Entity-Based Retrieval

Hadas Raviv
Faculty of Industrial Engineering and Management
Technion, Haifa 32000, Israel
hadasrv@tx.technion.ac.il

ABSTRACT

We address the core challenge of the entity retrieval task: ranking entities in response to a query by their presumed relevance to the information need that the query represents. As an initial research direction we explored two models for entity ranking that were evaluated using the INEX entity ranking dataset and which posted promising performance. A natural future direction to explore is how to generalize these models to address various types of information needs that are associated with entities.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords: entity retrieval, markov random fields, cluster ranking

1. INTRODUCTION

The ad hoc retrieval task in information retrieval (IR) is to answer a user's information need, represented by a query, by retrieving relevant information satisfying that need. Traditionally, this task is handled by ranking a list of documents in response to a query. Recently, it has been observed that for many queries, named entities (e.g. people, organizations) better satisfy the user's information need than full documents [4]. This observation drove forward work on the entity retrieval task. The task is focused on queries that are better addressed by entities than by documents. The goal is ranking entities in response to such queries.

In the last decade several entity retrieval tasks have been studied. Examples include the tasks of retrieving entities of specific types (e.g., experts [2]), from specific domains (e.g., enterprise [2] , Wikipedia [3]) and retrieving entities that satisfy specific types of information needs (e.g., relation retrieval [1]).

These entity retrieval tasks pose two major challenges. The first challenge is *entity representation.* Entities are complex objects that can be associated with a varying number of properties (e.g., type, name) as well as defined by different identifiers (e.g., Web homepage, Wikipedia page, URI) in various domains. As a result, there is no single definition of the object (entity) that should be retrieved.

The second challenge is devising generic *entity ranking models* that utilize various entity representation approaches and address various types of entity related information needs. Many entity ranking algorithms are less generic than document ranking algorithms, and are often adapted to a specific corpus, task or entity representation.

We define an entity oriented query as a query for which the underlying information need is better satisfied by entities than by documents. In our research we intend to develop retrieval models for addressing such queries. Specifically, we intend to address the second challenge discussed above and utilize a generic entity representation for developing generic methods for entity retrieval.

The research direction just described is very broad. As a first step, we focused on a specific task. The goal is to rank Wikipedia entities in response to a query by defining a criteria that these entities should fulfil. We developed two ranking models. The first is based on the Markov Random Fields (MRF) framework [6]. The second is based on a cluster-based retrieval approach [5]. These models were evaluated using the INEX entity ranking dataset [3], aimed at retrieving entities in Wikipedia, and showed promising results.

Our plans are to generalize the MRF and cluster-based models for addressing various types of entity oriented queries in various types of collections. Specifically, we intend to explore the connection between entity related information used for ranking and the type of query addressed.

Acknowledgements. This research is based upon joint work with Oren Kurland and David Carmel.

2. REFERENCES

[1] K. Balog, P. Serdyukov, and A. P. de Vries. Overview of the TREC 2011 entity track. In *Proc. of TREC*, 2012.

[2] K. Balog, P. Thomas, N. Craswell, I. Soboroff, P. Bailey, and A. P. De Vries. Overview of the trec 2008 enterprise track. Technical report, DTIC Document, 2008.

[3] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the inex 2009 entity ranking track. In *Proc. of INEX*, pages 254–264, 2010.

[4] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc object retrieval in the web of data. In *Proc. of WWW*, pages 771–780, 2010.

[5] H. Raviv, D. Carmel, and O. Kurland. A ranking framework for entity oriented search using markov random fields. In *Proc. of JIWES*, page 1, 2012.

[6] H. Raviv, O. Kurland, and D. Carmel. The cluster hypothesis for entity oriented search. In *Proc. of SIGIR*, pages 841–844, 2013.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.
ACM 978-1-4503-2257-7/14/07.
http://dx.doi.org/10.1145/2600428.2610378 .