# Constructing Query-Specific Knowledge Bases

Jeffrey Dalton
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA, U.S.A.
jdalton@cs.umass.edu

Laura Dietz
Center for Intelligent Information Retrieval
University of Massachusetts
Amherst, MA, U.S.A.
dietz@cs.umass.edu

## ABSTRACT

Large general purpose knowledge bases (KB) support a variety of complex tasks because of their structured relationships. However, these KBs lack coverage for specialized topics or use cases. In these scenarios, users often use keyword search over large unstructured collections, such as the web. Instead, we propose constructing a 'knowledge sketch' that leverages existing KB data elements and relevant text documents to construct query-specific KB data. A knowledge sketch is a distribution over entities, documents, and relationships between entities, all for a specific information need. In our experiments we construct knowledge sketches for queries from the TREC 2004 Robust track, which emphasizes complex queries which perform poorly with existing text retrieval approaches.

## 1. INTRODUCTION

Leading web search providers are increasingly incorporating richer knowledge base information into search results in order to more effectively satisfy a users' query intents. However, for complex 'tail' queries with specialized information needs it is unlikely that all of the important entities and relationships will be captured in a general purpose KB. This may be because the important entities or relationship are rare, the schema is not specific to the query domain, the KB contains incorrect information, or the KB is out-of-date because of new events. One step towards making knowledge base reasoning available for every information need is a method for constructing query-focused knowledge resources on-demand.

Starting from an information need represented as one or more text queries, the goal is to identify both relevant structured KB data and unstructured text resources that complement one another. We propose a new framework for constructing query-specific knowledge resources, a 'knowledge sketch'. A knowledge sketch is a set of distributions over entities, documents, and relationships between entities specific to the information need. These relationships allow the user (or application) to make sense of a topic by providing multi-modal search results, including both documents and entities with relations. From this representation the user can understand what entities are important, how entities and documents relate to one another, and how they relevant to the information need.

This work is on jointly modeling a query-specific knowledge sketch from a given general-purpose knowledge base and collection of documents. In contrast to general KB construction, where all documents and entities are equally important, user-focused KB construction is performed with respect to an information need. This implies distributions of relevance over documents, entities, relationships, and attributes. This work advocates unified reasoning on relevance of these elements with respect to the user intent. In particular, we incorporate bi-directional evidence between pertinent KB entities and respective mentions in documents. We believe that this has potential to not only improve document retrieval effectiveness, but yields a knowledge product that is of immediate interest to the user.

We illustrate this framework for the information need [what has been the experience of residential utility customers following deregulation of gas and electric], query #437 from the TREC Robust 2004 evaluation [17]. Ideally, a full knowledge representation would cover aspects of 'customer experience' that changed in response to deregulation, such as changes in price, service reliability, customer satisfaction, and abuses, across regions and time. A knowledge sketch provides simpler view with important entities (such people, companies, and government agencies), and loosely defined entity relationships (e.g. Stephen Littlechild, Director General, OFFER).

The sketch models we propose are applicable to a broad range of corpora and knowledge bases. Possible choices for knowledge bases are Freebase[3], YAGO[16], spreadsheets, and even well-structured domain specific websites (e.g. soccer players, historic incidents, or music albums). The only requirements are a set of names and snippets of text associated with each entity and relationships between entities. For the experiments in this work we use Wikipedia with metadata from Freebase as our KB and TREC collections for our corpora.

The remainder of the paper is structured as follows. In Section 2, we introduce a probabilistic model for knowledge sketch inference. Details of a concrete model instance are given in Section 3. In Section 4 we report experimental results on the TREC Robust 2004 test collection, before concluding in Section 5.


*CIKM'13,* October 27-28, 2013, San Francisco, California USA.

## 2. QUERY-SPECIFIC KNOWLEDGE SKETCH CONSTRUCTION

For a given question $Q$, the goal is to infer a knowledge sketch $S$. The sketch quantifies which information from the general-purpose knowledge base and corpus is relevant to the user information need. Formally, the sketch $S$ is represented by a set of multinomial distributions $\epsilon$, $\delta$, $\rho$: over entities $E$, documents $D$, and relationships $R$, respectively. Many more aspects could be included in the sketch, such as Wikipedia categories, relation types, or grammatical patterns, which we leave for future work.

We limit the scope of this work to binary relations modeled by the predicate on a pair of entities $(e', e'')$ meaning "$e'$ is related to $e''$". The distribution $\rho$ represents both existence and salience of the relationship. In the following we distinguish between directly relevant entities $E'$ and related entities $E''$.

### 2.1 Naive Sketch

A naive sketch model can identify the document distribution $\delta$ by issuing a text query, $Q$, against the text corpus, and let the retrieval probability represent the relevance $p(D|\delta)$ over documents. This is possible if probabilistic retrieval models are used, such as the query likelihood model. See below for retrieval model used in this work, we refer to it as $p_{\mathrm{IR}}$ henceforth.

A similar approach can be used to generate a distribution over entities $\epsilon$. We perform object retrieval [14] over the knowledge base using the text query; the result is a distribution over KB entries with their retrieval probability. A naive derivation over relations $\rho$ can be derived from the structure of the knowledge base, weighted by entity relevance.

There are several issues with this approach. The distributions over entities and relations are not necessarily reflected in the documents. Not all relations between entities are also pertinent to the information need. Further, the knowledge about the relevance of entities and relations is not leveraged to infer relevance of documents.

### 2.2 Entity Linking Sketch

Given a relevance distribution over documents $p(D|Q)$, we can link the highest probability documents to the knowledge base using an entity linking system, such as KB Bridge [4]. This gives rise to the salience distribution over entities, by building entity models for each document as $p(E|D) = \frac{\#(E \in D)}{\sum_e \#(e,D)}$.

$$p(E|\epsilon) = \int p(E|D) p(D|Q)\, dD \tag{1}$$

Likewise, we can extract a distribution over relations from co-occurrences of entity mentions in the documents.

The entity linking sketch model ensures that documents, entities, and relations provide one coherent picture. The entity linking strategy further allows to identify out-of-KB entities that are pertinent to the information need, but not linkable to the knowledge base.

On the downside, in cases where the document distribution does not reflect the user intent well, this approach is likely to arrive at a distribution over entities that are not relevant. This is the case when the question itself is not sufficient for retrieving documents of high relevance.

### 2.3 Entity Expansion Sketch

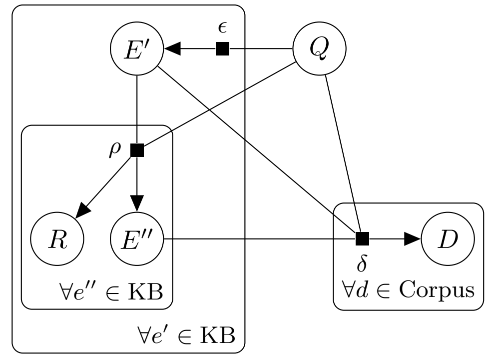


**Figure 1: Factor graph of the Joint Sketch Model.**

Alternatively, we can start with a distribution over entities $p(E'|Q)$, and expand the text query from the entity distribution. We use the RM3 variant of the relevance model[11] which combines the original query with a model from the highest probability $k$ entities $E'$.

$$p(D|Q, E') = \lambda p_{\mathrm{IR}}(D|Q) + (1-\lambda) \int p_{\mathrm{IR}}(D|E') p(E'|Q)\, dE' \tag{2}$$

The distribution over documents is modeled by a mixture weighted retrieval model based on $p_{\mathrm{IR}}$, given the trade-off parameter $\lambda$. We discuss query construction from entities $p(D|E')$ in Section 3. If available, the query can be further expanded accordingly with entities $E''$ that have salient relations.

This sketch approach provides robustness to the document relevance distribution by also leveraging the knowledge base as an external source. However, the query might be expanded with entities that are topically related, but not reflected in the relevant documents in the source corpus, e.g. "heroic acts" might retrieve entities from ancient Greek mythology, where the corpus contains references to modern heros from recent news articles.

### 2.4 Joint Sketch Model

We address the weaknesses of the previous sketch with a joint model of $D$, $E$, $R$ given $Q$ with the factorization given in in Figure 1 using the directed factor graph notation [5].

$$p(D, R, E|Q) = \underbrace{p(E'|Q)}_{\epsilon} \underbrace{p(E'', R|Q, E')}_{\rho} \underbrace{p(D|Q, E', E'', R)}_{\delta}$$

The first factor represents the prior distribution over entities given the question. It is estimated by retrieval against the KB index $p(E'|Q) \propto p_{\mathrm{IR}}(Q|E')$. The second factor expands towards with direct relations in the KB. For the Wikipedia KB inlinks, outlinks, and co-occurring links are considered as relations.

The last factor follows the approach of the entity expansion sketch in modeling $\epsilon$, see Equation 2. The relevance distribution over documents is modeled by the retrieval probability of the query expanded with entities according to their relevance distribution.

As this yields a generative model, we can perform inference in the style of a blocked Gibbs sampler as follows. Given a point estimate of a relevance distribution, we derive a likelihood distribution over entities given the documents.

Document-specific entity models can be extracted from each document $p(E'|d) = \frac{\#(E' \in d)}{\sum_e \#(e,d)}$, to compute the likelihood from the documents $p(E'|D,Q) = \int p(E'|D)p(D|Q)\, dD$. This can be achieved by counting matches of the expanded entity names in the IR query. Alternatively, we can follow ideas in the entity linking sketch, cf. Equation 1. The latter approach bears the potential to reveal new entities, which are not contained in the knowledge base, but are relevant to the user information need. The distribution over entities, $\epsilon$ can be updated to the posterior from prior and likelihood, $p(E'|\epsilon, Q, D) = \frac{1}{Z} p_{\text{IR}}(E'|Q)p(D|E',Q)p(E''|E',Q)$, where $Z$ is a factor to ensure proper normalization of the multinomial distribution.

With a similar approach, the prior on relations can be updated, leveraging colocations of entity mentions in the documents. If relation types are available in the knowledge base, it is immediately possible to learn which relation types are relevant, to be incorporated in identifying related entities.

### 2.5 Probabilistic Retrieval Models as a Factor

The joint sketch models relevance uses retrieval probabilities generated by an IR system. A simple probabilistic retrieval model is the query likelihood model[13] which scores indexed documents according to their likelihood of generating the terms $q_i$ in the query, assuming independence of terms, $p(Q|D) = \prod_{q_i \in Q} p(q_i|D)$. Application of Baye's rule yields a distribution over documents in the index given the query, $p(D|Q)$.

Better retrieval effectiveness has been achieved with the sequential dependence model[12] which combines the termwise model with a model over bigrams and orthogonal sparse bigrams.[1] Although the sequential dependence model has been introduced as a Markov Random Field, a rank equivalent generative model can be derived. Notice that the IR factor, $p_{\text{IR}}$, can be further nested with a multinomial mixture model and still govern a probability distribution over documents. Most IR systems are optimized to produce rank-equivalent scores in log-space. We approximate probabilities with highest probability $k$ documents, using exponentiation and renormalization as in Lavrenko and Croft [11].

## 3. EXPERIMENTS

We implement a prototype 'knowledge sketch' system for information needs using data from the TREC 2004 Robust adhoc retrieval task [17].

### 3.1 Data and processing

The text collection we use is the TREC 2004 Robust collection, which consists of TREC disks 4 and 5, minus the Congressional Record. It contains approximately 528,000 news documents from 1989 to 1994. We process the collection using the factorie[2] toolkit to annotate the documents. We perform tokenization, sentence detection, part of speech tagging, shallow parsing, and named entity recognition. After these steps the KB Bridge[3] system is used to perform entity linking on the documents.

We report results on two subsets of the robust queries. The first is a fourty two query sample, 42-rand, a random sample of the 250 queries, with topics that have at least twenty relevant documents. The second set, 20hard, is a set of challenging queries where current text retrieval approaches are ineffective and provide opportunity for significant improvements leveraging knowledge-based approaches. These queries have a mean average precision (MAP) of 0.02 and each query has an average precision score less than 0.05 with a strong retrieval baseline, the sequential dependence model[12]. We note that these queries are not significantly improved by the current state-of-the-art retrieval models (including weighted sequential dependence model [1] and multiple source expansion [2]).

For all of the experiments over both text and KB we use the Galago[4] search engine. We use the Markov Random Field retrieval model [12], specifically the Sequential Dependence retrieval model. For these experiments all terms are stopped using the Porter Stemmer and the Galago 418 word stop list is used. The retrieval parameters for KB retrieval were tuned on a subset of the TAC KBP [10] entity linking queries[4], with $mu = 96400$, $uniw = 0.29$,$odw = 0.21$, and $uww = 0.5$. For the document retrieval collection, the retrieval parameters for the corpora were turned using cross-validation as reported by Huston and Croft with $mu = 1269$, $uniw = 0.873$,$odw = 0.079$, and $uww = 0.048$.

### 3.2 KB Retrieval Setup

For ranking entities we use passage retrieval of sliding windows of 100 terms against the KB index to estimate $\epsilon$. The reasons is that some articles are long and cover diverse aspects of the entity. The relevance of the entity $\epsilon$ is represented by the retrieval probability of the highest scoring passage. For each Wikipedia entity we extract an entity representation consisting of the canonical name and a distribution over name variants from redirects, Freebase names, and Wikipedia-internal anchor text.

### 3.3 Evaluation

We evaluate how well the sketch satisfies user information needs. For first steps in this direction, we focus on the ability to identify relevant information sources and entities. These form the basis KB construction and extension. We directly evaluate the relevance of the documents using the TREC relevance judgments.

We do not have explicit judgments for entity relevance, but approximate them by linking all positively judged documents $D^*$ to Wikipedia. We build a relevance model of the entities from the relevant documents (because relevance is binary, each relevant document has the same weight) to construct a probability distribution over entities, $p(E|D^*)$. We take fifty entities with the highest probability and, after manual clean up, use these as a representation of the relevant entities for the topic.

### 3.4 Results

In this section we present the results comparing the different sketch distributions for documents and entities.

We first present results for the 20 most challenging queries for text-based IR approaches. In Figure 2 we see improvements with both the entity linking and entity expansion over the naive sketch. Furthermore, the joint sketch model, which takes the KB relations into account achieves further improvement. Finally, updating the posterior with the like-

[1] Both terms occurring within a window of eight words.
[2] http://factorie.cs.umass.edu/
[3] http://ciir.cs.umass.edu/~jdalton/kbbridge/
[4] http://www.lemurproject.org/galago

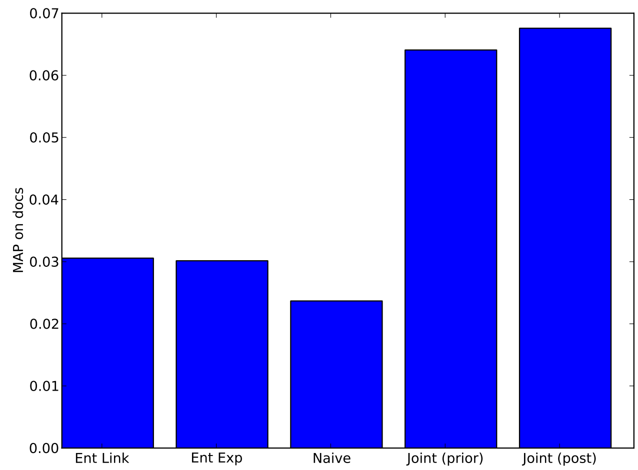


**Figure 2: Document retrieval effectiveness on 20hard queries using mean average precision.**

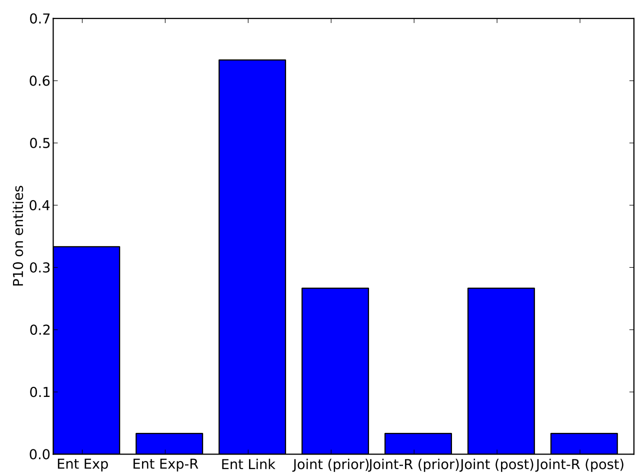


**Figure 3: Entity relevance on 42rand queries using Precision@10.**

lihood from the documents and running another iteration results in further small improvements.

In Table 1 we present document retrieval effectiveness on the random sample of robust queries. The table shows that the initial naive retrieval performs reasonably, returning five relevant documents in the top 10 on average. Performing entity expansion using the linked entities in the documents results in significant improvements in effectiveness. Similarly, using the joint sketches from the KB entities to retrieve documents results in gains in mean average precision.

Beyond document effectiveness, we also present preliminary results evaluating entity effiveness in Figure 3. We observe that the entity linking from retrieved documents performs well. We hypothesize that is due to strong initial retrieval effectiveness for some of the queries. In contrast, the entity expansion model, which uses the retrieved KB entities and relations retrieves on average three relevant entities in the top 10. The (-R) models do not use the in-KB relations. The results show that using the relations in Wikipedia results in significant effectiveness gains.

| Method | MAP | P@10 | nDCG@20 |
|---|---|---|---|
| Naive | 0.239 | 0.500 | 0.435 |
| Ent Exp | 0.286 | 0.562 | 0.450 |
| Joint (prior) | 0.287 | 0.524 | 0.436 |
| Joint (post) | 0.284 | 0.519 | 0.434 |

**Table 1: Document retrieval effectiveness on the 42rand queries.**

# 4. RELATED WORK

Several areas have focused on expanding or updating a knowledge base given large collections of documents. In the context of question answering, Schlaefer et al. use web retrieval to extend seed Wikipedia documents with content with extracted 'text nuggets' nuggets [15]. They find significant improvement in recall using these external sources. Similarly, the TREC Knowledge Base Acceleration track [8] performs filtering on a stream of news documents to identify documents that are citation worthy for entities and to detect changes in slot values over time in order to maintain a knowledge base. In both of these scenarios the focus is on a single entity and does not include a query topic. In contrast, in this work we focus on identifying documents that are both central to a group of relevant entities and more importantly to a user information need.

Recent research has shown that text query expansion using data extracted from Wikipedia can significantly improve retrieval effectiveness for a variety of information retrieval tasks [19, 7]. It has also been used to enrich keyword representations with explicit semantics (ESA) from Wikipedia [9] to improve clustering and classifications tasks. Egozi et al. [6] use pseudo-relevance feedback from ESA annotated text documents to identify concepts and also experiment with fusing text and concept-based scores. Instead of mapping all words to concepts, we link entity mentions explicitly.

Wick and McCallum [18] propose query-aware MCMC which focuses inference on a subset of variable in a graphical model. We similarly use the user information need to focus inference on relevant portions of the document and entity distributions. We use retrieval as a mechanism to measure dependence upon the query.

# 5. CONCLUSIONS

We presented a framework for query-specific knowledge base construction for specific and complex information needs. We introduced the notion of a 'knowledge sketch' as a multimodal representation containing both relevant KB data and unstructured documents. We presented several possible models for estimating sketches by exploiting relationships between entities, documents, and across modalities. We presented preliminary experiments on the TREC 2004 robust collection using Wikipedia as a knowledge base.

In future work, we plan to further explore the relationships between entities and unstructured documents. In particular, to focus on extracting entities, attributes, and relations that are not present in the general purpose KB.

## *Acknowledgements*

This work was supported in part by the Center for Intelligent Information Retrieval, in part under subcontract, prime contractor to DARPA contract #HR0011-12-C-0016. This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors' and do not necessarily reflect those of the sponsor.

## 6. REFERENCES


[1] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Learning concept importance using a weighted dependence model. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 31–40, New York, NY, USA, 2010. ACM.

[2] Michael Bendersky, Donald Metzler, and W. Bruce Croft. Effective query formulation with multiple information sources. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 443–452, New York, NY, USA, 2012. ACM.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[4] Jeffrey Dalton and Laura Dietz. A Neighborhood Relevance Model for Entity Linking. In *Proceedings of the 10th International Conference in the RIAO series (OAIR), 2013*, 2013.

[5] Laura Dietz. Directed Factor Graph Notation for Generative Models. Technical report, Max Planck Institute for Informatics, Saarbrücken, Germany, 2010.

[6] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch. Concept-based feature generation and selection for information retrieval. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*, AAAI'08, pages 1132–1137. AAAI Press, 2008.

[7] Jonathan L. Elsas, Jaime Arguello, Jamie Callan, and Jaime G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 347–354, New York, NY, USA, 2008. ACM.

[8] J. R. Frank, M. Kleiman-Weiner, D. A. Roberts, F. Niu, C. Zhang, C. Re, and I. Soboroff. Building an Entity-Centric stream filtering test collection for TREC 2012. In *Proceedings of the Text REtrieval Conference (TREC)*, 2012.

[9] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1606–1611, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[10] Heng Ji, Ralph Grishman, and Hoa Dang. Overview of the TAC2011 Knowledge Base Population Track. In *TAC 2011 Proceedings Papers*, 2011.

[11] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.

[12] Donald Metzler and W. Bruce Croft. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.

[13] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA, 1998. ACM.

[14] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval in the web of data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 771–780, New York, NY, USA, 2010. ACM.

[15] Nico Schlaefer, Jennifer C. Carroll, Eric Nyberg, James Fan, Wlodek Zadrozny, and David Ferrucci. Statistical source expansion for question answering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 345–354, New York, NY, USA, 2011. ACM.

[16] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[17] Ellen M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of TREC 2004*, 2004.

[18] Michael L. Wick and Andrew McCallum. Query-Aware MCMC. In *NIPS*, pages 2564–2572, 2011.

[19] Yang Xu, Gareth J. F. Jones, and Bin Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 59–66, New York, NY, USA, 2009. ACM.