Query Expansion Using Random Walk Models

Kevyn Collins-Thompson Jamie Callan
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA 15213-8213
{kct, callan}@cs.cmu.edu

ABSTRACT

It has long been recognized that capturing term relationships is an important aspect of information retrieval. Even with large amounts of data, we usually only have significant evidence for a fraction of all potential term pairs. It is therefore important to consider whether multiple sources of evidence may be combined to predict term relations more accurately. This is particularly important when trying to predict the probability of relevance of a set of terms given a query, which may involve both lexical and semantic relations between the terms.

We describe a Markov chain framework that combines multiple sources of knowledge on term associations. The stationary distribution of the model is used to obtain probability estimates that a potential expansion term reflects aspects of the original query. We use this model for query expansion and evaluate the effectiveness of the model by examining the accuracy and robustness of the expansion methods, and investigate the relative effectiveness of various sources of term evidence. Statistically significant differences in accuracy were observed depending on the weighting of evidence in the random walk. For example, using co-occurrence data later in the walk was generally better than using it early, suggesting further improvements in effectiveness may be possible by learning walk behaviors.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]

General Terms

Algorithms, Experimentation.

Keywords

Query expansion, pseudo-relevance feedback, random walk, semi-supervised learning.

1. INTRODUCTION

Associative models consider relationships between terms in addition to the terms themselves. They have been extensively considered and studied for information retrieval, e.g. by Bush [3], Stiles [29], van Rijsbergen [31] and Salton & Buckley [25] among many others. There are many lexical and semantic relations that may be considered for associating a pair of terms. For example: stemming, based on common morphology; synonymy, where aspects of meaning are shared; co-occurrence, in which both words tend to appear together; and general association, where a person is likely to give one word as a free-association response to the other.

Each relation may be thought of as an inference step, in which a source word v has some property $R(v)$, and a new word w can be inferred to have the property value $R(w)$ with probability $P_R(w|v)$, based on their shared relation. For example, if v is the word 'matrix' and the property $R(.)$ is 'relevancy to a query', then one possible way to calculate $P_R(w|v)$ is based on co-occurrence, so that, for example, the term 'row' also has some measure of relevance. Note that this is not symmetric: 'row' having more senses and being more common, it is less likely to imply relevance of 'matrix', unless another term is also present for context, such as 'column'.

While lexical and semantic relations may be useful individually, it is important to consider how they may be used in combination. One reason for this is a common problem in language processing called sparsity: for co-occurrence relations for example, even with a huge corpus, we only have reliable co-occurrence data for a fraction of all potential term pairs. External semantic resources such as WordNet or stemming dictionaries supply a broad set of terms but are limited in the depth and currency of their vocabulary. By combining multiple relations into chains of inference, we can help bridge the gaps that exist in the data.

A second reason is that the various relations between words represent potentially complimentary sources of evidence that may help to distinguish and disambiguate terms. For example, if 'bank' and 'merger' are known to be relevant to a query, then the following inference

CIKM'05, October 31-November 5, 2005, Bremen, Germany.

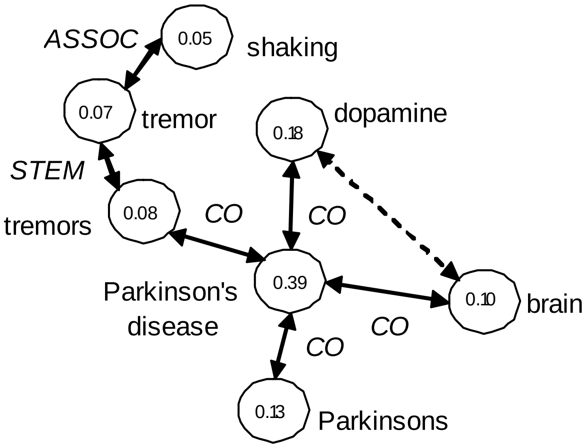

Figure 1. Simplified example of a stationary distribution induced by a random walk starting at the node ‘Parkinson’s disease’. The walk uses co-occurrence (CO) relations in early steps, then shifts to stemming (STEM) and association (ASSOC) for later steps in the walk. Values inside the nodes are example probabilities from the stationary distribution.

chains would provide evidence that ‘negotiations’ may also be relevant:

1. bank → agreement (C) → negotiate (C) → negotiations (M)
2. merger → talks (C) → negotiations (S)

where C, S, and M represent co-occurrence, synonymy, and morphology relations respectively. Note that chains can emphasize different types of evidence at different walk stages. In the above example, co-occurring terms are found first, followed by their synonyms or stems.

In this paper we propose and evaluate a Markov chain-based framework for modeling term relations that can perform such combination of behavior and apply this model to query expansion. Given a small set of initial query terms, we construct a term network and use a random walk to estimate the likelihood of relevance for potential expansion terms. The features used by the random walk can come from a variety of sources, such as term co-occurrence in an external corpus, co-occurrence in the top retrieved documents, synonym dictionaries, general word association scores, and so on.

Unlike many previous related models used for information retrieval, we define a much richer set of potential walk behaviors that support a variety of link types, where different combinations of evidence can be used at different stages of the walk. For example, co-occurrence may initially be given higher weight early in the walk, with synonyms weighted more highly in later steps. We also do not use a pre-defined network for all queries, but customize each network for each query.

We apply our model to the problem of query expansion in the language modeling approach to information retrieval. By estimating the probability that the various aspects of the query can be inferred from a potential expansion term, we essentially perform a form of ‘semantic smoothing’ of the query language model.

The main hypothesis of this paper is that combining query-specific term dependencies from multiple sources can lead to more accurate and/or robust expansion algorithms.

2. A MARKOV CHAIN FRAMEWORK FOR QUERY EXPANSION

The general motivation for using a Markov chain on a network of terms is that we want to infer a particular property (the *label*) of a target word given a set (usually small) of labeled source words.

In the case of query expansion, the target words are potential expansion terms, the source words are query terms, and the labels are probabilities of relevance. We then define a random process to propagate the label information through the graph. The stationary distribution of this process gives us a probability distribution over expansion terms. Figure 1 shows a portion of a term network for the query ‘Parkinson’s disease’. Solid connections denote explicit term associations, while the dashed line shows an implied connection inferred between ‘brain’ and ‘dopamine’ based on a short chain through the shared node ‘Parkinson’s disease’. We extend earlier work by Lafferty and Zhai [16] on using Markov chains for query expansion, by using a more flexible family of random walks similar to that described in Toutanova et al. [30], whose terminology we follow here.

2.1 Multi-stage Markov chain model

Let $\boldsymbol{W} = \{w_i\}$ be a vocabulary set of words. The relationship between words w_i and w_j is modeled as a combination of directional *links*, represented by *link functions* $\lambda_1, \ldots, \lambda_m$ Each link function $\lambda_m(w_i, w_j)$ represents a specific type of lexical or semantic relation between w_i and w_j, such as synonyms, stems, co-occurrence, and so on. Details on the specific link functions we used here are given in section 2.1.

We imagine a generative process where an author U creates a document of length N as follows:

Step 0: Choose an initial word w_0 with probability $P(w_0 \mid U)$ (If we have already generated N words, stop.)

Step i: Given we have chosen w_{i-1}, then with probability 1-α output the word corresponding to w_{i-1} and reset the process to step 0. Otherwise, with probability α sample a new word w_i according to the distribution:

$$p(w_i|w_{i-1}) = \frac{1}{Z}\exp\left(\sum_{m=0}^{L}\theta_m(i)\lambda_m(w_i, w_{i-1})\right)$$

where Z is the normalization quantity. This conditional probability may be interpreted as a mixture model in which a particular link type $\lambda_m(.)$ is chosen with probability $\theta_m(i)$ at timestep i. Note that the mixture is allowed to change at each timestep. For simplicity, we limit the number of such changes by grouping the timesteps of the walk into three *stages*: early, middle, and final. The function $\Gamma(i)$ defines how timestep i maps to stage s, where s is a value in {0, 1, 2}, and we now refer to $\theta_m(s)$ instead of $\theta_m(i)$.

Suppose we now have a query q consisting of the set of words $\{q_i\}$. For each link type $\lambda_m(.)$ we define a transition matrix $C(q,m)$ based on the query q. The reason q influences the transition matrix is that some link types, such as co-occurrence on top retrieved documents, are query-specific. Each stage s for a query q has an overall transition matrix $C(q,s)$ as the mixture of the individual $C(q,m)$:

$$C(q, s) = \sum_{m=1}^{M} \theta_m(s) C(q, m)$$

Combining the stages over k steps into a single transition matrix, which we denote C^k, we have:

$$C^k = \prod_{i=0}^{k} C(q, \Gamma(i))$$

Then for a query term q_i, the probability that a chain reaches q_i after k steps, starting at word w is:

$$p_k(q_i|w) = (1-\alpha)\alpha^k C^k_{w, q_i}$$

where C^k_{w, q_i} denotes the (w, q_i) entry in the matrix C^k.

The overall probability $p(q_i|w)$ of generating a query term q_i given a word w is therefore:

$$p(q_i|w) = \sum_{k=0}^{\infty} p_k(q_i|w) = (1-\alpha)\left[\sum_{k=0}^{\infty} \alpha^k C^k\right]_{w, q_i}$$

To ensure the Markov chain has a unique stationary distribution and avoid being trapped in short loops, we add a special last-stage walk step that has uniform transition probability to any node in the graph. This is implemented by using the ‘background smoothing’ link type as the final walk stage.

The walk continuation probability α can be viewed as a penalty for long chains of inference. In practice, we use a small number of steps (up to 4) on a sparse representation of the adjacency graph to perform the random walk steps.

In section 4, we discuss the specifics of how this model is used for query expansion, and in particular how the probabilities $p(q_i \mid w)$ are used.

2.2 Link types

We chose to include the following variety of semantic and lexical link types for our experiments. Each link type has a corresponding link function $\lambda_m(w_i, w_j)$:

- **Synonyms (SYN):** From Extended Wordnet [18]. At the moment, this only captures synonym information. The synsets from the top 5 senses were used. Transition probabilities were based on the weight of the node divided by the number of outgoing senses.
- **Stemming (STEM)**: Stems of a term v were generated by finding all words with a prefix of 3 or more letters in common with v which stemmed to the same root as v, using the Krovetz stemmer [12]. For this study we used uniform transition probabilities for the stems.
- **General word association (ASSOC):** A human association factor of target word w_i given cue word w_j, from the South Florida Word Association database [19]. This database has a wealth of statistics about the association strength between cue words and target words. The database contains 5,019 normed words with their 72,176 responses. The transition weights used were taken directly from the database and were the Forward Cue-to-Target Strength and Backward Cue-to-Target Strength respectively.
- **Co-occurrence in a large general Web corpus (CWEB):** Based on a corpus of about 700,000 WikiPedia articles (as of January 2005), words are considered highly related if they are highly predictive of each other based on average mutual information. The query terms are used to retrieve an initial set of documents; a set of highly associated terms is extracted; and then these terms are added to the network and the process is continued for a small number of steps. Transition probabilities were based on renormalized MI scores.
- **Co-occurrence in the top retrieved documents (CTOP):** Similar to above, but only using the top retrieved documents in the local TREC collection, based on the initial query.
- **Background smoothing (SM):** With uniform probability, links any single word w_i to all other words w_j.

Our goal is to train the link weights from training data, but for this study we hand-coded the weights $\theta_m(s)$ as described in the evaluation in Section 5.

3. RELATED WORK

The Markov chain approach for modeling term assocations is related to previous models based on term clustering and spreading activation networks, both of which have a long history that will only be briefly summarized here.

Stiles [29] described heuristics for using sets of term associations in improved indexing, and later Quillian [23] proposed a semantic network of concepts for binary relations between words. Gotlieb and Kumar [8] devised a semantic clustering of index terms using maximal complete subgraphs in a term network, although their method's effectiveness was never evaluated for retrieval. Early work in term clustering for query expansion by Sparck Jones [28] focused on constructing a similarity matrix of single index terms before any user queries were submitted. Wong and Raghavan [33] proposed the use of a matrix of term-term associations in ranking document vectors against query vectors, focusing on the special case of term correlation based on co-occurrence. Salton and Buckley [25], van Rijsbergen [31], and many others explored organizing of associations into networks for expanding the search vocabulary. These networks used various node activation heuristics and decay rules that were intuitively plausible but had limited retrieval success. Crestani [5] gives a summary of earlier work on spreading activation networks.

Stationary distributions have been used previously in information retrieval for 'influence weighting' schemes such as PageRank [1] and hub-authority [10][11], and also for query expansion [16]. Lafferty and Zhai considered a bipartite graph on query terms and documents [16] and calculated an approximate stationary distribution using a random walk. In their scheme, the random walk was defined in terms of words and documents, not words only. Our local co-occurrence link is calculated in a similar way, but our random walk framework is more general in that we can use multiple sources of lexical and semantic evidence, not just co-occurrence, with the potential to weight these sources differently at different stages of the walk.

With regard to other query expansion approaches, the idea that we present below of rewarding expansion terms reflecting multiple aspects of the original query was previously noted by Xu and Croft [34] for Local Context Analysis (LCA), which has shown good empirical performance. Their method uses an empirically derived formula to score potential expansion terms that is similar in effect to our probabilistic term scoring. A number of studies have used external resources for query expansion. For example, Voorhees used Wordnet with limited success [32]. Shah and Croft [26] used Wordnet synonyms to perform query expansion for high-precision retrieval, selecting terms with high clarity scores.

In contrast to many early spreading activation systems, the Markov chain approach to query expansion is relatively simple and offers a well-motivated probabilistic framework that fits well within the language modeling approach to information retrieval. Moreover, its close relationship with semi-supervised learning [27] means that we may make use of insights from that area to help illuminate the nature of query expansion and learn more robust expansion algorithms.

4. QUERY EXPANSION MODEL

To index and search the collection we used Indri [17], a new search engine in the Lemur toolkit [20]. Indri combines a language modeling approach with inference networks and supports an extended set of probabilistic structured query operators based on INQUERY [4].

4.1 Baseline expansion algorithm

For our baseline we chose an algorithm supplied with Indri. This hybrid method selects terms using a method described by Ponte [22], but assigns final term weights using Lavrenko's relevance model [14]. Specifically, a log-odds ratio is calculated for each potential expansion term *w* by calculating the log-odds ratio over all documents *D* containing *w*, with the document coming from collection *C*:

$$o(w) = \sum_{D} \log \frac{p(w|D)}{p(w|C)}$$

Next, the expansion candidates are sorted by descending $o(w)$, and the top *k* are chosen. Finally, the term weights $r(w)$ used in the expanded query are calculated based on Lavrenko's relevance model. A mu factor of 1000 is used for the Dirichlet smoothing of $p(w|D)$ in the relevance model:

$$r(w) = \sum_{D} p(q|D)p(w|D)\frac{p(w)}{P(D)}$$

The quantity $p(q|D)$ is the probability score assigned to the document in the initial retrieval. This expansion method appears competitive with other systems in practice on the same TREC collections [17].

The baseline unexpanded query for each topic used not only the original title terms, but also likely phrases, as determined with the Link Parser [15]. All terms were then combined with Indri's #combine operator. For example a typical baseline title query is:

```
#combine (ireland peace talks
#1(peace talks) #1(ireland peace talks) )
```

The query was expanded by adding a weighted combination of the expansion terms, with the original and expanded query weighted equally. For example:

```
   #weight( 0.50 #combine( ireland peace
   talks  #1(peace  talks)  #1(ireland
   peace talks) )

   0.50 #weight(
   0.00005345596124665 ireland
   0.00004102199003243 peace
   0.00004402604958434 talks
   0.00000954590561999 adams
   0.00001945337805346 fein
   0.00000204002264506 hume
   0.00002451393911871 ira
   0.00000893787186589 unionist
   0.00001955466404921 sinn
   0.00000422895825826 reynolds  ) )
```

4.2 Aspect-based expansion

We start with this hypothesis: a desirable property of good expansion terms is that they somehow reflect one,

ireland		peace		talks		peace talks		ireland peace talks	
w	$log\ p(A\|w)$	w	$log\ p(A\|w)$	w	$log\ p(A\|w)$	w	$log\ p(A\|w)$	w	$log\ p(A\|w)$
ireland	-0.01	peace	-0.01	talks	-0.01	negotiating	-5.1703	unionist	-8.0729
irelands	-6.1985	wing	-2.8843	adams	-5.8884	struggle	-6.739	ulster	-8.9783
republic	-8.5028	initiative	-4.2204	say	-6.0775	inquiry	-6.7404	peace	-9.0689
claim	-9.2827	commitment	-6.0857	referendum	-6.9954	talks	-9.1465	talks	-10.147
collapse	-9.3663	unionists	-8.9894	monday	-7.4853	fein	-9.5991	party	-11.133
unionist	-9.5016	struggle	-9.3153	inquiry	-7.5051	ira	-9.7275	progressive	-11.591
unionists	-9.6156	create	-9.5097	present	-8.5892	ulster	-10.523	adams	-12.583
june	-9.6906	patrick	-9.8341	unionist	-8.7776	dublin	-10.544	referendum	-13.684
catholic	-9.7521	having	-9.8575	patrick	-9.044	sinn	-10.945	inquiry	-14.189
omalley	-9.7704	fein	-9.9189	business	-9.2421	irelands	-10.959	monday	-14.553

Table 1. Sample aspect model log-probabilities using the multi-stage Markov chain method. The terms *w* are expansion candidates selected from the top 10 retrieved documents for topic 404. The terms are sorted by their log-likelihood 'distance' from the various query aspects shown in the heading. The random walk used link types CTOP+CWEB, (2 steps) and ASSOC+SYN+STEM (1 step).

and preferably more, aspects of the query. Xu & Croft [34] used this idea in their work on LCA.

Let $\boldsymbol{A}$ be a set of aspects associated with a query q. In our model, an aspect A_i in $\boldsymbol{A}$ is represented by one or more sets of words $A_i=\{t_j\}$ taken from the query. Assuming exchangeability of aspects, and of words within aspects, to evaluate a potential expansion term v, we calculate:

$$p(\boldsymbol{A}|v) = \prod_{A_i} p(A_i|v) = \prod_{A_i} \prod_{t_j \in A_i} p(t_j|v)$$

Taking logarithms:

$$\log p(A|v) = \sum_{A_i} \sum_{t_i \in A_i} \log p(t_j|v) = a(v)$$

The Markov chain model defined in section 2 now provides us with a method for estimating $p(t_j|v)$, the probability that the expansion term v will generate an aspect term t_j of the query. The value log $p(t_j|v)$ may be thought of as a semantic distance. An example of the language models for various query aspects generated from the Markov chain distribution is given in Table 1. Note that a term like 'unionist' may have high scores for many aspects but rank lower in the final expansion selection because its probability in the top documents is slightly lower than that of other good expansion terms.

Expansion terms are chosen by discounting the original log-odds with the combined aspect log-probability:

$$n(v) = \frac{o(v)}{a(v)}$$

This has the effect of rewarding terms that are closely related to the main aspects of the query, even if they may be less rare the collection than other expansion terms. The terms are sorted by $n(v)$ and the top k are chosen. The term weight assigned to term v in the expanded query is just a rescaled version of $n(v)$.

In addition to having expansion terms that reflect multiple query aspects, we also want documents that reflect all aspects of the query, not just a subset. An analysis at the RIA workshop [2] showed that a significant number of retrieval failures could be attributed to incomplete aspect coverage by the retrieval model. Kekäläinen and Järvelin [9] showed that one of the most effective structured query operators for query expansion was the probabilistic AND operator, in combination with maximally expanded query aspects. We therefore modified the expansion formula from the baseline to use #wand instead of #weight in the expanded query portion to combine the aspect-based expansion terms. The final query looks like:

```
#weight( 0.5 #combine( ireland peace
talks  #1(peace  talks)  #1(ireland
peace talks) )
0.5 #wand(
1.0211577500 ireland
0.4062418676 talks
0.2464178510 peace
0.1810290583 ira
0.1322647130 struggles
0.0740201537 armed
0.0442171104 political
0.0369219708 dominance
0.0279126814 continuation
0.02295469100 obstacles) )
```

5. EVALUATION

We examined the performance of our Markov chain-based expansion in three ways. First, we compared retrieval statistics to the Indri baseline and previously published results for the same topics and collections. Second, we compared different versions of the random walk that used different weightings of the evidence. Third, we compared the robustness of the expansion methods to Indri baseline expansion.

Our experiments are based on three different TREC datasets: the AP89 collection (topics 1-50), the TREC8 ad-hoc collection (disks 4&5 minus CR, topics 401-450), and the TREC 2001 wt10g (topics 501-550). These were chosen to vary the style and amount of content. All queries here use the "title" field of TREC topics only. In order to test the effectiveness of our modeling techniques we did not perform stemming.

Run	AP89 1-50		TREC8 401-450		TREC2001 501-550	
	MAP	*P10%*	*MAP*	*P10%*	*MAP*	*P10%*
No expansion	0.2143	0.3393	0.2807	0.4090	0.2095	0.3687
External w/ expansion	0.232 [16]	-	0.3063 [13]	-	0.2028 [24]	-
Indri expansion	0.2222	0.3283	0.2877	0.4344	0.2167	0.3436
Markov A: CWEB	0.2190	0.3003	0.2895	0.4631	0.2156	0.3485
Markov C: CTOP	0.2015	0.3198	0.2822	0.4555	0.2157	0.3347
Markov D: CTOP + [ASSOC, SYN, STEM]	0.1978	0.3078	0.2882	0.4598	0.2174	0.3412
Markov E: [CTOP, CWEB] +[ASSOC, SYN,STEM]	0.2105	0.3122	0.2935	0.4652	0.2273	0.3300
Markov F [ASSOC, SYN, STEM] + CTOP	0.2286	0.3153	0.2942	0.4637	0.2270	0.3445

Table 2. Comparison of expansion algorithm effectiveness.
(Title queries, top 50 terms from top 5 retrieved documents)

The main link types used in the random walk can be divided into two broad groups: links using co-occurrence data (CWEB, CTOP) and associative links (ASSOC, SYN, and STEM). To make it easier to compare the relative effect of these two groups, and to simplify our experiments, we kept these groups separate during different steps: a walk could use either an 'associative' step that combined all associative types, or a 'co-occurrence step' using co-occurrence data only. (See section 2.1 for the definition of link type names.)

The runs we chose are listed below. Links in square brackets indicate an equal mixture during a walk step, and the number in parentheses gives the maximum number of walk steps.

- **Markov A**: CWEB (3 steps)
- **Markov C**: CTOP (3)
- **Markov D**: CTOP (3) + [ASSOC, SYN, STEM] (1)
- **Markov E**: [CTOP, CWEB] (3)
  + [ASSOC, SYN, STEM] (1)
- **Markov F**: [ASSOC, SYN, STEM] (3) + CTOP (1).

We set the walk continuation probability $\alpha = 0.8$.

The top 5 documents retrieved, and top 50 expansion terms, were used for the expansion since this tended to give superior performance for both the baseline and our method. Significance testing was performed using the Wilcoxon matched-pair signed-ranks test.

The results for these runs are shown in Table 2. The External row gives comparative results for high-performing external systems on the same topics.

5.1 Comparing Markov chain with baselines

When measured by mean average precision (MAP), the best Markov chain results and Indri's baseline expansion were comparable for the 3 collections, with no statistically significant differences. For precision at 10% recall (P10%), a gain of 7% (significant at the 0.05 level) was obtained on TREC-8 using Markov E.

Compared to previous top external results (the External row), the Markov chain results were slightly better than the Okapi run at TREC 2001 [24], but slightly lower than the results for AP 1-50 reported by Lafferty & Zhai's bipartite Markov chain method [16]. Our best TREC-8 run, with MAP of 0.2942, was slightly lower than the 0.3063 score of one of the top TREC runs [13].

We examined results on 4 RIA 'failure topics' requiring coverage of multiple query aspects (355, 363, 372, 422). Markov expansion helped substantially for topic 372 (*Identify documents that discuss the growth of Native American casino gambling*), for which both 'gambling' and 'Native American' aspects needed to be present. Markov E obtained a MAP of 0.4621, a 30.8% improvement over the Indri baseline of 0.3532: slightly better than the best RIA system score of 0.4603, and far better than the median RIA MAP of 0.1607. The Markov E query was distinguished by its high weighting of several terms closely related to both aspects, such as specific tribes engaged in casino-building ("Pequots"). MAP for the other three topics was comparable to the Indri baseline, and more study is needed to understand when the aspect coverage of Markov expansion is most effective.

Local only	Web only (Wikipedia)	Combined
steel	steel	steel
production	production	production
iron	iron	iron
year	industry	***year***
industry	output	industry
output	**tonnes**	output
crude	consumption	**tonnes**
capacity	crude	crude
prices	prices	capacity
consumption	demand	prices
europe	capacity	consumption
krupp	market	***krupp***
market	drop	***europe***

Table 3. Comparison of top expansion terms for topic 413: 'Steel production' using local, global, and combined co-occurrence data. Words specific to each type of distribution are emphasized.

5.2 Early vs. late co-occurrence relations

We compared the Markov D and Markov F runs, which have identical walk parameters except that co-occurrence relations are emphasized late in the walk for Markov F, and early for Markov D. There was an improvement in MAP (in the case of AP89, more than 15%) when co-occurrence was used late in the random walk. This was true across all three collections. The reasons for this require further study, but it suggests that how the evidence is weighted by time-step does indeed matter. In this case, terms that co-occur with terms semantically close to the query appear to be more valuable than terms semantically close to many potential co-occurrence terms.

5.3 Local vs. Web co-occurrence evidence

After examining expansion terms for the various runs, we noted that the Wikipedia tends to act as a background 'topic' model for the query by emphasizing more general terms, while the local co-occurrence data acts to provide additional details on top of the topic model that reflect the corpus style and details, including specific names and places. As shown in Table 2, the addition of the Wikipedia evidence (Markov E) was marginally more effective than local evidence (Markov D).

A representative comparison of expansion terms is given in Table 3. While the lists tend to be fairly consistent in this case, the CTOP terms lean toward news-like terms that mention times, people and places, while the CWEB terms are more generic.

5.4 Robustness of expansion algorithms

We hypothesized that even in the cases where the overall accuracy of the Markov expansion algorithm was similar to existing methods, the bias in favor of adding more general but related terms would reduce the likelihood of query drift caused by choosing an off-topic term, which in turn would result in more robust expansion. The trade-off is that a related, more general term may also be less likely to significantly increase precision.

	AP89	TREC-8	TREC 2001
# queries with higher relative loss	Indri: 7 Markov: 5	Indri: 11 Markov: 4	Indri: 7 Markov: 6
# queries with higher relative gain	Indri: 7 Markov: 7	Indri: 6 Markov: 5	Indri: 7 Markov: 1

Table 4. The distribution of queries with significantly higher relative gain or loss after expansion, for both Indri and Markov chain expansion methods. (Queries where the difference was less than 1% were discarded.)

The data in Table 4 suggest this is actually happening. We examined all queries that satified one of two cases: 1) where expansion hurt for both methods and 2) where it helped for both methods, as measured by relative change in MAP. We compared the size of the relative error in both cases. For queries where expansion hurt, the Markov chain expansion had consistently more queries with lower relative error for all three collections (based on the Markov A run). Conversely, the Indri expansion method had consistently more queries with higher relative gains in cases where expansion helped.

5.5 Efficiency

While the number of potentially active term-nodes in the network is large (our vocabulary size was around 300,000), the Markov transition matrices are very sparse: a typical matrix has about 30,000 non-zero entries. Limiting a walk to a maximum of 4 steps typically results in less than 3000 (1%) of potential nodes becoming active for title-length queries. Even so, using the network efficiently requires some planning. For example, the first time a word is seen, its score is cached since the aspect probabilities $p(t_j|v)$ will not change over the lifetime of the query. The time to build the network can be reduced by doing off-line indexing to precompute a language model for each article.

6. DISCUSSION

The idea of using a spreading activation network on a network of terms has been re-discovered many times. While the intuition behind these heuristics was sound, there was limited understanding of the objective effect of these rules, for example, in terms of statistical language models. Other factors in the poor performance of spreading activation models may have been lack of training data, especially large, diverse external language resources like the Wikipedia that only recently have come into existence. Furthermore, the rules governing these past models were fairly rigid and did not generalize well, especially when the same network was applied to any query. Our model is a first step in a principled exploration of the properties that a flexible semantic kernel [6] should have to be most effective for query-specific tasks like relevance estimation.

7. CONCLUSIONS

We described a Markov chain model that allows chaining of multiple inference steps with different link types to perform "semantic smoothing" on language models, and applied this model to query expansion. A query is modeled as a combination of aspects, and expansion terms are favored that are not only more rare relative to the collection, but also semantically close to multiple query aspects. Our framework supports a richer set of potential behavior than past models, such as early, mid-, and late-stage variation in walk behavior, and arbitrary link weights.

Our initial results show that this model is comparable with the best results from other methods and can give modest improvements in precision, accuracy, and robustness for some test sets. Statistically significant differences in accuracy were observed depending on the weighting of evidence in the random walk. For example, using co-occurrence data later in the walk was generally better than using it early. This suggests that further improvements in accuracy are likely with more study of learned walk behaviors.

8. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their comments. This work supported by U.S. Dept. of Education grant R305G03123. Any opinions, findings, conclusions, or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

9. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proc. of 7th Int. WWW Conference,* pg. 107–117, 1998.

[2] C. Buckley. Why current IR engines fail. *SIGIR 2004,* pg. 584-585, 2004.

[3] V. Bush. As We May Think. *Atlantic Monthly*, July 1945.

[4] J. P. Callan, W. B. Croft and S. M. Harding, The INQUERY retrieval system, *DEXA-92*, pg. 78-83, 1992.

[5] F. Crestani, Application of spreading activation techniques in information retrieval. *AI Review* 11(6), pg. 453-482, 1997.

[6] N. Cristianini, J. Shawe-Taylor, H. Lodhi. Latent semantic kernels. *J. Intelligent Info. Systems*. 18(2-3), 127-152, 2002.

[7] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. Indexing by latent semantic analysis. *Journal of American Society for Information Science*. 41:391-407, 1990.

[8] C. C. Gottlieb, S. Kumar. Semantic clustering of index terms. *JACM* 15(4): 493-513, 1968.

[9] J. Kekäläinen, K. Järvelin, The impact of query structure and query expansion on retrieval performance. *SIGIR 1998,* pg. 130-137, Melbourne, Australia, 1998.

[10] J. Kleinberg. Authoritative sources in a hyperlinked environment. *JACM*, 46:604–632, 1999.

[11] J. Kleinberg and A. Tomkins. Applications of linear algebra in information retrieval and hypertext analysis. *18th ACM Symp. on Principles of Database Systems,* pg. 185-193, 1999.

[12] R. Krovetz. *Word sense disambiguation for large text databases*. Doctoral thesis, Univ. of Mass., Amherst, 1995.

[13] K.L. Kwok, L. Grunfeld, M. Chan. TREC-8 Ad-Hoc, Query and Filtering Track Experiments using PIRCS, NIST special publication, pg. 217, 2000.

[14] V. Lavrenko. *A generative theory of relevance*. Doctoral thesis, University of Mass. at Amherst. 2004.

[15] J. Lafferty, D. Sleator, and D, Temperley. Grammatical trigrams: a probabilistic model of link grammar. *Proc. of AAAI Conf. on Probabilistic Approaches to Natural Language*, Oct. 1992.

[16] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. *SIGIR 2001*. pg. 111-119.

[17] D. Metzler, T. Strohman, H. Turtle, and W. B. Croft. Indri at Terabyte Track 2004. *TREC 2004,* NIST special publication.

[18] D. Moldovan and V. Rus. Explaining answers with extended WordNet, *Proc. of ACL 2001*, Toulouse, France, 2001.

[19] D. L. Nelson, C. L. McEvoy, & T.A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

[20] The Lemur Toolkit. http://www.lemurproject.org

[21] J. Ponte and B. Croft. A language modeling approach to information retrieval. *SIGIR 1998*. pg. 275-281, 1998.

[22] J. Ponte. *A language modeling approach to information retrieval*. Doctoral thesis, Univ. of Mass. at Amherst. 1998.

[23] M. Quillian. Word concepts: a theory and simulation of some basic semantic capabilities. *Behav. Sci.* 12: 410-430, 1967.

[24] S.E. Robertson, S. Walker, and H. Zaragoza. Microsoft Cambridge at TREC-10: Filtering and web tracks. *TREC-10.* NIST special publication, 2002.

[25] G. Salton and C. Buckley. On the use of spreading activation networks in automatic information retrieval. *SIGIR 1988.* pg. 147-160.

[26] C. Shah and W. B. Croft. Evaluating high accuracy retrieval techniques. *SIGIR 2004*. pg. 2-9, 2004.

[27] A.J. Smola and R. Kondor. Kernels and regularization on graphs. *Proc. of COLT 2003*, Eds. B. Schölkopf and M. Warmuth, Lecture Notes in Computer Science, Springer.

[28] K. Sparck Jones and E.O. Barber. What makes an automatic keyword classification effective? *JASIS* 22(3) 166-175, 1971.

[29] H. E. Stiles, The association factor in information retrieval, *J. ACM*, 0004-5411 8:2(271-279), 1961.

[30] K. Toutanova, C. Manning, and A.Y. Ng. Learning random walk models for inducing word dependency distributions. *Proc. of ICML 2004*.

[31] K. van Rijsbergen. *Information Retrieval* (2nd edition). Butterworths, London, 1979.

[32] E. M. Voorhees. Query expansion using lexical-semantic relations. *SIGIR 1994.* pg. 61-69.

[33] S. K. M. Wong and V. Raghavan. Vector space model of information retrieval - a re-evaluation. *SIGIR 1984*. 167-185.

[34] J. Xu and W. B. Croft. Query expansion using local and global analysis. *SIGIR 1996*. pg. 4-11.