Towards entity search: Research roadmap

Michal Laclavík, Marek Ciglan

Institute of Informatics, Slovak Academy of Sciences,
Dúbravská cesta 9, 845 07 Bratislava
laclavik.ui@savba.sk, marek.ciglan@savba.sk

Abstract. In this paper we discuss the motivation and vision of better search, where search results are entities (things) and not documents. We describe work done in area of entity search (or semantic search) in the scientific research as well as working prototypes of big players such as Google, Facebook or Microsoft Bing. We discuss also work done at Institute of Informatics SAS in this area, possible research directions and applications, where entity search can be applied and deliver solutions even with state of the art approaches.

Keywords: entity search, semantic search, information retrieval, networks

1 Introduction

Today's prominent search engines such as Google or Bing return collection of relevant documents for the user intend represented by query, while user usually wants to receive information about entities (things). Entity search is one of prominent research topic in the Information Retrieval field. Early successful attempts such as results delivered based on Google Knowledge Graph[2] in Google search, Facebook beta search[3], or IBM Watson[4] starts to pop up. However we are still far away from reaching the goal of delivering piece of relevant information for user query instead of collection of relevant documents. Older attempts of Artificial Intelligence, Agent based Technology or Semantic Web failed to deliver concrete application results in this area. However situation changed recently with a growth of amount of available structured and unstructured information maintained with Human intervention such as Wikipedia, DBPedia[5], Freebase[6] or Linked Data[7] initiative.
Entity search is discussed deeply [1] on top Information Retrieval venues such as SIGIR 2013[8], where mainly the industry players from Bing and LinkedIn discuss new

[2] http://googleblog.blogspot.sk/2012/05/introducing-knowledge-graph-things-not.html
[3] http://tinyurl.com/GraphSearchFB
[4] http://static.usenix.org/event/lisa11/tech/slides/perrone.pdf
[5] http://dbpedia.org/About
[6] http://www.freebase.com/
[7] http://linkeddata.org/
[8] http://sigir2013.ie/

approaches for entity search. Also other talks focused on entities for enterprise search [3]. Google participated on SIGIR 2012 [2] focusing on *things, not strings* [1]. *The direction of search, at least in industry, is clear: entities are taking center stage* [1].

Research world is also active in this area, but mainly coming from semantic web[9] field, which is still tide-up in formal semantic models, but focus is being slightly shifted towards Linked Data and network structures. While in the past Question Answering[10] or Semantic Search[11] challenges were addressed mainly by researchers focusing on formal knowledge representation, nowadays many teams start to apply combinations of IR, semantic web and graph theory approaches.

2 Research Roadmap

Queries. In order to provide a satisfactory entity search system, the nature of entity oriented queries needs to be understand. First attempt at classification of ad-hoc semantic search queries was presented in [4]. Authors analyzed web search queries of a major search engine from semantic search point of view and have provided a classification of the queries. The largest part (40%) of queries were identified as *single entity queries*, where user is searching for a single specific entity. Other identified classes of queries included *type queries* (12%) where the goal is to retrieve set of entities (e.g. boroughs of nyc), *attribute queries* searching for attribute of a specific entity (e.g. Michael Jordan height), *relation queries* aiming at retrieval of relations among specified entities. The rest of the queries were classified as *other queries* (36%), comprising previous query classes modified by different context decorators, such as spatial or temporal (e.g. 'nightlife in Barcelona', or the 'theater plays Bratislava tomorrow'). The largest part of existing body of work on entity search is centered around the *single entity queries*, aimed at a retrieval of a single entity. Other classes of entity search queries are more complex and there is a large space for research still open. Large number of *type, attribute, relation* and *other* entity queries are indirect, that is, user is searching for entities/data related to a known entity (e.g. 'kurt vonnegut short stories'), where the target entities are one or more hops away from the specified entity or specified type in the underlying knowledge graph. (Example of multi-hop query might be 'spouse of Bill Clintons daughter', where the target entity is two hops away from starting entity 'Bill Clinton'). Semantically, complex entity search queries requiring several hops from starting entity or type are very similar to the task of question answering over structured data. The only difference is the form of queries: in ad-hoc entity search, keywords queries are expected whereas in question answering a grammatically well constructed question sentences are expected. In order to advance entity search and question answering, a high quality query segmentation and syntactic analysis is needed.

[9] http://en.wikipedia.org/wiki/Semantic_Web
[10] http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/
[11] http://semsearch.yahoo.com/

Additional challenges lies in the query understanding. During the query processing, entity retrieval system needs to distinguish distinct query segments between basic entities, type specifications, relation specifications and temporal clues. Important problem is, given an input query, to identify query class, query intent, target entity type. Multilinguality is also an important issue.

Knowledge bases creation. The most widely used open domain knowledge bases used for entity search tasks are DBPedia, Yago and Freebase. The major part of all three is data provided by Wikipedia. DBPedia extracts structured data about entities from Wikipedia pages infoboxes, Freebase (which is the base of Google's knowledge graph) relies on user entered / confirmed information. Although the scope of Wikipedia, and therefore also the knowledge bases derived from Wikipedia, is quite large, the extension of the knowledge bases beyond Wikipedia content is desirable.

Standard approach to the extension of knowledge base is to integrate other well structured and well described data sources. A real world example is the inclusion of IMDB and MusicBrainz data sets in Freebase. This approach of knowledge graph extension ensures high data quality. The main research problem is linking entities from different sources and identifying duplicate entities in the resulting entity graph.

Another approach to knowledge base creation, which is also a dominant research challenge, is to automatically extract entities, facts and relations directly from documents in natural languages. Here, the importance of high quality open domain information extraction methods is eminent. The knowledge base construction using open domain information extraction is extremely hard, when a reasonable data quality is expected. First attempts of such approaches are ReadTheWeb[12], KnowItAll[13] or already mentioned IBM Watson project.

3 Entity Search at IISAS

In this chapter we describe our effort done so far in area of entity search. After analysis of our past research in IKT research group[14], we believe most of our research fit well with the idea of entity search. In our research we use unique mix of following approaches which are quite important for entity search: Information Retrieval; Text Analysis; Semantic Web; Network and Graph Analysis; and Large scale data processing – Big Data.

In the field of Named Entity Recognition (NER), which is important task of identifying entities in unstructured texts, we have developed methods Ontea [9] and Annotowatch [8]. Ontea uses simple pattern based methods or connect existing NER tools but creates unique tree structures of annotations and network data when connecting multiple documents. Annotowatch is combination of NER tools using machine learning approach. With Annotowatch we have participate in MSM 2013 Information Extraction challenge[15], where we have finished at 2nd place, missing first

[12] http://rtw.ml.cmu.edu/rtw/
[13] https://github.com/knowitall
[14] http://ikt.ui.sav.sk/
[15] http://oak.dcs.shef.ac.uk/msm2013/challenge.html

place by 1%[16]. The challenge took place at MSM 2013 workshop, which was a part of WWW 2013 conference. The goal was to recognize entities such as people, organizations, locations and miscellaneous (e.g. names of the movies) in tweets.
We have applied NER approaches together with IR techniques (text parsing, indexing, full-text or faceted search) [10] on web data or large data sets using MapReduce paradigm on Hadoop [11]. This was enhanced also by other entity search features based on text graphs [12] created from unstructured data, where we can discover relations among entities using approach and tool called gSemSearch [7]. Interactive method of search, navigation and exploration of entities relations is used in the gSemSearch tool. Users can delete, merge, adjust entities while searching and navigating. Underling lightweight semantic network data are updated and method can use data from user interaction to learn and deliver better results.
We have also participated in the already mentioned Semantic Search Challenge 2011 in the track[17] on entity list search, where our SemSets [6] approach was the winning solution. SemSets reuse Wikipedia as knowledgebase, combining IR and network analysis approaches, where both text and structured data are used in the form of interlinked Wikipedia concepts (network) and concept properties.
Community detection in networks is also relevant task for entity search. In this field we have developed SCCD algorithm [13] with near linear complexity. Only algorithms with linear complexity can be applied on large networks, which are starting to pop-up in many applications such as telecommunications, banking or social network analysis. In addition, in our KDD 2013 paper [14], we have shown that current near linear time algorithms does not detect communities well on several types of real world networks. We have proposed modification on edges re-weighting in order to achieve better results. Recently we have also start to work on Query categorization, which is also entity search problem. We are processing and reusing available sources such as Wikipedia, DBPedia or Freebase and combining IR and network analysis methods for this task, where we map Wikipedia concepts (entities) on user query and then categorize identified Wikipedia pages (concepts) using combination of text categorization approaches (TF-IDF, LSI, n-gram matching) and reusing Wikipedia and DBPedia user defined categories.
While several, well funded, research groups are working on entity search problem, we believe we can deliver valuable results in this area, based on combination and tuning of the developed methods mentioned above, especially when focusing on concrete application which are discussed in next section.

4 Applications

While entity search is related mainly to todays's web search and its improvement, there are other tasks in the domains of online advertisement, business intelligence, multilingual text analysis or information security where entity search approaches can be applied.

[16] http://ikt.ui.sav.sk/index.php?n=Main.IEChallenge2013
[17] http://semsearch.yahoo.com/results.php#

Entity search approaches can be used in query categorization (QC) task. QC or query classification, is a task of identifying user search intent based on submitted query and mapping it to predefined categories. Query categorization can be used for several applications; the most notable one is probably the online advertising. In the online advertising domain, the QC can help to capture user interests and improve user modeling, which in turn can lead to increase in precision of user targeting with ads relevant to their interests and needs. Clearly in QC you are mapping one short string (query) on other short string (category) and you need to understand what entities query and category represents. We are developing competitive solution which achieve results comparable with state of the art approaches. Method is based on mapping both queries and categories on Wikipedia concepts (representing entities) as we have described briefly in previous section. Results are available online[18]. Similarly as QC, also the task of keyword expansion is the task very relevant for online advertisement. When having seed keywords such as *TV, 42, SONY*, we need to know what entity it represents, otherwise we would expand *TV* for *TV series* for example.

Similar approaches as for QC can be applied on more wide task of short text categorization and even on multilingual short text categorization since Wikipedia or Freebase data are available in many languages and language versions are interconnected over topics.

In the field of Business Intelligence (BI), the prominent task for entity search is an enterprise search [3], where it is demand for better search over both structured and unstructured data and entities such as customers, providers, products, services or business transactions are of primary importance. In this field we have worked with email analysis [7], relations search and enterprise search based on entities [15]. In today's BI it is also growing demand on Big Data analysis of social media, where NER tools such as Anotowatch [8] can be applied. In addition BI approaches need to do more and more with graph and network data processing, where we have developed fast graph database[19] or effective community detection algorithms [13, 14]. These can be applied on tasks like fraud detection or churn prediction, especially in the businesses, where network data are natively present such as telecom, social network providers or any business with online data coming from user interaction or online transactions.

In addition to mentioned domains entity search can be applied also in the field of information security, where we can look on the tasks of information leakage prevention or Data Leak Prevention (DLP[20]) as a problem of entity identification and definition of rules based on entity approach instead of deeply technical customization of DLP system by nearly manual definition of rules.

[18] http://ikt.ui.sav.sk/research/QC/

[19] http://ups.savba.sk/~marek/sgdb.html

[20] http://en.wikipedia.org/wiki/Data_loss_prevention_software

5 Conclusion

In is clear that we cannot reach the goal of complex Entity Search with our research in near future, however we describe our approach and achievement in the field. The approach is unique in combination of techniques coming from field of information retrieval, semantic web and network data analysis. We also describe applications, where current entity search approaches can bring innovative solutions with improved results compare to state of the art approaches.

Acknowledgment: This work is supported by projects: VEGA 2/0184/10, CLAN APVV-0809-11 and VENIS FP7-284984. It is also supported by Magnetic Inc.

References

1. Tunkelang, D. LinkedIn. Search: What's Cooking in the Lab. August 05, 2013 http://tinyurl.com/SIGIR13lab
2. Hofmann, T.: Towards Summarizing the Web of Entities. SIGIR 2012, 2012, http://tinyurl.com/WebOfEntitiesSIGIR12
3. Frost & Sullivan. Global enterprise search market to reach US$4.68bn by 2019, 25.1.2013, http://www.siliconrepublic.com/strategy/item/31182-global-enterprise-search-ma
4. Pound, J., Mika, P. Zaragoza., H. 2010. Ad-hoc object retrieval in the web of data. In Proceedings of WWW '10. ACM, NY, USA, 771-780.
5. Neumayer, R., Balog, K., Nørvåg, K. 2012. On the modeling of entities for ad-hoc entity search in the web of data. In ECIR'12,Springer-Verlag, Berlin, Heidelberg, 133-145.
6. Ciglan, M., Nørvåg, K, Hluchý, L. 2012. The SemSets model for ad-hoc semantic list search. In Proceedings of WWW '12. ACM, NY, USA, 131-140.
7. Laclavík, M., et al.. 2012. Emails as graph: relation discovery in email archive. In Proceedings of WWW '12 Companion. ACM, NY, USA, 841-846.
8. Dlugolinský, Š., Krammer, P., Ciglan, M., Laclavík M. 2013. MSM2013 IE Challenge: Annotowatch. In Concept Extraction Challenge at MSM workshop at WWW'13, 2013, Vol-1019, pages 21-26, 2013, http://ceur-ws.org/Vol-1019/paper_21.pdf
9. Laclavík, M., et al.: Email Analysis and Information Extraction for Enterprise Benefit; In Computing and informatics, 2011, vol. 30, no. 1, p. 57-87. ISSN 1335-9150
10. Dlugolinský, Š., Šeleng, M. Laclavík, M. Hluchý L.: Distributed Web-scale Infrastructure for Crawling, Indexing and Search with Semantic Support. In Computer Science Journal, Vol 13 No.4, pages 5-19, 2012, DOI: 10.7494/csci.2012.13.4.5
11. Laclavik, M., Seleng, M. Hluchy, .L: Towards Large Scale Semantic Annotation Built on MapReduce Architecture. In ICCS 2008 Proc., Part III, LNCS 5103, pp. 331-338, 2008.
12. Laclavík, M.: Discovering Entity Relations in Semantic Text Graphs. Habilitation thesis submitted for the Associate Professor degree, submitted in February 2013
13. Ciglan, M., Nørvåg K.: Fast detection of size-constrained communities in large networks, proceedings of WISE'10, LNCS Volume 6488/2010
14. Ciglan, M. Laclavík, M. Nørvåg, K.: On community detection in real-world networks and the importance of degree assortativity. In KDD '13. ACM, NY, USA, 1007-1015.
15. Laclavík, M., et al.: Lightweight Semantic Approach for Enterprise Search and Interoperability. In Interop-Vlab.It Workshop, CEUR WS, Vol-915, pages 35-42, 2012