Information Processing and Management 56 (2019) 1645–1666

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Relevance-based entity selection for ad hoc retrieval

Faezeh Ensan[*,a], Feras Al-Obeidat[b]

[a] Ferdowsi University of Mashhad, Mashad, Iran
[b] Zayed University, United Arab Emirates

ARTICLE INFO

Keywords:
Ad hoc retrieval
Semantic search
Entity-based retrieval
Entity selection

ABSTRACT

Recent developments have shown that entity-based models that rely on information from the knowledge graph can improve document retrieval performance. However, given the non-transitive nature of relatedness between entities on the knowledge graph, the use of semantic relatedness measures can lead to *topic drift*. To address this issue, we propose a relevance-based model for entity selection based on pseudo-relevance feedback, which is then used to systematically expand the input query leading to improved retrieval performance. We perform our experiments on the widely used TREC Web corpora and empirically show that our proposed approach to entity selection significantly improves ad hoc document retrieval compared to strong baselines. More concretely, the contributions of this work are as follows: (1) We introduce a graphical probability model that captures dependencies between entities within the query and documents. (2) We propose an *unsupervised* entity selection method based on the graphical model for query entity expansion and then for ad hoc retrieval. (3) We thoroughly evaluate our method and compare it with the state-of-the-art keyword and entity based retrieval methods. We demonstrate that the proposed retrieval model shows improved performance over all the other baselines on ClueWeb09B and ClueWeb12B, two widely used Web corpora, on the NDCG@20, and ERR@20 metrics. We also show that the proposed method is most effective on the *difficult* queries. In addition, We compare our proposed entity selection with a state-of-the-art entity selection technique within the context of ad hoc retrieval using a basic query expansion method and illustrate that it provides more effective retrieval for all expansion weights and different number of expansion entities.

1. Introduction

The growing availability of knowledge graphs has motivated researchers within the information retrieval community to consider exploiting knowledge graph entities within the ad hoc document retrieval process. Traditionally, retrieval techniques primarily focus on term matching and term proximity features (Paik, 2013; Robertson & Walker, 1994; Song & Croft, 1999) to connect query and document spaces. The use of entities can provide added-value to the retrieval process by offering access to auxiliary information embedded within the knowledge graph (Bagheri, Ensan, & Al-Obeidat, 2018; Dietz, Kotov, & Meij, 2017; Li, Xu et al., 2014; Schuhmacher & Ponzetto, 2014).

Earlier work that used entities relied primarily on *hard matching* between query and document entities (Xiong, Callan, & Liu, 2016). However, later approaches focused on the possibility of using entity relatedness measures learned from the context of entities within the knowledge graph to perform *soft matching* between queries and documents (Xiong, Power, & Callan, 2017). The soft

[*] Corresponding author.
E-mail address: ensan@um.ac.ir (F. Ensan).

https://doi.org/10.1016/j.ipm.2019.05.005
Received 11 August 2018; Received in revised form 30 April 2019; Accepted 8 May 2019

matching strategy has proven particularly helpful for alleviating the vocabulary mismatch problem, which occurs when the entities of a given input query do not directly appear in a highly relevant set of documents. The use of entity relatedness measures would facilitate the retrieval of relevant documents that are expressed in different terminological forms.

1.1. Research objectives and contributions

While the employment of the auxiliary information from the knowledge graph enhances the retrieval process, there are three major challenges, among others, that impede the performance of entity-based retrieval models:

(1) Entity-based retrieval models primarily rely on entity information for document retrieval; therefore, their performance is highly dependent on how well the query is represented as a set of entities. The employment of entity linkers for annotating queries faces limitations in terms of *precision* and *recall*. Although it is possible to apply a stringent confidence threshold for acceptable entities retrieved by the entity linker to maintain a high precision, this comes at the cost of recall. As an example, the average number of entities in the TREC Web 201-250 topics for ClueWeb12B, as explained in our experiments, is only 1.56. As such, it is important to systematically consider methods that enable the selection of additional entities for the representation of the query; hence, addressing the recall problem.
(2) While the use of knowledge graph entity relatedness measures can address issues such as *vocabulary mismatch*, it has been shown that given semantic relatedness is in essence not a *transitive* relation, its application can lead to *topic drift*. As such, it is important to discern between highly related entities on the knowledge graph that are relevant to the query and those that are related yet not relevant to the query.
(3) Considering the fact that queries might be represented with more than one entity, an entity-based retrieval model needs to consider the relationship between the entities observed in the query such that the retrieved results respect the relationship between the entities of the query. Often the entities are complementary and serve to qualify each other, e.g., the query 'Eggs Shelf Life' can be represented by two complementary entities `Egg as food` and `Shelf life`. Hence, such synergistic relationships between query entities need to be taken into consideration.

The objective of this paper is to address the above three challenges by enhancing the representation of a given input query through the selection of a set of relevant entities from the knowledge graph. In this paper, we introduce the Retrieval through Entity Selection (RES) method for finding and scoring entities that are related to a query, which can then be integrated within an entity-based retrieval model. RES models queries and documents in a graphical model, where nodes correspond to entities in them and links represent relatedness between entities. It also models candidate entities as another group of nodes that can have links to document and query nodes.

Let us motivate the importance of entity selection for effective entity-based retrieval. As an example, consider the topic 'Obama Family Tree', which is Topic #1 in TREC Web. The entity representation of this query, when ran through a linker such as TAGME (Ferragina & Scaiella, 2010), would only include `Barack Obama`, which essentially misses the important aspect of the query that relates to Obama's *family*. When performing either query expansion or document retrieval through *soft matching*, many entities would be considered as relevant based on meta-information from the knowledge graph. For example, a document that includes entities `John McCain`, `Barack Obama presidential campaign 2008`, and `Hillary Clinton` is highly relevant to the query entity `Barack Obama`, but it is irrelevant to the query at hand, hence, leading to topic drift. However, a document that includes `Ann Dunham` and `Barack Obama Sr.` is highly relevant to the query, even though knowledge graph-based relatedness methods may produce lower relatedness values for these entities compared to the less relevant entities mentioned earlier.

Our proposed relevance-based entity selection model addresses this challenge by joining the entity-based representation of the query and the pseudo-relevance feedback document collection. In our model, from among the candidate entities from the knowledge graph, we select those entities that result in higher retrieval effectiveness for the query. Employing a probabilistic graphical model, RES ranks candidate entities based on their relatedness to query entities and also based on their relatedness to entities in pseudo-relevant documents. In other words, a candidate entity that is semantically related to the query entity (e.g., `John McCain` in our example) receives a lower rank compared to an entity that is semantically related to both the query entity and the entities found in top-ranked pseudo-relevant documents (e.g., `Ann Dunham`). Using semantic relatedness, RES ensures those entities in pseudo-relevant documents that are not related to the query are ruled out. This approach ensures that only relevant entities are selected; hence, addressing both the *topic drift* and *precision/recall* challenges. Furthermore, the graphical model employed for representing the query, document and pseudo-relevant document spaces ensures that the interaction between entities is taken into consideration.

We will show in our experiments that our entity selection method can facilitate query expansion using relevant and effective entities that enhance retrieval effectiveness and as such addresses (1) the precision/recall, (2) the topic drift, and (3) the query entity interaction challenges introduced earlier. More concretely, the contributions of our work are as follows:

(1) We introduce a graphical probability model that captures dependencies between entities within the query and document spaces in the form of graph cliques, which is a richer form of query-document space integration compared to the state-of-the-art.
(2) We propose an *unsupervised* entity selection method based on the above graphical model for integration into the ad hoc document retrieval process. The entity selection model can facilitate query expansion and more effective retrieval.
(3) We evaluate our work based on different TREC datasets and show that our work outperforms state-of-the-art methods in ad hoc retrieval on several metrics.

From a theoretical viewpoint, the work in this paper distinguishes itself from the existing work in the literature in that it allows for the systematic selection of relevant knowledge graph entities from the set of pseudo-relevant documents. This is novel because our proposed model forms a graphical model composed of entities within query and document spaces that allow for the identification of cliques, which are in turn the basis for calculating entity relevance for entity selection. From a practical viewpoint, we show that the entities that are selected are instrumental for improving the performance of ad hoc document retrieval. We empirically show this by comparing our work with strong state-of-the-art methods on standard benchmark datasets.

The rest of this paper is structured as follows: Section 2 introduces related work in entity-based information retrieval. In Section 3, we provide the technical details of our proposed method, which is then followed by the evaluation of our work from the perspectives of ad hoc document retrieval via entity selection and query expansion. Section 5 concludes the paper.

2. Related work

The work presented in this paper is mainly related to two directions of work in the literature: searching and ranking entities for queries, and entity-based document retrieval.

There is a rich body of work that explores entity search and retrieval for web queries (Balog, Azzopardi, & de Rijke, 2009; Kaptein, Serdyukov, De Vries, & Kamps, 2010; Liu, Zheng, & Fang, 2013; Serdyukov, Rode, & Hiemstra, 2008). The task of entity retrieval has been formally defined in the literature as retrieving a ranked list of semantic web entities, or RDF resources, for a keyword query (Pound, Mika, & Zaragoza, 2010). This task has been evaluated by manually annotated lists of queries (Pound et al., 2010) or through a specific test collection (Balog & Neumayer, 2013) in which a number of queries from different query sets, e.g., the INEX 2009 Entity Ranking track (Demartini, Iofciu, & De Vries, 2010), are mapped to DBpedia entities. A new version of this test collection with a more recent DBpedia dump is introduced recently (Hasibi, Nikolaev et al., 2017). The work presented in Zhiltsov, Kotov, and Nikolaev (2015) uses this dataset to evaluate its model, which exploits term dependencies for ad hoc entity retrieval. Yahya, Barbosa, Berberich, Wang, and Weikum (2016) also include this dataset in their experiments for investigating *relationship* queries through casting the problem into a structured query language such as SPARQL.

Learning to rank methods, which are generally used for document ranking, are applied for learning the relevance of an entity to a web query (Chen, Xiong, & Callan, 2016), where the features are the ranking scores, e.g. BM25 and SDM (Metzler & Croft, 2005). Entity type and hierarchical type information has been investigated in entity retrieval Garigliotti, Hasibi, and Balog (2018) and exploited for defining a new smoothing method for entity retrieval language models (Lin & Lam, 2018). Finally, an open source toolkit for entity linking and entity retrieval is introduced, which implements a number of state-of-the-art methods (Hasibi, Balog, Garigliotti, & Zhang, 2017).

The main difference between these entity ranking methods and the work presented in this paper lies in the objective and hence, in the evaluation methods. While the main objective of our method is to find the most relevant documents to a web query by means of selecting related entities, the introduced methods aim at finding entities that can conceptualize the intent of queries. Consequently, while our method is evaluated with regards to document relevancy, entity selection methods are evaluated with regards to entity relevancy. Recently, a new entity selection and ranking method has been proposed, referred to as REWQ (Schuhmacher, Dietz, & Paolo Ponzetto, 2015), that aims at finding the set of entities that cover different aspects of the query instead of the dominant approach in the literature for finding a number of homogeneous entities. For example, for query 'Argentine British relations', it finds entities of different types such as '`Falklands War`' and '`Margaret Thatcher`' in order to conceptualize related facets into the query intent. This method provides a new evaluation dataset that maps entities to queries on this basis. We use REWQ as one of the baselines in our experiments.

The second direction of related work is the retrieval models that use knowledge graphs for searching and ranking documents (Balaneshinkordan & Kotov, 2016; Dietz et al., 2017; Egozi, Markovitch, & Gabrilovich, 2011). These works include research that introduce different features based on entity embeddings along with word and document embeddings and investigate their effectiveness in various learning to rank methods (Ensan, Bagheri, Zouaq, & Kouznetsov, 2017), and those that use Wikipedia and Freebase, as two important samples of knowledge graphs, for generating related terms to a query for query expansion (Keikha, Ensan, & Bagheri, 2017; Xiong & Callan, 2015; Xu, Jones, & Wang, 2009). Krishnan, Deepak, Ranu, and Mehta (2018) propose a method to address diversified query expansion, i.e., expanding queries with appropriate terms such that the top retrieved results cover diverse aspects of a query. Here, Wikipedia information and word embeddings are used to prioritize candidate terms, which are taken from the initial query search results. Wikipedia along with the document collection are used for *document expansion* (Sherman & Efron, 2017), instead of the usual query expansion, for better retrieval performance.

There are a number of works that model documents and queries as bag of entities where entities are usually found by automatic entity linking systems (Shen, Wang, & Han, 2015). Based on the bag of entities representation, the number of shared entities in query and document entity representations can be used for document ranking (Xiong et al., 2016). In another work based on a bag of entity representation, the relatedness between query and document entities are estimated based on a knowledge graph that is built using the Semantic Scholar search corpus and Freebase (Xiong, Power et al., 2017). In this retrieval model, the maximum relatedness between any document entity and all query entities are found first, and then the number of relatedness values in predefined ranges are counted and used for calculating the ranking score. The work presented in Raviv, Kurland, and Carmel (2016) defines a retrieval model based on the occurrence of query terms and query entities in documents.

Latent Entity Space (LES) model is proposed as a new retrieval approach according to which queries and documents are projected into a set of latent entities, and the relevance between a query and a document is estimated based on their projections in this latent entity space (Liu & Fang, 2015). In this work, the probability of an entity belonging to the latent representation is estimated by means of matching between the text surrounding an entity mention in documents in the collection (LES-COL) or in a knowledge base such as Freebase (LES-FB).

EQFE (Dalton, Dietz, & Allan, 2014), is a retrieval model that expands queries by name, anchors, and categories, among other information of related entities. The scores calculated for each document from expansion methods are used as features in a learning to rank system for estimating the final score of the document given a query. The Semantic Enabled Language Model (Ensan & Bagheri, 2017), SELM, is another knowledge-based retrieval method that models queries and documents as a graph of entities where the semantic relatedness between a document entity and a query entity is employed for document ranking. A more recent work by Xiong, Callan, and Liu (2017), referred to as *Duet* here, uses a neural attention model to identify and highlight important segments of the query, remove noisy entities and also rank documents. In this paper, we use LES (Liu & Fang, 2015), SELM (Ensan & Bagheri, 2017), EQFE (Dalton et al., 2014) and Duet (Xiong, Callan et al., 2017) as our baselines in the experiments.

In the work presented in this paper, we used a probabilistic graphical model to estimate the probability of observing query entities, given the document entities, where there may be semantic relatedness between query entities and between query and document entities. Probabilistic graphical models have been previously used for retrieval. Sequential Dependency Model (Metzler & Croft, 2005), SDM, is a well-recognized work that uses Markov Random Fields (MRFs) for modeling dependencies between query terms. MRFs are also used for generating one-term or multi-term concepts related to a query for the purpose of query expansion (Metzler & Croft, 2007). Our work differs from these works by focusing on entities and their semantic dependencies instead of terms. Here, the graphical model encodes a document as a set of nodes (contrary to one-node representation of a document in previous works Metzler & Croft, 2005; Metzler & Croft, 2007), where each document node, which represents an entity, may be connected to an arbitrary number of query nodes because of the semantic relatedness between entities. In our experiments, we use SDM as one of our baselines for the purpose of evaluation and comparison.

3. Proposed approach

In this section, we introduce our proposed approach, called Retrieval through Entity Selection (RES). In RES, queries and documents are represented as a set of entities such that $q = \{qe_1, qe_2, \ldots qe_m\}$ and $d = \{de_1, de_2, \ldots, de_n\}$ where qe_i and de_j represent query entities and document entities, respectively drawn from a knowledge graph. The objective of RES is to estimate $P(d|q)$, which is achieved in two steps: (1) expanding q based on the entities observed in the Pseudo-Relevance Feedback (PRF) document collection, and (2) ranking documents based on the expanded query. In order to expand the query, we estimate $P(c|q)$ for each candidate entity c as:

$$P(c|q) = \frac{\Sigma_{d\in D} P(c,\, q|d) P(d)}{\Sigma_{d\in D}\Sigma_{c\in C} P(c,\, q|d) P(d)} \tag{1}$$

where C, referred to as the set of candidate entities, includes the set of entities observed in the PRF document collection, and D is the set of all documents in the corpus. Analogous to widely-adopted relevance models (Lavrenko & Croft, 2001; Metzler & Croft, 2007), we approximate $P(c|q)$ in Eq. (1) by summing over $R \subset D$, which consists of the pseudo-relevant feedback documents for query q. Given the denominator in Eq. (1) is the same for all candidate entities, the ranking function, $f_{rank}(c|q)$, can be estimated as:

$$f_{rank}(c|q) \approx \Sigma_{d\in R} P(c,\, q|d) P(d) \tag{2}$$

Assuming that $P(d)$ is uniform over all documents in the collection, the main task of RES is to estimate $P(c,\, q|d)$, i.e., the joint conditional probability of a candidate entity and the set of query entities given entities observed in d. For this purpose, RES adopts an undirected graphical model for representing entities and their degrees of relatedness. In this graphical model, the set of nodes consists of the candidate entity being ranked as well as the entities in the query and document. Each edge represents the relatedness of two entities. There is body of work that focuses on finding relatedness between knowledge base entities (Feng, Bagheri, Ensan, & Jovanovic, 2017; Jiang, Bai, Zhang, & Hu, 2017; Strube & Ponzetto, 2006; Witten & Milne, 2008). In our work and in order to calculate the degree of relatedness between the document entity and the query and candidate entities, we employ the neural embedding-based representation of entities (Wang, Zhang, Feng, & Chen, 2014). In this approach, each entity is represented as a low dimensional dense vector where the cosine of the angle between two vectors measures their relatedness. In forming the graph, we assumed that there is an edge between two nodes where their cosine similarities is more than a threshold, which is set to 0.1 in our experiments. A clique in this graph is a fully-connected subgraph, which is a subset of entities such that there is an edge between every two of them, i.e. every two entities are similar based on their neural embedding vector representations. As we clarify in the following paragraphs, we are only interested in those cliques that has at least one entity from each of the following sets of entities: document entities, query entities, and the candidate set of entities, which is collected from the entities observed in the PRF document collection.

We use a variation of Conditional Random Fields (CRFs) (Lafferty, McCallum, & Pereira, 2001) for finding the conditional probability of the target variables (query and candidate entities), given the observed variables (document entities). Conditional Random Fields are usually applied to the supervised settings where the weights of different feature functions are learned based on available training data. On the contrary, our model is an unsupervised ranking model that defines features and their weights based on relatedness between entities in queries and documents. Here, the application of CRF in our work is very close to the dominant application of Markov Random Fields (MRF) (Metzler & Croft, 2005; 2007) in unsupervised retrieval systems. The only important difference is that CRFs, contrary to MRFs, do not encode the distribution over the observed variables, which are document entities in our case. It means that a document may consist of a set of entities with arbitrary number of dependencies while the probability distribution does not need to model these dependencies. The only important dependencies to be modeled are the ones that exist between (1) the candidate entity and the query entities, and (2) the candidate and query entities and the document entities. In fact, by using CRFs, we avoid encoding the distribution over the document entities whose dependencies may be very complex. RES differs from other work based on CRFs such as Ensan and Bagheri (2017) in that instead of defining a simple distribution probability over query-document entity pairs, RES considers inter-query entity dependencies as well as the dependency between a selected set of query-document entity cliques as discussed later.

The main objective of RES is to select entities that are jointly related to entities in the query and in the top-ranked documents. Hence, a candidate entity that is strongly related to all or most entities in the query while it is semantically related to a number of pseudo-relevant entities has a stronger chance of being selected by RES for query expansion compared to an entity that is related to just one or a limited number of query entities, or an entity that is strongly related to pseudo-relevant entities but has no semantic relevance to the query.

3.1. Explanatory example

In this section, we explain the main concepts related to RES through two examples. First, we illustrate the main components of the graphical model and second, we investigate the impact of dependencies between query entities. For the sake of the first example, let us assume that the TREC Web query #200, '*ontario california airport*' is represented by two entities: 'Ontario, California' and 'Airport'. Further assume that the following three entities are being considered as potential entities for expansion: 'Ontario International Airport', 'International Air Transport Association', and 'Corona, California'. We would need to find $f_{rank}\ (c_i|q)$ for each of these candidate entities. The ranking model operates over pseudo-relevance feedback document collection and for this purpose let us consider a document *d* from this set that has four concepts: 'Airport Terminal', 'California', 'Los Angeles International Airport', and 'American Airlines'. Fig. 1 depicts the entities and their relatedness for three candidate entities.

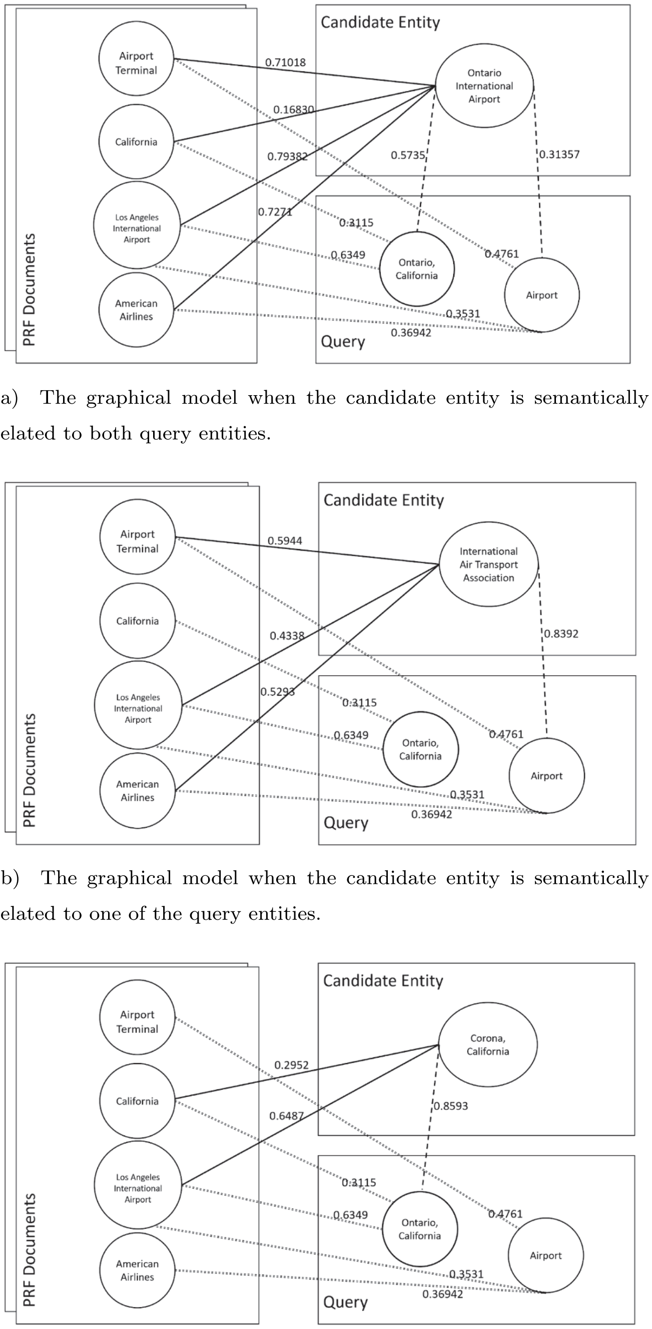
In RES, we capture three types of dependencies between the nodes in the graphical model: (1) dependency between query entities, which has not been shown in Fig. 1, because the two query entities are not related to each other; (2) dependency between a candidate entity and the query entities. For instance, 'Corona, California' is related to 'Ontario, California' (Fig. 1c) while 'Ontario International Airport' is related to both query entities (Fig. 1a), and (3) dependency between entities in the PRF collection and the query entities as well as candidate entities. For example, 'Ontario, California' is related to 'California' and 'Los Angeles International Airport' and 'Airport' is related to 'Airport Terminal', 'Los Angeles International Airport', and 'American Airlines'. As mentioned earlier, we avoid modeling dependencies between document entities. In Fig. 1, the labels on the edges are the cosine similarities between the vector representations of entities and show how strongly are two entities related. For example, Fig. 1b shows that 'International Air Transport Association' is strongly related to 'Airport' (0.8392 of 1). As we will see in Section 3.2, the proposed model uses these similarities for ranking candidate entities.

Fig. 2 shows the graphical model for the TREC Web query #200, '*sonoma county medical services*', where two entities are linked to the query, namely 'Sonoma County, California' and 'Health Care' and two candidate entities are depicted: 'Santa Rosa Memorial Hospital' (Fig. 2a) and 'Psychiatry' (Fig. 2b). In this example, query entities are related to each other (with the similarity of 0.2322). This example gives insight into the candidate ranking process whereby those entities that are closer to a higher number of query and PRF document entities would be ranked higher. For instance, in this example, 'Santa Rosa Memorial Hospital' would need to be ranked higher than 'Psychiatry' given it is strongly related to both query entities as well as a larger number of PRF document entities. In the following sections, we will refer to this example to explain about the graphical model and the ranking method in more details.

3.2. Candidate entity ranking for expansion

The first step of RES is to rank candidate entities for query expansion based on the dependencies in the graphical model. Let $G=(V,E)$ be an undirected graph, where $V=d\cup qc$ such that *d* is the set of random variables whose values are derived from the entities observed in the representation of any input document and *qc* is the set of random variables corresponding to the union of query entities and candidate entity, whose values need to be estimated by RES. Let *N* be the total number of entities in the knowledge graph, *d* and *qc* would each have *N* variables that take binary values of (1,0), corresponding to existence or non-existence of that entity in the document, and in the union of the query and candidate entity sets, respectively. At least one random variable in *d* must be 1. This means that any input document needs to have at least one entity for it to be considered by RES. In addition, at least two random variables in *qc* need to be 1, because there has to be at least one query entity and exactly one candidate entity in *qc*. In this graph, *E* is the set of undirected edges that connect related entities. Related entities are determined based on their degree of relatedness within the knowledge graph and are independent from the document and query collections. This representation of nodes and edges ensures that the structure of the graph *G* is fixed for all documents and queries, while the values of the random variables will form different variations for the graph.

In order to rank each candidate entity, we need to compute the joint probability distribution of d given qc over *G* based on the original formulation of CRFs as defined in (Sutton, McCallum et al., 2012; Wallach, 2004):



(a) The graphical model when the candidate entity is semantically related to both query entities.

(b) The graphical model when the candidate entity is semantically related to one of the query entities.

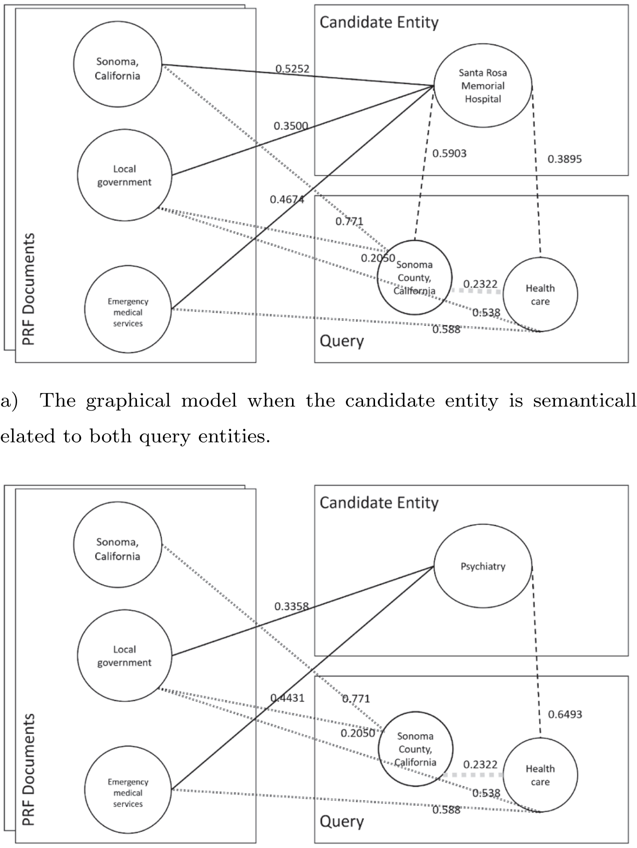
(c) The graphical model when the candidate entity is semantically related to one of the query entities.

Fig. 1. Explanatory example for the proposed approach: main components of the graphical model.

$$P(qc|d) = \frac{1}{Z(d)} exp\left(\sum_{i=1}^{i=k} f_k(Cl_i, qc, d) \right) \tag{3}$$

where Cl_i is the ith clique where there are exactly k cliques and f_k is a feature function defined over the kth clique.

Referring to our example (Fig. 2) in Section 3.1, '`Psychiatry`', '`Local Government`' and '`Health Care`' form a clique of size 3. This clique has one concept from each category (query, document and candidate set). On the other hand, '`Santa Rosa Memorial Hospital`', '`Local Government`', '`Sonoma County, California`', and '`Health Care`' form a clique of size 4, that includes two



(a) The graphical model when the candidate entity is semantically related to both query entities.

(b) The graphical model when the candidate entity is semantically related to one of the query entities.

Fig. 2. Explanatory example for the proposed approach: query entities dependencies.

query entities. For ranking two candidate entities, ‘`Psychiatry`’ and ‘`Santa Rosa Memorial Hospital`’, Eq. (4) finds all cliques to whom each of this candidate entities belong and makes a summation over the features defined over theses cliques.

In Eq. (4), $Z(d)$ is a normalization constant:

$$Z(d) = \sum_{qc \in \mathcal{P}} exp\left(\sum_{i=1}^{i=k} f_k(Cl_i, qc, d) \right) \tag{4}$$

where $\mathcal{P}$ is the power set of query entities and candidate entities. In Eq. (4), we set $f_k(Cl_i, qc, d) = 0$ when at least one of the random variables in Cl_i is 0. This means that we define features over the cliques of G, when all entities corresponding to the random variables in the cliques exist collectively in the query entity and candidate entity sets. Based on this, each feature is defined as follow:

$$f_k(Cl_i, qc, d) = \sum_{d_j \in d} ef(d_j, d) \times Sim(Cl_i, qc, d_j) \tag{5}$$

where d_j is an entity in document d and $ef(d_j, d)$ is the frequency of the entity d_j in d. $Sim(Cl_i, qc, d_j)$ denotes the relatedness between the document entity d_j and the query and candidate entities in Cl_i. As we mentioned earlier, we employed the embedded vector representation of entities for finding their relatedness. Neural embeddings have shown to have interesting geometric properties, e.g., the representation of a bag of entities could be calculated by averaging over the vectors of its constituting entities forming the centroid for that bag of entities.

As such, we can calculate $Sim(Cl_i, qc, d_j)$ as follows:

$$Sim(Cl_i, qc, d_j) = \quad cosine(\overrightarrow{d_j}, \overrightarrow{Cl_i}) \tag{6}$$

where $\overrightarrow{Cl_i}[k]$ is defined as:

$$\vec{Cl_i}[k] = \frac{\sum_{a \in Cl_i} \vec{a}\,[k]}{|Cl_i|} \tag{7}$$

Also, $\vec{Cl_i}[k]$ denotes the kth dimension of the embedding representing the centroid for Cl_i and $\vec{a}\,[k]$ is the kth dimension of the vector for entity a. In our experiments, we use the entity embeddings provided in the literature (Li et al., 2016), which have shown to provide strong performance on a number of competitive tasks. Note that in Eq. (3), given we use an unsupervised application of graphical probabilistic models, no weights are learnt and instead the feature function consists of similarities between the concepts in the cliques.

Returning back to our running example, in the case depicted in Fig. 2, For the 3-clique '`Psychiatry`', '`Local Government`' and '`Health Care`', the feature function is defined based on the cosine similarity between the centroid vector for '`Psychiatry`' and '`Health Care`' and the vector that represents the document entity '`Local Government`'. Also, for the 4-clique '`Santa Rosa Memorial Hospital`', '`Local Government`', '`Sonoma County, California`', and '`Health Care`', the feature function is estimated based on the cosine similarity between the centroid vector for '`Santa Rosa Memorial Hospital`', '`Sonoma County, California`', and '`Health Care`' and the vector representation for '`Local Government`'. In this example, '`Santa Rosa Memorial Hospital`' belongs to both 3-cliques and 4-cliques and can have higher chance to be ranked higher depending on the similarities to the document entities.

We define the entity ranking model as follows:

$$f_{RES}(c, q|d) = logP(qc|d) \tag{8}$$

where c is the candidate entity and q is the set of query entities. Also, $P(qc|d)$ is estimated as follows:

$$P(qc|d) \approx exp\left(\sum_{i=1}^{i=k} f_k(Cl_i, qc, d)\right) \tag{9}$$

Note that in Eq. (9), the normalization constant is dropped, because computing Z for an exponential number of possible query entities is computationally expensive. Recalling that $Z(d)$ is a document-dependent constant and does not relate back to either the query or the candidate entities; as such, we assume a uniform distribution for $Z(d)$ across pseudo-relevant documents and hence remove it. The final entity selection score is defined as follows:

$$\begin{aligned} Score_{RES}(c, q) &= \sum_{d \in R} log(f_{RES}(c, q|d)) \\ &= \sum_{d \in R} log\left(\sum_{i=1}^{i=k} f_k(Cl_i, qc, d)\right) \\ &= \sum_{d \in R} log\left(\sum_{Cl_i} \sum_{dj \in d} ef(d_j, d) \times Sim(Cl_i, qc, d_j)\right) \end{aligned} \tag{10}$$

In Eq. (10), summing over the logarithmic form of feature summation makes the score of a candidate entity with similarities across a number of documents higher than an entity with strong similarities to just one or few documents.

3.3. Entity-based retrieval

In the second step of our work, we employ the candidate entity rankings to perform query expansion before document ranking. We follow the popular query re-weighting approach (Carpineto & Romano, 2012) for entity expansion as follows:

$$w'_{e,q'} = (1 - \alpha)w_{e,q} + \alpha Score^{n}_{RES}(e, q) \tag{11}$$

where $Score^{n}_{RES}(e, q)$ is $Score_{RES}(e, q)$ normalized with respect to the maximum and minimum scores obtained for candidate entities, q' is the expanded query, q is the original query, and $w_{e,\ q}$ and $w'_{e,q'}$ shows the weight of an entity in the query and the expanded query, respectively. We estimate $w_{e,\ q}$ as follows:

$$w_{e,q} = Sim(e, q) \tag{12}$$

Sim(e, q) denotes the similarity between an entity and the set of all entities in the query, which is computed by the cosine similarity between the embedding vector for e and the centroid of the embedding vectors of the entities in q. Given these new weights, any baseline retrieval model such as a standard language model or BM25 can be applied for document ranking and retrieval. In our experiments, we used BM25.

3.4. Interpolation with keyword-based retrieval models

Combining entity-based retrieval with keyword systems is a standard approach in models that benefit from a knowledge graph. It has been reported that entity-based retrieval can enhance keyword-based systems (Liu & Fang, 2015; Raviv et al., 2016). One of the important reasons for this could be that some queries do not have an entity representation to capture their full meaning. For example, for the TREC Web #138: 'jax chemical company', the entity linking system finds '`Chemical industry`', but there is no entity in the knowledge graph for 'jax' or 'jax company'. In such queries, term matching for 'jax' is more effective than similarity based on entities.

Different works have reported a linear combination of entity-based retrieval with other retrieval systems (Bagheri et al., 2018; Liu & Fang, 2015; Raviv et al., 2016; Xiong, Power et al., 2017), where the entity retrieval score is linearly interpolated with the baseline retrieval system with a coefficient that is learned on training data. In this work, we use a similar strategy and linearly interpolate the normalized scores of RES with a keyword-based baseline using a coefficient that is learned using cross-validation, as explained later. In other words, the final score obtained for a document d given the query q is obtained as follows:

$$Score(d, q) = (1 - \lambda_{RES})Score_{KW}(d, q) + \lambda_{RES}Score_{RES}(d, q) \tag{13}$$

where $Score_{KW}(d, q)$ is the normalized score found by the baseline keyword-based system, $Score_{RES}(d, q)$ is the normalized score obtained by RES through query expansion (as explained in Section 3.3), and λ_{RES} is a coefficient that balances the impact of keyword-based versus RES in the final scoring.

4. Experiments

The work presented in this paper includes a stage of entity selection and ranking for ad-hoc queries for the purpose of document retrieval. In order to evaluate the work proposed in this paper, we conducted two sets of experiments, namely (*i*) retrieval via query expansion and (*ii*) retrieval via entity selection. In the former set of experiments, we evaluate the performance of the proposed retrieval model; which includes the selection and ranking of entities for queries, using the selected entities for ranking documents, and finally interpolating entity-based retrieval with a baseline keyword-based system for the purpose of more comprehensive ranking; and compare it with the performance of a variety of keyword-based and entity-based retrieval systems. In the second set of experiments, we focus on the first stage of the proposed approach, which is the process of entity selection and entity ranking for ad-hoc queries. The purpose of this set of experiments is to evaluate the quality of entity ranking algorithm and comparing it with the state-of-the-art solutions assuming that the document retrieval method that uses these entities for ranking are identical. For this purpose, we used a basic entity frequency-based retrieval method and compare its performance when the entities provided by the ranking algorithm proposed in this paper and when they are provided by the state-of-the-art entity selection method. More details on experimental setups, baselines, and results are reported in the following sections.

4.1. Retrieval via query expansion

4.1.1. Baselines

For the sake of comparison, we choose two keyword-based retrieval systems, Sequential Dependency Model (SDM) (Metzler & Croft, 2007) and the RM3 variant of the Relevance Model (Lavrenko & Croft, 2001). SDM is a state-of-the-art retrieval model that uses Markov Random Fields for modeling dependencies between query terms. RM3 is also a strong baseline that finds relevant terms to a query and expands the original query with the expanded terms. We also use five entity-based retrieval systems LES-FB and LES-COL (Liu & Fang, 2015), SELM (Ensan & Bagheri, 2017), EQFE (Dalton et al., 2014) and Duet (Xiong, Callan et al., 2017) introduced in Section 2.

In order to keep our experiments comparable to these methods, we used the parameter settings reported in Dalton et al. (2014); Ensan and Bagheri (2017); Liu and Fang (2015) and Xiong, Callan et al. (2017) for the baseline methods. In Liu and Fang (2015), pertaining to LES-COL and LES-FB, the available runs are reported for only 20 documents per query.

4.1.2. Experimental setup

We use ClueWeb09 Category B dataset (ClueWeb09B), which consists of the first 50 million English Web pages of ClueWeb09, and ClueWeb12 Category B (ClueWeb12B) dataset, which is a subset of over 50 million documents from ClueWeb12 in our experiments. Two of our baselines, namely LES-COL and LES-FB, reported their results over ClueWeb09 Category B, but did not report results for the ClueWeb12B dataset. As such in our evaluation, we included LES-COL and LES-FB in ClueWeb09B but not in ClueWeb12B.

The queries that are used include TREC Web track topics 1-200 for ClueWeb- 09B, and Web track topics 201-250 for ClueWeb12B. We used a locally installed version of TAGME (Ferragina & Scaiella, 2010) for entity linking. This is the most widely used strategy for obtaining entities in entity-based ranking models, *cf.*, Raviv et al. (2016) and Xiong et al. (2016); Xiong, Callan et al. (2017). One of the reasons for adopting this strategy by the related literature has been the findings by Dalton et al. (2014) that show FACC1 (Gabrilovich, Ringgaard, & Subramanya, 2013) does not necessarily contain annotations for the majority of Wikipedia articles in the ClueWeb corpora. As suggested in Dalton et al. (2014), we built a pool of documents consisting of top-100 documents from the baselines (top-20 for the LES variants) for each query. We use the publicly available runs provided by these baselines.[1] Based on Xiong, Callan et al. (2017), all ClueWeb documents were parsed using Boilerpipe (Kohlschütter, Fankhauser, & Nejdl, 2010) where 'KeepEverythingExtractor' was used to maintain as much of the document content as possible. Document pools, entities found by TAGME, along with the results of our runs and employed qrels are publicly available.[2] In terms of evaluation metrics, we report

[1] http://ciir.cs.umass.edu/downloads/eqfe/runs/, http://xtliu.com/data/les/ and https://github.com/SemanticLM/SELM, http://boston.lti.cs.cmu.edu/appendices/SIGIR2017_word_entity_duet/.

[2] https://github.com/EntityBasedIr/RES-IR.

Table 1
Results of the comparative performance of RESS with different baselines. Values denoted by † show cases where RESS has a statistically significant better performance according to paired *t*-test at *p*-value < 0.05.

		MAP@20	ΔMAP@20	NDCG@20	ΔNDCG@20	ERR@20	ΔERR@20
ClueWeb09B	RM	0.1994†	−0.0260 (−13.06%)	0.2554†	−0.0723 (−28.29%)	0.1504†	−0.052 (−25.41%)
	SDM	0.1916†	−0.0339 (−17.69%)	0.2488†	−0.0789 (−31.70%)	0.1387†	−0.064 (−31.28%)
	EQFE	0.1814†	−0.0440 (−24.26%)	0.2384†	−0.0893 (−37.48%)	0.1419†	−0.062 (−30.64%)
	LES-COL	0.1053†	−0.0273 (−25.88%)	0.2834†	−0.0442 (−15.61%)	0.1735†	−.031 (−15.19%)
	LES-FB	0.1129†	−0.0196 (−17.36%)	0.2998†	−0.0278 (−9.29%)	0.2006	−0.003 (−1.92%)
	SELM	0.2002†	−0.0253 (−12.63%)	0.2691†	−0.0586 (−21.79%)	0.1494†	−0.0553 (−26.94%)
	Duet	0.1797†	−0.0458 (−25.49%)	0.3213	−0.0064 (−1.99%)	0.2026	−0.002 (−0.9%)
	RESS	**0.2255** (0.1326**)		**0.3277**		**0.2046**	
ClueWeb12B	RM	0.0357†	−0.0215 (−60.16%)	0.1085†	−0.0670 (−61.80%)	0.0776†	−0.501 (−39.23%)
	SDM	0.0417†	−0.0155 (−37.24%)	0.1239†	−0.0516 (−41.66%)	0.09231†	−0.0353 (−27.71%)
	EQFE	0.0454†	−0.0118 (−25.99%)	0.1430†	−0.0325 (−22.75%)	0.1064†	−0.0203 (−16.6%)
	SELM	0.0443†	−0.0129 (−29.12%)	0.1315†	−0.0440 (−33.49%)	0.0995	−0.0282 (−22.08%)
	Duet	0.0472†	−0.01 (−21.08%)	0.1724	−0.0031 (−1.77%)	0.1213	−0.0064 (−5.01%)
	RESS	**0.0572**		**0.1756**		**0.1277**	

NDCG@20 and ERR@20 where statistical significance is determined and reported using a paired *t*-test with a *p*-value < 0.05.

Pseudo-relevance feedback documents needed by our approach were obtained based on top-k documents retrieved by SDM for each query. We used a single pass approach for parameter tuning. We performed *five-fold cross validation* on queries where each fold consisted of 20% of the queries. We used 80% of the queries for training the parameters in each iteration, where the trained parameters are exploited for answering the remaining 20% of the queries through the system. Repeating in five iterations, we make sure that the reported results for each fold are obtained using the parameters that are trained over the remaining four folds of queries. For parameter setting, a combination of all possible values for the parameters are calculated and used for parameter setting. Four parameters, namely, the interpolation co-efficient, the expansion co-efficient (α in Eq. (11)), the number of expansion entities, and the value for k in top-k documents retrieved by SDM for pseudo-relevance feedback, were set based on this approach. Parameter are tuned to optimize NDCG@20. The interpolation coefficient and the expansion coefficient tuned over a range of values between 0 and 1 with the interval of 0.1 $(0.1, 0.2, \dots, 0.9)$. The number of expansion entities tuned over a range of values between 10 and 100 with the interval of 10. Finally, the k variable is tuned over 5, 10, 20, and 100 documents. Based on our parameter tuning method, we set the following values for the parameters: for both datasets, α is set to 0.1 and k is set to 5. For the ClueWeb09 the interpolation variable is set to 0.5 and the number of expansion entities is set to 100, while in ClueWeb12B dataset these parameters are set to 90 and 0.3, respectively. The results that are reported are those that found using the tuned parameters.

4.1.3. Results

In this experiment, RESS denotes the interpolation of RES with SDM according to Section 3.4. The reason we chose SDM for interpolation was because (1) SDM is a purely keyword-based model unlike other baselines such as EQFE, SELM and Duet, which consider entities, and (2) It does not perform query expansion and deals with the query as-is unlike RM and SELM. The results are shown in Table 1 where NDCG@20 and ERR@20 values are reported for each baseline as well as for RESS. As mentioned earlier, the authors of the two LES variants have not published runs for the ClueWeb12B corpus and as such the table does not include LES for the ClueWeb12B dataset. As shown in Table 1 RESS shows improved performance compared to all the baselines on both of the corpora for the NDCG@20 and ERR@20 metrics.

4.1.3.1. Success/failure analysis. In Figs. 3 and 4, we show how RESS performs compared to each of the baselines on a per-query basis. In the figures, the relative improvement of MAP over each baseline has been reported, i.e., a higher mass on the top-left compared to the bottom right shows that a higher number of queries have been helped by RESS compared to the baseline. We also report the actual number of queries helped by RESS (improved MAP) and hurt by RESS (reduced MAP) compared to the baseline in each chart denoted by (*a*, *b*). It should be noted that $a + b$ does not always add up to the total number of queries as there are cases where the performance of RESS is tied with the baseline. Compared to all of the baselines and for both corpora, RESS helps a larger

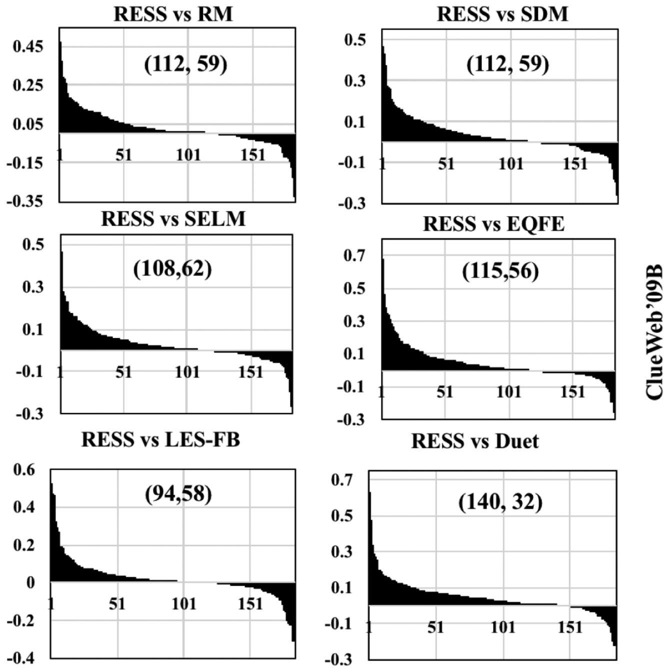


Fig. 3. Delta of MAP RESS over the baselines on ClueWeb09B. Positive values show improvement. LES-COL not plotted due to weaker performance to LES-FB.

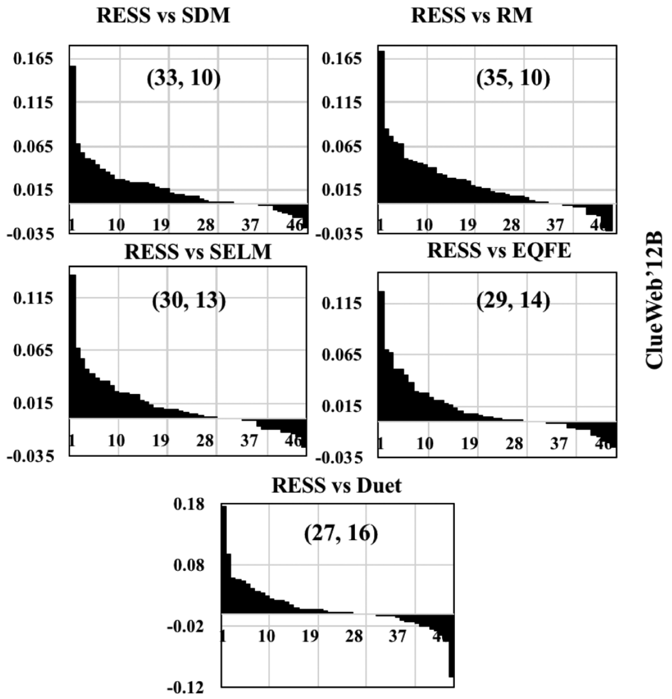


Fig. 4. Delta of MAP of RESS over the baselines on ClueWeb12B. Positive values show improvement.

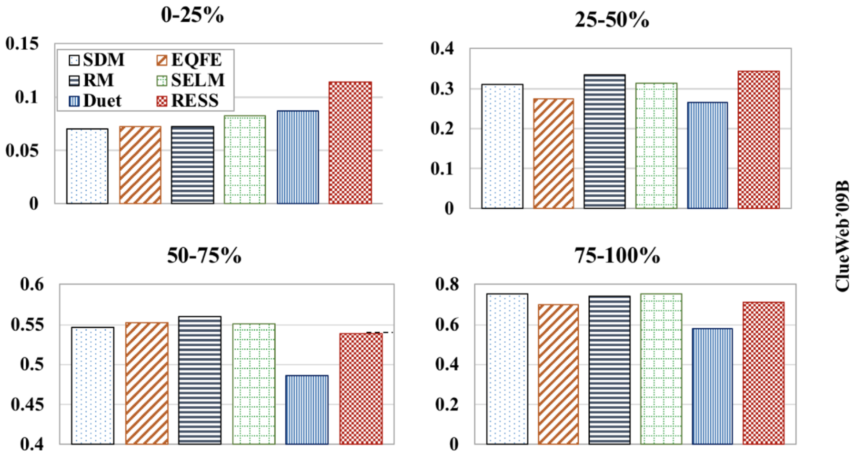


Fig. 5. Mean retrieval effectiveness for different query difficulties, measured on the percentile of SDM on CW09B.

number of quires than it hurts. The help-hurt ratio is between 1.62*x* and 4.38*x* on ClueWeb09B and between 1.69*x* and 3.5*x* on ClueWeb12B. This means that in the worst case 1.62*x* and 1.69*x* more queries were helped by RESS compared to the baselines.

4.1.3.2. Query difficulty. We also analyze the impact of RESS based on the difficulty of the queries. As suggested in Ensan and Bagheri (2017), we classify queries into four groups based on the performance of SDM (SDM MAP) where the queries in the bottom 0–25% are considered to be the most difficult and the queries in the 75–100% range are considered to be easier queries. Fig. 5 shows the performance of each baseline compared to RESS for different query difficulties on the ClueWeb09B corpus. As seen in the figure, the major strengthen of RESS is on the most difficult queries (0–25%) where the difference between the MAP of RESS compared to the other baselines is consistently statistically significant. In the other three difficulty ranges, the performance of RESS is either similar or weaker than the baselines but the differences are not statistically significant. Fig. 6 reports performance on ClueWeb12B. Here, RESS performs better than the baselines for the first three difficulty ranges while in the softest queries in the 75–100% range it shows weaker performance compared to SELM but the difference is not statistically significant. Our observations show that the strength of RESS is on improving retrieval performance for queries that are more difficult for SDM to retrieve.

4.2. Retrieval via entity selection

The second experiment focuses on comparing our proposed entity selection approach with a state-of-the-art entity selection technique within the context of ad hoc retrieval.

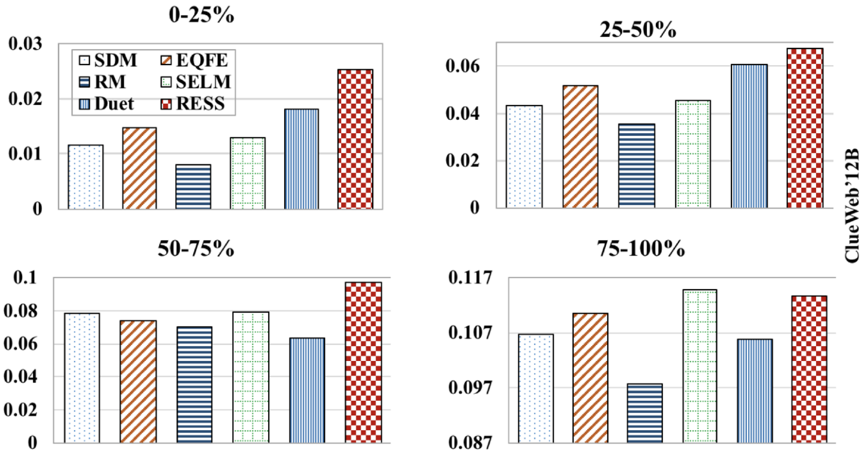


Fig. 6. Mean retrieval effectiveness for different query difficulties, measured on to the percentile of SDM on CW12B.

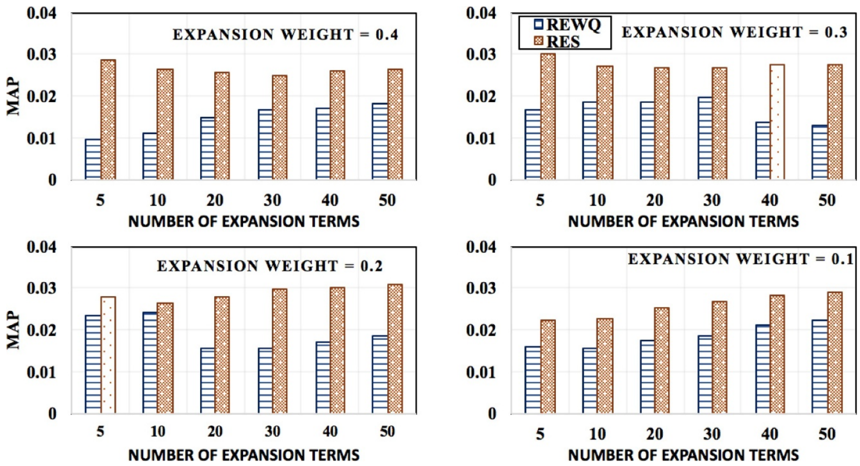


Fig. 7. The comparative analysis of REWQ and RES methods on REWQ ClueWeb12B Dataset. Darker shades on RES bars show statistical significance at 0.05 for paired *t*-test.

4.2.1. *Baselines*

We use REWQ (Schuhmacher et al., 2015), which has been shown to have strong performance on entity selection, as our baseline. REWQ selects a list of candidate entities from high ranking documents relevant to the input query. It defines *mention, query-mention, query-entity*, and *entity-entity* features and uses learning to rank methods over the candidate entities and the dataset documents for finding the most appropriate entities for a query. We compare the document retrieval performance when entities are found by REWQ and when they are found by our method.

4.2.2. *Experimental setup*

In REWQ (Schuhmacher et al., 2015), TREC Robust'04 and ClueWeb12B are used as datasets and a ranked list of 50 related entities are provided for 25 queries from TREC Topics 301-450, and 601-700 in the Robust'04 dataset, and 22 queries from TREC Web2013/2014 topics in the ClueWeb12B dataset. We used the same queries and dataset in this experiment. In order to weight expansion entities in REWQ, we used the normalized scores provided for the selected entity lists in http://mschuhma.github.io/rewq/ . We used RES *without* interpolation in this experiment, because the retrieval method is the same for both algorithms and they differ only in the entity selection method.

4.2.3. *Results*

The results of ad hoc retrieval based on the selected entities of the two method are compared on both ClueWeb12B and Robust'04 and reported in Figs. 7 and 8. In both figures, we report the performance of RES and REWQ for different number of expanded entities

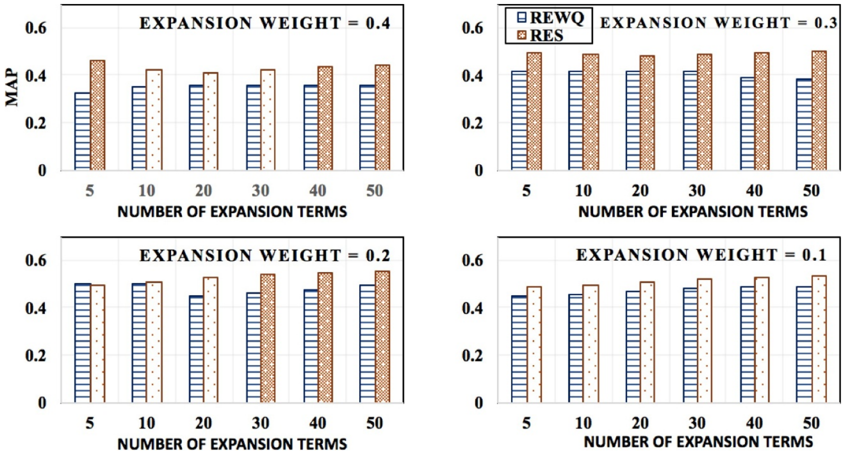


Fig. 8. The comparative analysis of REWQ and RES methods on the REWQ Robust'04 Dataset. Darker shades on RES bars show statistical significance at 0.05 for paired *t*-test.

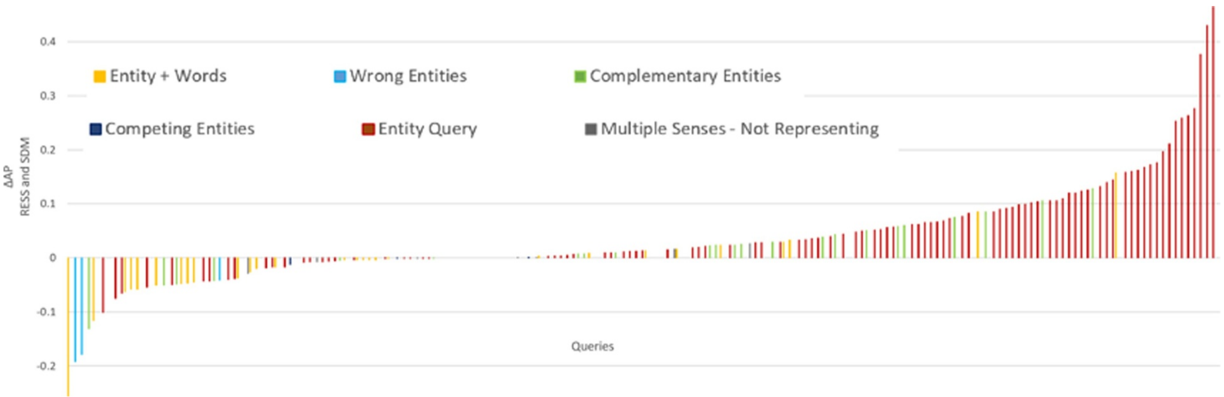


Fig. 9. The comparative analysis of RESS and SDM overClueWeb09B Dataset through different categories of queries.

from 5 to 50 as well as different expansion weights in {0.1, 0.2, 0.3, 0.4}. The figures show that RES provides more effective retrieval compared to REWQ for all expansion weights and different number of expansions where in 32 of the variants the observed improvement is statistically significant.

The best performance by REWQ on the ClueWeb12B was observed when an expansion weight of 0.2 was applied and 10 entities were used in expansion. Within the same setting RES produced a result of 0.0263, which was statistically significant over REWQ. The most effective variant of RES was observed at an expansion weight of 0.2 and the number of expansions of 50. This produced a MAP of 0.031 by RES while REWQ reported at statistically significant lower value of 0.0186.

Within the Robust'04 dataset, the best retrieval performance for REWQ was obtained when 5 entities were used for expansion with a weight of 0.2 resulting in a MAP of 0.4993 whereas the same setting provided an improved performance of 0.5101 by RES, which was not statistically significant. In contrast, the best performing variation for RES is at the expansion weight of 0.2 with 50 additional entities resulting in a statistically significant better MAP of 0.5516 compared to 0.4955 reported by REWQ. Summarily, we find that the entities selected by RES are more effective in improving the performance of the ad hoc document retrieval task.

4.3. Discussion

In this section, we provide more insight into the performance of RESS and the conditions under which it performs differently. For this purpose, we analyzed the queries of the experiment datasets, their linked entities, and how these entities cover the query text. Tables A1 and A2 in Appendix list all queries, the entities extracted by the entity linking system, and a label that we assign to them according to the extracted entities. These labels are as following:

- *Entity Query*: Where the query is about a Wikipedia Entry and this Entry is correctly found by the linking system. Examples include '*Yahoo*', and '*Atari*'.
- *Entity + Words*: Where the query is about one or more entities but also has some extra texts that are not linked to any entity. For examples '*source of the nile*', which is linked to entity '*Nile* and '*dog clean up bags*', which is linked to the entity '*Dog*'.
- *Complementary Entities* and *Competing Entities*: Queries that are linked to complementary entities, entities that complement each other for describing the user intent; and queries that are linked to competing entities, entities that compete for describing the meaning of a query. '*dutchess county tourism*', which is linked to '`Dutchess County, New York`' and '`Tourism`' Entities is an example of the former one, while '*Website design hosting*', which is linked to two entities, namely'`Web design`' and '`Web hosting service`', is a sample of the later one.
- Finally we have a set of *Multiple Senses Entities* queries, those that could mean different senses (such as '*Kiwi*' that could refer to a bird or a fruit) and *Wrong Entities*, those queries that are linked to wrong entities by the linking system.

Figs. 9 and 10 show Δ*AP*, which is the difference between the average precision achieved by RESS and the one achieved by SDM, for the queries sorted by their Δ*AP* values. Also, different categories of queries, according to our classification, are depicted by different colors in the figures.

As it can be seen in these figures, RESS is working much better than SDM in *Entity Queries* in both datasets. More specifically, RESS outperform SDM in 73 out of 96 Entity queries in ClueWeb09B dataset and 13 out of 14 queries in ClueWeb12B dataset. It shows that

Table 2
Example queries, their entity representation as well as the entities selected by our proposed approach.

Query	From TAGME	From KG embeddings	From pseudo-relevance	Our method
‘madam cj walker’	Madam C. J. Walker	Negro Academy Plantation tradition, A’Lelia Bundles, Henry Box Brown,	Americans, Woman, Social status, Denver, African Americans	African Americans, Business, Irvington, New York Villa Lewaro Hair Care, Indianapolis, Philanthropy
‘dutchess county ‘tourism’	Dutchess County Tourism	Ulster County Albany County Schenectady County Columbia County Economy Industry Infrastructure Agriculture	Television station City block WRNQ Sales Hudson Valley	Poughkeepsie (town) New York Farm Hudson Valley WKIP (AM) Hudson River
‘Website design’ hosting’	Web design Web hosting service	Desktop publishing Microsoft Office Intranet Content management Border Gateway Protocol Whitelisting Application layer	Graphical user interface Computer Server (computing) HTML	Online advertising Cloud computing Internet service provider

RESS can efficiently find and rank documents that have query entities. Here, RESS performance can be justified as follows: Contrary to keyword-based methods such as SDM, RESS conduct search over documents that are represented as sets of entities. For example, for the query *atari*, RESS looks for all documents that have one or more links to the Wikipedi Entry #2234, ‘*Atari*’. RESS also expand the query with the related entities, but as our experiments suggest, the best performance is achieved when the expansion coefficient is much less than the original entity coefficients. In other words, for Entity queries, a document that has no occurrence of the query entity should link to a lot of strongly related entities in order to be ranked high in the list. Also, according to RESS method, a document that includes query entities and a set of strongly related entities get higher score than one that has only the query entities surrounded by non-related context.

Another important observation from the RESS performance over the *Entity Queries* is that the expansion method, when the query is correctly linked to entities and there is no text without links, does not lead to topic drift. This fact can be justified by the entity selection method that selects expanding entities based on two sources: the knowledge graph (KG) embeddings and the pseudo-relevance feedback documents. For instance, for the query ‘*madam cj walker*’ TAGME retrieves an accurate entity link to the DBpedia entity dedicated to Madam Walker. Based on this identified entity, it is possible to retrieve semantically similar entities from the knowledge graph based on the similarity of their embeddings. As evident in Table 2, from the list of entities selected based on this approach, while the entities can be considered relevant, they clearly introduce *topic drift*, which is undesirable for document retrieval. For instance, Henry Box Brown is a 19th century slave who managed to escape for his freedom and as such relates to the suffering of African-Americans but does not relate to our query. Additionally, when looking at the most frequent entities in the pseudo-relevant documents, one can see that related yet non-specialized entities are retrieved such as social status, woman, and African Americans. However, while our method relies on pseudo-relevant documents and KG embeddings, it is able to address these two issues in that it does not lead to *topic drift* and it does not select *generic entities*. The entities selected by our method include Madam Walker’s hometown (New York), her business (hair care), location of business operations (Indianapolis) and the reason she is well known for, which is philanthropy.

Figs. 9 and 10 also show that RESS outperforms SDM in queries that have Complementary Entities. More concretely, RESS achieves higher average precision in 17 queries out of the total of 23 complementary queries in ClueWeb09B dataset and 11 queries out of the total of 16 queries in ClueWeb12B. RESS enjoys an entity selection and ranking method that finds expanding entities based on their relatedness to *all* entities in the query. For instance, for the TREC query #127 ‘*dutchess county tourism*’, that is linked to two complementary entities, two entity subsets emerge, each of which is related to one of the entities of the query (See Table 2). Also, the top entities of the pseudo-relevant documents are also quite generic to be informative or helpful. However, RESS performs entity selection by imposing that the entities need to be closely related to all of the entities of the query through the graph cliques. Therefore, the selected entities are closely related to both of the entities of the query. The entities include the location of Dutchess County (`Hudson Valley` in `Poughkeepsie, NY`), the main attractions of this area (`Farms` and the `Hudson river`) as well as the local radio station for this area (`WKIP AM`).

Our analysis show that RESS does not outperform SDM in queries that has entities and words (Only 9 queries out the total of 28 have been improved by RESS in ClueWeb09B). As an example, '*dogs clean up bags*', which is linked to the entity '*Dog*', cannot be appropriately handled by RESS. This challenge is also recognized by other knowledge-graph based search systems (Ensan & Bagheri, 2017). One solution could be to use a query performance prediction method tailored for entity-based systems that predicts which queries cannot be efficiently answered by entity-based systems, e.g. those in which the unlinked parts of the query texts are important in interpreting the query intent, and dynamically adjust the interpolation coefficient for different queries according to these predictions. We leave more analysis and works on this subject for future work.

RESS is also vulnerable at processing queries with wrongly linked entities and also queries with competitive entities. In ClueWeb09B dataset, RESS perform worse than SDM for all three queries with wrongly detected entities. In addition, in ClueWeb12B RESS performs worse than SDM in 2 out 3 of such queries. This fact highlights the importance of the entity linking performance in the success of our ranking method. For competing entities queries, RESS is outperformed by SDM in 2 out of 3 queries in ClueWeb09B. (We did not find such queries in ClueWeb12 dataset). This fact can be explained by RESS entity expansion method. RESS requires all the entities representing the query to be present in the graph cliques, in such cases, the final selected entities for query expansion would be those that are able to capture the commonalities between the query entities. This can be seen as a drawback of our approach that primarily relates to the way entity linkers relate entities to queries. One possible solution for this issue is to design a weighting strategy (such as the attention model Xiong, Callan et al., 2017) for determining entity importance when two or more competing entities are present in the representation of the query.

In summary, the strengths of our proposed entity selection method stem from how the graph representation is constructed from knowledge graph entities and entities derived from pseudo-relevance feedback documents. It specially benefits from the formed graph clique structures to address the three challenges that were introduced in the introduction section of this paper and (i) act as a mechanism to address *topic drift*, (ii) identify a suitable number of relevant yet not too generic entities, and (iii) capture entity interactions within those queries that are composed of more than one complementary entities. We also point out that an area for future improvement would be to address queries with the text that is not linked to any entity and queries that consist of more than one entity that are semantically very close.

5. Concluding remarks

In this paper, we presented an entity selection method for ad hoc document retrieval. The model performs document ranking through query entity expansion, i.e., expanding entities in the query with related entities from pseudo relevant documents. Given the fact that an entity in a knowledge graph can be related to numerous other entities from different aspects, the main objective of the proposed method is to find and score a subset of relevant entities that can more effectively contribute to the document retrieval process. For a set of expansion candidate entities, the proposed method models' dependencies between query entities, between query and candidate entities, and between document entities and a union of entity and query entities, where dependencies between entities are obtained from the knowledge graph. Based on the graphical model, our method estimates the probability of the union of query entities and a candidate entity and document entities. In our experiments, we showed that the retrieval model based on the proposed approach outperforms state-of-the-art keyword-based and entity-based retrieval models. We also showed that the entities found by our method are more effective than a state-of-the-art entity selection baseline for improving retrieval performance.

We also demonstrate that the retrieval model is mostly effective for *entity queries* and for queries with *complementary entities*, while it cannot effectively answer queries that include unlinked text and also queries with *competing* entities. The introduced entity selection method tends to lean towards more generic entities that serve as the common denominator for the two or more query entities, which would by nature lack specificity. For future work, we would like to work on two important directions: first, predicting the entity-based retrieval performance for different types of queries for appropriately adjusting the keyword-semantic interpolation coefficient and second, investigating a weighting strategy for prioritize entities in queries with competing entities.

Appendix A

Table A1
ClueWeb09B dataset queries, their entities, and their classifications.

Query number	Query text	Label	Entities (Wikipedia Entries)
1	obama family tree	Entity + Words	17775180
2	french lick resort and casino	Complementary Entities	112521, 8511510
4	toilet	Entity Query	19167644
5	mitchell college	Entity Query	502360
6	kcs	Entity Query	345688
7	air travel information	Entity Query	51215, 36674345
8	appraisals	Entity Query	871336
10	cheap internet	Entity + Words	14539
11	gmat prep classes	Entity + Words	255232
13	map	Entity Query	19877
14	dinosaurs	Entity Query	8311
15	espn sports	Entity Query	77795
16	arizona game and fish	Complementary Entities	21883824, 7113815, 4699587
17	poker tournaments	Complementary Entities	23014, 141837
18	wedding budget calculator	Entity + Words	32893, 377116
19	the current	Entity Query	440603
20	defender	Multiple Senses - Not Representing	649702
21	volvo	Entity Query	32412
22	rick warren	Entity Query	735151
23	yahoo	Entity Query	188213
24	diversity	Multiple Senses - Not Representing	51885
25	euclid	Entity Query	9331
26	lower heart rate	Entity + Words	304942
27	starbucks	Entity Query	178771
28	inuyasha	Entity Query	113028
29	ps 2 games	Complementary Entities	3266317, 1336512
30	diabetes education	Entity + Words	40017873
31	atari	Entity Query	2234
32	website design hosting	Competing Entities	34035, 157465
33	elliptical trainer	Entity Query	1393614
34	cell phones	Entity Query	19644137
35	hoboken	Multiple Senses - Not Representing	125235
36	gps	Entity Query	11866
37	pampered chef	Entity Query	888155
38	dogs for adoption	Entity + Words	258700
39	disneyland hotel	Entity Query	6175201
41	orange county convention center	Entity Query	6961997
42	the music man	Entity Query	97723
43	the secret garden	Entity Query	410873
44	map of the united states	Complementary Entities	3434750, 19877
45	solar panels	Entity Query	3507365
46	alexian brothers hospital	Entity Query	5198401,
47	indexed annuity	Entity + Words	22046794
48	wilson antenna	Entity + Words	187317
49	flame designs	Wrong Entities	11145, 21732545
50	dog heat	Multiple Senses - Not Representing	4269567, 19593167
51	horse hooves	Entity Query	5433125
52	avp	Entity Query	2603563
53	discovery channel store	Entity + Words	77807
54	president of the united states	Entity Query	24113
55	iron	Entity Query	14734
56	uss yorktown charleston sc	Complementary Entities	216058, 2366794
57	ct jobs	Entity + Words	314993
58	penguins	Entity Query	23878
59	how to build a fence	Entity + Words	42273
60	bellevue	Entity Query	137979
61	worm	Multiple Senses - Not Representing	19180096
62	texas border patrol	Complementary Entities	29810, 567453
63	flushing	Entity Query	267693
64	moths	Entity Query	66633
65	korean language	Entity Query	16756
66	income tax return online	Competing Entities	50845, 514183
67	vldl levels	Entity + Words	502410
68	pvc	Entity Query	24458
69	sewing instructions	Entity + Words	92295
70	to be or not to be that is the question	Entity Query	729006
71	living in india	Entity + Words	14533
73	neil young	Entity Query	87985

(*continued on next page*)

Table A1 (*continued*)

Query number	Query text	Label	Entities (Wikipedia Entries)
74	kiwi	Multiple Senses - Not Representing	17362
75	tornadoes	Entity Query	37530
76	raised gardens	Entity + Words	42139
77	bobcat	Entity Query	171820
78	dieting	Entity Query	8460
79	voyager	Entity Query	47795
80	keyboard reviews	Entity + Words	18842281
81	afghanistan	Entity Query	737
82	joints	Entity Query	210242
83	memory	Entity Query	31217535
84	continental plates	Entity + Words	24944
85	milwaukee journal sentinel	Entity Query	1272811
86	bart sf	Entity Query	60340
88	forearm pain	Entity + Words	237647
89	ocd	Entity Query	20082214
90	mgb	Entity Query	1426566
91	er tv show	Entity Query	177153
93	raffles	Entity Query	768522
94	titan	Entity Query	47402
96	rice	Entity Query	36979
97	south africa	Entity Query	17416221
99	satellite	Entity Query	27683
101	ritz carlton lake las vegas	Complementary Entities	9428452, 94988, 2237980
102	fickle creek farm	Wrong Entities	18842308, 59790
103	madam cj walker	Entity Query	472573
104	indiana child support	Complementary Entities	21883857, 7178087
105	sonoma county medical services	Complementary Entities	82117, 261925
106	universal animal cuts reviews	Wrong Entities	170326, 2056466, 150374
107	cass county missouri	Entity Query	94674,
108	ralph owen brewster	Entity Query	30873342
109	mayo clinic jacksonville fl	Complementary Entities	160843, 60613
111	lymphoma in dogs	Entity + Words	3813982
112	kenmore gas water heater	Complementary Entities	138004, 18993869, 521801
113	hp mini 2140	Multiple Senses - Not Representing	20972581
114	adobe indian houses	Complementary Entities	682, 21217, 13590
115	pacific northwest laboratory	Entity + Words	78147
116	california franchise tax board	Entity Query	13718746
117	dangers of asbestos	Complementary Entities	24462958, 21492663
118	poem in your pocket day	Multiple Senses - Not Representing	22926
120	tv on computer	Complementary Entities	29831, 7878457
122	culpeper national cemetery	Entity Query	4480425
123	von willebrand disease	Entity Query	311436
124	bowflex power pro	Entity + Words	11990673
125	butter and margarine	Complementary Entities	46183, 193276
126	us capitol map	Complementary Entities	31979, 19877
127	dutchess county tourism	Complementary Entities	50528, 29789
128	atypical squamous cells	Multiple Senses - Not Representing	377933, 483490
129	iowa food stamp program	Complementary Entities	26810748, 659087
130	fact on uranus	Entity + Words	44475
131	equal opportunity employer	Entity Query	4922510
132	mothers day songs	Entity + Words	46276
133	all men are created equal	Entity Query	331170
135	source of the nile	Entity + Words	21244
136	american military university	Entity + Words	3884115
138	jax chemical company	Entity + Words	58721
139	rocky mountain news	Entity Query	1897579
141	va dmv registration	Complementary Entities	32432, 4993736
143	arkadelphia health club	Entity + Words	106883
144	trombone for sale	Entity + Words	29837
145	vines for shade	Entity + Words	66607
146	sherwood regional library	Complementary Entities	2524043, 17727
147	tangible personal property tax	Complementary Entities	24695, 373814
148	martha stewart and imclone	Complementary Entities	190995, 70145
149	uplift at yellowstone national park	Complementary Entities	1415891, 34340
150	tn highway patrol	Complementary Entities	30395, 318666
151	403b	Entity Query	689685
152	angular cheilitis	Entity + Words	3392594
153	pocono	Entity Query	1180662
154	figs	Entity Query	57893

(*continued on next page*)

Table A1 (*continued*)

Query number	Query text	Label	Entities (Wikipedia Entries)
155	last supper painting	Entity Query	30667
156	university of phoenix	Entity Query	489589
157	the beatles rock band	Entity Query	29812,
158	septic system design	Entity + Words	217773
159	porterville	Entity Query	108303
160	grilling	Entity Query	52987
161	furniture for small spaces	Entity + Words	48597
162	dnr	Entity Query	166811
163	arkansas	Entity Query	1930
164	hobby stores	Entity Query	311886
165	blue throated hummingbird	Entity Query	2442673
166	computer programming	Entity Query	5311
167	barbados	Entity Query	3455
168	lipoma	Entity Query	288150
169	battles in the civil war	Entity + Words	863
170	scooters	Entity Query	23809410
171	ron howard	Entity Query	58928
172	becoming a paralegal	Entity Query	236584
173	hip fractures	Entity Query	1706838
174	rock art	Entity Query	928469
175	signs of a heartattack	Complementary Entities	562958, 20556798
176	weather strip	Entity Query	8208783
177	best long term care insurance	Entity + Words	1160191
178	pork tenderloin	Entity Query	7440150
179	black history	Entity Query	1142431
180	newyork hotels	Entity + Words	14276
181	old coins	Entity + Words	, 7558
182	quit smoking	Entity Query	289607, 12254052
183	kansas city mo	Entity Query	, 17454
184	civil right movement	Entity Query	49001
185	credit report	Entity Query	1476274
186	unc	Entity Query	77940
187	vanuatu	Entity Query	32443
188	internet phone service	Entity Query	75028
189	gs pay rate	Wrong Entities	2532789, 304942
190	brooks brothers clearance	Entity + Words	802150
191	churchill downs	Entity Query	955377
192	condos in florida	Complementary Entities	375303, 18933066
193	dog clean up bags	Entity + Words	4269567
194	designer dog breeds	Competing Entities	825162, 79676
195	pressure washers	Entity Query	2748878
196	sore throat	Entity Query	310094
197	idaho state flower	Complementary Entities	14607, 3328431
198	indiana state fairgrounds	Entity Query	1318490
199	fybromyalgia	Entity Query	318049
200	ontario california airport	Complementary Entities	108010, 37575, 22218

Table A2
ClueWeb12B dataset queris, their entities, and their classifications.

Query number	Query text	Label	Entities (Wikipedia Entries)
201	raspberry pi	Entity Query	21111, 21555729
202	uss carl vinson	Entity Query	31692117
203	reviews of les miserables	Complementary Entities	9288
204	rules of golf	Entity Query	271126, 368048
206	wind power	Entity Query	85533
207	bph treatment	Complementary Entities	13690575, 384701
208	doctor zhivago	Entity Query	11955, 809197
209	land surveyor	Entity Query	505878, 782824
210	golf gps	Wrong Entities	70896, 208092
211	what is madagascar known for	Entity Query	22093, 4380391
212	home theater systems	Entity Query	197352
213	carpal tunnel syndrome	Entity Query	16556402
214	capital gains tax rate	Complementary Entities	5004226
215	maryland department of natural resources	Entity Query	129368
216	nicolas cage movies	Complementary Entities	3014744
217	kids earth day activities	Entity + Words	183370
218	solar water fountains	Complementary Entities	51946
219	what was the name of elvis presley's home	Entity + Words	863, 4181, 27956
220	nba records	Complementary Entities	737, 11424
221	electoral college 2008 results	Entity + Words	20107078, 19344418, 682482
222	male menopause	Entity + Words	32851
223	usda food pyramid	Complementary Entities	88164, 13311819
224	making chicken soup from scratch	Entity + Words	182188, 417370
225	black and gold	Entity Query	439075
226	traverse city	Entity Query	32817449, 50482, 125715
227	i will survive lyrics	Entity + Words	9228, 4269567
228	hawaiian volcano observatories	Complementary Entities	5265384
229	beef stroganoff recipe	Entity + Words	13270, 32571, 58968
230	world's biggest dog	Entity + Words	46461
232	hurricane Irene flooding in manville nj	Complementary Entities	19196010, 600368
233	hair dye	Entity + Words	11065202, 33029735
234	dark chocolate health benefits	Complementary Entities	10683, 88486
235	ham radio	Entity Query	150550
236	symptoms of mad cow disease in humans	Complementary Entities	48726
237	lump in throat	Entity Query	97758
238	george bush sr bio	Wrong Entities	, 158548
239	frank lloyd wright biography	Complementary Entities	277289
240	presidential middle names	Wrong Entities	6672660, 808818
241	what is a wiki	Entity + Words	13595572
242	cannellini beans	Entity Query	56462
243	afghanistan flag	Complementary Entities	18964
244	old town scottsdale	Complementary Entities	60891
245	roosevelt island	Entity Query	2019834, 2019834
246	civil war battles in South Carolina	Complementary Entities	922583
247	rain man	Entity Query	, 49611
248	eggs shelf life	Complementary Entities	18940583, 106659
249	occupational therapist	Entity Query	23275402
250	ford edge problems	Entity + Words	1436561

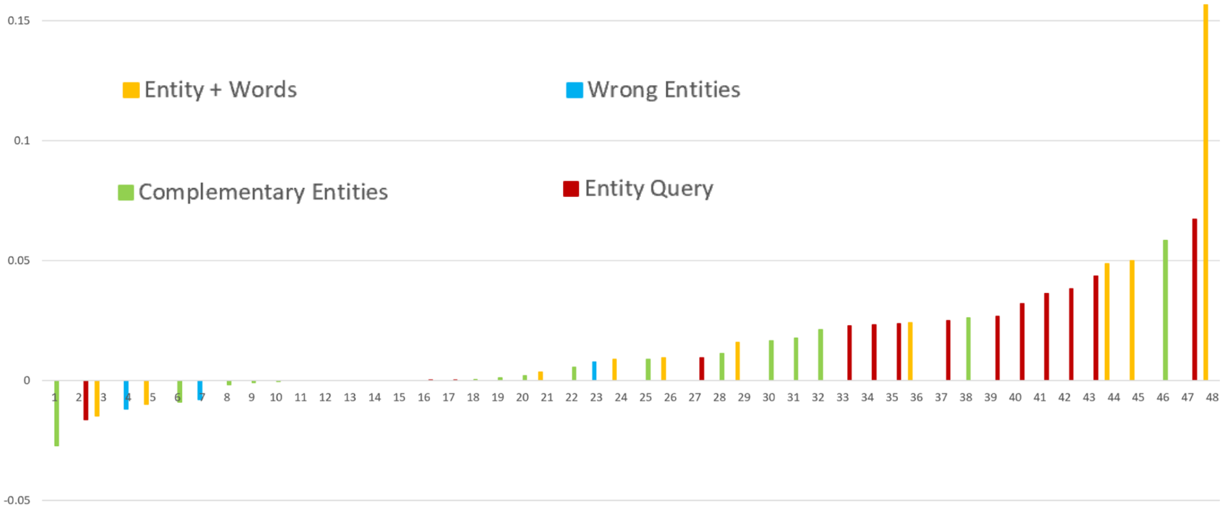


Fig. 10. The comparative analysis of RESS and SDM overClueWeb12B Dataset through different categories of queries.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ipm.2019.05.005.

References

Bagheri, E., Ensan, F., & Al-Obeidat, F. (2018). Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management, 54*(4), 657–673.

Balaneshinkordan, S., & Kotov, A. (2016). An empirical comparison of term association and knowledge graphs for query expansion. In N. Ferro,, F. Crestani,, M.-F. Moens,, J. Mothe,, F. Silvestri,, & G. M. Di Nunzio, (Eds.). *Advances in information retrieval* (pp. 761–767). Springer International Publishing.

Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management, 45*(1), 1–19.

Balog, K., & Neumayer, R. (2013). *A test collection for entity search in DBpedia. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrievalSIGIR '13*ACM737–740. https://doi.org/10.1145/2484028.2484165.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR), 44*(1), 1.

Chen, J., Xiong, C., & Callan, J. (2016). *An empirical study of learning to rank for entity search. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrievalSIGIR '16*ACM737–740.

Dalton, J., Dietz, L., & Allan, J. (2014). *Entity query feature expansion using knowledge base links. Proceedings of the 37th international ACM SIGIR conference on research; development in information retrievalSIGIR '14*ACM365–374.

Demartini, G., Iofciu, T., & De Vries, A. P. (2010). *Overview of the INEX 2009 entity ranking track. Proceedings of the focused retrieval and evaluation, and 8th international conference on initiative for the evaluation of XML retrievalINEX'09*Springer-Verlag254–264.

Dietz, L., Kotov, A., & Meij, E. (2017). *Utilizing knowledge graphs in text-centricinformation retrieval. Proceedings of the tenth ACM international conference on web search and data mining.* ACM815–816.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM TOIS, 29*(2), 8.

Ensan, F., & Bagheri, E. (2017). *Document retrieval model through semantic linking. Proceedings of the tenth ACM international conference on web search and data miningWSDM '17*ACM181–190.

Ensan, F., Bagheri, E., Zouaq, A., & Kouznetsov, A. (2017). *An empirical study of embedding features in learning to rank. Proceedings of the 2017 ACM on conference on information and knowledge managementCIKM '17*ACM2059–2062.

Feng, Y., Bagheri, E., Ensan, F., & Jovanovic, J. (2017). The state of the art in semantic relatedness: A framework for comparison. *The Knowledge Engineering Review,* 1–30.

Ferragina, P., & Scaiella, U. (2010). *Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities). Proceedings of the 19th ACM international conference on information and knowledge managementCIKM '10*ACM1625–1628.

Gabrilovich, E., Ringgaard, M., & Subramanya, A. (2013). *FACC1: Freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0)* http://lemurproject.org/clueweb09/FACC1/, Cited by 5

Garigliotti, D., Hasibi, F., & Balog, K. (2018). Identifying and exploiting target entity type information for ad hoc entity retrieval. *Information Retrieval Journal,* 1–39.

Hasibi, F., Balog, K., Garigliotti, D., & Zhang, S. (2017). *Nordlys: A toolkit for entity-oriented and semantic search. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval.* ACM1289–1292.

Hasibi, F., Nikolaev, F., Xiong, C., Balog, K., Bratsberg, S. E., Kotov, A., et al. (2017). *DBpedia-entity v2: A test collection for entity search. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrievalSIGIR '17*ACM1265–1268. https://doi.org/10.1145/3077136.3080751.

Jiang, Y., Bai, W., Zhang, X., & Hu, J. (2017). Wikipedia-based information content and semantic similarity computation. *Information Processing and Management, 53*(1), 248–265.

Kaptein, R., Serdyukov, P., De Vries, A., & Kamps, J. (2010). *Entity ranking using Wikipedia as a pivot. Proceedings of the 19th ACM international conference on information and knowledge managementCIKM '10*ACM69–78.

Keikha, A., Ensan, F., & Bagheri, E. (2017). Query expansion using pseudo relevance feedback on Wikipedia. *Journal of Intelligent Information Systems,* 1–24.

Kohlschütter, C., Fankhauser, P., & Nejdl, W. (2010). *Boilerplate detection using shallow text features. Proceedings of the third ACM international conference on web search and data miningWSDM '10*ACM441–450.

Krishnan, A., Deepak, P., Ranu, S., & Mehta, S. (2018). Leveraging semantic resources in diversified query expansion. *World Wide Web, 21*(4), 1041–1067.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). *Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the eighteenth international conference on machine learningICML '01*Morgan Kaufmann Publishers Inc.282–289.

Lavrenko, V., & Croft, W. B. (2001). *Relevance based language models. Proceedings of the 24th annual international ACM SIGIR conference on research and development in* *information retrievalSIGIR '01*ACM120–127.

Li, H., Xu, J., et al. (2014). Semantic matching in search. *Foundations and Trends® in Information Retrieval, 7*(5), 343–469.

Li, Y., Zheng, R., Tian, T., Hu, Z., Iyer, R., & Sycara, K. (2016). *Joint embedding of hierarchical categories and entities for concept categorization and dataless classification. COLING 2016*2678–2688.

Lin, X., & Lam, W. (2018). *Entity retrieval via type taxonomy aware smoothing. European conference on information retrieval*. Springer773–779.

Liu, X., & Fang, H. (2015). Latent entity space: A novel retrieval approach for entity-bearing queries. *Information Retrieval Journal, 18*(6), 473–503.

Liu, X., Zheng, W., & Fang, H. (2013). An exploration of ranking models and feedback method for related entity finding. *Information Processing & Management, 49*(5), 995–1007.

Metzler, D., & Croft, W. B. (2005). *A Markov random field model for term dependencies. Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrievalSIGIR '05*ACM472–479.

Metzler, D., & Croft, W. B. (2007). *Latent concept expansion using markov random fields. Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrievalSIGIR '07*ACM311–318.

Paik, J. H. (2013). *A novel TF-IDF weighting scheme for effective ranking. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrievalSIGIR '13*ACM343–352.

Pound, J., Mika, P., & Zaragoza, H. (2010). *Ad-hoc object retrieval in the web of data. Proceedings of the 19th international conference on world wide webWWW '10*ACM771–780.

Raviv, H., Kurland, O., & Carmel, D. (2016). *Document retrieval using entity-based language models. Proceedings of the 39th international ACM SIGIR conference on research and development in information retrievalSIGIR '16*ACM65–74.

Robertson, S. E., & Walker, S. (1994). *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrievalSIGIR '94*Springer-Verlag New York, Inc.232–241.

Schuhmacher, M., Dietz, L., & Paolo Ponzetto, S. (2015). *Ranking entities for web queries through text and knowledge. Proceedings of the 24th ACM international on conference on information and knowledge managementCIKM '15*ACM1461–1470.

Schuhmacher, M., & Ponzetto, S. P. (2014). *Knowledge-based graph document modeling. Proceedings of the 7th ACM international conference on web search and data miningWSDM '14*ACM543–552.

Serdyukov, P., Rode, H., & Hiemstra, D. (2008). *Modeling multi-step relevance propagation for expert finding. CIKM'08*. ACM1133–1142.

Shen, W., Wang, J., & Han, J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering, 27*(2), 443–460.

Sherman, G., & Efron, M. (2017). *Document expansion using external collections. Proceedings of the 40th international ACM SIGIR conference on research and development in information retrievalSIGIR '17*ACM1045–1048.

Song, F., & Croft, W. B. (1999). *A general language model for information retrieval. Proceedings of the eighth international conference on information and knowledge managementCIKM '99*ACM316–321.

Strube, M., & Ponzetto, S. P. (2006). *Wikirelate! computing semantic relatedness using Wikipedia. Proceedings of the 21st national conference on artificial intelligence - volume 2AAAI'06*AAAI Press1419–1424.

Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning, 4*(4), 267–373.

Wallach, H. M. (2004). Conditional random fields: An introduction. *Technical Reports (CIS),* 22.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). *Knowledge graph and text jointly embedding. EMNLP14. EMNLP* 1591–1601.

Witten, I. H., & Milne, D. N. (2008). *An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proceedings of the aaai 2008 workshop on Wikipedia and artificial intelligence (wikiai 2008)*. AAAI Press.

Xiong, C., & Callan, J. (2015). *Query expansion with freebase. Proceedings of the 2015 international conference on the theory of information retrievalICTIR '15*ACM111–120.

Xiong, C., Callan, J., & Liu, T.-Y. (2016). *Bag-of-entities representation for ranking. Proceedings of the 2016 ACM international conference on the theory of information retrievalICTIR '16*ACM181–184.

Xiong, C., Callan, J., & Liu, T.-Y. (2017). *Word-entity duet representations for document ranking. SigirSIGIR '17*ACM763–772.

Xiong, C., Power, R., & Callan, J. (2017). *Explicit semantic ranking for academic search via knowledge graph embedding. Www'17*. International World Wide Web Conferences Steering Committee1271–1279.

Xu, Y., Jones, G. J., & Wang, B. (2009). *Query dependent pseudo-relevance feedback based on Wikipedia. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrievalSIGIR '09*ACM59–66.

Yahya, M., Barbosa, D., Berberich, K., Wang, Q., & Weikum, G. (2016). *Relationship queries on extended knowledge graphs. Proceedings of the ninth ACM international conference on web search and data miningWSDM '16*ACM605–614.

Zhiltsov, N., Kotov, A., & Nikolaev, F. (2015). *Fielded sequential dependence model for ad-hoc entity retrieval in the web of data. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrievalSIGIR '15*ACM253–262.