Tapping into Knowledge Base for Concept Feedback: Leveraging ConceptNet to Improve Search Results for Difficult Queries

Alexander Kotov
akotov2@illinois.edu

ChengXiang Zhai
czhai@illinois.edu

Department of Computer Science
University of Illinois at Urbana-Champaign
201 N Goodwin Ave, Urbana, IL, 61801, USA

ABSTRACT

Query expansion is an important and commonly used technique for improving Web search results. Existing methods for query expansion have mostly relied on global or local analysis of document collection, click-through data, or simple ontologies such as WordNet. In this paper, we present the results of a systematic study of the methods leveraging the ConceptNet knowledge base, an emerging new Web resource, for query expansion. Specifically, we focus on the methods leveraging ConceptNet to improve the search results for poorly performing (or difficult) queries. Unlike other lexico-semantic resources, such as WordNet and Wikipedia, which have been extensively studied in the past, ConceptNet features a graph-based representation model of commonsense knowledge, in which the terms are conceptually related through rich relational ontology. Such representation structure enables complex, multi-step inferences between the concepts, which can be applied to query expansion. We first demonstrate through simulation experiments that expanding queries with the related concepts from ConceptNet has great potential for improving the search results for difficult queries. We then propose and study several supervised and unsupervised methods for selecting the concepts from ConceptNet for automatic query expansion. The experimental results on multiple data sets indicate that the proposed methods can effectively leverage ConceptNet to improve the retrieval performance of difficult queries both when used in isolation as well as in combination with pseudo-relevance feedback.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, query formulation, relevance feedback*

WSDM'12, February 8–12, 2012, Seattle, Washington, USA.

General Terms

Algorithms, Experimentation

Keywords

Query Analysis, Query Expansion, Knowledge Bases, ConceptNet

1. INTRODUCTION

The information needs of the searchers, concisely formulated as keyword queries, can greatly vary in complexity, which is determined by the number of concepts that constitute the need. For this reason, the quality of search results largely depends on how completely and accurately all the concepts constituting the information need are translated into the query terms. It is often the case that the users of Web search systems tend to minimize their effort by intentionally posing very short, under-specified queries. As a result, many documents representing certain aspects of the information need will be missing in the search results. In addition to that, synonymy gives rise to the problem of *vocabulary mismatch*, which occurs when the authors of relevant documents and the searchers use different terms to designate the same concepts. This problem arises particularly often when non-professional users perform domain-specific searches and are not closely familiar with the vocabulary of the domain of the search problem. The most common examples of such domains are legal and medical.

Query expansion is a standard technique allowing to mitigate the problems of differing vocabularies and partially specified information needs by selecting and adding the related terms and phrases to the initial query. The main difficulty in effective application of automatic query expansion lies in correct identification of underrepresented aspects of the information need and selecting the right expansion terms with the right weights. Typical sources of term associations for query expansion can be either static and exist at the time of query (such as the search logs, ontologies, encyclopedias, manual or statistical thesauri constructed from the corpus) or dynamic, such as the top-ranked initially retrieved documents, from which the expansion terms can either be selected automatically by the system through pseudo-relevance feedback (PRF) or by asking the users to designate the relevant documents through explicit relevance feedback. All approaches using the dynamic

sources of expansion terms rely on the assumption that the initially retrieved results include some relevant documents, which can be used as a source of expansion terms. It is often the case, however, that the top-ranked search results for a query include very few or no relevant documents and neither the search systems nor the users can communicate the positive relevance signals back to the search system through feedback mechanisms. Such queries are often referred to as *difficult* for a particular search system and underspecified queries as well as vocabulary mismatch are some of the main reasons for search systems failures. In most cases, the users are unaware of the underlying problems of poorly performing queries, the search systems currently offer no support to them in trying to improve the search results. Although many users are aware that the quality of search results can be improved by reformulating a query, finding the right query formulation can be a fairly difficult and time consuming process. While some of the static sources of expansion terms, such as the query logs and statistical co-occurrence thesauri constructed through the global collection analysis, allow to avoid the dependence on the quality of initial results, the coverage of these resources is limited and they may simply not contain effective expansion terms, which are broadly or conceptually related to a particular query.

In this work, we systematically and comprehensively explore the potential of *concept feedback*, a set of different strategies for leveraging ConceptNet [22] [1] as a source of expansion terms for difficult queries. ConceptNet is presently the largest commonsense knowledge base, consisting of more than 1.6 million assertions about the world. Similar to Wikipedia, ConceptNet reflects the "wisdom of the crowds" and was constructed by gathering a large number of sentence-like assertions about the real world from a large number of on-line collaborators. ConceptNet uses semantic network as a knowledge representation model. The nodes in its semantic network correspond to semi-structured natural language fragments (e.g., "food", "grocery store", "buy food", "at home") and represent the real world concepts. An edge between the two nodes represents a semantic relationship between the two concepts. A fragment of the concept graph of ConceptNet is shown in Figure 1. As opposed to ontologies, such as WordNet, ConceptNet is not limited to hyponym/hypernym relations and features a more diverse relational ontology of twenty relationship types, such as causal, spatial and functional. As opposed to on-line encyclopedias, such as Wikipedia, the network structure of ConceptNet does not require any additional analysis to establish the relations between the concepts.

The network-based structure of ConceptNet in combination with its rich relational ontology opens up possibilities for making more complex, multi-step inferences. For example, from Figure 1 it follows that the concepts "morning" and "stomach" are related via the following chain of inferences: "wake up in the morning"→"eat breakfast"→"full stomach". The key idea behind this work is that the network-based structure of ConceptNet can be leveraged to make similar complex inferences to identify the effective expansion terms that are *broadly* related to the query, when the initially retrieved results are of poor quality and, consequently, cannot be used as a source of expansion terms. Although other lexico-semantic resources, such as WordNet, can be used

[1] `http://conceptnet5.media.mit.edu/`

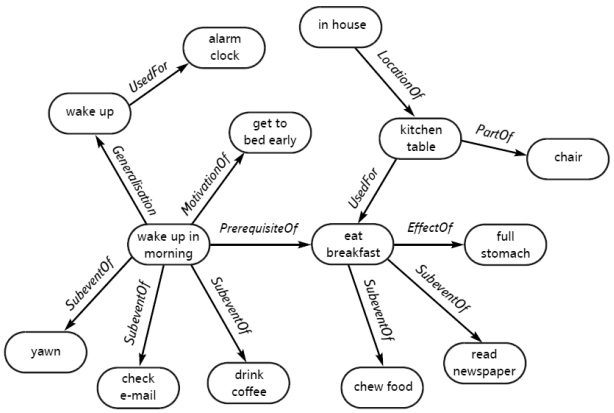


Figure 1: Fragment of the concept graph of ConceptNet (adopted from [22])

[23] [29] to help address the issue of vocabulary divergence between the queries and relevant documents, effective expansion terms may have much broader conceptual relations to the query terms than the tight semantic coherence of WordNet synsets may allow.

Concept	AP
target	0.2011
threat	0.1864
machine	0.1698
explosive	0.1684
tighten	0.1564
afghanistan	0.1559
worldwide	0.1495
chain	0.1464
measure	0.1462
link	0.1450

Table 1: Top 10 expansion concepts from ConceptNet for the TREC query 341 "airport security" along with the average precision (AP) of the query expanded with each of the concepts, when the expanded queries are run on the AQUAINT dataset. AP of the original query is 0.0657.

Figure 1 provides an example of the expansion concepts from ConceptNet identified using the simulation experiment presented in Section 4.2. As can be seen from this example, some effective expansion concepts have very broad and complex semantic relationships with the original query terms. Establishing such relationships may require making several inference steps, for which the graph-based knowledge representation model of ConceptNet is particularly well-suited. On the other hand, the hierarchical structure of WordNet and vector-space knowledge representation models for Wikipedia [14] present certain difficulties for making complex and multi-step inferences.

In this work, we explore different methods of leveraging the semantic network of ConceptNet to select a *small set of highly effective expansion concepts* to improve the performance of difficult queries. In particular, we address the following two research questions. The first question is whether ConceptNet can in principle be leveraged to improve the retrieval results of *difficult queries*? To answer this question, we conducted a simulation experiment, in which for each

query we measured the retrieval performance of that query extended with each single concept within a certain distance from its terms. The retrieval results of the best performing expansion concept for each query were used to determine the overall upper bound for the potential retrieval effectiveness of concept feedback on each data set. The results of this experiment are presented in Section 4.2. The second question is how to design the methods to automatically select a small number of effective expansion concepts? To answer this question, in Sections 3.1 and 3.2 we propose the heuristic and machine learning-based methods for selecting expansion concepts and present the results of an experimental evaluation of those methods in Section 4.

The main contribution of the present work is in systematic and comprehensive exploration of the heuristic and machine learning-based methods for selecting the query expansion terms from ConceptNet and comparing the effectiveness of those methods with the effectiveness of automatic query expansion based on pseudo-relevance feedback.

2. RELATED WORK

The problem of vocabulary mismatch is a fundamental problem in information retrieval. It has long been recognized that capturing lexico-semantic relationships can reduce the negative impact of this problem on the quality of retrieval results. Early attempts to utilize lexico-semantic relationships include introduction of the concept of associative retrieval. Associative retrieval is based on the association hypothesis formulated by van Rijsbergen [28], which states that "if an index term is good at discriminating relevant from irrelevant documents, then any closely associated index term is also likely to be good at this". In associative retrieval, knowledge about associations among information items (terms, concepts or documents) is represented as a network, in which the information items correspond to the nodes and associations between them to the links connecting the nodes. Constrained spreading activation [12] [26] [9] is a typical processing paradigm used in associative retrieval. The main difficulty faced by the early attempts to apply associative retrieval was a labor-intensive process of manual construction of the association network between the documents and concepts. As a result, the focus of query expansion research has shifted to using local and global document analysis to automatically extract the expansion terms. Early global expansion techniques [15] [33] [30] aim to determine the strength of semantic relatedness between the terms based on the term co-occurrence statistics obtained through the analysis of the entire document collection and use the terms that are strongly associated with the initial query terms for expansion. More complicated global-analysis models than those involving simple term co-occurrence statistics have been proposed later [5]. In our previous work [20] [19], we proposed *sense feedback* and *question-based feedback*, two complimentary collection analysis-based methods to interactively improve the performance of difficult queries, when the poor retrieval performance is caused by underspecified and ambiguous queries. Local collection expansion techniques [31] generally follow a two-stage process often called pseudo-relevance feedback. First, a query is issued to retrieve the initial set of results and then a certain number of the top-ranked documents is used to extract the expansion terms. Xu and Croft [30] proposed a local analysis method, in which the candidate expansion terms are ranked by their co-occurrence with the original query terms and weighted by a constant according to their rank. They also compared the performance of the local and global document analysis and concluded that the local analysis is generally more effective than the global analysis. However, a major deficiency of the local analysis is that it is based on the assumption that the initially retrieved results include at least some relevant documents, which is violated in case of difficult queries.

The emergence of the Web and hand-crafted general purpose or domain-specific ontologies provided access to the new sources of high quality term associations for query expansion. In particular, three major types of external resources have been explored: external corpora (including the Web), general-purpose and domain-specific ontologies and Wikipedia. For example, Diaz and Metzler [13] demonstrated in the context of relevance models that using a high quality external corpus that is comparable to the target corpus can be as, if not more, effective than using the Web for pseudo-relevance feedback. Yin et al. [32] proposed an expansion method based on using random walk on the query-URL graph generated from the web query logs and snippets provided by an external search engine. Their main assumption is that users submit various queries to express the same information need and, therefore, the query can be expanded using related query formulations. Several researchers have experimented with heuristic methods based on WordNet and mixed results have been reported. Voorhees [29] experimentally determined the upper bound for the effectiveness of different WordNet-based query expansion strategies by manually choosing the query terms for expansion and annotating the query topics with the WordNet synsets. The reported results indicate that while query expansion makes little difference in the retrieval effectiveness, if the original queries are relatively complete descriptions of the information need, lexico-semantic relations have the potential to significantly improve less well-formulated initial queries. Liu et al. [23] proposed several heuristic methods for disambiguating and selecting the candidate expansion terms using adjacent query terms and WordNet. Only the candidate terms that are globally correlated with the query terms were used for expansion. Shah and Croft [27] proposed heuristic methods for query term re-weighting and locating the query terms to expand with WordNet synonyms with the goal of improving precision in the top document ranks. Meij et al. [25] showed that discriminative semantic annotations of documents using domain-specific ontologies, such as MeSH, can be effectively used to improve retrieval. Li et al. [21] experimented with using Wikipedia articles retrieved with the original query as a source of the expansion terms for PRF and observed that the queries, for which the standard PRF failed and which were improved using Wikipedia-based PRF, were the ones that did not perform well after the initial retrieval. Hsu and Chen [16] investigated the utility of commonsense knowledge in ConceptNet for image retrieval by focusing on finding concepts related to the original query through a set of spatial relationships and found that commonsense knowledge is deeply context-sensitive and effective for precision-oriented tasks.

In addition to experimenting with individual lexico-semantic resources for query expansion, several methods combining multiple resources to overcome the problem of data sparsity have also been proposed for both the vector-space and language modeling-based retrieval models. In the context of

vector-space models, Mandala et al. [24] proposed a method to combine three different thesaurus types for query expansion: manually constructed (WordNet), automatically constructed based on term co-occurrence, automatically constructed based on head-modifier relations and found out that improvements in retrieval performance can be achieved by combining all three types of lexical resources. Bodner et al. [4] conducted similar experiments by combining WordNet and co-occurrence based thesauri for query expansion. In the context of language modeling approach, Bai et al. [3] proposed a method for query expansion by integrating term relationships (documents co-occurrence, HAL scores, globally and locally computed information flows) explicitly into the query language model. In Cao et al. [7] term relationships from co-occurrence statistics and WordNet were used to smooth the document language model, so that the probabilities of the related terms in the document model are increased. Collins-Thompson and Callan [10] proposed a Markov chain framework for query expansion, combining multiple sources of knowledge on term associations, such as synonyms from WordNet, terms that share the same prefix when stemmed to the same root, terms co-occurring in a large Web corpus and terms co-occurring in the top retrieved documents. Given a small set of initial query terms, they constructed a term network and used a random walk to estimate the likelihood of relevance for potential expansion terms. Hsu et al. [17] compared the effectiveness of using WordNet and ConceptNet for query expansion. The experimental results indicated that WordNet and ConceptNet can complement each other, since the queries expanded using WordNet have higher discrimination ability (i.e., expansion concepts from WordNet are usually more specific than those from ConceptNet), whereas the queries expanded using ConceptNet have higher concept diversity (i.e., expansion concepts from ConceptNet usually co-occur with the topical terms in relevant documents). They also demonstrated that the retrieval performance improves when the expansion concepts are manually filtered to remove noise, but did not propose any algorithm for that. In general, to the best of our knowledge, an extensive and systematic study of the feasibility of using ConceptNet for query expansion has not yet been conducted.

3. CONCEPT FEEDBACK

This work follows the language modeling approach to information retrieval, specifically the KL-divergence retrieval model [34], according to which the retrieval task involves estimating a query language model, Θ_q, for a given keyword-based query q and the document language models Θ_{D_i} for each document D_i in the collection $\mathcal{C} = \{D_1, \ldots, D_m\}$. The documents in the collection are scored and ranked according to the Kullback-Leibler divergence:

$$\mathrm{KL}(\Theta_q||\Theta_D) = \sum_{w \in V} p(w|\Theta_q) \log \frac{p(w|\Theta_q)}{p(w|\Theta_D)}$$

Within the KL-divergence retrieval model, relevance feedback [35] is considered as the process of updating the query language model Θ_q, given the feedback obtained after the initial retrieval results are presented to the users. Such feedback may be explicitly provided by the users or implicitly derived from the top-ranked retrieval results. Following this approach, a concept expansion language model, $\hat{\Theta}_q$, derived for a given query q from ConceptNet can be used for updating the original query language model Θ_q through linear interpolation:

$$p(w|\tilde{\Theta}_q) = \alpha p(w|\Theta_q) + (1-\alpha) p(w|\hat{\Theta}_q)$$

where α is the interpolation coefficient between the concept expansion language model and the original query model.

The two major challenges for query expansion methods are in identifying as many effective expansion terms as possible and adding those terms to the original query with the weights that would accurately reflect the degree of their effectiveness for improving the quality of retrieval results. If a limited number of automatically identified expansion terms are added to the query, there is a possibility that effective expansion terms will be missed and the results are unlikely to be substantially improved. On the other hand, when the query vocabulary is substantially altered, the advantages gained from effective expansion terms may be lost due to the query topic drift. Within the language modeling context, selecting the right number of terms becomes less important than the right allocation of weights. In this section, we propose heuristic and learning-based approaches for selecting the expansion concepts from ConceptNet and assigning the weights to them. Both methods utilize the concept relations graph of ConceptNet. The main idea behind the first method is that effective expansion concepts should be along the paths connecting the original query terms in the concept relations graph. The second method is based on a finite-step random walk on the concept relations graph, which starts from the query terms. Before discussing the proposed methods in more detail, we need to provide several important definitions.

DEFINITION 1. QUERY TERM CONTEXT C_q^r *of radius r for a given query term q includes all the concepts in ConceptNet that are at most r edges away from q.*

For example, the query term context of radius 2 includes all the concepts that are connected with the given query term (query term neighbors) and all the concepts that are connected with query term neighbors.

DEFINITION 2. QUERY CONCEPT GRAPH $G_q^r = (V, E)$ *of radius r for a given query $q = \{q_1, q_2, \ldots, q_n\}$ is a weighted sub-graph of the entire concept relations graph of ConceptNet $\mathcal{C} = (\mathcal{V}, \mathcal{E})$, such that $V = \bigcup_{i=1}^{n} C_{q_i}^r, V \subseteq \mathcal{V}$ and $E = \{(c_1, c_k, w_{1k}), \ldots, (c_m, c_n, w_{mn})\}, \forall i, j : (c_i, c_j) \in \mathcal{E}$.*

When constructing a query concept graph, all the concepts represented by a phrase were split into concept terms. For example, given a pair of concepts "telescope" and "astronomical tool" and a relation "IsA" between them, the concept "telescope" in the resulting query concept graph will be connected with weighted edges to two separate concept nodes (concept terms) "astronomical" and "tool". In addition to that, at most 100 neighboring concept terms with the highest IDF ($IDF(t) = N/log(c(t,d))$, where N is the total number of documents in the collection and $c(t,d)$ is the number of documents containing the concept term t) were considered for each concept term, excluding very common concept terms (that occur in more than 10% of the documents in the collection).

Since the relations between the concepts in ConceptNet do not have explicit weights, we designed an empirical pro-

cedure to calculate them, which is presented in detail in Section 4.3.

3.1 Heuristic methods

3.1.1 Path finding

DEFINITION 3. PATH $\rho(c_i \rightarrow c_j)$ *between the two concepts* c_i *and* c_j *in the query concept graph* G_q^r *of radius* r *corresponds to a set of concepts and their associated weights* $\{(c_n, w_n), \ldots, (c_m, w_m)\}$ *along any non-repetitive sequence of edges* $((c_i, c_n), \ldots, (c_m, c_j))$ *connecting* c_i *and* c_j*.*

Given a query $q = \{q_1, q_2, \ldots, q_n\}$ and a query concept graph G_q^r of radius r, this method determines all possible unique paths between the query terms and uses a set of concepts $\bigcup_{i=1}^{n} \bigcup_{j=i}^{n} \rho(c_i \rightarrow c_j)$ corresponding to those paths as expansion concepts.

3.1.2 Random walk

Given a query $q = \{q_1, q_2, \ldots, q_n\}$ and a query concept graph G_q^r of radius r, this method first constructs the concept matrix $\mathbf{C}$ (i.e., adjacency matrix of G_q^r) and performs a k-step random walk on that matrix. The weight of the expansion concept c for a query term q_i is determined as follows:

$$p(c|q_i) = (1-\beta)\beta^k \mathbf{C}^k_{c,q_i}$$

where β is the probability of continuing the random walk.

3.2 Learning-based method

The learning-based expansion method is based on training a regression model, in which the independent variables are the features of an expansion concept and the response variable is a measure of performance (we used mean average precision or MAP) of an original query expanded with the given concept. After estimating the parameters of a regression model based on the actual MAP values of the queries expanded with the concept terms in the training data set, new expansion concepts can be ranked based on the MAP values predicted by the model and a certain number of the highest ranked concepts can then be used to expand the query. The expanded query can then be used to retrieve a new improved set of results, which can be either presented to the users or used for pseudo-relevance feedback to further improve the performance of the query. The learning-based expansion method is further referred to as LR and the combined leaning-based expansion and pseudo-feedback method as LR-PF.

3.3 Model

Due to its computational efficiency, we used the generalized linear regression model (GLM) as the learning algorithm. We also experimented with logistic regression and found it to be consistently worse than the GLM. Since both models are very similar, due to space limitations, we do not provide the experimental results for logistic regression. We also leave experimentation with other methods (e.g., learning-to-rank, such as ListNet [8]) as future work. Given a vector of features $\bar{x}$, the GLM estimates a vector of feature weights $\bar{w}$ during training, and generates the output as a linear combination of the feature and weight vectors, $f(\bar{x}) = \bar{x}\bar{w}$, during testing. Another advantage of GLM over other methods is that it allows to easily interpret the feature weights in order to determine the important properties of effective expansion concepts.

3.4 Features

The set of features used in the experiments is presented in Table 2. This feature set reflects the properties of queries, expansion concepts and expansion concepts with respect to queries. It extends the set of features used in [18] (designated by bullets in the BL column) and includes 7 new features, focused on the structural properties of the expansion concepts with respect to the query terms in the query concept graph: CONFANOUT, RNDWALKSCORE, PATHFINDSCORE, AVGQDIST, MAXQDIST, AVGPWEIGHT, MAXPWEIGHT. Since PATHFINDSCORE and RNDWALKSCORE correspond to the score of an expansion term using the heuristic methods presented in Sections 3.1.1 and 3.1.2 respectively, learning-based method unifies and extends the heuristic methods.

4. EXPERIMENTS

In this section, we present the results of an experimental evaluation of unsupervised and supervised methods for query expansion with concepts from ConceptNet. We first discuss our experimental setup and datasets.

4.1 Experimental setup and datasets

All experiments in this work were conducted on three standard TREC collections: ROBUST04, which was used in TREC 2004 Robust Track [1]; AQUAINT, which was used in TREC 2005 HARD [2] and Robust Tracks; and GOV, which was used in TREC 2004 Web Track [11]. AQUAINT and ROBUST04 datasets consist of the newswire documents, while GOV consists of the Web documents. Various statistics for experimental datasets are summarized in Table 3.

Corpus	#Docs	Size(Mb)	#Topics	Avg. topic
AQUAINT	1,033,461	3042	50	2.56
ROBUST04	528,155	1910	250	2.65
GOV	1,247,753	18554	225	3.04

Table 3: Statistics of the datasets used for experiments.

In this work, we focus on studying the effectiveness of expansion using ConceptNet with respect to difficult queries. We define a difficult query as a query, for which either the average precision of the retrieved results is less than 0.1 or the top 10 results are non-relevant (i.e. $Pr@10 = 0$). In all the experiments in this work, we used the same suggested settings of Dirichlet priors for both the baselines (KL-divergence retrieval model and model-based pseudo-relevance feedback) and concept feedback methods: 2000 for AQUAINT and ROBUST04 and 500 for GOV.

4.2 Upper-bound performance

In order to determine the upper bound for the potential effectiveness of using ConceptNet for query expansion, we conducted a simulation experiment, in which for each query term it was first checked if there exists a concept in ConceptNet that matches it. If such node was found, then all the concepts within the query term context of certain radius were identified. We experimented with the contexts of one, two and three edges from the query terms. If a concept was

Feature	BL	Description
Features of the query		
NumQryTerms	•	number of query terms
TopDocScore	•	retrieval score of the top-ranked document for the initial query
Features of the expansion concept		
ExpTDocScore	•	retrieval score of the top-ranked document for the initial query expanded with the concept
TopTermFrac	•	ratio of the number of occurrences of the expansion concept over all the terms in the top 10 retrieved documents
NumCanDocs	•	number of the top 10 documents containing the expansion concept
AvgCDocScore	•	average retrieval score of the documents containing the expansion concept
MaxCDocScore	•	maximum retrieval score of the documents containing the expansion concept
ConIDF	•	IDF of the expansion concept
ConFanOut		number of nodes adjacent to the expansion concept node in the query concept graph
SpActScore	•	spreading activation score of the expansion concept in the query concept graph
SpActRank	•	rank of the expansion concept after spreading activation in the query concept graph
RndWalkScore		weight of the expansion concept by using the Finite Random Walk method (Section 3.1.2)
PathFindScore		weight of the expansion concept by using the Path Finding method (Section 3.1.1)
Features of the expansion concept with respect to the query terms		
AvgColCor	•	average co-occurrence of the expansion concept with the query terms in the collection
MaxColCor	•	maximum co-occurrence of the expansion concept with the query terms in the collection
AvgTopCor	•	average co-occurrence of the expansion concept with the query terms in the top 10 retrieved documents
MaxTopCor	•	maximum co-occurrence of the expansion concept with the query terms in the top 10 retrieved documents
AvgTopPCor	•	average co-occurrence of the expansion concept with pairs of query terms in the top 10 retrieved documents
MaxTopPCor	•	maximum co-occurrence of the expansion concept with pairs of query terms in the top 10 retrieved documents
AvgQDist		average distance of the expansion concept to the query terms in the query concept graph
MaxQDist		maximum distance of the expansion concept to the query terms in the query concept graph
AvgPWeight		average weight of the paths to the expansion concept from the query terms in the query concept graph
MaxPWeight		maximum weight of the paths to the expansion concept from the query terms in the query concept graph

Table 2: Features for ranking the expansion terms. Baseline feature set is designated in the BL column with a bullet (•).

designated by a phrase, it was split into individual concept terms and very popular concept terms (the ones that occur in more than 10% of the documents in the collection) were not considered.

First, each query was expanded with each of the neighboring concepts (expansion context of size one) of the query terms by simply adding each concept term to the query with equal weight $1/|q|$ (where $|q|$ is the length of an expanded query) and comparing the average precision of the original query with the average precision of the expanded query. Then in each dataset we counted the number of queries, for which there was at least one expansion concept that improved the query performance (Improved), the number of queries for which all expansion concepts degraded the query performance (Hurt) and the number of queries, for which none of the expansion concepts worsened or improved the query performance (Neutral). The results for this experiment along with the total number of queries and difficult queries in each dataset are presented in Table 4.

	Total	Diff	Improved	Hurt	Neutral
AQUAINT	50	17	42	8	0
ROBUST04	250	75	232	14	4
GOV	225	147	161	8	56

Table 4: Number of improved, hurt and neutral queries when simulating expansion using the expansion context of size one.

As follows from Table 4, for most queries (not only the difficult ones) in all datasets there exists at least one effective expansion concept among the immediate context of the query terms. For difficult queries in each dataset, we also determined the upper-bound effectiveness of concept expansion with respect to the mean average precision (MAP), geometric mean average precision (GMAP), the total number of relevant documents retrieved (RR) and precision at top 10 retrieved documents (P@10) by varying the radius of the expansion context. The results of this experiment are presented in Table 5.

As follows from Table 5, in the upper bound, concept feedback (CF) significantly improves the performance of the KL-divergence retrieval model and also outperforms the baseline (model-based pseudo-relevance feedback), even when the context of size 1 (column CF-1) is used for expansion. In other words, *for each difficult query there exists at least one highly effective expansion concept*. Using the most effective concept within the context of radius one on average doubles the performance of the baseline KL-divergence retrieval for difficult queries on all datasets. Using the contexts of larger radius improves the performance even more, with the context of radius 3 (column CF-3) having about *triple* (for the newswire datasets) and *six times* (for the Web dataset) the performance of the KL-divergence retrieval model without feedback. This simulation experiment clearly illustrates that using conceptually related terms for query expansion has a tremendous potential for improving the performance

		KL	KL-PF	CF-1	CF-2	CF-3
AQUAINT	MAP	0.0521	0.0429	0.1247	0.1622	0.1880
	GMAP	0.0414	0.0214	0.1033	0.1438	0.1707
	RR	567	519	671	730	772
	P@10	0.1176	0.1000	0.3765	0.5412	0.6059
ROBUST04	MAP	0.0509	0.0788	0.1061	0.1539	0.1823
	GMAP	0.0268	0.0225	0.0718	0.1162	0.1464
	RR	2078	2573	2560	2826	3102
	P@10	0.1467	0.1587	0.2893	0.3973	0.4280
GOV	MAP	0.0748	0.0447	0.1830	0.3481	0.4326
	GMAP	0.0243	0.0166	0.0668	0.1627	0.2403
	RR	760	742	811	793	800
	P@10	0.0347	0.0197	0.0823	0.1190	0.1558

Table 5: Comparison of the upper-bound performance of concept feedback (CF) with KL-divergence retrieval model (KL) and model-based pseudo-relevance feedback (KL-PF) on difficult topics by varying the the radius of the expansion context.

of difficult queries. However, how to automatically identify those few highly effective expansion concepts is unclear. In the rest of this work, we propose and study heuristic and learning-based methods for selecting and assigning the importance weights to the expansion concepts.

4.3 Edge weighting

Since our automatic query expansion procedure selects many expansion terms, we need to design a method to allocate the importance weights to them. In particular, for this task we can use several properties of the expansion terms, such as the length of the paths, as well as the types and weights of ConceptNet relations connecting them to the query terms. Given a concept graph constructed from ConceptNet for a particular query, we used the following empirical procedure to assign the weights to its edges:

1. First, before query processing we used the results of simulation experiment presented in Section 4.2 on the dataset with the largest number of queries (ROBUST04) to count the number of times the best expansion concept was connected to the original query term with the ConceptNet relation of each type.

Relation	Count	Group
IsA	132	1
HasProperty	72	1
CapableOf	65	1
AtLocation	40	1
ConceptuallyRelatedTo	35	1
UsedFor	35	1
HasA	27	2
DefinedAs	26	2
ReceivesAction	21	2
PartOf	15	2
CausesDesire	8	2
LocatedNear	5	2
Causes	5	2
HasPrerequisite	2	3
Desires	2	3
InstanceOf	2	3
MadeOf	2	3
MotivatedByGoal	2	3
HasFirstSubevent	1	3
SimilarSize	1	3

Table 6: Number of times the best expansion term was connected to the expanded query term with the relation of each type.

We then sorted the relations according to those counts and divided them into three groups of the same size, which are presented in Table 6.

2. Second, we constructed a term relationship graph for all the terms in the vocabulary of the collection. Term relationship graph is a weighted graph, in which the set of vertices corresponds to the terms in the collection and the edges correspond to the semantic relationships between the terms. The weight of an edge represents the degree of semantic relatedness of between the two terms measured by the Hyperspace Analog to Language (HAL) model [6]. Constructing HAL space for an n-term vocabulary involves traversing a sliding window of width w over each word in the corpus, ignoring punctuation, sentence and paragraph boundaries. All the words within the sliding window are considered as the local context of the term, over which the sliding window is centered. Each word in the local context receives a score, according to its distance from the center of the sliding window (words that are closer to the center receive higher scores). After traversing the entire corpus, an $n \times n$ HAL space matrix $\mathbf{H}$, which aggregates the local contexts for all the terms in the vocabulary is produced. In this matrix, the row vectors encode the preceding word order and the column vectors encode the posterior word order. An example of the HAL space for the sentence "the effects of pollution on the population" constructed using the sliding window of size 10 (5 words before and after the center word) is shown in Table 7.

	the	eff	of	poll	on	pop
the	1	2	3	4	5	
eff	5					
of	4	5				
poll	4	5				
on	2	3	4	5		
pop	5	1	2	3	4	

Table 7: HAL space for the sentence "the effects of pollution on the population".

The final HAL weights are produced by merging the row and column corresponding to each term in the HAL space matrix. Each term t in the vocabulary $V, |V| = n$ corresponds to a row in the HAL space matrix:

$$\mathbf{H}_t = \{(t_1, c_1), \ldots, (t_n, c_n)\}$$

where each $c_1, \ldots, c_n$ is the number of co-occurrences of the term t with other terms in the vocabulary V. After the

merge, each element in the HAL matrix $\mathbf{H}$ is normalized to estimate the strength of semantic relatedness between the terms $\mathbf{H}_{ti} = c_i / \sum_{j=1}^{n} c_j$. Unlike the mutual information, in which the entire document is used as the context to calculate the number of co-occurrences between the terms, HAL uses the contexts of smaller size (we used a sliding window of 10 words before and after the center word) and has been shown in our previous work [20] to produce less noisy term relationship graphs.

3. Third, for each query we constructed the query concept graph and performed two passes over its edges. In the first pass, if an edge between the concept terms in the query concept graph also existed between the same terms in the term relationship graph, its ConceptNet relation type and weight in the term relationship graph was used to calculate the average weight of all relations in the query concept graph, which belong to the same relation group according to Table 6. In the second pass, the weight of an edge between the concept terms in the query concept graph was set to the weight of an edge between the same terms in the term relationship graph, if such an edge existed. Otherwise, its weight was set to the average weight of relations in the same relation group, determined in the previous pass. Given the weighted query concept graph, the weight of a concept within the context of a certain radius from the query term was determined as the product of the weights of all the edges and IDFs of all the concepts (including the target one) along the shortest path from the query term.

4.4 Learning-based expansion

We used 5-fold cross validation to train and test the linear regression model. During testing we selected 100 top-scoring concepts and used them for expansion. In order to determine the optimal setting for learning-based concept expansion (LR) and the combined method (LR-PF), we experimented with different feature sets and contexts of different size (2 and 3). Performance of different configurations of the learning-based expansion method on different datasets is presented in Figure 2.

Several interesting observations can be made based on the analysis of Figure 2. First, for both the learning-based concept expansion and the combined method, using the extended feature set (FULL) generally results in better performance than using the baseline (BASE) feature set, which empirically demonstrates the benefits of exploiting the graph-based properties of the expansion concepts with respect to the query terms. Second, selecting candidate concept terms from the query concept graph of larger radius generally results in better performance for both feature sets, which is consistent with the results of the simulation experiment presented in Section 4.2. Similar conclusions can also be drawn by analyzing the behavior of different configurations of the learning-based (LR) and the combined (LR-PF) methods on the ROBUST04 and GOV datasets, although on those datasets the different methods within each group behaved very similar to each other. In addition to that, applying pseudo-feedback on top of the learning-based method not only did not further improve its performance on GOV (as opposed to both ROBUST04 and AQUAINT), but even significantly decreased it. This can be attributed to the fact that GOV queries are highly focused and have fewer relevant documents associated with them, thus applying pseudo-feedback on top of concept-feedback may cause the query drift.

In order to better understand the properties of effective expansion concepts, we averaged across the splits the feature weights from the model learned on the ROBUST04 dataset when the size of the expansion context is 3. Table 8 shows the weights of some of the features of the expansion concepts.

Feature	Weight
TopDocScore	0.0944
AvgTopCor	0.0518
TopTermFrac	0.0119
AvgColCor	0.0078
FanOut	0.0052
AvgPWeight	-0.0012
SpActRank	-0.0087
AvgQDist	-0.0267
ConIDF	-0.1004

Table 8: Feature weights for the expansion concepts.

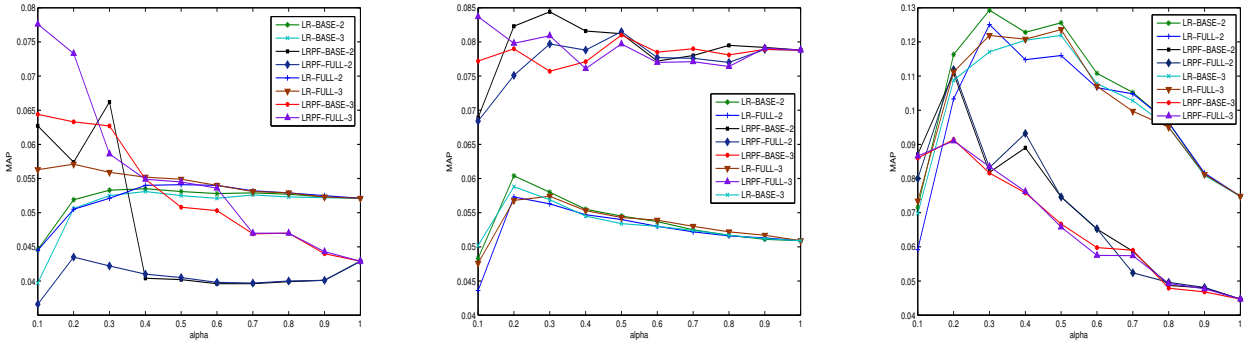
As follows from Table 8, effective expansion concepts are those concepts, which both frequently occur in the top 10 retrieved documents (as follows from the high positive weight of the TopTermFrac feature) and frequently co-occur with the query terms in the entire collection and in the top 10 retrieved documents (AvgColCor and AvgTopCor). The high negative weight for the ConIDF feature indicates that effective expansion concepts are not rare in the collection (can't have high IDF). Positive weight for the FanOut and negative weight for the SpActRank features both indicate that effective expansion concepts are typically connected with edges of high weight to many other concepts in ConceptNet and are not far from the query terms (as evidenced by the negative weights for AvgQDist and AvgPWeight).

4.5 Comparison of methods

Having determined the best performing configuration of the learning-based methods, we compare them with the heuristic methods and the baselines in Figure 3 and Table 9.

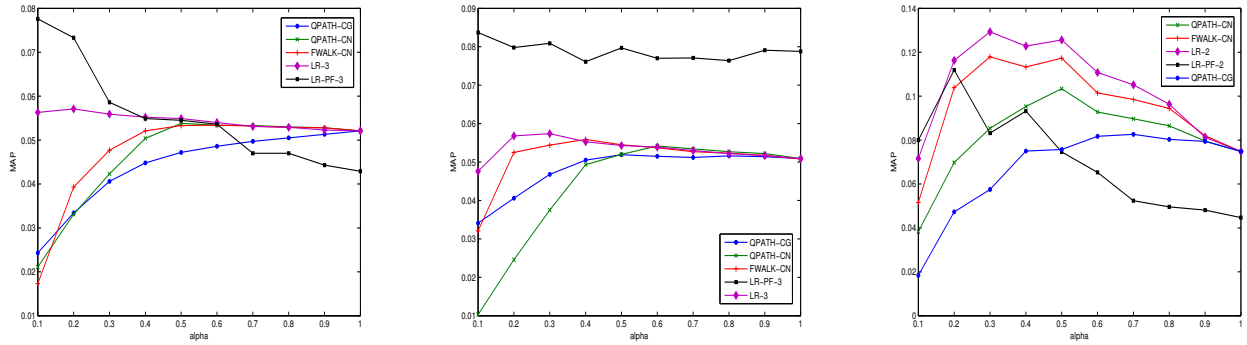
As follows from Figure 3, the learning-based method predictably outperforms both heuristic methods on all datasets, since it uses the weights of expansion concepts generated by both heuristic methods as features. Moreover, as the weight of the original query language model in the mixture increases, the performance of the learning-based and the combined methods drops, which clearly illustrates the effectiveness of expansion concepts. The best performance of heuristic, learning-based and combined methods is summarized in Table 9

Several important conclusions can be drawn based on the analysis of Table 9. First, heuristic, learning-based and combined methods all improve over the KL-divergence baseline (KL) on all datasets. Secondly, the combined method (LR-PF) consistently outperforms pseudo-feedback (KL-PF) on all datasets. Moreover, as opposed to pseudo-feedback (KL-PF), the combined method actually improves the performance relative to the KL baseline for both expansion contexts of radius 2 and 3 on AQUAINT and GOV datasets. Consequently, we can conclude that learning-based method is an effective strategy to combine multiple signals for robust query expansion and pseudo-feedback in case of difficult queries, when the initially retrieved documents cannot be considered as a reliable source of effective expansion terms.



(a) **AQUAINT dataset** (b) **ROBUST04 dataset** (c) **GOV dataset**

Figure 2: Comparison of performance of learning-based concept feedback methods with different feature sets and different values of context radius by varying the interpolation coefficient α.



(a) **AQUAINT dataset** (b) **ROBUST04 dataset** (c) **GOV dataset**

Figure 3: Comparison of performance of heuristic to learning-based methods by varying the interpolation coefficient α

		KL	KL-PF	QPATH	RWALK	LR-2	LR-PF-2	LR-3	LR-PF-3
AQUAINT	MAP	0.0521	0.0429	0.0538	0.0534	0.0535	0.0662	0.0571*	0.0776*†
	P@10	0.1176	0.1	0.1353	0.1118	0.0941	0.1059	0.1294	0.1471
ROBUST04	MAP	0.0509	0.0788	0.0542	0.0559	0.0604^{+}	0.0844	0.0588^{+}	0.0837
	P@10	0.1467	0.1587	0.1413	0.1707	0.1747	0.1747	0.1707	0.1627
GOV	MAP	0.0748	0.0447	0.1034	0.1179	0.1293*	0.1119*†	0.1236*	0.0914*†
	P@10	0.0347	0.0197	0.0401	0.0673	0.066	0.0517	0.0551	0.049

Table 9: Comparison of the best performance of heuristic (QPATH and RWALK), learning-based (LR-2 and LR-3) and the combined (LR-PF-2 and LR-PF-3) methods with the KL-divergence retrieval model (KL) and model-based pseudo-relevance feedback (KL-PF) on difficult topics. $^{+}$ and * indicate statistical significance relative to KL (95% and 99% confidence levels, respectively) according to the Wilcoxon signed-rank test. † indicates statistical significance relative to KL-PF (99% confidence level) according to the Wilcoxon signed-rank test. Significance testing was performed on the results of the learning-based and combined methods only.

5. SUMMARY AND CONCLUSIONS

In this work, we presented the results of the first systematic exploration of the potential for utilizing the knowledge base of ConceptNet to improve the retrieval results of *difficult queries* and overcome the problem of the lack of relevant documents for such queries in the initial search results. In particular, we:

1. conducted a simulation experiment to determine the upper bound for the effectiveness of query expansion with the related concepts from ConceptNet, which demonstrated that there exists a small number of highly effective expansion concepts for difficult queries;

2. proposed several heuristic and learning-based methods for automatically selecting effective expansion concepts, which leverage the graph-based knowledge representation structure of ConceptNet, and empirically compared the proposed methods on different datasets. Our results indicate

that learning-based expansion methods can effectively leverage the common sense knowledge in ConceptNet to improve the search results of difficult queries both through query expansion alone and in combination with the traditional model-based pseudo-relevance feedback;

3. analyzed the obtained learning-based expansion models to determine the properties of effective expansion terms.

Designing the methods to improve the search results of difficult queries is a challenging and very important practical and theoretical information retrieval problem. Our results and findings indicate that ConceptNet has a great potential to be effectively used as an external source of expansion terms, when the initial search results are of very bad quality and other techniques such as relevance feedback and pseudo-relevance feedback become ineffective.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Number NSF-1027965, AFOSR MURI Grant FA9550-08-1-0265, and a gift grant from Microsoft. We would like to thank the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] J. Allan. Overview of the trec 2004 robust retrieval track. In *Proceedings of TREC 13*, 2004.

[2] J. Allan. Hard track overview in trec 2005 - high accuracy retrieval from documents. In *Proceedings of TREC 14*, 2005.

[3] J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao. Query expansion using term relationships in language models for information retrieval. In *Proceedings of CIKM*, pages 688–695, Bremen, Germany, 2005.

[4] R. C. Bodner and F. Song. Knowledge-based approaches to query expansion in information retrieval. *Advances in Artificial Intelligence*, pages 146–158.

[5] P. Bruza and D. Song. Inferring query models by computing information flow. In *Proceedings of CIKM*, pages 260–269, 2002.

[6] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences and discourse. *Discourse Processes*, 25:211–257, 1998.

[7] G. Cao, J.-Y. Nie, and J. Bai. Integrating word relationships into language models. In *Proceedings of SIGIR*, pages 298–305, 2005.

[8] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of ICML*, pages 129–136, 2007.

[9] P. R. Cohen and R. Kjeldsen. Information retrieval by constrained spreading activation in semantic networks. *Information Processing and Management*, 23(4):255–268, 1987.

[10] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *Proceedings of CIKM*, pages 704–711, 2005.

[11] N. Craswell and D. Hawking. Overview of the trec 2004 web track. In *Proceedings of TREC 13*, 2004.

[12] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[13] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, pages 154–161, 2006.

[14] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, pages 1606–16011, 2007.

[15] S. Gauch and J. Wang. A corpus analysis approach for automatic query expansion. In *Proceedings of CIKM*, pages 278–284, 1997.

[16] M.-H. Hsu and H.-H. Chen. Information retrieval with commonsense knowledge. In *Proceedings of SIGIR'06*, pages 651–652, 2006.

[17] M.-H. Hsu, M.-F. Tsai, and H.-H. Chen. Query expansion with conceptnet and wordnet: An intrinsic comparison. In *Proceedings of AIRS*, pages 1–13, 2006.

[18] M.-H. Hsu, M.-F. Tsai, and H.-H. Chen. Combining wordnet and conceptnet for automatic query expansion: A learning approach. In *Proceedings of AIRS*, pages 213–224, 2008.

[19] A. Kotov and C. Zhai. Towards natural question-guided search. In *Proceedings of WWW*, pages 541–550, 2010.

[20] A. Kotov and C. Zhai. Interactive sense feedback for difficult queries. In *Proceedings of CIKM*, pages 163–172, 2011.

[21] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using wikipedia as external corpus. In *Proceedings of SIGIR*, pages 797–798, 2007.

[22] H. Liu and P. Singh. Conceptnet-a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):146–158.

[23] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of ACM SIGIR'04*, pages 266–272, 2004.

[24] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of ACM SIGIR'99*, pages 191–197, 1999.

[25] E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Conceptual language models for domain-specific retrieval. *Information Processing and Management*, 2:25–45, 2000.

[26] G. Salton and C. Buckley. On the use of spreading activation methods in automatic information retrieval. In *Proceedings of SIGIR*, pages 147–160, 1988.

[27] C. Shah and W. B. Croft. Evaluating high accuracy retrieval techniques. In *Proceedings of SIGIR*, pages 2–9, 2004.

[28] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, UK, 1979.

[29] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of ACM SIGIR'93*, pages 61–69, 1994.

[30] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of ACM SIGIR'96*, pages 4–11, 1996.

[31] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18:79–112, 2000.

[32] Z. Yin, M. Shokouhi, and N. Craswell. Query expansion using external evidence. In *Proceedings of ECIR*, pages 362–374, 2009.

[33] J. Yufeng and W. B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO*, pages 146–160, 1994.

[34] C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pages 111–119, 2001.

[35] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM'01*, pages 403–410, 2001.