The First International Workshop on Entity-Oriented Search (EOS)

Krisztian Balog
NTNU, Norway
krisztian.balog@idi.ntnu.no

Arjen P. de Vries
CWI/TU Delft
arjen@acm.org

Pavel Serdyukov
Yandex, Russia
pavser@yandex-team.ru

Ji-Rong Wen
Microsoft Research Asia
jrwen@microsoft.com

Abstract

The First International Workshop on Entity-Oriented Search (EOS) workshop was held on July 28, 2011 in Beijing, China, in conjunction with the 34th Annual International ACM SIGIR Conference (SIGIR 2011). The objective for the workshop was to bring together academic researchers and industry practitioners working on entity-oriented search to discuss tasks and challenges, and to uncover the next frontiers for academic research on the topic. The workshop program accommodated two invited talks, eleven refereed papers divided into three technical paper sessions, and a group discussion.

1 Introduction

Many user information needs concern entities: people, organizations, locations, products, etc. These are better answered by returning specific objects instead of just any type of documents. Both commercial systems and the research community are displaying an increased interest in returning "objects," "entities," or their properties in response to a user's query. While major search engines are capable of recognizing specific types of objects (e.g., locations, events, celebrities), true entity search still has a long way to go.

Entity retrieval is challenging as "objects" unlike documents, are not directly represented and need to be identified and recognized in the mixed space of structured and unstructured Web data. While standard document retrieval methods applied to textual representations of entities do seem to provide reasonable performance, a big open question remains how much influence the entity type should have on the ranking algorithms developed.

The objective of the workshop was to provide a forum to discuss entity-oriented search, without restricting to any particular data collection, entity type, or user task, and to solicit research contributions on topics including entity mining, entity ranking, query log analysis, or user context. In sum, the workshop tried to uncover the next research frontiers in entity-oriented search.

We accepted a total of 11 papers out of 19 submissions. Each was reviewed by at least two members of the program committee, consisting of 27 researchers from academia and industry,

representing a broad range of disciplines. Accepted contributions were presented either as short (8+2 min) or regular (20+5 min) oral presentations. A best paper award was given out based on a secret ballot voting conducted among workshop participants. In addition, the EOS program featured two invited talks by Paul Ogilvie (LinkedIn) and Andrey Plakhov (Yandex). The day was concluded with a discussion session.

In the following section we give an overview of the workshop program. The papers are available online at `http://research.microsoft.com/sigir-entitysearch/`.

2 Workshop Program

The day was divided into two invited talks, three technical paper sessions, and a final discussion period. We summarize each below.

2.1 Invited Talk 1

The first invited talk, entitled *Anchoring Relevance with Entities*, was given by Paul Ogilvie (LinkedIn). The following entity-oriented information need was used as an example during the first part of talk: "Who bought Skype?". Paul discussed three different solutions to answering this question. A QA approach is to retrieve candidate passages, extract entities from them, and perform inference across the results. Alternatively, a typical TREC/INEX approach to entity search would first extract entities from the collection, build textual representations for them, which, in turn, can be ranked. Finally, one could try to find the answer in knowledge bases, such as Wikipedia or Freebase, by formulating a structured query. Typically, the first two families of approaches would result in high recall at the cost of precision, while with knowledge bases it is the other way around. There is no one-size-fits-all approach; it depends on the actual information need which strategy would perform best. No matter which retrieval technique is used, the textual context of the extracted entities is important.

The second part of the talk focused on LinkedIn ("the Facebook of professionals"). LinkedIn is a massive network of entities of types that are specific to the LinkedIn ecosystem: people, companies, industries, locations, skills, etc. Entities are created either (1) by users (e.g., people), (2) using curation or external resources (e.g., industries and locations), (3) algorithmically (e.g., skills), or (4) a mixture of the above (e.g., companies). Linking entities to other entities it mostly done manually (by users), while linking content to entities is mainly performed algorithmically. When performing a search task, in addition to the textual context, the context may be relationships to the user. The importance of the context varies across information needs, but this can be detected. To date, the user's network of interest and connections are largely unexploited in search beyond degree distance. To be able to advance there, it is important to do the entity recognition early on in the processing pipeline.

Next in the presentation, Paul discussed challenges, chief of which are entities themselves. Using existing off-the-shelf solutions for named entity recognition limits the research questions we can ask. The Stanford NER system, for example, identifies people, locations, and organizations. But what about products, music groups, architecture, important concepts, fields of study, etc.? To tap into new entity-oriented search tasks, we need to be more engaged in the task of entity recognition itself. A very related problem is entity resolution (or normalization). We would first need to understand when ambiguity happens; one could

take, for example, Wikipedia disambiguation pages to identify classes of ambiguous entities. Another major challenge is the issue of test collections that raises a number of questions: How do we capture user context for reusable text collections? How do we collect relevance judgments in this context? Can people provide meaningful relevance judgments given another user's context? How can industry and academia cooperate without violating privacy concerns?

Paul closed his talk with a proposal for using Wikipedia as a test collection. Wikipedia has been used extensively before (for example at the INEX Entity Ranking track), but research has primarily focused on the entity network of content. Wikipedia as a social network is still untapped. There is a vibrant community of editors, with over 370 thousand users with five or more edits in the Articles namespace per month. Talk pages, that are like a wall, email, or public chat, give us an explicit social network of users. Moreover, the edit history provides a way to connect content entities with editors, i.e., this way we have connections between the two networks. Finding information about entities (not just the entity itself) on Wikipedia is difficult, therefore, it might be a good task for investigating the role of entities and for identifying query types where the role of context varies.

2.2 Short Paper Session

The paper by Liu et al. [10], *High Performance Clustering for Web Person Name Disambiguation Using Topic Capturing*, addresses the web people search task of the Web People Search (WePS) evaluation campaign: clustering web search results corresponding to different namesakes. The authors extend the standard hierarchical agglomerative clustering method by considering an additional centroid vector that captures the topicality of clusters. The proposed solution has shown to outperform all WePS participating systems.

In their paper, entitled *Extracting Dish Names from Chinese Blog Reviews Using Suffix Arrays and a Multi-Modal CRF Model*, Tsai and Chou [13] present an interesting sub-task within the problem of mining restaurant reviews: extracting (Chinese) dish names. It is a challenging task for two main reasons: (1) there are no comprehensive dictionaries or lists of dish names available, and (2) in Chinese culture, the dish names can be extremely creative (e.g., "Buddha jumps over the wall"). The authors propose a two-step approach that consists of an unsupervised candidate identification step (using suffix arrays of n-grams that appear at least twice in the reviews), followed by supervised candidate validation phase (using superstring/substring relationships simultaneously using CRF).

Lin et al. [9] study the Related Entity Finding (REF) task of the TREC Entity track in their paper entitled *LADS: Rapid Development of a Learning-To-Rank Based Related Entity Finding System using Open Advancement*. Using a modular architecture that consists of interchangeable components with common interfaces, the authors discuss and extensively evaluate six components: query analysis, document retrieval, entity extraction, feature extraction, entity ranking, and homepage retrieval.

In their paper, entitled *Differences in Document Retrieval and Entity Retrieval: Finding Support Documents with a Learning-to-Rank Approach*, Li and He [8] address the task of finding supporting documents for entities. They show that although sounds similar, this task is different from conventional document retrieval. The authors address the supporting document finding task as a learning to rank problem and propose a logistic regression method based on four types of features: query features, document features, rank features, and similarity features. Using 70 topics from the 2009 and 2010 editions of the TREC Entity

track, the logistic regression method has shown to outperform three baselines approaches.

The paper *The Sindice-2011 Dataset for Entity-Oriented Search in the Web of Data* by Campinas et al. [5] describes a large new entity-oriented data collection. The authors argue that the Billion Triple Challenge 2009 (BTC-2009) dataset, used in both editions of the Semantic Search Challenge and also at the 2010 TREC Entity track, is not representative anymore of what can be found on the Web. The new data collection contains a total of 11 billion RDF statements about 1.7 billion entities, and is an accurate reflection of the current Web of Data. Along with the dataset, an open source search infrastructure is provided with functionalities to process, index, and retrieve entities in this collection.

2.3 Paper Session 1

The paper by Sun and Grishman [12] entitled *Cross-Domain Bootstrapping for Named Entity Recognition* addresses the domain adaptation problem of named entity recognizers: systems trained on one domain perform poorly out-of-domain. The authors propose a cross-domain bootstrapping method with two novel features: training seeds are generalized with word clusters (using the Brown word clustering algorithm) and automatically classified instances are selected based on multiple criteria (novelty, confidence, density, and diversity).

In their paper entitled *Semi-supervised Statistical Inference for Business Entities Extraction and Business Relations Discovery*, Lau and Zhang [7] present a two-step approach for extracting business networks, i.e., entities and their relationships, from Web 2.0 data (financial news, reviews, blogs, forums, etc.). Step one concerns the extraction of entities; it is performed using an extended binary mutual information measure in combination with a backward searching algorithm to semi-automatically expand the gazetteer of known companies. Step two addresses the discovery of business relations, with two specific type of relations in the focus: collaboration and competition. The authors develop a statistical inference method to automatically extract a set of domain-specific relationship indicators based on some seeding relation lexicon.

Vechtomova [14] in her paper entitled *Unsupervised Related Entity Finding* propose an unsupervised approach to the related entity finding task, using NLP tools. An initial candidate list of entities is extracted from top ranked documents retrieved using a web search engine. This list is then refined based on the similarity of candidate entities and so-called seed entities, which are instances of the target category. Category names are extracted by POS tagging and NP-chunking of the topic narrative. To compute distributional similarity, entities are represented by feature vectors created out of grammatical relationships. Experimental evaluation is performed on the TREC Entity 2010 data set as well as on QA list questions from TREC 2005.

2.4 Invited Talk 2

The second invited talk was given by Andrey Plakhov (Yandex) with the title *Entity-oriented Search Result Diversification*. He discussed the approach to search results diversification currently used at Yandex, a system called Spectrum. Spectrum performs reformulation-driven search results diversification using a greedy algorithm (similar to xQuAD) that optimizes an IA-type diversity metric, pFound-IA (that is similar to other IA metrics [1]). This metric requires the fraction of search intents amongst all query instances; Spectrum uses query expansions to understand query intents and/or different query meanings, like in [11].

Building on the observation that there are not that many types of frequent intents (but there is a long tail of rare intents), Spectrum focuses on queries that fall into a predefined list of frequent categories: movies, books, people, gadgets, cars, diseases, etc. The system first identifies entities in queries. Each entity is then classified into one or more categories. For each category, there is a limited number of aspects/facets that users typically look for when searching for entities of that type. For example, for cars: compare, review, images, info, parts, etc.; for diseases: symptoms, treatment, epidemiology, textbook, etc. Spectrum distinguishes between three specific types of queries: (1) entity (e.g., "el gauchito," "bmw x5"), (2) entity + indicator, where the indicator could be an entity class or just a clue (e.g., "el gauchito restaurant," "bmw x5 dealers"), and (3) entity + explicit search intent (e.g., "el gauchito driving directions," "bmw x5 used"). Besides the category-specific indicators and intents, there are so-called "universal intents." These are query expansions that should always be considered as explicitly stated intents (even if we don't know the entity's category), e.g., "images/photos," "list," "price," "translation", etc. These universal intents come from a (language-specific) dictionary.

Putting the above components together, for an ambiguous query Spectrum knows (1) what categories it falls into, (2) what search intents we see in its expansions (along with frequency counts), and (3) what universal search intents are present in its expansions (again, with counts). This is enough information to generate a ranking that maximizes the target search effectiveness metric (pFound-IA). Spectrum alters the search engine result page for 15-20% of all search queries which amounts to a total of about 15 million a day, to date. It resulted in 1% less abandonment on popular queries, and CTR for positions from 2 to 10 are went up by 2-5%. Also, it allows to easily implement search intent highlighting in snippets.

Future work concerns automated synonymic intents detection (e.g., "download" vs. "download for free" and "trailer" vs. "movie trailer online") and answering "natural language" queries. It is in between quotes, because these queries are not actually in Russian (or in any other spoken language); they lack complex grammar features, but emerge and evolve in a natural way as users learn frequent patterns (typically containing entity, clarification, and search intent in some order). Expressive constructs become more frequent, while ineffective constructs perish. Given the simple grammar and the availability of full usage statistics, it is a fertile ground for research.

2.5 Paper Session 2

In their paper entitled *Learning to Rank Homepages For Researcher-Name Queries*, Das et al. [6] propose a learning to rank approach to identify homepages of researchers. Two scenarios are considered: query-independent, to find all researcher homepages, and query-dependent, to find the homepage for a given researcher name query. The authors study a number of features to represent the content and structure of homepages; many of these features are based on topic modeling. Evaluation is performed on two datasets: DBLP and WebKB.

Amigó et al. [2] in their paper entitled *An Evaluation Framework for Aggregated Temporal Information Extraction* propose a framework for the evaluation of temporal cross-document information extraction. The task being looked at is the following: given a particular attribute of an entity, find the most accurate temporal boundaries that can be collected from a set of relevant documents. One contribution of the paper is a model for representing temporal information that has been derived from the task requirements. Another contribution is the definition of an evaluation metric for this task grounded in a set of formal constraints.

The paper entitled *Entity Search Evaluation over Structured Web Data* by Blanco et al. [4] provides an overview of the 2011 Semantic Search Challenge. Two tasks are investigated: entity search, where each query refers to one particular entity, and list search, where queries target a group of entities that match certain criteria. The dataset for both tasks is the Billion Triple Challenge 2009 (BTC-2009) collection. Queries (a total of 50 for each task) are hand-picked from Yahoo! search logs (as well as from TrueKnowledge 'recent' queries, for the list search task). Relevance assessments are obtained using Amazon's Mechanical Turk.

2.6 Discussion

The discussion session was structured around the following points:

- Entities, what are they? Specific sub-questions: representation (i.e., how to match and display them) and identity (e.g., how to refer to them).
- Evaluation campaign desiderata.
- Entities in complex search and analysis tasks.
- Semantic (or) web (or, simply put, RDF vs. HTML5)?
- Research EOS in industry, academia, or both?

The definition question quickly converged into a quite pragmatic one, albeit a bit abstract: an entity is a 'proper noun,' 'something that is referred to.' The group's opinion was that entity identity is a very hard question to answer, and disambiguating what is being referred to remains a challenge. Trying to make the above definition more precise turns out to be difficult—for example, when 'unique in the world' was considered, a participant immediately gave 'Mickey Mouse' as an example causing problems in defining 'the' world. 'Songs by Bob Dylan' was mentioned as a counter-example to the limitation imposed in for example the TREC Entity track (assuming entities would have 'a homepage').

We identified several evaluation campaigns used in entity oriented search studies: the INEX 2007-2009 and TREC 2009-2011 Entity tracks, TAC's Knowledge Base Population track (2009-2011), WePS-1, 2 and 3, and the Semantic Search Challenge series (2010 and 2011). Participants expressed for TREC a desire to encourage work on more specific types, mentioning that the product target type seemed more challenging than finding people and organisations. The importance of reputation management in Web people search was also brought forward. A nice result at the end of the workshop has been that organisers from WePS, Semantic Search, and the TREC Entity track started to make plans for working together more closely in the future, perhaps even joining forces.

Moving on to the final discussion topic, the industry participants made it clear that they value the efforts in academia, even if these only study abstractions of the 'real' problems they have to deal with in practice. They do use the outcomes of evaluation campaigns to compare their own work to. Also, the importance of training young researchers was explicitly put forward by one industry representative, emphasizing that entity oriented work in academia teaches new researchers the right background, encourages creative thinking, and how this on its own is a great contribution even if the practical impact on 'real' search engines may turn out to be limited. When one workshop organiser then made the remark 'Then give us data!', industry participants highlighted that a lot of data cannot leave the company boundaries, but that they have by now good experience with internships—the interns often have access to real data during their research visit.

2.7 Best Paper Award

A best paper award was given out based on a secret ballot voting conducted among workshop participants. Papers that were presented as regular (20 minutes) talks were considered as candidates (6 papers in total). Each workshop participant was asked to rank the top three papers he/she considered best both in terms of content and in terms of presentation, using a three-point assessment scale. Because of the anonymous voting, people were allowed to vote for their own paper. Based on the the 24 voting forms we received, the best paper award went to Olga Vechtomova (University of Waterloo) for her paper "Unsupervised Related Entity Finding" [14]. The award came with a cash price of $300 generously sponsored by Yandex. The two runners-ups who came very close were "Cross-Domain Bootstrapping for Named Entity Recognition" [12], presented by Ang Sun, and "Entity Search Evaluation over Structured Web Data" [4], presented by Peter Mika. We congratulate to them.

3 Conclusions

The EOS program featured two excellent keynotes and a broad range of interesting academic papers, covering many different aspects of entity-oriented research. The workshop was successful in bringing people from research communities (Information Retrieval, Natural Language Processing, and Semantic Web) as well as from industry together, and offered a highly interactive environment with lively discussions throughout the whole day. We are planning a follow-up workshop next year, possibly at a Semantic Web venue, to underline our aim of building bridges between different research communities working on entity-related search problems.

Acknowledgments We would like to thank ACM and SIGIR for hosting the workshop. We are grateful for the sponsorship received from Yandex to award the best workshop paper.

We would also like to thank the members of the program committee for their efforts: Wojciech M. Barczynski (SAP Research, Germany), Indrajit Bhattacharya (Indian Institute of Science, Bangalore, India), Roi Blanco (Yahoo! Research Barcelona, Spain), Paul Buitelaar (DERI - National University of Ireland, Galway, Ireland), Wray Buntine (NICTA Canberra, Australia), Jamie Callan (Carnegie Mellon University, USA), Gianluca Demartini (L3S, Germany), Lise Getoor (University Maryland, USA), Harry Halpin (University of Edinburgh, UK), Michiel Hildebrand (VU University Amsterdam, NL), Prateek Jain (Wright State University, USA), Arnd Christian Knig (Microsoft, USA), Jisheng Liang (Microsoft Bing, USA), Peter Mika (Yahoo! Research Barcelona, Spain), Marie Francine Moens (Katholieke Universiteit Leuven, Belgium), Iadh Ounis (Univsersity of Glasgow, UK), Ralf Schenkel (Max-Planck Insitute for Computer Science, Germany), Shuming Shi (Microsoft Research Asia, China), Ian Soboroff (NIST, USA), Fabian Suchanek (INRIA, France), Duc Thanh Tran (Karlsruhe Institute of Technology, Germany), Wouter Weerkamp (University of Amsterdam, NL), and Jianhan Zhu (Open University, UK).

We extend our sincere gratitude to all the authors and presenters as well as to our invited speakers for their contributions to the material and productive discussions that formed an outstanding workshop.

References

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14. ACM, 2009.

[2] E. Amigó, J. Artiles, Q. Li, and H. Ji. An evaluation framework for aggregated temporal information extraction. In Balog et al. [3], pages 59–64.

[3] K. Balog, A. P. de Vries, P. Serdyukov, and J.-R. Wen, editors. *Proceedings of the First International Workshop on Entity-Oriented Search (EOS)*, 2011.

[4] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and T. Tran Duc. Entity search evaluation over structured web data. In Balog et al. [3], pages 65–71.

[5] S. Campinas, D. Ceccarelli, T. E. Perry, R. Delbru, K. Balog, and G. Tummarello. The Sindice-2011 dataset for entity-oriented search in the web of data. In Balog et al. [3], pages 26–32.

[6] S. Das, P. Mitra, and C. Lee Giles. Learning to rank homepages for researcher-name queries. In Balog et al. [3], pages 53–58.

[7] R. Y. Lau and W. Zhang. Semi-supervised statistical inference for business entities extraction and business relations discovery. In Balog et al. [3], pages 41–46.

[8] Q. Li and D. He. Finding support documents with a logistic regression approach. In Balog et al. [3], pages 20–25.

[9] B. Lin, K. Dela Rosa, R. Shah, and N. Agarwal. LADS: Rapid development of a learning-to-rank based related entity finding system using open advancement. In Balog et al. [3], pages 14–19.

[10] Z. Liu, Q. Lu, and J. Xu. High performance clustering for web person name disambiguation using topic capturing. In Balog et al. [3], pages 1–6.

[11] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 881–890. ACM, 2010.

[12] A. Sun and R. Grishman. Cross-domain bootstrapping for named entity recognition. In Balog et al. [3], pages 33–40.

[13] R. T.-H. Tsai and C.-H. Chou. Extracting dish names from chinese blog reviews using suffix arrays and a multi-modal CRF model. In Balog et al. [3], pages 7–13.

[14] O. Vechtomova. Unsupervised related entity finding. In Balog et al. [3], pages 47–52.