Towards an Understanding of Entity-Oriented Search Intents

Darío Garigliotti and Krisztian Balog(✉)

University of Stavanger, Stavanger, Norway
{dario.garigliotti,krisztian.balog}@uis.no

Abstract. Entity-oriented search deals with a wide variety of information needs, from displaying direct answers to interacting with services. In this work, we aim to understand what are prominent entity-oriented search intents and how they can be fulfilled. We develop a scheme of entity intent categories, and use them to annotate a sample of queries. Specifically, we annotate unique query refiners on the level of entity types. We observe that, on average, over half of those refiners seek to interact with a service, while over a quarter of the refiners search for information that may be looked up in a knowledge base.

1 Introduction

A large portion of information needs in web search look for specific entities [11]. Entities are natural units for organizing information, and can provide not only more focused responses, but often immediate answers [9]. Another type of entity-bearing queries is more transaction-oriented. Either trying to book a flight or looking for tickets for an concert, just to mention two popular examples, users are often engaged to fulfill information needs by interacting with a third-party service or application. There has been an increasing focus on supporting task-based search [7], and on modeling actionable knowledge; see, e.g., the dedicated vocabulary for actions in the schema.org ontology, and the NTCIR AKG task.[1] These developments display the interest and efforts towards transforming search engines into actions-guided task completion assistants [1]. In this work, we are interested in studying one particular type of information needs, namely, entity-oriented searches. Specifically, we want to answer a question arising from this web landscape: *what do entity-oriented queries ask for?* Furthermore, which of those searches can be fulfilled by looking up direct answers from a knowledge base, and which would require to interact with external services?

Most entity-oriented queries consist of an entity name, complemented with context terms, i.e., *refiners*, to express the underlying intent of the user [11]. Examples of these queries are *"the rock movies"* and *"london book a hotel"*. Our main objective is to understand entity-related search intents by studying those refiners. Specifically, we represent refiners on the level of entity types.

[1] http://ntcirakg.github.io/tasks.html.

G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, pp. 644–650, 2018.
https://doi.org/10.1007/978-3-319-76941-7_57

Just like entity types boost the disambiguation of known entities and the grouping of emerging ones [10], these type-level characterizations of entity refiners would favor knowledge abstraction and generalization. As an example, by representing with *[city]* any entity of the type city, we want to categorize a refiner, e.g., "rentals", in the type-level query *"[city] rentals"*. Then, we categorize these type-level refiners using an intent classification scheme. Our classification scheme comprises four main categories: property, website, service, and other.

We perform this study without having direct access to past usage data or query logs. To overcome the absence of such data, we exploit query suggestions from a major search engine API. This strategy has been employed successfully in previous work for various applications [3,5]. After acquiring query suggestions for entities of a given type, they are aggregated to extract type-level refiners. Then, for a representative sample of 50 Freebase types, we collect human annotations for those refiners with respect to the classification scheme we developed.

Our main findings show that, on average, more than a half of all unique type-level refiners correspond to interacting with external services, while over a quarter of them look for information that may be looked up in a knowledge base. Another contribution of this work is a large collection of type-level refiners, annotated with intent categories. The resources developed within this paper are made available at http://bit.ly/ecir2018-intents.

2 Related Work

Broder's categorization of information needs is broadly accepted and is the most commonly used one for web search [4], with further refinements, e.g., in [6,13]. We strive for a similar high-level categorization of intents, but specifically for entity-oriented search queries. Previous work has identified high-level patterns from web search queries. For example, according to Lin et al. [8], a query can be classified as an entity, an entity plus a refiner (e.g., *"emma stone 2017"*), a category, a category plus a refiner (e.g., *"doctors in barcelona"*), a website, or other sort of query. Such classification relies merely on lexico-syntactic forms and lacks a more semantically-grounded distinction.

Search intents have been studied in previous work. Reinanda et al. [12] explore entity aspects in user interaction log data. Beyond finding aspects by comparing clustering methods over refiners, they address the tasks of ranking the intents for a given entity independently from a query and recommending aspects. Unlike them, we (i) operate with individual query refiners (i.e., without clustering them together), (ii) model entity intents at the level of types, (iii) always consider entities in queries, and (iv) perform our study in the absence of search logs.

3 Approach

This section describes the process we followed for understanding entity-oriented search intents. An *entity-oriented* or *entity-bearing query* is a query that consists

of an entity name possibly complemented with a refiner, usually as a suffix. Here, by *entity* we mean an individual with its own independent existence, uniquely identified in a knowledge base [2]. More than just a syntactic complement, a *refiner* is a complementary surface form expressing an underlying user *intent* in relation with the entity. As an example, consider the entity *keens steakhouse* (a restaurant) in the search query *"keens steakhouse menu."* The refiner "menu" expresses the intent of reading the restaurant's menu. To understand what these entity-bearing queries ask for, we characterize the refiners on the level of *entity types*, where an entity type is a semantic class that groups entities together with common characteristics. For example, one of the types of *Albert Einstein* in Freebase is `award_winner`.

Our approach, to be detailed in the next subsections, can be summarized as follows. We collect refiners for a set of prominent entities, and aggregate them across entity types to obtain type-level refiners. Next, we develop a classification scheme of *intent categories*, with a focus on how to fulfill the intent expressed by a type-level refiner. Finally, we annotate a representative sample of entity types with intent categories, and obtain a corpus of prominent type-level refiners assigned to those categories.

3.1 Collecting Refiners

We use the type system of Freebase. It is a two-layer categorization system, where types on the leaf level are grouped under high-level domains. Specifically, we use the latest public Freebase dump (2015-03-31), discarding domains meant for administering the Freebase service itself (e.g., `base`, `common`).

We focus on prominent entities, since in this way we benefit from observing a larger and more representative selection of information needs. As the criterion of an entity prominence, we rely on Wikistats page views.[2] This dataset registers the number of times its English Wikipedia article has been requested. We set empirically a prominence threshold of 3,000 page views per article over a span of one year (from June 2015 to May 2016). Given a Freebase type, we select it if it covers at least 100 entities with a prominence above the threshold. Applying these criteria, the selected set contains 634 types.

In a second step, we collect query suggestions from the Google Suggestions API for at most top 1,000 entities per type according to the above prominence criteria. Then, we replace the name of the entity by its type in each query suggestion. This can be viewed as getting queries where a refiner complements the type. For example, the type-level query *"[travel destination] map"* is obtained from all queries for popular travel destinations, e.g., *"sydney map"* and *"paris map"*. Finally, we retain only those refiners that occur in at least 5 suggestions for the given type. This leads to a total of 2,688 distinct type-level refiners for 631 types.

[2] https://dumps.wikimedia.org/other/pagecounts-ez/.

3.2 Classification Scheme

To address our main goal of understanding entity-related search intents, we need a suitable scheme to classify the entity intents. After a close inspection of the type-level refiners, we define the following scheme of *intent categories*. These categories are focused on how (and from which type of source) the information need can be fulfilled.

- **Property**: The refiner looks for a specific entity property or attribute that can be looked up in a knowledge base. For example, "children" in the query *"angelina jolie children"* or "opening times" in *"at&t stadium opening times"*.
- **Website**: The refiner is about reaching a specific website or application. For example, "twitter" in the query *"karpathy twitter"*. This category is a rough equivalent of navigational queries in [4].
- **Service**: The refiner expresses the need to interact with a service, possibly by redirecting to an external site or app. For example, "menu" in the query *"keens steakhouse menu"* would indicate the need for accessing to an external site for reading the restaurant's menu. As another example, "new album" in *"eric clapton new album"* looks for a service to read about, or listen to, or buy the new album.
- **Other**: None of the previous ones is applicable. For example, "batman" in the query *"christian bale batman"* serves to disambiguate the person's role of interest.

3.3 Annotation

We need to sample a set of representative types, since it is unfeasible to annotate all types in the knowledge base. From the set of 631 types, we perform stratified sampling as follows. We sort the types by the total aggregated frequencies of refiners. We delimit 5 roughly equally-sized intervals by the splitting values of 1,500, 3,000, 6,000, and 8,500 refiners per type; we randomly pick 10 types from each interval. We annotate data for this final set of 50 representative Freebase types.

We used crowdsourcing to annotate type-level refiners with intent categories. Specifically, using the Crowdflower platform, for each annotation instance we presented workers with the query, indicating its entity type and refiner, and asked them to select one of the four intent categories. A total of 5,301 unique instances (type-level refiners) were annotated, each by at least 3 judges (5 at most, if necessary to reach a majority agreement, using dynamic judgments). We paid ¢5 per batch, comprising 11 annotation instances. We ensured quality by requiring a minimum accuracy of 80%, a minimum time of 20 s per batch, and a minimum confidence threshold of 0.7. For each type, we only retain an annotated refiner if at least three annotators agreed on the majority category. This leads to a total of 2,313 unique refiners.

4 Results and Analysis

Figure 1 presents the number of refiners classified per each category, for the 50 sampled types, grouped in one plot for each of the 5 intervals of the stratified sampling. Since the final set of types was sampled from types with prominent entities, this ordering, given by the number of refiners, in a way also reflects the prominence of types.

We obtain a distribution of entity intent categories per type after normalizing the frequency of each category by the total of refiners for that type. From the average proportions in these distributions, we can answer our initial questions. A 54.06% of unique entity-oriented queries are to be fulfilled by interacting with some external service or app, meanwhile, 28.6% look for direct answers from a knowledge base. Further, 5.34% of the type-level refiners represent an attempt to reach a website, while 12.08% of them do not fit into any of the previous three categories.

The types with the largest proportion of *service* intents are `netflix_genre` (with refiners, e.g., "videos", "live"), `election` ("map", "polls"), `football_match` ("video", "highlights"), and `music_album`. The *property* intent category covers refiners that are of a more static nature, e.g., `chemical compound` (with refiners like "structural formula", "molecular weight"), `political_party` ("slogan", "president"), `star` ("type of star", "temperature"), or `tower` ("hours", "height"); only the first one is a very prominent type. Most of the entity types exhibit a non-empty proportion of *website* intents. Among all the types, this category exceeds the average proportion, e.g., for `organization`, `business_operation`, `hotel` and `blogger`. The most frequent website refiners in the whole corpus are "wikipedia", "twitter", "facebook", and "youtube". For a few types like `muscle`, `election`,

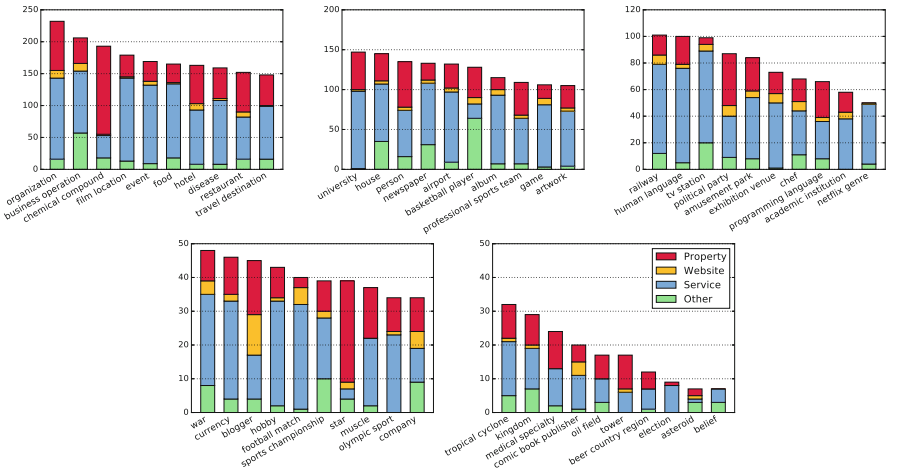

Fig. 1. Distributions of intent categories for the sampled types. Note that the y-axis scales differ.

Table 1. Examples of refiners for each intent category, for each (stratified) type group.

Entity type	Intent category			
	Property	Website	Service	Other
comic_book_publisher	logo, address	wiki, website, twitter	submissions, publishing, comics	movies
tower	height, address, opening hours	wiki	tickets, restaurant	collapse
war	deaths, results, cause	youtube, wikipedia, reddit, quizlet	video, uniforms, pictures, documentary	ap euro, in hindi
academic_institution	logo, email, notable alumni	wiki, login, twitter, portal	scholarships, ranking, map, library, jobs	baseball
automotive_company	stock, logo, ceo, address	wikipedia, website, linkedin, facebook	parts, careers, investor relations	india, inc
programming_language	syntax, ide	wikipedia, wiki, github	jobs, examples, interview questions	3, 2017
restaurant	phone number, owner, location	yelp, twitter, app, tripadvisor, groupon	wine list, vouchers, recipes, menu prices	sf, nj, nyc
music_album	value, cast, release date	youtube, wikipedia, amazon, imdb	zip download, video, ukulele chords, tracklist	2015, lp
person	son, salary, real name	youtube, instagram, snapchat	tour, quotes, photos, new album	sr, now, ww2
travel_destination	zip code, train station	craigslist	weather radar, vacation, tours, things to do	today, nj

belief, or medical_speciality, all in the lowest populated groups, no website refiner is present. A marginal proportion of refiners are classified as having the *other* intent. A few exceptional cases with large proportions of other intents are, e.g., business_operation and house (where the refiner is usually a location), or basketball_player (for which many refiners refer mostly to an NBA franchise, e.g., "lakers"). Table 1 provides additional examples for a selection of types.

5 Conclusions and Future Work

The study performed in this work has lead to a better understanding of what entity-oriented queries ask for. We have developed a classification scheme to categorize entity-oriented search intents and annotated a representative sample of type-level refiners using this scheme. We have found that, on average, more than a half of those are to be fulfilled by interaction with services; another large proportion of information needs look for direct answers from a knowledge base. Several lines of future work arise from our study. One of them is to develop a method for automatic intent categorization. Another direction is the clustering of refiners which express the same underlying intent. Finally, we seek to extend our approach to be able to capture tail entities and intents.

References

1. Balog, K.: Task-completion engines: a vision with a plan. In: Proceedings of the 1st International Workshop on Supporting Complex Search Tasks (2015)
2. Balog, K.: Entity retrieval. Encyclopedia of Database Systems, pp. 1–6. Springer, New York (2017). https://doi.org/10.1007/978-1-4899-7993-3
3. Benetka, J.R., Balog, K., Nørvåg, K.: Anticipating information needs based on check-in activity. In: Proceedings of WSDM, pp. 41–50 (2017)
4. Broder, A.: A taxonomy of web search. SIGIR Forum **36**(2), 3–10 (2002)
5. Fourney, A., Mann, R., Terry, M.: Characterizing the usability of interactive applications through query log analysis. In: Proceedings of CHI, pp. 1817–1826 (2011)
6. Jansen, B.J., Booth, D.L., Spink, A.: Determining the informational, navigational, and transactional intent of web queries. Inf. Process. Manage. **44**(3), 1251–1266 (2008)
7. Kelly, D., Arguello, J., Capra, R.: NSF workshop on task-based information search systems. SIGIR Forum **47**(2), 116–127 (2013)
8. Lin, T., Pantel, P., Gamon, M., Kannan, A., Fuxman, A.: Active objects: actions for entity-centric search. In: Proceedings of WWW, pp. 589–598 (2012)
9. Mika, A.: Entity search on the web. In: Proceedings of WWW, pp. 1231–1232 (2013)
10. Nakashole, N., Tylenda, T., Weikum, G.: Fine-grained semantic typing of emerging entities. In: Proceedings of ACL, pp. 1488–1497 (2013)
11. Pound, J., Mika, P., Zaragoza, H.: Ad-hoc object retrieval in the web of data. In: Proceedings of WWW, pp. 771–780 (2010)
12. Reinanda, R., Meij, E., de Rijke, M.: Mining, ranking and recommending entity aspects. In: Proceedings of SIGIR, pp. 263–272 (2015)
13. Rose, D.E., Levinson, D.: Understanding user goals in web search. In: Proceedings of WWW, pp. 13–19 (2004)