NAGA: Searching and Ranking Knowledge
Authors: Gjergji Kasneci Fabian M. Suchannek Georgiana Irfim Maya Ramanath Gerhard Weikum

Language-model-based Ranking for Queries on RDF-Graph
Authors: Gjergji Kasneci Fabian M. Suchannek Georgiana Irfim Maya Ramanath Gerhard Weikum

Presented by: Shasha(Amy) Liu
Motivation

Example queries

- Which politicians are also scientists?
- Which gods do the Maya and the Greeks have in common?
Motivation

We've also added our first Greek Legend: The Labors Of Heracles. More myths and legends coming ... Did you know the Maya Gods are obsessed with football? ...
www.godchecker.com/- 27k - Cached - Similar pages

Mayan Mythology: Gods, Goddesses, Spirits, Legends of the Maya
Mayan Mythology: Meet the Gods of Meso-america. ... The Gods Of Mayan Mythology.
Mayan Gods The current Top Ten: 1st: CHAC 2nd: AH-PUCH 3rd: IXCHEL ...
www.godchecker.com/pantheon/mayan-mythology.php - 24k - Cached - Similar pages
More results from www.godchecker.com »

Pictures Of Greek Gods And Goddesses | Pictures Of Greek Symbols
CW As the Greek gods goddesses to maintain. Chingachgook gazed at his word for ... Aye I would have announced it Aileen put some pictures of mayan gods on ...
8.ppslooqax.com/m - Similar pages

Thousands of NAMES OF GODS, GODDESSES, DEMIGODS, MONSTERS, SPIRITS ...
List of Greek Gods & Goddesses with Roman names in parentheses .... MAYAN GODS & GODDESSES Several gods who played significant roles in the Post classic ...
www.lowchensaustralia.com/names/gods.htm - 64k - Cached - Similar pages

Greek Gods, River Styx, River Acheron, Hades, and Death
This brought about his mythological relationship to the Greek god Hades. Because the mythology of the gods is more known than the actual religious roles of ...
www.river-styx.net/greek-gods-hades.htm - 42k - Cached - Similar pages

Mayan Gods Deity Depicting Kings Related Articles
They are united by their common faith in Islam, which is the second largest .... The stories of ancient Greek mythology tell about Greek gods and heroes, ...

Ancient Artifacts: Egyptian Statues & Reliefs - Greek Gods ...
We are proud of our line of Egyptian statues & reliefs, Greek Gods & Goddesses, ... Greeks, the Orient Hindus, Buddhist, Aztecs, and Mayan cultures. ...
www.ancientartifactstoday.com/- 37k - Cached - Similar pages
Motivation

- Keyword queries are too weak to express advanced user intentions such as
  - concepts,
  - entity properties
  - relationships between entities

- Data is not knowledge.
  - Data extraction and organization needed
Outline

- Framework
  - Data model
  - Query language
  - Ranking model

- Evaluation
  - Setting
  - Metrics
  - Results
Framework (Data model)

- Entity-relationship (ER) graph
  - Node label : entity
  - Edge label : relation
  - Edge weight : relation “strength”
Framework (Data model)

- **Entity-relationship (ER) graph**
  - Node label: entity
  - Edge label: relation
  - Edge weight: relation “strength”

- **Fact**
  - Represented by an edge

- **Evidence pages for a fact** $f$
  - Web pages from which $f$ was derived

- **Computation of fact confidence** (i.e. edge weights): $\text{conf} (f) = \frac{1}{n_f} \sum_{i=1}^{n_f} \text{ExtrConf} (f, P_i) \cdot \text{Trust}(P_i)$
Framework (Query language)

- $R$: set of relationship labels
- $\text{RegEx}(R)$: set of regular expressions over $R$-labels
- $E$: set of entity labels
- $V$: set of variables

Definition (fact template)

A fact template is a triple $<e_1 \ r \ e_2>$ where $e_1, e_2 \in E \cup V$ and $r \in \text{RegEx}(R) \cup V$.

Examples:

- Liu $\xrightarrow{\text{givenNameOf} \ | \ \text{familyNameOf}}$ $x$
- Albert Einstein $\xrightarrow{x}$ Mileva Maric
Framework (Query language)

- **Definition (NAGA query)**
  
  A *NAGA query* is a connected directed graph in which each edge represents a fact template.

- **Examples**

  1) Which physicist was born in the same year as Max Planck?

  ![Diagram](Diagram1.png)

  2) Which politician is also a scientist?

  ![Diagram](Diagram2.png)

  3) Which scientist are called Liu?

  ![Diagram](Diagram3.png)

  4) Which mountain is located in Africa?

  ![Diagram](Diagram4.png)

  5) What connects Einstein and Bohr?

  ![Diagram](Diagram5.png)
Framework (Query language)

- **Definition (NAGA answer)**
  
  A *NAGA answer* is a subgraph of the underlying ER graph that matches the query graph.

- **Examples**

1) Which physicist was born in the same year as Max Planck?

2) Which mountain is located in Africa?

3) What connects Einstein and Bohr?
Framework (Ranking model)

- **Question**
  How to rank multiple matches to the same query?

- **Ranking desiderata**

  **Confidence**
  Correct answers
  - Certainty of IE
  - Trust/Authority of source

  **Informativeness**
  Prominent results preferred
  - Frequency of facts

  **Compactness**
  Prefer “tightly” connected answers
  - Size of the answer graph

  "Max Planck born in Kiel"
  `bornIn (Max_Planck, Kiel)` (Source: Wikipedia)

  “They believe Elvis hides on Mars”
  `livesIn (Elvis_Presley, Mars)` (Source: The One and Only King's Blog)
Framework (Ranking model)

- **Question**
  How to rank multiple matches to the same query?

- **Ranking desiderata**
  
  **Confidence**
  Correct answers
  - Certainty of IE
  - Trust/Authority of source

  **Informativeness**
  Prominent results preferred
  - Frequency of facts

  **Compactness**
  Prefer “tightly” connected answers
  - Size of the answer graph

NAGA exploits *language models* for ranking.
Framework (Ranking model)

Statistical Language Models for Document IR


- each doc has LM: generative prob. distr. with parameters $\theta$
- query q viewed as sample
- estimate likelihood that q is sample of LM of doc d
- rank by descending likelihoods (best „explanation“ of q)

$$s(d, q) = P[q \mid d] = P[q \mid \theta] = P[q_1...q_m \mid \theta] \approx \prod_i P[q_i \mid \theta]$$

MLE: sparseness

$$s(d, q) = P[q \mid \theta] = \lambda P[q \mid d] + (1 - \lambda) P[q]$$

mixture model

Background model (smoothing)
Framework (Ranking model)

- Scoring answers
  
  Query \( q \) with templates \( q_1q_2 \ldots q_n \), e.g. Albert Einstein \( \rightarrow \) $x$
  
  Given \( g \) with facts \( g_1g_2 \ldots g_n \), e.g. Albert Einstein \( \rightarrow \) Physicist

  We use *generative mixture models* to compute \( P[q | g] \)

  \[
  s(g,q) = P[q | g] = \prod_{i=1}^{n} \left( (1 - \alpha) \cdot P[q_i | g_i] + \alpha \cdot P[q_i] \right)
  \]

\[
\beta \cdot P_{\text{conf}}[q_i | g_i] + (1 - \beta) \cdot P_{\text{inform}}[q_i | g_i]
\]

based on IE accuracy and authority analysis

\[
P(\text{Physicist} | \text{Albert Einstein}, \text{isA}) = \frac{P(\text{Albert Einstein}, \text{isA}, \text{Physicist})}{P(\text{Albert Einstein}, \text{isA})} = \frac{P(\text{Albert Einstein}, \text{isA}, \text{Physicist})}{\sum_\ast P(\text{Albert Einstein}, \text{isA}, \ast)}
\]

estimated using knowledge base graph structure

estimated by correlation statistics
Framework (Ranking model)

- Estimating Confidence

The maximum likelihood estimator for confidence is:

\[ P_{\text{conf}}(q_i | g) = \prod_{f \in \text{match}(q_i, g)} P(f \text{ holds}) \]

where \( P(f \text{ holds}) \) is estimated by \( c(f) \). The likelihood of that sequence being true is the product of the confidences of the single facts, assuming that the facts are independent.

Confidence value

\[ c(f) = \frac{1}{n} \sum_{i=1}^{n} \text{acc}(f, p_i) \cdot \text{tr}(p_i) \]

provided by the extraction mechanism, similar to PageRank.
Framework (Ranking model)

- Estimating Informativeness

Consider

Consider

Possible results

Possible results

NAGA Ranking (Informativeness)

\[
P(\text{Physicist} \mid \text{Albert Einstein, isA}) = \frac{P(\text{Albert Einstein, isA, Physicist})}{\sum_{\ast} P(\text{Albert Einstein, isA, } \ast)}
\]

\[
\approx \frac{\#\text{GoogleHits(Albert Einstein, Physicist)}}{\#\text{GoogleHits(Albert Einstein)}} > \frac{\#\text{GoogleHits(Albert Einstein, Vegetarian)}}{\#\text{GoogleHits(Albert Einstein)}}
\]

\[
\approx P(\text{Vegetarian} \mid \text{Albert Einstein, isA}) = \frac{P(\text{Albert Einstein, isA, Vegetarian})}{\sum_{\ast} P(\text{Albert Einstein, isA, } \ast)}
\]
Framework (Ranking model)

- **Estimating Informativeness**

Consider:

```
Albert Einstein isA $x
```

Possible results:

```
Albert Einstein isA Physicist
Albert Einstein isA Vegetarian
```

**BANKS Ranking** (Bhalotia et al. ICDE 2002)

- Relies only on underlying graph structure
- Importance of an entity is proportional to its degree

⇒ Vegetarian more important than Physicist
Framework (Ranking model)

- **Background model**

- $P(q_i)$, which plays the role of giving different weights to different fact templates in the query. This is similar in spirit to the idf style weights for weighting different query terms in traditional LMs.

  - For example, consider the query $Q$ with two fact templates $q_1 = \$y \text{bornIn} \text{Ulm}$ and $q_2 = \$y \text{isA scientist}.$

  - If there are many people born in Ulm, but there are only few scientists overall, this suggests giving a higher weight to $q_2.$

Traditionally, the more important condition is the more specific one – the one that is expected to have fewer matches, i.e., higher idf.
Framework (Ranking model)

- Estimating Compactness

  The more facts in an answer graph, the lower its likelihood and thus its compactness.

  *Eg: for the query Margaret Thatcher connect Indra Gandhi: the answer graph stating that they are both prime ministers, is more compact than the answer that they are both prime-ministers of English-speaking countries.*
Evaluation (Setting)

- Knowledge graph YAGO (Suchanek et al. WWW 2007)
  - 16 Million facts

- 85 NAGA queries
  - 55 queries from TREC 2005/2006
  - 12 queries from the work on SphereSearch (Graupmann et al. VLDB 2005)
  - We provided 18 regular expression queries
Evaluation (Setting)

- The queries were issued to

  - Google,
  - Yahoo! Answers,
  - START (http://start.csail.mit.edu/),
  - NAGA (Banks scoring)
    - relies only on the structure of the underlying graph. 
      (see Bhalotia et al. ICDE 2002)
  - NAGA (NAGA scoring)
Evaluation (Setting)

- The queries were issued to
  - Google,
  - Yahoo! Answers,
  - START (http://start.csail.mit.edu/),
  - NAGA (Banks scoring)
    - relies only on the structure of the underlying graph.
      (see Bhalotia et al. ICDE 2002)
  - NAGA (NAGA scoring)

- top-10 answers assessed by 20 human judges as relevant (2), less relevant (1), and irrelevant (0).
# Evaluation (Setting)

## Benchmark

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Question with NAGA translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC</td>
<td>When was Shakespeare born?</td>
</tr>
<tr>
<td></td>
<td>Shakespeare bornInYear $x</td>
</tr>
<tr>
<td></td>
<td>In what country is Luxor?</td>
</tr>
<tr>
<td></td>
<td>Luxor locatedIn $x</td>
</tr>
<tr>
<td></td>
<td>$x isA country</td>
</tr>
<tr>
<td>SphereSearch</td>
<td>In which movies did a governor act?</td>
</tr>
<tr>
<td></td>
<td>$y isA governor</td>
</tr>
<tr>
<td></td>
<td>$y actedIn $z</td>
</tr>
<tr>
<td></td>
<td>$z isA movie</td>
</tr>
<tr>
<td>OWN</td>
<td>What was discovered in the 20th century?</td>
</tr>
<tr>
<td></td>
<td>$x discoveredInYear $y</td>
</tr>
<tr>
<td></td>
<td>$y after 1900</td>
</tr>
<tr>
<td></td>
<td>$y before 2000</td>
</tr>
<tr>
<td></td>
<td>Who produced or directed the movie</td>
</tr>
<tr>
<td></td>
<td>”Around the World in 80 Days”?</td>
</tr>
</tbody>
</table>
|           | $x produced|directed\
|           | Around_the_World_in_80_Days    |
|           | What do Albert Einstein and Niels Bohr have in common? |
|           | Albert_Einstein connect Niels_Bohr |
**Evaluation (Metrics & Results)**

- **NDCG (normalized discounted cumulative gain)**
  - rewards result lists in which relevant results are ranked higher than less relevant ones
  - Useful when comparing result lists of different lengths

- **P@1**
  - to measure how satisfied the user was on average with the first answer of the search engine

<table>
<thead>
<tr>
<th>Benchmark</th>
<th># Q</th>
<th># A</th>
<th>Metrics</th>
<th>Google</th>
<th>Yahoo! Answers</th>
<th>START</th>
<th>BANKS scoring</th>
<th>NAGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC</td>
<td>55</td>
<td>1098</td>
<td>NDCG P@1</td>
<td>75.88%</td>
<td>26.15%</td>
<td>75.38%</td>
<td>87.93%</td>
<td>92.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67.81%</td>
<td>17.20%</td>
<td>73.23%</td>
<td>69.54%</td>
<td>84.40</td>
</tr>
<tr>
<td>SphereSearch</td>
<td>12</td>
<td>343</td>
<td>NDCG P@1</td>
<td>38.22%</td>
<td>17.23%</td>
<td>2.87%</td>
<td>88.82%</td>
<td>91.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19.38%</td>
<td>6.15%</td>
<td>2.87%</td>
<td>84.28%</td>
<td>84.94</td>
</tr>
<tr>
<td>OWN</td>
<td>18</td>
<td>418</td>
<td>NDCG P@1</td>
<td>54.09%</td>
<td>17.98%</td>
<td>13.35%</td>
<td>85.59%</td>
<td>91.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27.95%</td>
<td>6.57%</td>
<td>13.57%</td>
<td>76.54%</td>
<td>86.56</td>
</tr>
</tbody>
</table>
Shortcoming

- NAGA can rank only exact matches to a given query
- Ranking is helpful only for the too-many-answers case but not for the too-few-answers problem

- ?Any improvement → Language-model-based Ranking for Queries on RDF-Graph
Knowledge Graph

\[ G = <V, A, l_V, l_A, L> \]

- L: a set of labels
- V: a set of nodes
- A \subseteq V \times V \times L \text{ a set of labeled arcs}
- l_V : V \rightarrow L, an injective function that returns the label of a node
- l_A : A \rightarrow L, an injective function that returns the label of an arc such that l_A ((u, v, l)) = l for any (u, v, l) belongs to A
- A knowledge Graph G can be represented as a set of RDF triples \( T(G) = \{t_1, \ldots, t_{|A|}\} \)
Framework in Brief(2)

- A key – augmented knowledge graph G
  - Derived by enriching the knowledge graph with a function \( KG: A \rightarrow 2\text{KEY} \) assigning a finite set of keywords to an arc of \( G \)

- Witness count \( c(t) \)
  - Indicates the number of times the triple was seen and extracted from the corpus and gives a measure of importance

- Accuracy
  - Each extracted triple could be associated with a confidence value reflecting the accuracy of the employed extraction method and authenticity and authority of the data sources
Framework in Brief(3)

- Purely Structured Graph Queries
  - == NAGA
  - E.g.

- Keyword-augmented Structured Queries
  - Allows keyword to be associated
  - E.g.

- Relaxed Queries
  - Allows for approximation matching of queries
  - -- alleviates the problem of “too few results”
  - E.g.
Ranking Model in Brief

- Ranks results based on Kullback-Leibler divergence with respect to the query model
- Different from traditional
  - No notion of a document in the setting – large graph of facts from which sub-graphs can be constructed
  - Queries are made up with triples patterns, while results are made up of triples
Result Comparison in Brief

<table>
<thead>
<tr>
<th>Purely structured queries with relaxation</th>
<th>OWN</th>
<th>WOR</th>
<th>BANKS</th>
<th>NAGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>OWN</td>
<td>WOR</td>
<td>BANKS</td>
<td>NAGA</td>
</tr>
<tr>
<td>IMDB</td>
<td>0.880</td>
<td>0.751</td>
<td>0.777</td>
<td>0.798</td>
</tr>
<tr>
<td>LT</td>
<td>0.876</td>
<td>0.787</td>
<td>0.721</td>
<td>0.869</td>
</tr>
<tr>
<td>Keyword-augmented queries with relaxation</td>
<td>OWN</td>
<td>WOR</td>
<td>BANKS</td>
<td>NAGA</td>
</tr>
<tr>
<td>Dataset</td>
<td>OWN</td>
<td>WOR</td>
<td>BANKS</td>
<td>NAGA</td>
</tr>
<tr>
<td>IMDB</td>
<td>0.884</td>
<td>0.722</td>
<td>0.782</td>
<td>0.776</td>
</tr>
<tr>
<td>LT</td>
<td>0.853</td>
<td>0.835</td>
<td>0.690</td>
<td>0.782</td>
</tr>
</tbody>
</table>

Table 12: Avg. NDCG for all evaluation queries