E

Entity Retrieval

Krisztian Balog
University of Stavanger, Stavanger, Norway

Synonyms

Entity search

Definition

Entities are uniquely identifiable objects or things (such as persons, organizations, and places), characterized by their types, attributes, and relationships to other entities. *Entity retrieval* refers to a variety of search tasks where the user is presented with specific entities, or properties of entities, instead of documents, in response to a search query.

The area of entity retrieval has a number of unique characteristics that makes it challenging and sets it apart from standard document retrieval. First, unlike documents, entities are not necessarily directly represented as retrievable units but need to be recognized and identified through occurrences in documents. Information about a given entity has to be collected and aggregated from multiple documents and even multiple collections and potentially combined with structured data sources. Second, there is more structure (or references to structures) available than in standard document retrieval: entity types (from some taxonomy), attributes (from some ontology), and relationships to other entities ("typed links"). Exploiting these structures is a fertile area of ongoing research.

Historical Background

Up until the mid-1990s, information retrieval (IR) research has focused primarily on document retrieval. This document-oriented view was challenged by researchers working on *question answering* (QA) who claimed that "users want answers, not documents." It was QA that first considered entities as the unit of retrieval. Specifically, many factoid questions (e.g., "Who invented the paper clip?") and list questions (e.g., "Which countries have won the FIFA world cup?") can be answered by named entities such as people, locations, dates, etc. The area received a big boost with the arrival of the TREC Question Answering track in 1999 [28].

The *expert finding* task, studied at the TREC Enterprise track (2005–2008) [11], focused on a single type of entity: people. Given a crawl of an organization's intranet and a list of candidate experts, the task is to find the experts (i.e., a ranked list of people) on a particular query topic. Expert finding received considerable attention from the IR research community, and rapid progress was made, in modeling, algorithm design, and evaluation [6].

L. Liu, M.T. Özsu (eds.), *Encyclopedia of Database Systems*,
DOI 10.1007/978-1-4899-7993-3_80724-1

The INEX Entity Ranking task (2007–2009) [12] broadened the scope of retrieval to include arbitrary types of entities. The document collection is (a dump of the English) Wikipedia, and entities are represented by their corresponding Wikipedia articles. Two tasks are presented: entity ranking and list completion. Both seek a ranked list of entities in response to a free-text query (e.g., "Nordic authors who are known for children's literature") but differ in how the textual query is supplemented with additional clues, in order to better specify the underlying information need. In *entity ranking*, the topic statement specifies a small number of target categories, according to the category system of Wikipedia (e.g., "writers"), while in *list completion*, a small number of example entities are provided (e.g., "Hans Christian Andersen," "Astrid Lindgren," and "Tove Jansson"). The rich structure of the collection (category information and links between entities) can be exploited to improve retrieval performance over plain document retrieval [15].

Next, the TREC Entity track (2009–2011) [4] introduced the *related entity finding* task: given a source entity (e.g., "Michael Schumacher"), a relation (e.g., "His teammates while he was racing in Formula 1"), and a target type (e.g., "people"), identify entities that stand in the specified relation with the source entity and that satisfy the target type constraint. This task aims at making arbitrary relations between entities searchable in unstructured Web data, where answer entities are not restricted anymore to having their own Wikipedia page.

The Semantic Search Challenge (2010–2011) [9] addressed entity retrieval over structured data, represented in RDF, by means of keyword queries. Two tasks are investigated: *entity search* queries refer to one particular entity, albeit often an ambiguous one (e.g., "Ben Franklin"); *list search* queries target a group of entities that match certain criteria (e.g., "Axis powers of World War II"). Similarly, the 2011 edition of the INEX Data Centric track [30] also addressed *known item search* and *list search* as separate tasks, but the data collection is the Internet Movie Database (IMDB) in XML format, and entities are restricted to movies and persons. Most recently, the INEX Linked Data track (2012–2013) [29] considered *ad hoc entity search* over Wikipedia, but complemented the textual content with RDF properties from both DBpedia and YAGO2, "with the goal to bring together different communities and to foster research at the intersection of Information Retrieval, Databases, and the Semantic Web" [29].

Scientific Fundamentals

The main research questions in entity retrieval, similarly to other retrieval tasks, can be organized around three main themes: (i) How to represent entities? (ii) How to represent information needs? (iii) How to match these representations? These three questions are closely interconnected, and solutions to them should not be sought in isolation. This entry focuses on models and approaches that answer information needs with a ranked list of entities.

Characteristics of Entities

Entities are characterized by having the following properties:

- Unique identifier(s)
- Name(s)
- Type(s)
- Attributes (or description)
- (Typed) relationships to other entities

Entities need to be uniquely identifiable; for example, the TREC Enterprise track used email addresses to identify experts, the INEX Entity Ranking track relied on Wikipedia page IDs, and the Semantic Search Challenge used URIs. Within a given data collection, there is a one-to-one correspondence between entity IDs and the real-world objects they represent. The same entity, however, may be recorded under different identifiers in other collections; reconciling these records is an important information integration problem. Entities are typically proper

nouns that are known (and referred to) by one or more names (often referred to as "surface forms"). Type can be defined using a set of possible values or associated with some hierarchical type classification system (i.e., a taxonomy). In the context of structured data (relational databases or RDF), attributes are a collection of property-value pairs, where properties come from some controlled vocabulary or ontology. Similarly, relationships are typed using a controlled-vocabulary scheme. When working with unstructured data, entities and relations are commonly represented by their associated "language usage," as observed around entity occurrences in documents.

Ranking Term-Based Entity Representations

To be able to retrieve entities by means of keyword queries, textual representations need to be constructed. We distinguish between two main settings: (i) when there are ready-made entity descriptions available and (ii) when information about entities needs to be collected and aggregated from unstructured data.

With ready-made entity descriptions. It is not unrealistic to assume that entity descriptions are made readily available in an organized form. Examples include entity profile homepages (e.g., a movie page in IMDB, a LinkedIn profile, or a Wikipedia article), database records or RDF triples describing an entity. The basic approach is to ignore any structural clues or elements in these entity descriptions and construct a bag-of-words document representation for each entity. These entity profile documents can then be ranked using standard document retrieval techniques, such as language models [5, 15] or BM25 [22]. In practice, such entity descriptions are rarely just flat text. The predominant approach to incorporating structure is to represent each entity using a set of fields, where fields correspond to properties (predicates) of the entity. These representations can then be ranked using fielded extensions of standard document retrieval models, such as the Mixture of Language Models (MLM) [21] or BM25F [22]. One pragmatic issue in heterogeneous data collections is that the number of possible document fields is huge and the optimization of field weights becomes intractable. A commonly used solution is *predicate folding*: reducing the number of fields by grouping them together. This grouping can be based on the type of predicates (attributes, in/out relations, etc.) [21, 22] or on their (manually determined) importance [8]. Instead of assigning a fixed weight to each field, the Probabilistic Retrieval Model for Semistructured Data (PRMS) allows for a dynamic mapping of query terms to document fields on a term-by-term basis [17]. (It has to be noted that the PRMS model assumes a homogenous collection and that fields have distinctive term distributions [1].)

Without ready-made entity descriptions. When entity descriptions are not readily available, they need to be constructed by examining the contents of documents in which the entities are mentioned. Much of the modeling in this area has been developed in the context of expert finding [6]. Since these models do not directly model the concept of expertise but, instead, estimate the strength of the association between a person and the query topic, they are generally applicable to entity retrieval, assuming that documents have been annotated with entities (e.g., using entity linking techniques). Two principal approaches are distinguished: *candidate-centric approaches* build a term-based representation of entities, from terms co-occurring with entity mentions, and then rank these representations; *document-centric approaches* first locate documents that are relevant to the query and then consider the entities associated with these documents. The language modeling framework offers a theoretically sound statistical formulation of these ideas that is coupled with good empirical performance [3]. Alternative approaches include discriminative probabilistic models [14], voting models [20], and graph-based models [25].

Incorporating Type Information

One of the distinctive characteristics of entities is that they are typed. Target type(s) may be pro-

vided explicitly by the user, for example, through query suggestion or faceted search services. Alternatively, target type(s) may be inferred automatically, by considering the types associated with the top-ranked entities [15] or entities in relevant sentences [26]. Types may also be ranked much like entities, i.e., by applying the candidate-centric and document-centric strategies [2].

Type-based similarity is commonly incorporated as an additional component to the retrieval model (through weighted addition or multiplication with the term-based component). In practice, the type information stored for entities is often incomplete or inconsistent (due to human annotators). In large hierarchical type system, further issues arise related to granularity, i.e., too generic or too specific types may be assigned. Therefore, target types should not be treated as a strict filter but used for estimating a type-based similarity score. Possibilities include set-based similarity [27], distance in the hierarchy [15], term-based similarity (based on type labels) [13], or distributional similarity [5]. Further, target types may be expanded based on the type hierarchy (e.g., by considering subcategories or siblings) [13] or using feedback techniques [5]. Instead of performing type detection and entity ranking as two subsequent steps, a recent study proposes to complete them jointly and shows that it outperforms the two-step approach [24].

Exploiting and Searching Relationships

The link structure of entities, as defined by their relationships, can be exploited for retrieval as relevance propagation [25] or as entity priors (with different flavors of centrality measures, such as in-degree or PageRank) [15].

Searching for related entities is an area of ongoing research. One line of work focuses on purpose-built interfaces that assist the user in formulating complex queries; see, e.g., [7, 18]. Another line of research studies the "single search box" paradigm, where users express their information need in unconstrained natural language. The TREC Entity track provided an evaluation platform for the abstraction of this second setting, where the source entity and target type were pre-annotated in the query. A typical pipeline for this tasks consist of candidate entity identification (based on co-occurrences with the source entity), filtering (based on target type), entity ranking, and (optionally) entity homepage finding [10,16]. Knowledge bases (like Wikipedia or DBpedia) can be utilized in any of these steps [4].

Key Applications

Entity retrieval is becoming increasingly important in all information access domains. General Web search is one of the flagship applications; it has been estimated that over 40% of Web search queries target entities [19,23]. Major Web search engines rely heavily on large-scale knowledge bases to effectively answer these requests. Vertical search is focused on a single or, at most, a handful of entity types in a given domain (such as shopping, travel, or scholarly literature). Similarly, enterprise search also focuses on entities within a single domain, that is, the organization at hand. Social networks represent yet another application area, where a greater emphasis is placed on relationships between entities than on the attributes of entities.

Entities are a key enabling component for *semantic search*; they can be used to enrich search result pages to enable direct answers, contextual information, or serendipitous results. Further, entities provide a means to bridge the gap between unstructured and structured data.

Future Directions

There is a clear trend in entity retrieval toward increasing reliance on structured data sources. Commercial providers devote significant resources to populating graph-based entity repositories, such as Google's Knowledge Graph and Microsoft's Satori. Academic efforts include the TAC Knowledge Base Population and TREC Knowledge Base Acceleration tracks. On top of identifying properties and relationships, entities may also be characterized in terms of actions that can be performed on them [19].

Entity Retrieval, Table 1 Entity retrieval test collections. The last column indicates whether the natural language (keyword) query is complemented with additional components

Evaluation campaign	Data collection	Additional query component(s)
TREC Enterprise track [11]	Intranet crawl (W3C, CSIRO)	– / example documents (in 2007)
INEX Entity Ranking [12]	Wikipedia	Target types or example entities
TREC Entity [4]	Web crawl (ClueWeb09)	Source entity, target type
Semantic Search Challenge [9]	Semantic Web crawl (BTC2009)	–
INEX Data Centric [30]	XML (IMDB)	–
INEX Linked Data [29]	Wikipedia+DBpedia+YAGO2	–

Data Sets

A number of entity retrieval test collections have been developed within the context of various evaluation campaigns. Each comprises a data collection, a set of queries, and corresponding relevance judgments. Table 1 presents an overview.

Cross-References

- Concept-Based Information Retrieval
- Field-Based Information Retrieval Models
- Keyword Search in Databases
- Knowledge Bases/Web of Data
- Semantic Search

Recommended Reading

1. Balog K. Semistructured data search. In: Ferro N, editor. PROMISE winter school 2013. Lecture notes in computer science, vol. 8173. Berlin/Heidelberg: Springer; 2014. p. 74–96.
2. Balog K, Neumayer R. Hierarchical target type identification for entity-oriented queries. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM; 2012. p. 2391–4.
3. Balog K, Azzopardi L, de Rijke M. A language modeling framework for expert finding. Inf Process Manag. 2009;45:1–19.
4. Balog K, de Vries AP, Serdyukov P, Thomas P, Westerveld T. Overview of the TREC 2009 entity track. In: Proceedings of the 18th Text REtrieval Conference. NIST Special Publication; 2010.
5. Balog K, Bron M, De Rijke M. Query modeling for entity search based on terms, categories, and examples. ACM Trans Inf Syst. 2011;29(4):22:1–22:31.
6. Balog K, Fang Y, de Rijke M, Serdyukov P, Si L. Expertise retrieval. Found Trends Inf Retr. 2012;6(2–3):127–256.
7. Bast H, Bäurle F, Buchhold B, Haussmann E. A case for semantic full-text search. In: Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search. ACM; 2012.
8. Blanco R, Mika P, Vigna S. Effective and efficient entity search in RDF data. In: Proceedings of the 10th International Conference on The Semantic Web. Springer; 2011. p. 83–97.
9. Blanco R, Halpin H, Herzig DM, Mika P, Pound J, Thompson HS, Tran T. Repeatable and reliable semantic search evaluation. Web Semant Sci Serv Agents World Wide Web. 2013;21(0):14–29.
10. Bron M, Balog K, de Rijke M. Ranking related entities: components and analyses. In: 19th ACM International Conference on Information and Knowledge Management. ACM; 2010. p. 1079–88.
11. Craswell N, de Vries A, Soboroff I. Overview of the TREC-2005 enterprise track. In: Proceedings of the 14th Text REtrieval Conference. NIST Special Publication; 2006.
12. de Vries A, Vercoustre A-M, Thom JA, Craswell N, Lalmas M. Overview of the INEX 2007 entity ranking track. In: Focused access to XML documents, vol. 4862. Berlin/Heidelberg: Springer; 2008. p. 245–51.
13. Demartini G, Firan CS, Iofciu T, Krestel R, Nejdl W. Why finding entities in Wikipedia is difficult, sometimes. Inf Retr. 2010;13(5):534–67.
14. Fang Y, Si L, Mathur AP. Discriminative probabilistic models for expert search in heterogeneous information sources. Inf Retr. 2011;14(2):158–77.
15. Kaptein R, Kamps J. Exploiting the category structure of Wikipedia for entity ranking. Artif Intell. 2013;194:111–29.
16. Kaptein R, Serdyukov P, De Vries A, Kamps J. Entity ranking using Wikipedia as a pivot. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM; 2010. p. 69–78.
17. Kim J, Xue X, Croft WB. A probabilistic retrieval model for semistructured data. In: Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval. Springer; 2009. p. 228–39.

18. Li X, Li C, Yu C. Entityengine: answering entity-relationship queries using shallow semantics. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM; 2010. p. 1925–6.
19. Lin T, Pantel P, Gamon M, Kannan A, Fuxman A. Active objects. In: Proceedings of the 21st International Conference on World Wide Web. ACM; 2012. p. 589–98.
20. Macdonald C, Ounis I. Voting for candidates: adapting data fusion techniques for an expert search task. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. ACM; 2006. p. 387–96.
21. Neumayer R, Balog K, Nørvåg K. On the modeling of entities for ad-hoc entity search in the web of data. In: Proceedings of the 34th European Conference on Advances in Information Retrieval. Springer; 2012. p. 133–45.
22. Pérez-Agüera JR, Arroyo J, Greenberg J, Iglesias JP, Fresno V. Using BM25F for semantic search. In: Proceedings of the 3rd International Semantic Search Workshop. ACM; 2010. p. 1–8.
23. Pound J, Mika P, Zaragoza H. Ad-hoc object retrieval in the web of data. In: Proceedings of the 19th International Conference on World Wide Web. ACM; 2010. p. 771–80.
24. Sawant U, Chakrabarti S. Learning joint query interpretation and response ranking. In: Proceedings of the 22nd International Conference on World Wide Web; 2013. p. 1099–109.
25. Serdyukov P, Rode H, Hiemstra D. Modeling multi-step relevance propagation for expert finding. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM; 2008. p. 1133–42.
26. Vallet D, Zaragoza H. Inferring the most important types of a query: a semantic approach. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2008. p. 857–8.
27. Vercoustre A-M, Pehcevski J, Thom JA. Using Wikipedia categories and links in entity ranking. In: Focused Access to XML Documents: 6th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2007). Lecture notes in computer science, vol. 4862. Springer: Heidelberg; 2008. p. 321–35.
28. Voorhees EM. The TREC question answering track. Nat Lang Eng. 2001;7(4):361–78.
29. Wang Q, Kamps J, Camps GR, Marx M, Schuth A, Theobald M, Gurajada S, Mishra A. Overview of the INEX 2012 linked data track. In: Forner P, Karlgren J, Womser-Hacker C, editors. CLEF 2012 evaluation labs and workshop, online working notes; 2012.
30. Wang Q, Ramírez G, Marx M, Theobald M, Kamps J. Overview of the INEX 2011 data centric track. In: Geva S, Kamps J, Schenkel R, editors. Focused Retrieval of Content and Structure: 10th International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2011). Lecture notes in computer science, vol. 7424. Springer; 2012.