The First Joint International Workshop on Entity-oriented and Semantic Search (JIWES)

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

David Carmel
IBM Research Haifa
carmel@il.ibm.com

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

Daniel M. Herzig
Karlsruhe Institute of Technology
herzig@kit.edu

Peter Mika
Yahoo! Research, Barcelona
pmika@yahoo-inc.com

Haggai Roitman
IBM Research Haifa
haggai@il.ibm.com

Ralf Schenkel
Saarland University/MPII
schenkel@mmci.uni- saarland.de

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

Thanh Tran Duc
Karlsruhe Institute of Technology
duc.tran@kit.edu

Abstract

The First Joint International Workshop on Entity-oriented and Semantic Search (JIWES) workshop was held on Aug 16, 2012 in Portland, Oregon, USA, in conjunction with the 35th Annual International ACM SIGIR Conference (SIGIR 2012). 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, eight refereed papers divided into two technical paper sessions, and a group discussion.

1 Introduction

This First Joint International Workshop on Entity-oriented and Semantic Search (JIWES) came into existence as a result of a merger between two workshops, both of which were accepted at SIGIR 2012. One group of researchers proposed to organize the second edition of the SIGIR 2011 workshop on Entity-Oriented Search (EOS) [1]; this proposal was already a joint effort between the organizers of EOS and that of the Search and Mining Entity-Relationship Data (SMER) at CIKM 2011 [7]. Another group proposed to organize the fifth edition of Semantic Search Workshop series (SemSearch) that has previously run at the WWW conference [5]. For the sake of not (further) dividing, but rather uniting two

communities that work on very similar problems, we decided to join efforts and to organize a single joint workshop instead of two (that is the reason why this workshop ended up having an unusually large number of organizers). Our hope was that the merger would help further to bring people from different communities (IR, SW, DB, NLP, HCI, etc.) and backgrounds (both academics and industry practitioners) together, to identify and discuss emerging trends, tasks and challenges in entity-oriented and semantic search.

In general, the objective of the workshop was to encompass various tasks and approaches that go beyond the traditional bag-of-words paradigm and incorporate an explicit representation of the semantics behind information needs and relevant content. This kind of semantic search, based on concepts, entities and relations between them, has attracted attention both from industry and from the research community. Particularly, the workshop aimed to gather all works that discuss entities along three dimensions: tasks, data and interaction. Under tasks we meant entity search (search for entities or documents representing entities), relation search (search entities related to an entity), as well as more complex tasks (involving multiple entities, spatio-temporal relations inclusive, involving multiple queries). In the data dimension, we considered (web/enterprise) documents (possibly annotated with entities/relations), Linked Open Data (LOD), as well as user generated content. The interaction dimension was to provide room for research into user interaction with entities, also considering how to display results, as well as whether to aggregate over multiple entities to construct entity profiles. The workshop especially encouraged submissions on the interface of IR and other disciplines, such as the Semantic Web, Databases, Computational Linguistics, Data Mining, Machine Learning, or Human Computer Interaction. Examples of topic of interest included:

- Data acquisition and processing (crawling, storage, and indexing)
- Dealing with noisy, vague and incomplete data
- Integration of data from multiple sources
- Identification, resolution, and representation of entities (in documents and in queries)
- Retrieval and ranking
- Semantic query modeling (detecting, modeling, and understanding search intents)
- Novel entity-oriented information access tasks
- Interaction paradigms (natural language, keyword-based, and hybrid interfaces) and result representation
- Test collections and evaluation methodology
- Case studies and applications

We accepted a total of 8 papers out of 17 submissions. Each was reviewed by at least two members of the program committee, consisting of 22 researchers from academia and industry, representing a broad range of disciplines. Accepted contributions were presented either as short (15+5 min) or regular (20+5 min) oral presentations. A best contribution award was given out based on a secret ballot voting conducted among workshop participants. In addition, the JIWES program featured two invited talks by John Shafer (Microsoft Research) and Kaushik Chakrabarti (Microsoft Research). 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 in the ACM Digital library[1] and at the workshop website.[2] The latter also contains copies of the slides for most presentations.

2 Workshop Program

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

2.1 Invited Talk 1

The first invited talk was given by John Shafer, principal researcher at the Search Labs of Microsoft Research, with the title *The Lincoln Project: Building a Web-Scale Semantic Search Engine*. The Lincoln project was founded in early 2009 with the goal of building a working end-to-end experimental commerce search engine; ideas that have been proven to work here can then be integrated into Bing. The presentation started with a brief overview of the user interface and main functions of the system, illustrated with screenshots. Shafer then continued with introducing the core components of the backend: query analysis (commercial intent detection, query classification and annotation), query execution (matching and ranking products to be returned to the user), content service (fetching and returning display data to the UI), prodcast service (historical pricing and price-trend detection for a given product), and front door service (coordinating requests for the UI).

Next, Shafer discussed the query processing and execution components in detail. First, *query classification* finds the most specialized and relevant categories with probability scores. Challenges here include the followings: (1) queries can be vague and match multiple categories, (2) there are too many categories for a single multi-class classifier (6000 node taxonomy), and (3) difficult to find training data that reflects real user queries. The approach taken is to build a separate classifier for each node; training data is inferred from click logs (clicks to shopping sites are strong signals). Second, *query annotation* produces all relevant structured interpretations of a given query with scores. A few key observations here: (1) use the structured data to extract query structure, (2) annotations reinforce each other, and (3) un-annotated words can still help disambiguate. A generative model is used to score queries; model parameters are estimated using structured data statistics and via EM over web queries. Third, *query execution* finds the most relevant results for an annotated query and groups these results by category. Challenges that need to be dealt with in this area: (1) distinction between matching, ranking and selection, (2) users may not know the domain, much less the DB, and (3) must handle over-specified and ill-specified queries. The ranking function used has three main components: textual, structural, and domain relevance; structured relevance takes precedence over textual.

The presentation then continued with an overview of the product catalog that integrates three main functionalities: (1) classification of new offers and products to the taxonomy, leveraging title text, source taxonomy, and image similarity, (2) matching offers to structured product records, leveraging implicit structure in offer titles; category-specific scoring model

[1] `http://dl.acm.org/citation.cfm?id=2379307`
[2] `http://km.aifb.kit.edu/ws/jiwes2012/`

determines the most likely interpretation of structure, and (3) synthesizing products by clustering unmatched offers and extracting a structured product record from each cluster; it leverages the structured HTML pages from the merchant URLs. The underlying DBMS is basically a triple-store (with entityID, attribute name, and attribute value), with no fixed schema; an attempt is made to normalize attributes.

Shafer closed his talk with highlighting open future challenges. (1) There is more to semantics than structure; (2) Entity search is a process that can span multiple sessions and days; distinct phases require different user experience; (3) As systems become more complex, can we keep the UI intuitive? Do we need new paradigms? Explainability of results, i.e., "Why did you show me this?" (4) Ultimately, search should not be siloed; web and structured entity data support each other and search must smoothly blend web and entity content.

Publications related to the Lincoln project can be found at `http://research.microsoft.com/en-us/groups/searchlabs/pubs.aspx`.

2.2 Paper Session 1

The paper by Raviv et al. [6] entitled *A ranking framework for entity oriented search using Markov random fields* studies the problem of entity ranking for entity-oriented search. The authors suggest a new ranking approach based on a mixed Markov Random Fields model. First, each entity is represented by a document which relevance to the query is determined based on an MRF model. Second, the proximity between the query and each entity type is determined based on the semantic distance induced by minimal paths within Wikipedia's category graph. Finally, the proximity between the query terms and each entity's name is determined in two levels: locally, by considering proximity within the top-ranked documents and globally, by considering the whole collection. Experiments using the INEX 2007-2009 Entity ranking task benchmark showed promising results, demonstrating the importance of entity types and names for entity-oriented search tasks. Interestingly, the MRF dependence models (SD, FD) have not shown improvement over the independence model (ID).

In her paper, entitled *A semi-supervised approach to extracting multiword entity names from user reviews*, Vechtomova [11] described a semi-supervised approach to extracting multiword units that belong to a specific semantic class of entities. The proposed approach uses a small set of seed words representing the target class and calculates the distributional similarity between the candidate and seed words. She adapted a well-known document ranking function, BM25, to the task of calculating the similarity between vectors of context features representing seed words and candidate words, and performed a systematic comparison to a number of other distributional similarity measures. The proposed BM25-based measure proved to be competitive and showed statistically significant improvements over other measures. Finally, she introduced a method for ranking multiword units by the likelihood of belonging to the target semantic class.

In their paper *A case for semantic full-text search*, Bast et al. [3] point out the limitations of pure keyword-based search and pure ontological search. They opt for a combined approach that integrates keyword and semantic search in a single search system. The paper identifies four main challenges of this scenario: recognizing entities in the text, determining meaningful semantic units of the text, building an integrated index for entities and keywords, and designing an easy-to-use user interface. For each of these challenges, the paper presents the solution chosen for the Broccoli search engine, a semantic full-text search engine devel-

oped by the authors. The talk included an impressive demo of this engine, showing both the expressiveness of the query language and the efficiency of the underlying processing engine.[3]

2.3 Invited Talk 2

Kaushik Chakrabarti, researcher at the Data Management, Exploration and Mining Group of Microsoft Research gave the second keynote of our day, entitled *Simple Models, Lots of Data: Mining semantics about entities using Web-Scale Data*. Chakrabarti walked us through a number of entity-mining tasks that he worked on in the past years. Common to these projects is that in each case the models are relatively simple, but exploit large amounts of data, which he also refers to as the idea of "statistical semantics." The common thread across these projects is they start by assigning a weak interpretation to data elements and then amplify it by propagation through large similarity graphs.

The first project tackled the problem of mining entity synonyms, i.e., alternative ways of referring to the same entity, with the goal of finding a scalable, domain independent, yet precise approach [4]. The method he co-developed involves looking at co-clicks in query logs, that is, finding pairs of queries where the clicked results largely overlap. Click logs, however, are too sparse when it comes to the long tail. For this reason, co-clicks are complemented with a measure of similarity of the clicked documents. Last, they also check that the mined synonyms are of the same type by comparing their query context, i.e., the words appearing before or after the suggested synonyms.

Next, Chakrabarti presented their work on entity attribute discovery (finding the right attribute given a set of values) and entity augmentation (filling in missing attribute values). Here, Chakrabarti and colleagues exploit web tables (HTML tables) as a readily available information source [13]. A common approach in this setting is to consider tables from the Web that match the query table on other known attributes. Novel to the approach of Chakrabarti and colleagues is that they also consider indirectly matching tables, i.e., pairs of tables that can be indirectly matched through an intermediate table or chain of tables.

The last project he presented addressed the problem of entity linking, also known as Named Entity Disambiguation (NED), in the particular context where the target entities are of the same type (e.g., a list of company names), but lack further descriptions, specifically, do not have associated documents. In this case they exploit contextual features such as homogeneity (similarity among the contexts of identical references), co-mentions (co-occurrences within sentences), and interdependency between sentences [12].

2.4 Paper Session 2

Urbain [10] in his paper entitled *User-driven relational models for entity-relation search and extraction* presented a novel user-driven approach for integrating extracted entities and their relational dependencies for ranking sentences containing the most relevant entities and relational dependencies to a query. The retrieval process can be described as follows: (1) The user presents a natural language query; (2) The NLP engine parses the query, extracts candidate entities, relations, and textual context; (3) A relational query model is generated from the evidence the NLP engine is able to extract from the query; (4) The relation query model is used to rank sentences from the dimensional index; (5) The user can provide relevance feedback to the system. Preliminary results over the ACE 2005 newspaper data set

[3] `http://broccoli.informatik.uni-freiburg.de:6222/Broccoli/`

demonstrate the efficiency of the approach using several relatively basic retrieval models. The retrieval model capturing compatible entities and relations significantly outperforms models using entities, terms, or relations alone, or only entities and terms in combination.

In their paper entitled *Semantic preference retrieval for querying knowledge bases*, Scheel et al. [8] study the problem of personalized recommendations over structured knowledge bases. This problem is decomposed into the two tasks of inferring the user interests to construct a user profile and creating queries to retrieve recommendations. The authors propose to apply preference learning over structured entities captured by the knowledge base to infer weighted predicate-object relations as the main elements of the user model. Then, a solution is presented for automatically computing queries from these relations to retrieve recommended items.

Finally, in the paper entitled *Taxonomy-based query-dependent schemes for profile similarity measurement*, Tuarob et al. [9] present an approach to measure the similarity between two author profiles with respect to a given query. Profiles and queries are modeled as a list of weighted topics. The topics and a corresponding taxonomy are extracted from Wikipedia. Profiles are populated by taking the research interests from `ArnetMiner.org` and by matching them to topics of the taxonomy using Wikipedia Miner. Analogously, keyword queries are matched to topics. Several different schemes to compute the similarity between two profiles for a given query are discussed and "anecdotal" results of an initial evaluation are presented, along with possible applications for the approach.

2.5 Discussion

Before starting the actual discussion, we initiated a poll to have an idea about the research communities with which participants identify themselves. More than 20 people considered themselves to be part of the IR community, 15 thought about themselves as NLP researchers, and around 7 considered themselves to belong to the Data Mining, Machine Learning, or Database communities; only 3 participants named the Semantic Web among their research interests (this might not come as a surprise given that this was a workshop at SIGIR). Even though everyone seemed to have some background relevant to SIGIR, at least half of the audience expressed interest in having a separate conference or single-day venue dedicated to the topics of entity-oriented and semantic search, i.e., would travel just for that.

Much of the discussion centered around finding new challenges and tasks for the entity search and mining community. Getting more representative information needs, favoring long queries over short ones, and limiting search to a smaller, fixed set of entity types (as opposed to arbitrary types of entities) were proposed as possible action points. In general, people would like clear task definitions for which type-specific entity properties can be exploited; on the other hand, research should yield methods that generalize. While using specific verticals (such as products, locations, movies) seemed like a promising direction to go, the idea did not receive overall agreement from participants. It was generally agreed, however, that neither purely text-based nor purely ontology-based entity ranking is truly interesting and that it is much more challenging when the collection integrates both structured and unstructured information about entities.

Another main discussion point concerned entity-oriented benchmarking campaigns, like the INEX and TREC Entity Ranking tracks and the Semantic Search Challenge. These platforms are indeed popular among researchers and many build on them, albeit people often (maybe too often) choose to deviate from the original task definitions or do not use

the "official" relevance assessments. For example, many people do not mind to restrict the set of documents/entities to Wikipedia, as it provides a good mixture of structured and unstructured data and covers sufficiently many entities. Others perform additional judgments to compensate for the fact that many of their retrieved documents/entities are not judged. The reason behind this is that the assessed pools were not always of very good quality due to the limited number of participants and runs received. (Actually, that was the main reason why the INEX and TREC Entity Ranking benchmarks were stopped at some point.) Another difficulty is that some tasks (e.g., the Related Entity Finding task at TREC) have become too complex (this is informally defined as "too much work for a single PhD student"). As an effective treatment to this situation, participants suggested to lower the "cost" of participation in these benchmarks. For example, by providing baseline implementations or pre-processed input data for certain components.

2.6 Best Contribution Award

A best contribution award was given out based on a secret ballot voting conducted among workshop participants. All presented papers (both short and regular) were considered as candidates. 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. The best contribution award went to Hannah Bast et al. (University of Freiburg) for their paper *A case for semantic full-text search* [3]. The award came with a cash price of $300 generously sponsored by Yandex. We congratulate to the winner!

3 Conclusions

The JIWES 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 as well as from industry together, and offered a highly interactive environment with lively discussions throughout the whole day. We are planning the continuation of the workshop at SIGIR next year.

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), Roi Blanco (Yahoo! Research), Pablo Castells (Universidad Autnoma de Madrid), Gianluca Demartini (University of Fribourg), Michiel Hildebrand (VU University Amsterdam), Arnd Christian Knig (Microsoft Research), Oren Kurland (Technion, Israel Institute of Technology), Edgar Meij (University of Amsterdam), Einat Minkov (University of Haifa), Kavitha Srinivas (IBM Research), Martin Theobald (Max-Planck-Institut Informatik), Sivan Yogev (IBM), and Ilya Zaihrayeu (Universit degli Studi di Trento).

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] K. Balog, A. P. de Vries, P. Serdyukov, and J.-R. Wen. The First International Workshop on Entity-Oriented Search (EOS). *SIGIR Forum*, 45(2):43–50, December 2011.

[2] K. Balog, D. Carmel, A. P. de Vries, D. M. Herzig, P. Mika, H. Roitman, R. Schenkel, P. Serdyukov, and T. Tran Duc, editors. *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search*, JIWES '12, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1601-9.

[3] H. Bast, F. Bäurle, B. Buchhold, and E. Haussmann. A case for semantic full-text search. In Balog et al. [2], pages 4:1–4:3. ISBN 978-1-4503-1601-9.

[4] K. Chakrabarti, S. Chaudhuri, T. Cheng, and D. Xin. A framework for robust discovery of entity synonyms. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1384–1392, New York, NY, USA, 2012. ACM.

[5] M. Grobelnik, P. Mika, T. Tran Duc, and H. Wang, editors. *Proceedings of the 3rd International Semantic Search Workshop*, SEMSEARCH '10, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0130-5.

[6] H. Raviv, D. Carmel, and O. Kurland. A ranking framework for entity oriented search using markov random fields. In Balog et al. [2], pages 1:1–1:6. ISBN 978-1-4503-1601-9.

[7] H. Roitman, R. Schenkel, and M. Grobelnik, editors. *Proceedings of the 1st international workshop on Search and mining entity-relationship data*, SMER '11, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0957-8.

[8] C. Scheel, A. Said, and S. Albayrak. Semantic preference retrieval for querying knowledge bases. In Balog et al. [2], pages 6:1–6:6. ISBN 978-1-4503-1601-9.

[9] S. Tuarob, P. Mitra, and C. L. Giles. Taxonomy-based query-dependent schemes for profile similarity measurement. In Balog et al. [2], pages 8:1–8:6. ISBN 978-1-4503-1601-9.

[10] J. Urbain. User-driven relational models for entity-relation search and extraction. In Balog et al. [2], pages 5:1–5:6. ISBN 978-1-4503-1601-9.

[11] O. Vechtomova. A semi-supervised approach to extracting multiword entity names from user reviews. In Balog et al. [2], pages 2:1–2:6. ISBN 978-1-4503-1601-9.

[12] C. Wang, K. Chakrabarti, T. Cheng, and S. Chaudhuri. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 719–728, New York, NY, USA, 2012. ACM.

[13] M. Yakout, K. Ganjam, K. Chakrabarti, and S. Chaudhuri. Infogather: entity augmentation and attribute discovery by holistic matching with web tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 97–108, New York, NY, USA, 2012. ACM.