Graph-Based Entity-Oriented Search: Imitating the Human Process of Seeking and Cross Referencing Information

José Devezas
INESC TEC and FEUP
jld@fe.up.pt

Sérgio Nunes
INESC TEC and FEUP
ssn@fe.up.pt

In an information society, people expect to find answers to their questions quickly and with little effort. Sometimes, these answers are locked within textual documents, which often require a manual analysis, after being retrieved from the web using search engines. At FEUP InfoLab, we are researching graph-based models to index combined data (text and knowledge), with the goal of improving entity-oriented search effectiveness.

We live in a world where an evergrowing web is able to deliver a large body of knowledge to virtually anyone with an internet connection. At the same time, the high availability of content has morphed human information seeking behavior [1] — people expect to find answers to their questions quickly and with little effort. The quality of the answers is frequently tied with the search engine ability to understand query intent, using information from curated knowledge bases to provide direct answers, based on identified entities and relations, alongside the traditional textual document results. Search engines are greatly dependent on the inverted index, inspired by the back-of-the-book index of printed manuscripts, to rank documents with matching keywords, but they are also increasingly dependent on knowledge bases. While there are automatic methods for knowledge base construction, most search engines still depend on manual curation for this task. On one side, there is the error associated with automatic knowledge base construction and, on the other side, there is the time constraint and domain expertise of manually curating a knowledge base. We propose an intermediate solution based on a novel graph-based indexing structure, with the goal of combining the power of the inverted index with any available and trustworthy information, through established knowledge bases.

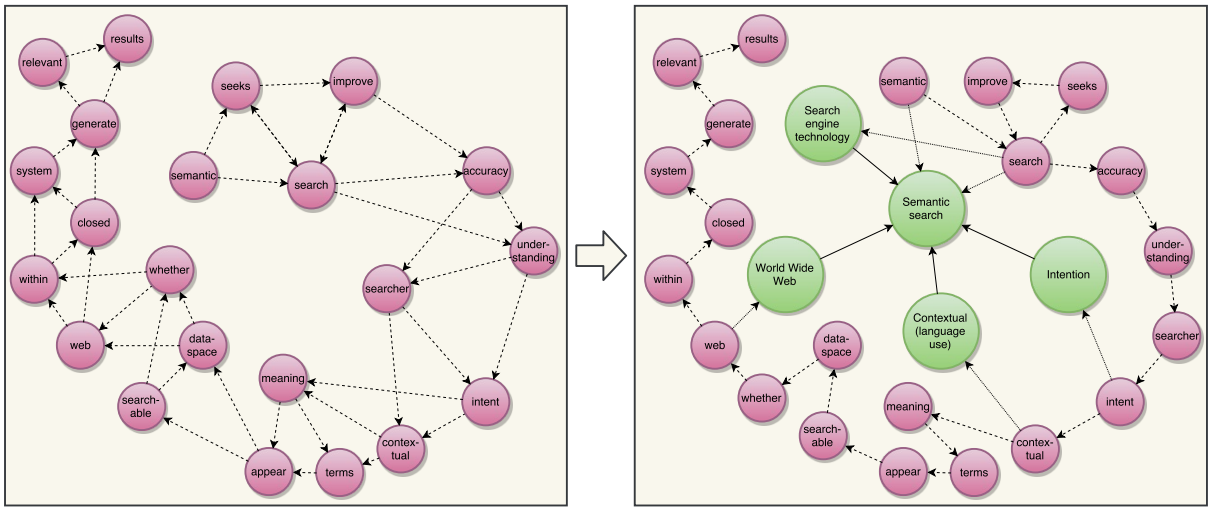

Figure 1: From the graph-of-word to the graph-of-entity; a representation of an example text document (left and right) and associated knowledge (right only).

Our current research is focused on finding novel ways of integrating text and information through a graph, without losing the properties of terms in an inverted index, and entities and relations in a

triplestore. The goal is to be able to retain the characteristics of text and knowledge, while combining them, through a unified representation, to improve retrieval. The hypothesis is that by establishing potentially weak links between text and entities, we might be able to support and imitate the human process of seeking and cross referencing information: knowledge-supported keyword-based search. As humans, each of us compiles knowledge through multiple sources (e.g., the world, books, other people), establishing relations between entities and continuously correcting for consistency, based on concurrent information and its trustworthiness. When we have an information need, we either ask someone or consume some sort of media (e.g., a book, a video, an audio lesson) to obtain answers. Let us take, for instance, the task of searching within a book to solve an information need. Specifically, let us assume a back-of-the-book index search for a given set of terms (analogous to the traditional keyword query). Let us then assume that we skip to a page indicated by one of those terms and read a textual passage. How do we determine whether or not it is relevant for our information need? While we already know it contains one of the terms we seek, we must use existing knowledge (ours or otherwise) to assess the relevance of the text.

There is an obvious connection between text and knowledge that isn't being captured by existing search technologies. While there is a clear and growing integration of text-based search and entity-based decorations (e.g., an infobox about the most relevant entity, or a list of entities for the given entity type expressed by the query), the inverted index still exists separately from the knowledge base and vice-versa. Our goal is to explore the opportunity of improving retrieval effectiveness based on a seamless integration of text and knowledge through a common data model, while proposing one or several unifying ranking functions that only decide based on the maximum available information.

We have based our work on the graph-of-word [2], a document representation and retrieval model that defies the term independence assumption of the traditional bag-of-words approach used in inverted indexing. Figure 1 (left) shows the graph-of-word representation for an example document (the first sentence of the Wikipedia page for "Semantic Search"). Each term links to the following two terms, as a way to use indegree to establish the context of a word. We propose the graph-of-entity (Figure 1; right), where we link each term (in pink) only to the following term (dashed line), but also include entity nodes (in green), basic *contained_in* edges between term and entity nodes (dotted line; weak relation based on substring matching), and edges between entity nodes (solid line), representing relations between entities in a knowledge base or, in this case, indirectly based on the hyperlinks for the Wikipedia article. The objective is to unify text and knowledge retrieval as a combined task, in order to use structured and unstructured data to provide better answers for the information needs of the users.

Links

[L1] `http://infolab.fe.up.pt`

[L2] `http://ant.fe.up.pt`

References

[1] S A Knight and Amanda Spink. Toward a web search information behavior model. In *Web search*, pages 209–234. Springer, 2008.

[2] François Rousseau and Michalis Vazirgiannis. Graph-of-word and tw-idf: new approach to ad hoc ir. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 59–68. ACM, 2013.

Please contact:

José Devezas
INESC TEC and Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
jld@fe.up.pt

Sérgio Nunes
INESC TEC and Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto PORTUGAL
ssn@fe.up.pt

José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).