What People Ask About: Mining Entity Search Intents in Community Question Answering Sites

Abstract

In this work we propose a novel search intent-based representation for named entities that is based on the questions people ask about them in a community question answering (CQA) site. Entity search intents (ESI) are represented by entity related questions, answered by community members, which depict a meaningful search intent about the entity.

We show that ESI representation can effectively be used for measuring entity relatedness, and evaluate its performance by measuring its correlation with human relatedness judgment over a dataset of entity pairs. The high correlation with human judgments confirms its applicability for this task. We additionally compare the ESI-based relatedness measurement with two strong baselines based on Wikipedia data. We show that combining ESI with the baselines significantly improve relatedness measurement accuracy, an indication that the two approaches complement each other while using two orthogonal domains.

1 Introduction

Recent studies show that about 71% of Web search queries contain named entities (*e.g.* people, locations, organizations, products) (Guo et al., 2009; Pound et al., 2010; Yin and Shah, 2010). Addressing such queries effectively, *i.e.* providing rich information that can satisfy diversified entity related search intents, is a major challenge faced by search engines today.

General Web search engines exploit structured data sources such as Wikipedia[1] and Freebase[2] to enrich the search results for an entity query by presenting related information including a short description of the entity, known attributes (*e.g.* a person's age, a company's annual income), images and videos, and much more. Additionally, related queries to the entity are also suggested for query reformulation on the search result page. Such queries provide an interesting perspective about the target entity – the crowd perspective. When properly selected, they summarize what people usually ask about the entity, or in other words, what the common search intents with respect to this entity are.

The concept of *search intent* have been used in many different contexts in the past (Broder, 2002; Yin and Shah, 2010; Pound et al., 2010; Cheung and Li, 2012; Li et al., 2013). In this work, we focus on information need, *i.e.*, the user is looking for specific information *related* to the entity. We refer to the specific information need, associated with the entity, as Entity Search Intent (ESI).

Several previous works generalized the notion of entity related search intents. They attempt to extract generic search intents that are related to a class of entities, such as musicians, actors, etc. (Yin and Shah, 2010; Jain and Pennacchiotti, 2010; Xue and Yin, 2011; Cheung and Li, 2012; Li et al., 2013). Once extracted, these generic search intents can assist in organizing typical entity related search tasks. For example, when searching for "*Bob Dylan*", typical information regarding musicians, including albums, concerts and lyrics, can be presented to users to help them narrow down their search. Generic search intents are typically identified by extracting concepts or patterns that frequently co-occur within queries that are related to the target entity class.

While many search intents are shared between entities of the same class, there are also important search intents that are very specific to each entity. For example, '*accomplished achievements*'

[1] wikipedia.org
[2] freebase.com

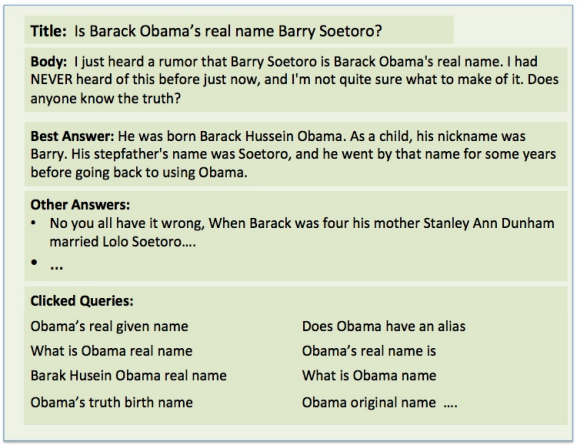

Figure 1: An example of a QA object from Yahoo Answers representing the search intent about the real name of '*Barack Obama*'.

is a common search intent for American presidents in general. Yet, '*birth place*' is much more related to "*Barack Obama*" than to other US presidents, due to the controversy surrounding it. As another example, when analyzing class-based intents, "*Britney Spears*" tends to be similar to other singers such as "*Celine Dion*" and "*Bruce Springsteen*". However, when looking at her idiosyncratic search intents, she also has a lot in common with other overly gossiped celebrities such as "*Paris Hilton*" and "*Serena Williams*" (Jain and Pennacchiotti, 2010).

In this paper we propose to identify search intents that are unique to each entity in order to further enrich its representation. We analyze a novel resource for mining search intents that are specific for an entity: Community-based Question Answering (CQA) sites such as Yahoo Answers, Baidu Zhidao and StackOverflow. A community question in these sites, together with its metadata, can often be viewed as representing a meaningful search intent around the entity it is focused on. Figure 1 presents a question from Yahoo Answers, representing a specific information need that relates to "*Barack Obama*".

We explore the hypothesis that an entity can be represented by extracting its related questions from the CQA archive, considering each question as a unique entity search intent. We argue that representing an entity by its specific related questions, and not by its category's related queries as was done before (Yin and Shah, 2010; Xue and Yin, 2011; Cheung and Li, 2012; Li et al., 2013), captures much more specific search intents and enables detecting interesting relations between entities, even from different classes.

Entity related queries that are suggested by commercial search engine typically represent frequent information needs that are associated with an entity. In contrast, CQA questions that relate to an entity cover diversified related topics such as opinions, relationships with other entities, events, etc. One reason for this difference is that common information needs for an entity can be discovered using commercial search engines, and therefore many times questions that are asked in CQA sites regarding the target entity address information needs that are not easily retrieved via search engines. In addition, the search intent behind a CQA question can be identified with ease, since the question object contains a detailed title and body, some human answers to the question, and a set of clicked queries (*i.e.*, Web queries that landed on the CQA page). This is in contrast to Web queries who are usually short and ambiguous (Liu et al., 2012).

To assess the value of our novel representation we conducted two experiments. First, we define *intent-based similarity* measures between pairs of questions that are associated with some entity or with two different entities. The various similarity functions are evaluated on a test set consisting of entities and their related community questions. In this dataset the related questions per entity were manually clustered into different search intents.

In a second experiment, we describe how semantic relatedness between two entities can be estimated based on their intent-based representations. Specifically, we utilize the proposed intent-based similarity measures to evaluate entity relatedness. In this evaluation, intent-based similarity achieves very high correlation with the average relatedness score given by human annotators. Moreover, integrating CQA-based relatedness score with Wikipedia relatedness score significantly outperform both approaches, pointing at a complementary information in these resources.

2 Previous Work

2.1 Entity Search Intent

Our work is focused on identifying fine-grained entity search intents that relate to a specific entity. Previous work on such search intent identification were mostly focused on generic search intents for a class of entities (musicians, actors, etc.).

Yin and Shah (Yin and Shah, 2010) proposed an approach for building a hierarchical taxonomy of generic search intents for a specific entity class by gathering and analyzing queries that are related to entities in the class. Frequent words and phrases, co-occuring with the entities in the user queries, represent the generic intents. The intents were organized into a hierarchical taxonomy, where the relationships between them were inferred based on user clicks. Xue et al. (Xue and Yin, 2011) further organized the co-occurring terms into topics using a named entity topic model. These topics help to recognize major search intents for entities in the class.

Cehung et al. (Cheung and Li, 2012) took another approach for search intent representation by extracting patterns consisting of a sequence of semantic concepts or lexical items in the entity queries. For example, a pattern that describes "*Harry Potter showtime Boston*" would be represented by '*[movie-title] showtime [city]*'. Similar queries annotated with the same pattern might include "*Madagascar showtime Sydney*". Entity related queries are clustered and summarized into an intent pattern per cluster. Li et al. (Li et al., 2013) extended this pattern-based approach by identifying synonymous intent templates.

The entity search intent identification methods described above are all based on analyzing a large set of entity queries related to the same entity class. While there are many shared intents related to a class of entities, there are also many important search intents that are specific to a particular entity. Our work, in contrast, analyzes search intents of a specific entity, which may differentiate it from other entities in its class. In addition, previous work is focused on identifying a pattern, or a term, that represents a unique search intent. We view a CQA object, i.e., the question, its associated answers, and other associated metadata, as a unique entity related ESI.

The work which is most closely related to ours is of Jain and Pennacchiotti (Jain and Pennacchiotti, 2010), who explored features of named entities using a search engine query log. Entities in this work are represented by analyzing the contexts in which they appear in related user queries, where each context is weighted based on its Corrected Pointwise Mutual Information (CPMI) with the entity. This representation was used for clustering the entities into entity classes.

There are several major differences between our representation and the context-based representation which is based on query log analysis. First, Web queries are quite different from CQA question objects which we use for search intent representation. Web queries are usually short and ambiguous, while a CQA question, in general, is rich and comprehensive enough to clearly represent a precise search intent (for example, see Figure 1). Second, for measuring entity relatedness, context-based similarity is based on the amount of shared concepts that are jointly related to the two entities. In contrast, we do not expect to find shared questions between entities, since any question object represents a specific search intent that is only related to the target entity. Instead, we estimate the semantic relatedness between entities by measuring the **similarity** between their ESI-based representations.

2.2 Entity Relatedness

Measuring the semantic relatedness between entities is of great importance for many search applications including entity retrieval (Balog et al., 2010), query suggestion (Boldi et al., 2009), related entity recommendation (Blanco et al., 2013), and many more. There are several approaches for estimating the semantic relatedness between entities. Many works follow the explicit semantic analysis (ESA) paradigm (Gabrilovich and Markovitch, 2009) in which texts are projected on a KB concept space (e.g. Wikipedia) and the semantic similarity is estimated based on the cosine similarity between the concept-based representations. Liu and Birnbaum (Liu and Birnbaum, 2007) measured the semantic similarity between named-entities using the Open Directory project (ODP). Entities are represented by their relevant ODP pages and the entity profile is constructed according to the directory assignments of the pages which capture various entity's features.

Several other works exploit the graph structure of Wikipedia. Witten and Milne (Witten and Milne, 2008) measured the relatedness between general terms using the links found within their corresponding Wikipedia articles rather than their textual content. Liu and Chen (Liu and Chen, 2010) applied the same approach in which the relatedness between entities is measured through Wikipedia pages related to the named entities. The authors selected a set of hyperlinks from the en-

tity's related articles and estimated their relatedness by calculating the similarity between the two sets of hyperlinks.

A different approach for estimating semantic relatedness between entities is based on analyzing user information needs that are associated with the two (Blanco et al., 2013). Entities are considered related when many users search for information about them simultaneously, *i.e.* they search for information that is related to the two entities or search for one entity and then for the second one in the same search session. This co-query measurement approach essentially learns the relatedness between two entities by utilizing user actions.

As we will show in this paper, our search-intent based representation can successfully be used to identify interesting, sometimes surprising, relationships between entities, even from different classes.

2.3 Yahoo Answers

In this work we focus on the Yahoo Answers site, since it is the largest CQA site to date with many references to different entity types. In Yahoo Answers, a user posts a question that consists of a short summary (the title) and an optional detailed description (the body). Each question can be answered by any other user. The asker may then choose the best answer, but if she does not, the task of selecting a best answer is delegated to the community for an indefinite time. Once a best answer is chosen, the question is said to be "resolved".

We accumulate the information regarding a question in a data structure called a Question-Answers (QA) object. The QA object contains the text of the question being asked including the title and the body of the question. It also includes all answers provided for this question, and if a best answer was chosen for the question, it is appropriately marked. Finally, in a QA object we also maintain a set of Web search queries that resulted in a click on the corresponding Web page of this question, which we term here *clicked queries*. Figure 1 presents some parts of a QA object representing the search intent about president *Barack Obama's* real name.

The dataset used in this study contains 50 million English QA objects of Yahoo Answers sampled between the years 2006 and 2013.

3 Search Intent Representation

We now describe our entity representation model that is based on the entity's related questions extracted from a CQA cite. We define *search intent* $SI(e)$ as a specific information need that is associated with an entity e. Our goal is to mine such intents from a given CQA archive. Specifically, we utilize Question-Answers (QA) objects (Section 2.3) that are associated with e for this task. Examining such objects reveals that, unlike standard Web queries, the question title and body, as well as the answers and the clicked queries, provide an enriched representation of the underlying search intent. Figure 1 exemplifies the comprehensive presentation captured in the QA object focused on the alleged "real" name of president '*Barack Obama*'.

To create our proposed representation for a target entity e, we retrieve a set of QA objects from Yahoo Answers that are focused on the given entity. Specifically, we search for the entity name, or the entity identifier assigned by an entity linking tool, in the title field and in the clicked queries field of the QA objects. The assumption behind this retrieval strategy is that if the entity is mentioned in the question title or in a related query, then it is the focus of the QA object with a high probability. The retrieved set is denoted

$$ESI(e) \stackrel{\text{def}}{=} \{QA_i | i = 1...k\},$$

where ESI stands for Entity Search Intents.

3.1 Similarity between QA Objects

To estimate the intent-based similarity between QA objects that are associated with an entity e or with two different entities, e and e', we utilize four QA object fields: title ($title$), combination of title and body ($title + body$), best answer ($banswer$) and clicked queries (cq). In the following, we describe several unsupervised similarity measures that may be applied to each of these fields.

Surface level similarity This similarity function, denoted $TfIdf_f(\cdot,\cdot)$, measures for each field f the cosine between the two tf-idf vectors representing the content of f of the two compared objects. Formally, given a a term t in a field f, $tf\text{-}idf_f(t) \stackrel{\text{def}}{=} \log(freq(t,f)+1) \cdot idf_f(t)$ where $freq(t,f)$ is the frequency of t in f; $df_f(t)$ counts the number of entity associated QA objects whose field f contains the term t; $idf_f(t) =$

$\log(k/df_f(t))$, where k is the number of QA objects retrieved for entity e. The terms we consider are unigrams and bi-grams. Using the entity associated QA objects for calculating the idf statistics, rather than the general corpus statistics, enables us to discriminate between terms that represent specific entity-related intents from terms that are common to many related intents.

Entity-based similarity The set of entities that appear in the context of a target entity e implicitly represents search intents. For example, an occurrence of the entity '*Democratic Party*' in a QA object that is related to *Barack Obama* indicates a question concerned with Obama's political views. Therefore, this similarity function, denoted $Entity_f(\cdot,\cdot)$ calculates the weighted Jaccard similarity between the vectors representing all entities appearing within field f in the compared objects. Entities are extracted using the state-of-the-art entity linking tool TagMe (Ferragina and Scaiella, 2010). The target entity on which each QA objects is focused is filtered out from their representations.

LDA-based similarity To go beyond surface level representation, we learn an LDA model (Blei et al., 2003) for each target entity e, using its set of retrieved QAs as pseudo documents (taking all such "documents" as a training-set). Specifically, we utilize all textual fields in each QA object by concatenating their texts to form a pseudo document. One outcome of the LDA model learning phase is an inferred entity specific topic distribution for each QA object. The LDA-based similarity function (denoted $LDA(\cdot,\cdot)$) measures the Jensen-Shannon similarity between pairs of such topic distributions.

It is important to emphasis that unlike standard uses of LDA, we learn topics for each entity separately and not for the general corpus. We note that entity related topics can also be considered as a soft clustering representation of the entity.

ESA-based similarity This measure, denoted by $ESA_f(\cdot,\cdot)$, also goes beyond surface level word comparison. For any field f, it measures the cosine similarity between two vectors that are composed of Wikipedia concepts representing the content of field f in each of the compared objects. Specifically, we use the Explicit Semantic Analysis (ESA) method (Gabrilovich and Markovitch, 2009) to create such a representation.

Clicked-query similarity The clicked queries field is a very valuable source of information when analyzing the search intent of a QA. It basically contains many different reformulations of the same search intent that are covered by the QA object, queried by many different users. To utilize this information we create a histogram of the clicked queries, counting the number of times each query has led to a click on the QA object. The clicked-query similarity between two objects, denoted by $CommonCQ(\cdot,\cdot)$, is computed by the Jensen-Shannon similarity between the two corresponding histograms.

3.2 Supervised Similarity Measure

The similarity measures presented above capture different aspects of intent-based similarity between two QA objects. Still, their combination may result in better estimation of the overall similarity. To this end, we view a similarity function as a binary classifier, with a pair of QA objects as input and binary output indicating similar/dissimilar pair. Given a training set of manually annotated QA pairs, described in details in Section 4.1, the classifier learns to predict this output.

We employ Gradient Boosted Decision Trees (GBDT) (Friedman, 2002) as our classifier. GBDT solves the classification problem by creating an ensemble of regression trees with a logistic regression loss function, using an iterative boosting procedure. One advantage of GBDT is that it is quite robust when combining features of different scales and types, as is the case of our various unsupervised similarity measures.

In our setting, at prediction time, GBDT assigns a similarity score ranging between 0 and 1 for an input pair of QA objects based the values of the unsupervised similarity features. Specifically, we learn two different classifiers using two different combinations of the similarity measures described above. The first classifier, denoted $Comb(\cdot,\cdot)$, utilizes all the similarity measures. The second classifier, denoted $Comb_{\text{-}cq}(\cdot,\cdot)$, is trained by considering all measures except those that are based on the clicked queries field (including LDA which concatenates all textual fields). Such a classifier is important in practice because many QA objects were rarely or even never been clicked, and therefore this field is often sparse or empty.

4 Evaluation

4.1 Experimental Settings

4.1.1 CQA collection

Our CQA data collection is composed of 50M resolved English QA objects, sampled from Yahoo Answers, from the years 2006 to 2013,. Entities in each QA object were extracted using the TagMe system[3], which links mentioned named entities to Wikipedia entries. We annotated all the textual fields of the QA objects using TagMe. To ensure reasonable entity quality, we filtered out all entities who's Tagme annotation confidence score is below 0.1. The textual data, as well as the annotated entities, were indexed using the Lucene[4] search engine. All texts were pre-processed using Lucene services, including tokenization, stopword removal and Porter stemming.

4.1.2 Dataset Creation

We explore three main assumptions: (1) Each entity has many specific related search intents and (2) Each QA object represents a unique entity related search intent (3) A few QA objects can often represent the same search intent. To verify these assumptions, and to evaluate the feasibility of intent-based representation for an entity, we constructed a gold-standard dataset[5] for a sample of 40 named entities. These entities are either people (12), organizations (12), locations (10), and others (6). Entities of each type were ranked according to their frequency occurrence in the Yahoo Answers collection. Then, from each frequency level, one entity of each type was chosen. All selected entities appear in at least 500 questions in our dataset. We aware that long-tailed entities which are not very popular in Yahoo Answers are not represented in the dataset. We leave the representation of long-tailed entities for future work.

For each entity we constructed the basic ESI representation by retrieving the top associated QA objects from our corpus. The full technical details are provided below. Then, two human annotators clustered the related QA objects of each target entity, grouping together objects that share the same search intent. In addition, each QA object was classified to whether it has a specific or common search intent. A common search intent was detected based on the observation whether the specific entity can be replaced by other entities in its class and whether the question relates to an entity attribute inherited from the general class.

4.1.3 Entity related questions retrieval

We used BM25 (Robertson and Zaragoza, 2009) as our retrieval function for obtaining the top QA objects that are focused on a specific entity. Specifically, given an entity e, we search for the entity name and the entity TagMe ID within the title and clicked-queries fields of the QA objects. The ranking scores of the fields were combined with equal weights to obtain a single retrieval score for each QA object. The retrieval score itself is not used on our model, since we assume that all top retrieved objects are focused on the entity.

There are two main considerations for choosing k, the number of top scored QA objects to retrieve. On one hand, the retrieved set should be large enough to cover a variety of search intents. On the other hand, the larger k is, the more irrelevant documents, *i.e.* objects not focused on the target entity, are includes in the retrieved list. We empirically set $k = 100$ to balance both considerations.

4.1.4 Search Intent Similarity Quality Estimation

The manually created clusters induce similarity labels for all 4,950 unordered pairs of objects, constructed from the 100 objects of each entity representation. A pair is labelled *similar* if its objects were assigned to the same intent cluster by the annotators. The label *non-similar* was given to pairs of objects who were assigned to different intent clusters.

For most entities, the number of non-similar pairs is an order of magnitude higher than the number of similar pairs (thousands vs. hundreds). Therefore, to test the similarity measure quality in a balanced way, we randomly sampled 250 similar pairs and 250 non-similar pairs for each entity. Named entities were divided into a training set and a test set, 14 and 26 in size respectively. Overall, 3500 similar pairs and 3500 non-similar pairs were sampled as a training set, and an equal amount was sampled in a non-repetitive way to create the test set. The training set was used to learn the parameters of the supervised similarity function (see Section 3.2).

To estimate the quality of each intent similarity measure we computed the area under the ROC

[3] tagme.di.unipi.it

[4] lucene.apache.org

[5] This dataset will be publicly available.

Entity	Common ESI	Specific ESI
Albert Einstein	What is the contribution of AE to science?	Why did AE regret the development of the atomic bomb?
Britney Spears	Why is BS so popular?	Why did BS shave her head?
George W. Bush	Is GWB good president?	GWB and 9/11.
Facebook	Why is FB called FB?	Ideas for funny FB statuses?
Mac Cosmetics	Is MC a good brand?	Question about back to MAC program.

Table 1: Common and specific ESIs

curve (AUC) for the purpose of quantitative comparison. We calculated both *MacroAUC* and *MicroAUC* where *Macro* is the average of AUC values computed separately for each entity, while *Micro* measures the AUC value while considering the pairs of all entities together. Statistically significant differences of similarity measures quality were determined using the Wilcoxon signed-rank test using a 95% confidence level.

4.2 Experimental Results

4.2.1 Search Intent Specificity Evaluation

Using the manually annotated dataset, we first analyzed the percentage of specific search intents for an entity. We found that the percentage of such specific questions is non-negligible and cannot be ignored; at least 11% of the questions for each entity are specific, while the average of specific questions per entity is 40.5%. Moreover, the percentage of specific questions vary among entities; The lowest value was 11%, while the highest 82%.

The average number of entity specific questions is higher for people (49.5%) and organizations (53.2%) while much lower for location (19%) entities. Table 1 exemplifies common and specific search intents by presenting the title of related questions for several entities. As can be seen, specific questions relate to very specific events or properties associated with the entity. In contrast, common questions relate to attributes inherited from the entity class.

4.2.2 Similarity Measure Evaluation

Table 2 presents the effectiveness of each of the intent similarity measures defined in Section 3 assessed using *MacroAUC* and *MicroAUC*. Several interesting findings are observed from the results. First, the clicked queries field is the most useful one for estimating intent similarity of two QA objects. For most similarity types, the highest effectiveness is achieved by utilizing the clicked query data. Second, using the title and body information of the QA object together is more effective than using the title alone. We did not experiment with the body alone, since it is missing in many QA objects. Third, LDA similarity which is based on latent topics presentation, performs very well in capturing intent-based similarity, and outperformed ESA measures which use Wikipedia concepts for textual representation. Finally, combining all measures in a supervised way achieves an improvement over each unsupervised measure. Specifically, the improvement for $Comb\text{-}_{cq}$ is statistically significant over all measures that do not exploit the clicked queries field, showing that different measures capture somewhat different notions of intent-based similarity. The $Comb$ function, on the other hand, outperforms all other measures, however, it is not significantly better than $TfIdf_{cq}$, implying that the clicked query data signal is hard to beat.

Measure	MacroAUC	MicroAUC
$TfIdf_{title}$	0.640	0.639
$TfIdf_{title+body}$	0.682	0.682
$TfIdf_{banswer}$	0.632	0.629
$TfIdf_{cq}$	0.765	0.755
$Entity_{title}$	0.535	0.522
$Entity_{title+body}$	0.550	0.560
$Entity_{banswer}$	0.569	0.574
$Entity_{cq}$	0.693	0.681
LDA	0.670	0.629
ESA_{title}	0.590	0.556
$ESA_{title+body}$	0.609	0.583
$ESA_{banswer}$	0.626	0.617
ESA_{cq}	0.604	0.560
$CommonCQ$	0.566	0.569
$Comb$	$\mathbf{0.781}^*$	$\mathbf{0.760}$
$Comb\text{-}_{cq}$	$\mathbf{0.726}^*$	$\mathbf{0.714}^*$

Table 2: The effectiveness of the intent-based similarity measures measure by AUC. '$*$' mark statistically significant improvement of the supervised measure over composing unsupervised measures.

5 Entity Relatedness

Measuring semantic relatedness between two entities is an important task with many practical applications such as relationship extraction, entity linking (Ferragina and Scaiella, 2010), entity ranking (Pound et al., 2010), and more. In this section we investigate whether our novel intent-based entity representation can be used for this task. We suggest to estimate the relatedness between two entities based on the similarity between their associated search intents. This measure relies on the hypothesis that *people ask similar questions about*

	Starbucks	KFC
Similar Community Questions	Starbucks **Interview** Questions?	**Interview** with KFC next Sunday, What should I expect and do?
	Calories, Starbucks cappuccino with whipped cream?	How many **calories** are in a KFC Toonie Tuesday meal?
	Which coffee is better **McDonald**'s or Starbucks?	**Mcdonald**s vs KFC - which is better?
	Your favorite cakes at Starbucks?	What's **your favorite** KFC combo or meal?
	Why am I **sick** after drinking Starbucks?	**Food poisoning** from KFC? PLEASE HELP?
	STARBUCKS: **Overpriced** or worth it?	Why is KFC **so expensive** now?
Shared Wikipedia Concepts	AmRest, Restaurant Brands, Fast food restaurant, PepsiCo, List of The Chaser's War, Burger King franchises Cranium Inc., Hopwood Park services, Membury services Burger King products	

Figure 2: Similar questions representing '*Starbucks*' and '*KFC*'. Top-shared Wikipedia concepts, are shown at the bottom.

strongly related entities. For example, the relationship between '*Starbucks*' and '*KFC*' may be revealed due to similar related questions concerning recipes, job opportunities, food quality, etc., as demonstrated in Figure 2.

Several state-of-the-art methods for estimating entity semantic relatedness are based on existing knowledge-bases. Specifically, Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2009), a high performance model for measuring semantic similarity between texts, represents the meaning of the texts in a high dimensional space of concepts derived from Wikipedia entries. Estimating the semantic similarity between the two texts in this space is performed by computing the cosine between their corresponding ESA vectors. ESA can be directly applied to entity relatedness estimation by measuring the similarity between the entities corresponding ESA vectors. Figure 2 shows, at the bottom, some Wikipedia top-shared concepts of '*Starbucks*' and '*KFC*'.

Our proposed measure is substantially different from ESA due to two main reasons. First, the representation spaces of the two measures are different and have rather complementary properties. On one hand, Wikipedia is an encyclopedic, semi-structured information resource, containing relatively high quality articles that span over a wide variety of topics. The information in CQA sites, on the other hand, is very informal and sometimes unorganized (*e.g.* redundant questions). In addition, users tend to ask questions in CQA sites whose information is not easily available on the internet and in knowledge resources. Finally, many Wikipedia concepts are static and are not updated very frequently, while CQA users tend to ask questions related to on-going events which makes some of the questions highly time dependent.

Moreover, ESA assumes that related entities share many common Wikipedia concepts. On the contrary, in our measure, two entities are usually represented by disjoint sets of QA objects. Therefore, relatedness has to be estimated by measuring the intent-based similarity between the representative sets. Specifically, similarity between entities is calculated based on the intent-based similarity between the QA objects associated with each entity.

Formally, our method for estimating entity relatedness works as follows. Given a pair of entities, (e_1, e_2), we first create an $ESI(e)$ representation for each of the entities in the pair. Then, we create a bipartite graph where nodes represent the related QA objects and edges connect all nodes belonging to $ESI(e_1)$ with all nodes belonging to $ESI(e_2)$. The edges are weighted by applying the intent-based similarity measurement between QA object pairs (see Section 3.1; for $idf_f(t)$ and LDA computation we used the statistics gathered from the union of QA objects of the two entities.)

Then, a match is found in the graph in a greedy manner by sorting the edges in decreasing order and repeatedly selecting the top-scored edge whose two nodes have not been covered already. The selection process stops when all nodes are covered[6]. Finally, the relatedness between e_1 and e_2 is computed by averaging the weights of the edges in the match.

5.1 Experimental Setup

A dataset which is extensively used for semantic relatedness evaluation between words is WordSim-353 (Gabrilovich and Markovitch, 2009). This dataset contains 353 word pairs together with an average relatedness score assigned by humans. However, this dataset does not fit our work on entity relatedness since it contains only a few of them. Therefore, we constructed a gold-standard dataset[7] of human relatedness estimations for pairs of entities. We manually created a list of 150 entity pairs, composed of entities that

[6]The structure of the graph guarantees the existence of a perfect match that can be found by e.g. the Hungarian method (Kuhn, 1955). However, for practical reasons, we settle in this work with the match found by the suggested greedy process.

[7]This dataset will be publicly available.

were selected from a closed group of types: people (38), organizations (38), locations (18), products (12), movies (8) and diseases (11). Two types of entity pairs were inserted into the list. First are pairs of entities of the same type (104) and second are pairs of entities of different types (56). To ensure variability, each entity in the list was assigned to at least two pairs but many of them were assigned to four pairs. In general, an entity e was randomly assigned with other entities in the list, while keeping the ratio of 2/3 pairs of the same entity type and 1/3 pairs of different entity types.

Relatedness estimation for the entity pairs was provided by 30 human annotators. Each annotator was presented with a random sample of the 50 pairs and was asked to assign a relatedness score that ranges between 1 to 5, were 1 denotes an unrelated pair and 5 a highly related pair. Annotators were instructed to follow the same instructions used for the WordSim-353 database construction, which are to give an intuitive score they have in mind for each pair, ignoring pairs containing unfamiliar entities. Overall, we collected 5-15 annotations (9.5 on average) for each pair and the final relatedness score was chosen to be the average annotator scores. The mean of the average human annotated score per pair is 2.92 and the average standard deviation is 0.699.

The intent-based similarity functions we experimented with are those described in Section 3.1. Following common practice (Gabrilovich and Markovitch, 2009; Witten and Milne, 2008), we measured for each tested model the Spearman correlation between the automatically assigned scores for the entity pairs in the dataset and the human average scores. As baselines, we implemented two state-of-the-art Wikipedia-based relatedness estimators: ESA (Gabrilovich and Markovitch, 2009) and WLM (Witten and Milne, 2008). In addition, we measured the Spearman correlation between our our entity pairs score combined with the ESA score and the human average scores. To compare two correlation values we approximated their sampling distribution using the bootstrap method (Efron and Tibshirani, 1994). Statistically significance differences between the distributions were determined using the paired t-test with 95% confidence level.

5.2 Results

Table 3 presents the performance of the baselines and of our entity relatedness model while using the various intent similarity functions. We report the Spearman correlation between the scores assigned by the measures and the human relatedness scores.

Several interesting findings are observed from the results. First, we can see that both baseline measures are highly effective for entity relatedness estimation, with Spearman correlation with human judgement above 0.74 in our dataset. Yet, for most similarity measures we used, a high correlation with human relatedness scores was also obtained. This finding supports our hypothesis that similar questions about two entities imply that they are semantically related. LDA performed badly in this task compared to other measures, due the difficulty in measuring similarity between the two unrelated topic distributions of the two entities. In addition, the supervised similarity measures, $Comb$ and $Comb_{\text{-}cq}$, did not outperform the unsupervised methods as happened in the clustering case, as they were both tuned with pairs of QAs that are related to the same entity while in our case they relate to different entities.

Second, the performance of the ESA_f based measures, is very high, for all fields, almost compared to that of state-of-the-art Wikipedia-based baselines, even though these measures compare entity related QA fields while the Wikipedia based measures compare the named entities. Interestingly, the ESA similarity between best answers outperformed the question-based similarity.

An appealing property of the CQA representation of entities would lie in its complementary information with respect to Wikipedia. To test this hypothesis we combined the Wikipedia-based ESA measure and each of our measures as a geometric average of the scores provided by the two models. The right column in Table 3 shows the results of these joint measure approach. This results show that the joint measure improves the correlation compared to each of the combined models when used alone. Furthermore, all measures outperform the baselines (excluding $CommonCQ$) systematically, and the improvement is statistically significant. We think that this performance gain is a good indication that the knowledge captured by the two representations, ESA and ESI, is to some extent complementary.

Similarity Method	Spearman Correlation	ESI * ESA Correlation
ESA	0.742	
WLM	0.740	
$TfIdf_{title+body}$	0.623	0.758^*
$TfIdf_{banswer}$	0.616	0.752^*
$TfIdf_{cq}$	0.584	0.761^*
$Entity_{title+body}$	0.642	0.761^*
$Entity_{banswer}$	0.622	0.755^*
$Entity_{cq}$	0.525	0.752^*
LDA	0.367	0.749^*
$ESA_{title+body}$	0.637	0.747^*
$ESA_{banswer}$	0.705	0.763^*
ESA_{cq}	0.723	0.757^*
$CommonCQ$	0.525	0.575
$Comb$	0.658	0.752^*
$Comb\text{-}_{cq}$	0.631	0.749^*

Table 3: Spearman correlation of the semantic relatedness measures. The right column shows the multiplication of each measure with ESA based relatedness score. '$*$' mark statistically significant higher correlation value in comparison to the ESA baseline.

Finally, it is important to note that the relative effectiveness of the similarity measures we use is different in this task compared to their relative effectiveness for clustering. The entity based measures, for example, which count on the number of shared entities mentioned in both QA objects, are more effective here than the $TfIdf$ based measures. This is not surprising taking into consideration that the highly effective measures based on knowledge resources (ESA, WLM) also count on shared concepts.

6 Summary

In this work we proposed a novel search intent-based representation for named entities, denoted ESI, which is based on the questions people ask about them in CQA sites. We studied several ESI-based similarity measurements between entity's related community questions, and evaluated their quality using manually annotated data.

We then described how ESI-based representation can be effectively used for measuring entities relatedness, and evaluated its performance by measuring its correlation with human relatedness judgement over a dataset of entity pairs. As a baseline, we used two strong KB-based relatedness measurements. While KB-based approaches outperform the ESI-based approach, its high correlation with human judgements confirmed its applicability for this task. Moreover, we showed that combining the KB-based approach with the ESI-based approach outperformed both of them, a result that indicate that the two complement each other while measuring entity relatedness using two orthogonal domains.

References

Krisztian Balog, Pavel Serdyukov, and Arjen P de Vries. 2010. Overview of the TREC 2010 entity track. In *Proceedings of TREC*.

Roi Blanco, Berkant Barla Cambazoglu, Peter Mika, and Nicolas Torzec. 2013. Entity recommendations in web search. In *The Semantic Web–ISWC 2013*, pages 33–48. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, and Sebastiano Vigna. 2009. Query suggestions using query-flow graphs. In *Proceedings of WCSD*, WSCD '09, pages 56–63. ACM.

Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September.

Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. In *Proceedings of WSDM*, WSDM '12, pages 383–392. ACM.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*, volume 57. CRC press.

Paolo Ferragina and Ugo Scaiella. 2010. Fast and accurate annotation of short texts with wikipedia pages. *arXiv preprint arXiv:1006.3498*.

Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Evgeniy Gabrilovich and Shaul Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34(1):443–498, March.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of SIGIR*, SIGIR '09, pages 267–274. ACM.

Alpa Jain and Marco Pennacchiotti. 2010. Open entity extraction from web search query logs. In *Proceedings of COLING*, COLING '10, pages 510–518. Association for Computational Linguistics.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Yanen Li, Bo-June Paul Hsu, and ChengXiang Zhai. 2013. Unsupervised identification of synonymous query intent templates for attribute intents. In *Proceedings of CIKM*, CIKM '13, pages 2029–2038. ACM.

Jiahui Liu and Larry Birnbaum. 2007. Measuring semantic similarity between named entities by searching the web directory. In *Proceedings of Web Intelligence*, WI '07, pages 461–465. IEEE Computer Society.

Hui Liu and Yuquan Chen. 2010. Computing semantic relatedness between named entities using wikipedia. In *Proceedings of AICI*, AICI '10, pages 388–392. IEEE Computer Society.

Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. When web search fails, searchers become askers: understanding the transition. In *Proceedings of SIGIR*, SIGIR '12, pages 801–810. ACM.

Jeffrey Pound, Peter Mika, and Hugo Zaragoza. 2010. Ad-hoc object retrieval in the web of data. In *Proceedings of WWW*, WWW '10, pages 771–780. ACM.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

I Witten and David Milne. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30.

Xiaobing Xue and Xiaoxin Yin. 2011. Topic modeling for named entity queries. In *Proceedings of CIKM*, CIKM '11, pages 2009–2012. ACM.

Xiaoxin Yin and Sarthak Shah. 2010. Building taxonomy of web search intents for name entity queries. In *Proceedings of WWW*, WWW '10, pages 1001–1010. ACM.