How NAGA Uncoils: Searching with Entities and Relations

Gjergji Kasneci
Max-Planck-Institut
Saarbruecken / Germany
kasneci@mpii.mpg.de

Fabian M. Suchanek
Max-Planck-Institut
Saarbruecken / Germany
suchanek@mpii.mpg.de

Maya Ramanath
Max-Planck-Institut
Saarbruecken / Germany
ramanath@mpii.mpg.de

Gerhard Weikum
Max-Planck-Institut
Saarbruecken / Germany
weikum@mpii.mpg.de

ABSTRACT

Current keyword-oriented search engines for the World Wide Web do not allow specifying the *semantics* of queries. We address this limitation with NAGA[1], a new semantic search engine. NAGA builds on a large semantic knowledge base of binary relationships (facts) derived from the Web. NAGA provides a simple, yet expressive query language to query this knowledge base. The results are then ranked with an intuitive scoring mechanism. We show the effectiveness and utility of NAGA by comparing its output with that of Google on some interesting queries.

Categories and Subject Descriptors:
H.3.3 [Information Search and Retrieval]: General

General Terms: Graph Queries, Knowledge Base, Semantic Search

Keywords: Search, Relation, Entities, Ranking

1. INTRODUCTION

The World Wide Web is the world's largest knowledge base but we are far from exploiting its full potential. In order to effectively exploit it, we need to extract and structure the information it makes available and provide expressive and efficient ways of querying it. Current keyword-oriented search engines merely provide "best-effort" heuristics to find relevant needles in this humongous haystack.

As a concrete example, suppose we want to find out which other physicists were born in the same year as Max Planck. First, it is close to impossible to formulate this query in terms of keywords. Second, the answer to this question is probably distributed across multiple pages, so that no state-of-the-art search engine will e able to find it.

There are systems that provide graph-oriented keyword search. However, they focus only on implicit semantic relations such as foreign key references among database tuples (e.g. [4]) or references among XML elements (e.g. [3, 2]). Other approaches (e.g. [7]) rely on manual semantic tagging of Web data while our work aims at a completely automatized system. Some approaches (e.g. [1]) use automated information extraction, but do not reconcile the collected facts into a unified and consistent knowledge base and provide only standard querying capabilities. The ones that do provide unified and consistent models (based on OWL or SPARQL) usually lack the means to express uncertainty, which is crucial in the context of automated knowledge extraction and representation.

In this paper, we present NAGA, a new semantic search engine which addresses these problems in an intuitive and comprehensive way. NAGA builds on a graph-oriented knowledge base, which consists of facts extracted from the Web with certain confidence values. NAGA's expressive query language can be used to formulate precise queries, enabling the user to provide detailed information about his or her information need. The query results are then ranked according to a scoring mechanism that takes their certainty values into account.

2. DATA MODEL

NAGA's data model is a graph, in which nodes are labeled with *entities* (e.g. `Max Planck`) and edges are labeled with relationships (e.g. BORNINYEAR). Each edge, together with its end nodes, represents a fact, e.g. <`Max Planck`, BORNINYEAR, `1858`> or <`Max Planck`, TYPE, `physicist`>. Since these facts are derived from Web pages using possibly unreliable Information Extraction techniques, we attach a *certainty value* to each fact.

Formally, our data model is a directed, labeled multigraph (V, E, L_v, L_e), where V is a set of nodes, $E \subseteq V \times V$ is a multi-set of edges, L_v is a set of node labels and L_e is a set of edge labels. Each node $v \in V$ has a label $l(v) \in L_v$ and each edge $e \in E$ has a label $l(e) \in L_e$. We compute the certainty value $c(e) \in [0, 1]$ as:

$$c(e) = \sum_{i=1}^{n} C(e, P_i)T(P_i)$$

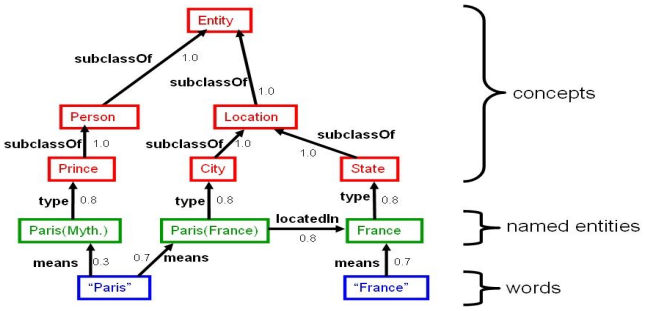
where P_i denotes one of the n pages from which the fact corresponding to e was derived. The trust value $T(P_i)$ represents the authority of page P_i and can be computed by PageRank or similar algorithms. We assume $\sum_i T(P_i) = 1$. $C(e, P_i)$ is the confidence with which the fact corresponding to e was extracted from page P_i. Thus, the certainty value for e accumulates trust and extraction quality values across all pages in which the corresponding fact was found.

NAGA's knowledge base is a projection of YAGO [6]. It contains about 1 million entities and 6 million facts, partly

[1]In the mythologies of Hinduism and Buddhism, NAGA is a huge snake. Here, it stands for the size and diversity of the Web.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

extracted by LEILA [5]. A sample of this knowledge-base is shown below:



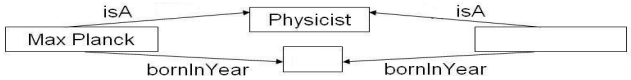
3. QUERY AND ANSWER MODEL

In the spirit of the example query in the introduction we present a taxonomy of queries supported by NAGA. Basically, a NAGA query is a directed graph $G(V, E)$ where V is a set of possibly labeled nodes and E a set of possibly labeled edges.

EVIDENCE QUERIES (EQ) An evidence query is a connected directed graph the nodes and labels of which are labeled.

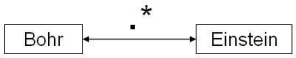
In this case, the user wants to know whether there is any evidence for a certain hypothesis.

DISCOVERY QUERIES (DQ) A discovery query is a connected directed graph the nodes and edges of which may be unlabeled.



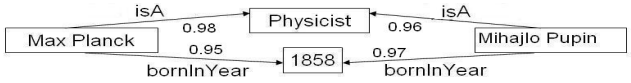
Here, the user wants to discover pieces of missing information such as entities or relations, represented by unlabeled nodes or edges respectively. The above query is the query example from the introduction.

RELATEDNESS QUERIES (RQ) A relatedness query is a connected directed graph the nodes and edges of which may be unlabeled and at least one of the edges is labeled with a regular expression over relationship labels.



In this setting, the user is interested in the broad relation between pieces of information. Obviously, it holds **EQ** $\subseteq$ **DQ** $\subseteq$ **EQ**.

The answer to a query is a subgraph A of the knowledge graph that matches the query. For example, NAGA could answer the above discovery query as follows:



The numbers on the edges of an answer graph represent the certainties of the facts. In order to compute the overall certainty of an answer, the certainties of the edges have to be accumulated. We think of the certainty value as the probability that a fact is correct. Since the facts have been extracted independently, the probability that the complete answer A is correct is just the product of the certainties of the edges:

$$c(A) = \prod_{e \text{ edge in } A} c(e)$$

In case there are multiple answers to a query, we return the top-k answers ranked by their certainty.

4. PRELIMINARY RESULTS

NAGA is implemented in Java on top of an Oracle data base. The query engine translates the user queries into SQL and performs graph searches to solve relatedness queries. We conducted several preliminary experiments to assess the quality of NAGA's answers. Below, we present some sample queries to illustrate the difference between NAGA and Google:

Query	Google	NAGA
When was BBC News established?	Answer found on following link.	1922.
Which other physicists were born in the same year as Albert Einstein?	Fails. Gives lots of links to Einstein's biography.	von Laue, Pfund, Burton, and several others.
What is the connection between Einstein and Bohr?	Links to the debate between Bohr and Einstein on quantum theory.	Both are scientists. There are Moon craters and asteroid belts named after them. Tom Cruise connects them by being a vegetarian (as Einstein) and by being born in 1962 (when Bohr died).
What is the connection between Indira Gandhi and Margaret Thatcher?	Links of events involving the two women	Both are female heads of government, both attended Oxford, both were politicians in English-speaking countries (India and UK, respectively)

5. CONCLUSION AND OUTLOOK

We presented NAGA, a new semantic search engine with a query language that can express complex queries by means of graph structures and regular expressions over relations. For future work, we plan to exploit logical inferences in the knowledge graph and to predict not only the correctness, but also the interestingness of answers. NAGA can be tried out online at: `http://www.mpii.mpg.de/~suchanek/naga`.

7. REFERENCES

[1] M. Cafarella, C. Re, D. Suciu, and O. Etzioni. Structured querying of web text data: A technical challenge. In *CIDR*, 2007.

[2] Sara Cohen, Jonathan Mamou, Yaron Kanza, and Yehoshua Sagiv. Xsearch: A semantic search engine for xml. In *VLDB*, pages 45–56, 2003.

[3] J. Graupmann, R. Schenkel, and G. Weikum. The spheresearch engine for unified ranked retrieval of heterogeneous XML and web documents. In *VLDB*, 2005.

[4] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, and H. Karambelkar. Bidirectional expansion for keyword search on graph databases. In *VLDB*, 2005.

[5] Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. Combining linguistic and statistical analysis to extract relations from web documents. In *KDD*, 2006.

[6] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, New York, NY, USA, 2007. ACM Press.

[7] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. Semantic wikipedia. In *WWW*, 2006.