Connecting Knowledge for A New Kind of Search

Andias Wira-Alam
GESIS – Leibniz Institute for the Social Sciences
Unter Sachsenhausen 6-8
Cologne, Germany
Email: andias.wira-alam@gesis.org

Abstract
With the ever increasing information existing on the World Wide Web, the importance of developing comprehensive methods and applications to make use of this continuously growing information has intensified. In this research work, our main aim is to build a new kind of search engine that discovers "hidden" connections and generates paths between knowledge, as well as attemps to give the big picture of complex story developments.

Research Statement
In the last four years, the intention to build modern search engines has raised dramatically. Wolfram Alpha[1], started in 2009, introduced a new way of finding and presenting information. Google Knowledge Graph[2] offers structured information gathered from different knowledge sources. In addition to the retrieved documents, an *infobox* containing some extra information related to search queries, especially those that matches named-entities in their database, e.g. for persons or places having Wikipedia entries (see Figure 1). For the part "People also search for" in the infobox, other prominent persons are also provided. However, the users might ask e.g. why "Isaac Newton" or "Thomas Edison" are also searched by most users? Or whether "Einstein" and "Isaac Newton" share common interests in physics, if yes in which research fields? Or even they were againts each other? Here the context of the connections is missing and therefore not easy to be figured out. Providing connections and context will help users to understand such relationships in an efficient way.

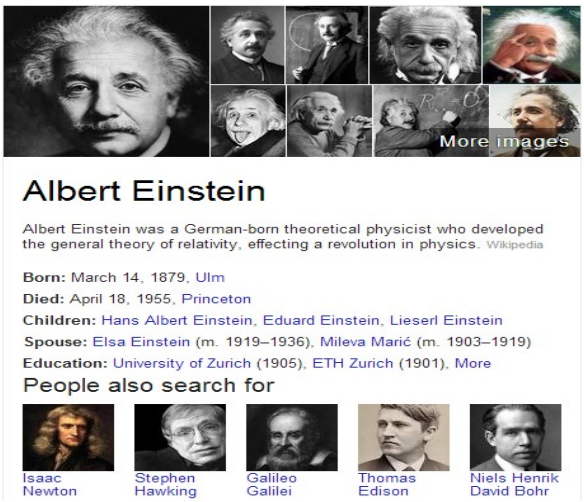

Figure 1. Infobox provided by Google for the query "Einstein"

However, both products are proprietary and therefore they are not reproducible for further investigations. Recently, Shahaf et al. [3,4,5] introduced methods with theoretical guarantees for creating chains automatically showing the coherence between two connected documents or articles. The methods embrace e.g. a notion of coherence with respect to documents influence, coverage, and random walks with implicit links generated by important words. The first step we plan to do is to implement those methods using real data from the Web, e.g. Wikipedia, news articles, and scientific corpora. After this implementation step, we evaluate the methods that generate the chains and determine the coherence in a matter of time. According to our initial experiments, the methods are not efficient, therefore we plan to do some feasible improvements. The first improvement we plan to do is to make the graph algorithms to generate the chains more efficient. However, we could not precisely test the methods since our current hardware resource is quite limited. Moreover, we also plan to improve the results by giving users text summaries instead of set of documents as

the current methods do. For this purpose we need to test some algorithms, e.g. topics model and text summarization[9], and evaluate the outcomes. Naturally, we are carrying on many possible improvements throughout the experiments.
In conjunction with the aforementioned plans, we have already done some initial experiments with some data from English Wikipedia and news articles. We imported the articles and links provided by Wikipedia as dump files. We use Debian/GNU Linux running on dual Intel R CoreTM2 CPUs with 2TB Harddisk and 3GB RAM to store the data. Besides, we also collected news articles from Spiegel magazine (online and print edition) with over 400 thousands news articles. A working prototype has been developed as proof of concept, which currently focused on the links within Wikipedia articles (see Figure 2). Some previous supporting this work have also been published [6,7,8,9].

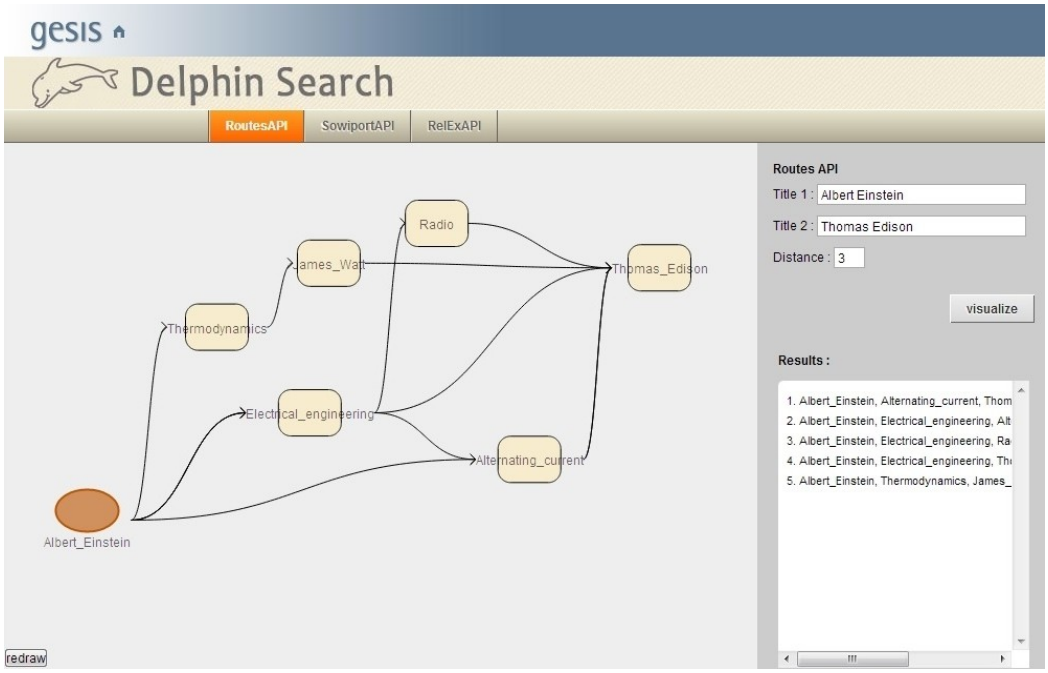

Figure 2. A working prototype of the desired application, showing possible paths between Albert Einstein and Thomas Edison[1]

References

[1] Wolfram Alpha. http://www.wolframalpha.com/
[2] Google Knowledge Graph. http://www.google.com/insidesearch/features/search/knowledge.html
[3] Dafna Shahaf, Carlos Guestrin: Connecting the Dots between News Articles. KDD 2010.
[4] Dafna Shahaf, Carlos Guestrin, Eric Horvitz: Trains of thought: generating information maps. WWW 2012: 899-908
[5] Dafna Shahaf, Carlos Guestrin, Eric Horvitz: Metro maps of science. KDD 2012: 1122-1130
[6] Brigitte Mathiak, Víctor Manuel Martínez Peña, Andias Wira-Alam: Extracting term relationships from Wikipedia. Lecture notes in business information processing, 140, Berlin: Springer, S. 267-280.
[7] Andias Wira-Alam, Farag Saad, Peter Muschke: Query expansion based on conceptual and contextual term relationships in Wikipedia. In: Hobohm, Hans-Christoph (Hrsg.): ISI 2013, Potsdam
[8] Andias Wira-Alam, Brigitte Mathiak: Mining Wikipedia's Snippets Graph - First Step to Build a New Knowledge Base. In: Völker, J., Paulheim, H., Lehmann, J., and Niepert, M. (eds.) Proceedings of the First International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data. pp. 43–48.
[9] Andias Wira-Alam, Matthäus Zloch: Improving Web Search Results with Explanation-Aware Snippets: An Experimental Study. WEBIST 2013. Aachen
[10] Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein: Improving Search Result Summaries By Using Searcher Behavior Data. SIGIR 2013. Dublin, Ireland.

1 This prototyp can be accessed online at http://multiweb.gesis.org/delphinSearch/ on tab "RoutesAPI".