ENTITY SEARCH IN TEXT GRAPHS

Michal LACLAVÍK, Štefan DLUGOLINSKÝ, Marek CIGLAN
Martin ŠELENG, Ladislav HLUCHÝ

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

Abstract. Entity Search is becoming a popular alternative for full text search. Recently Google released its entity search based on confirmed, human-generated data such as Wikipedia. In spite of these developments, the task of entity discovery, search, or relation search in unstructured text remains a major challenge in the fields of information retrieval and information extraction. This paper briefly summarizes research addressing that challenge, focusing specifically on entity relation discovery. This is achieved by processing unstructured text using simple information extraction methods, building lightweight semantic graphs and reusing them for entity relation discovery by applying algorithms from Graph Theory. An important part is also user interaction with semantic graphs, which can significantly improve information extraction results and entity relation search.

1. Introduction

Entity Search is becoming a popular alternative for full text search. Recently Google released its entity search [8] based on confirmed, human-generated data such as Wikipedia and Freebase, and Facebook is experimenting with graph search over its user generated context. New types of question answering systems such as IBM Watson, based on structured and unstructured data [9], are being developed, but the task of entity discovery, search, or relation search in unstructured text still remains a major challenge in information retrieval and information extraction fields.

2. Semantic Text Graphs and Entity Relation Search

Graphs or networks often appear as a natural form of data representation in many applications like social tetworks, call networks of communicating people, Internet, Wikipedia, LinkedData or emails. The analysis of email communication allows the extraction of social networks with links to people, organizations, locations, topics or time information. Social networks included in email archives are becoming increasingly valuable assets in organizations, enterprises, and communities, though to date they have been little explored.

Unstructured text is still the most common medium for information sharing and communication. While it is available on the web, in emails, or within new social media like Facebook, Twitter or LinkedIn, it is also present in enterprises' analytical data like document repositories or even database text fields. All of these web, media communication, and organizational resources preserve a large part of their knowledge in unstructured textual form. In addition, such data is connected with graph/network data through web links, communication links, transactions, or social links and tags (lightweight semantics) in social media and is shared among many users and resources.

It has been proved on Web 2.0 (or social web) that the lightweight semantics (tags) and social networks (graph data) give additional value to knowledge sharing, reuse, recommendation and analytics. The text can be transformed to trees or graph/network structures [5], which have a similar property to the information networks mentioned above. In our work we have examined properties of such networks [1, 4].

Searching, analyzing, accessing, and visualizing information and knowledge hidden in such network structures are becoming increasingly important tasks in the area of data analytics [6], but different algorithms must be used for unstructured data processing such as text. We have tried to create network structures from unstructured text [4] data similar to those of structured data [2].

Email communication is unique in this respect because it connects social networks (communication) with information networks, which can be extracted from text. We believe that email communication and its links to other organizational as well as public resources (e.g. LinkedData) can be a valuable source of information and knowledge for knowledge management, business intelligence, better enterprise, and personal email search. The future of email [7] is in interconnecting email with other resources, services (like social networks or collaboration tools), or data and entities which are present in email. This was also the main motivation and drive for our work, but we have discovered that the approach could be applied to any unstructured text data, and not only to email communication.

In our work [4] we have also tried to experiment with structured [2] and unstructured data [1, 4]. We examined same algorithms for entity relation discovery on the information network from ACM publications [2] as well as other unstructured text. Text based information networks were extracted using Ontea [5] information extraction tool and then we searched [1] and interacted [3] with these networks in different applications [4] using gSemSearch tool utilizing optimized spreading activation algorithm [4].

3. Conclusions

In this paper we have briefly summarized R&D work done on entity relation discovery from unstructured text, where text was transformed into an information network with similar properties to other social or information networks. We have conducted experiments on several networks and graphs extracted from diverse text resources as well as on structured data such as ACM publications. We have shown and evaluated an interactive method of relation discovery available in the gSemSearch prototype.

We believe the information networks, such as the graph in our experiments [4], can help to interconnect unstructured and structured data such as text documents (web pages, emails, documents) with the structured data (hyper text networks, social networks,

LinkedData). When structured and unstructured data possess similar properties of small world information networks, we can apply common algorithms for search and exploration of entities and their relations.

Acknowledgement: This work was partially supported by TRADICE APVV-0208-10.

References

to other papers publishing the results that are summarized here

[1] Laclavík,M., Dlugolinský, Š., Šeleng, M., Ciglan, M. Hluchý, L.: Emails as graph: relation discovery in email archive. In *Proceedings of the 21st international conference companion on World Wide Web (WWW '12 Companion)*. ACM, New York, NY, USA, 841-846. 2012, http://doi.acm.org/10.1145/2187980.2188210

[2] Mojžiš, J., Laclavík, M.: Navigácia v zjednodušenom LinkedData grafe. In *7th Workshop on Intelligent and Knowledge Oriented Technologies*. - Bratislava : Nakladateľstvo STU, 2012, p. 15-18, ISBN 978-80-227-3812-5, 2012

[3] Laclavík, M.: Improving entity and relation discovery by user interaction with semantic graphs. In *7th Workshop on Intelligent and Knowledge Oriented Technologies*: P. 161-164. - Bratislava: Nakladateľstvo STU, 2012. ISBN 978-80-227-3812-5.

[4] Laclavík, M.: Discovering Entity Relations in Semantic Text Graphs. *Habilitation thesis* submitted for the Associate Professor degree, 2013

[5] Laclavík,M., Dlugolinský, Š., Šeleng, M., Kvassay, M., Gatial, E., Balogh, Z., Hluchý: Email analysis and information extraction for enterprise benefit. In *Computing and informatics*, 2011, vol. 30, no. 1, p. 57-87. ISSN 0232-0274.

Other references

[6] Aggarwal, C. C. (Ed.): *Social network data analytics*. Springer. 502p, SBN 978-1-4419-8462-3, 2011

[7] Fauscette, M.: The Future of Email Is Social. White Paper; IBM IDC report; ftp://ftp.lotus.com/pub/lotusweb/232546_IDC_Future_of_Mail_is_Social.pdf, 2012

[8] Ulanoff L.: Google Knowledge Graph Could Change Search Forever. http://mashable.com/2012/02/13/google-knowledge-graphchange-search/, 2012

[9] Ferrucci, D. A.: IBM's Watson/DeepQA. SIGARCH *Comput. Archit. News* 39, 3 (June 2011), DOI=10.1145/2024723.2019525 http://doi.acm.org/10.1145/2024723.2019525