiate a conceptual space using Hyperspace Analogue to Language (HAL)[3] to generate higher order concepts which are later used for ad-hoc retrieval. In this poster, we propose an alternative implementation of the conceptual space by using a probabilistic HAL space (pHAL). To evaluate whether converting to such an implementation is beneficial we have performed an initial investigation comparing the concept combination of HAL against pHAL for the task of query expansion. Our experiments indicate that pHAL outperforms the original HAL method and that better query term selection methods can improve performance on both HAL and pHAL.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval —*Retrieval Models*

General Terms

Theory, Experimentation

1. CONCEPTUAL SPACES FOR IR

A Conceptual Space is a model of cognition that views symbolic processing on three levels; Symbolic, Conceptual and Associationist (see [4] for full details). In the context of Information Retrieval, Song and Bruza[6, 2] offer an implementation of the Conceptual Space using the Hyperspace Analogue to Language[3]. Concepts within this HAL space are combined (concept combination) and important concepts are selected as query expansion terms[1]. The intuition underlying the HAL space is that when a human encounters a new concept they derive its meaning from accumulated experience in the context in which the concept appears[2]. The meaning of a concept can be inferred from its usage with other concepts within the same context.

HAL The HAL Space is constructed automatically from a high dimensional semantic space over a corpus of text[3], and is defined as follows: each term t in the vocabulary T is composed of a high dimensional vector over T, resulting in a $|T| \times |T|$ HAL matrix, where $|T|$ is the number of terms in the vocabulary. A window of length K is moved across the corpus of text at one term increments ignoring punctation, sentence and paragraph boundaries. All terms within this window are said to co-occur with the first term in the window with strengths inversely proportional to the distance between them. The weighting assigned to each co-occurrence of terms is accumulated over the entire corpus. The HAL weighting for a term t and any other term t' is given by:

$$HAL(t'|t) = \sum_{k=1}^{K} w(k)n(t,k,t') \qquad (1)$$

where $n(t,k,t')$ is the number of times term t' occurs a distance k away from t, and $w(k) = K - k + 1$ denotes the strength of relationship between the two terms given k. Concept combination is performed through a series of steps, which attempt to mimic the actual cognitive process according to Information Flow Theory[1]. For instance, if term t co-occurs with both query term q_1 and q_2, then the weight for t is doubled (see [6] for full details of concept combination). After concept combination a vector over the vocabulary defines the weighting assigned to each term. In [2], they then convert this vector into a probability distribution by taking its norm, and hence a probabilistic HAL Space. Then, the top η most weighted terms are used as the expanded query. The entire process is rather ad-hoc in nature, but it provides a novel pre-retrieval mechanism for query expansion.

Instead, we propose a fully integrated probabilistic alternative and compare the performance with the original HAL space within the language modeling framework. A major advantage of using our probabilistic HAL space is that a clear and intuitive interpretation of the space is obtained as the probabilities represent the degree of co-occurrence. This also allows us to employ more theoretically principled query term selection methods. Furthermore, it can be naturally used within the Language Modeling framework using the KL divergence measure.

pHAL As suggested in [5] a probabilistic term co-occurrence matrix can be defined by normalizing the count of terms. We extend this approach to represent the HAL space by encoding the prior $p(k)$ to denote strength of the co-occurrence

[*]This author is partly funded by Memex Technology Ltd. (www.memex.com)

[1]It is assumed that the concepts in the space are equivalent to terms.

[2]HAL was originally developed as a representational model of semantic memory.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

between terms given k. The pHAL space is defined as:

$$p_h(t'|t) \;=\; \sum_k p(k)p(t'|t,k) \qquad (2)$$

where $p(t'|t,k) = \frac{n(t,k,t')}{\sum_t n(t,k,t')}$. By marginalizing over all possible values of k, we obtain the conditional probability $p_h(t'|t)$. The pHAL vector is re-normalized to ensure that it is a proper probability distribution. This is required because there may not be any occurrences of t and t' for a particular k, however smoothing techniques could be employed such as those in [7].

To build a query model θ_Q, from which to select query terms, we use a simple mixture model to combine terms. The dominance of each query term is encoded using the prior λ_i, such $p(t|\theta_Q)$ is:

$$p(t|\theta_Q) = \lambda_0 p(t|Q) + \sum_{i=1}^{l} \lambda_i p_h(t|q_i) \qquad (3)$$

where $p(t|Q)$ is the maximum likelihood estimate of the probability of term t given the query Q. The constraints $\sum_{i=0}^{l} \lambda_i = 1$ where $\lambda_i > \lambda_{i+1}$ and so on for all i are imposed. λ_i denotes the dominance of the ith term within the query according to the its I.D.F value[6]. Note Equation 3 strips most of the cognitive rationale away from the original combination process, and is relatively naive in comparison. Hence, we apply a boolean filter where terms are kept if they co-occur in with at least two of the query terms, and then we employed the log likelihood ratio (LLR)[5] to determine whether there was any statistical association between the terms in the query model. Query terms were drawn from the subset which were significantly associated. We shall refer to this query model as pHAL-2, and without the boolean condition as pHAL-1.

2. EMPIRICAL STUDY

To compare whether similar performance could be obtained by the new conceptual space we performed the follow experiment: We indexed 40000 documents from the Wall Street Journal Collection, where we removed standard stop words and applied Porter Stemming ($|T|$= 30239). The HAL and pHAL space was constructed with window size from one to five[3] We used two different weighting schemes; Original $w(k) = p(k) = \frac{K-k+1}{2(\sum_{k'}(K-k+1))}$ and uniform, $w(k) = p(k) = \frac{1}{2K}$. The suggested parameter values for combining concepts within the HAL space were used [6] and we shall refer to this model as HAL-0.

The titles of the TREC Topics 101-150 provided the initial terms to perform concept combination. And the top 5 to 100 terms were selected to submitted as a query. All querying was then performed using a standard language modeling approach with Bayes smoothed document language models[7].

3. RESULTS AND DISCUSSION

The results achieved from the best configurations of different (p)HAL spaces are shown in Table 3. All (p)HAL methods significantly outperformed the baseline of no expansion (* Wilcoxon Rank Sum Test at 5% significance).

	Baseline	HAL-0	pHAL-1	pHAL-2
mAP	25.3%	26.3%*	27.0%*	**27.9%***

Table 1: The mean Average Precision (mAP) from query expansions derived from the Conceptual Spaces. Results shown are when $K = 5$ and $\nu = 100$ given the (p)HAL space.

Empirically, we found that: (1) the best IR performance was obtained at highest window ($K = 5$), though as the window size increased there was a diminishing return on performance; (2) that queries of length 85-100 returned the best performance; (3) the uniform weighting function applied to either HAL or pHAL had a slight improvement to performance and (4) that using the LLR to select expansion terms also ameliorated the performance. Both (1) and (2) confirm past findings, whilst (3) and (4) show that greater improvements may be obtain through more appropriate weighting functions or selection techniques. Indeed, when we applied the later two techniques to HAL-0, the performance improved to 27.4% mean Average Precision[4].

We have independently replicated the HAL space for IR proposed by Song and Bruza[6] confirming that significant increases in performance can be achieved through employing the concept combination process. Furthermore, we have shown that our probabilistic interpretation of the HAL Space delivers comparable IR performance. This motivates further research and development into a fully integrated probabilistic variant of the Conceptual Space by including the Information Flow Component.

4. REFERENCES

[1] J. Barwise and J. Seligman. *Information Flow: The Logic of Distributed Systems*. Number 44 in Cambridge Tracts in Theoretical Computer Science. 1997.

[2] P. D. Bruza and D. Song. A comparison of various approaches for using probabilistic dependencies in language modelling. In *The 26th ACM SIGIR*, pages 419–420. ACM Press, 2003.

[3] B. C., K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25(2-3):211–257, 1998.

[4] P. Gardenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[5] W. Lowe and S. McDonald. The direct route: Priming in semantic space. In *Proceedings of the Seventeen Annual Meeting of the Cognitive Science Society*, pages 600–665, 2000.

[6] D. Song and P. D. Bruza. Discovering information flow using a high dimensional conceptual space. In *The 24th ACM SIGIR*, pages 327–333, New Orleans, LO, 2001.

[7] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *The 24th ACM SIGIR*, pages 49–56, New Orleans, LO, 2001.

[3] We were restricted to an upper limit to 5 due to memory limitations, however since we used a bi-directional window, the effective window size was 10).

[4] A technical report with full experimental details is available from http://cis.paisley.ac.uk/research/reports/