NAGA: Searching and Ranking Knowledge

Authors: Gjergji Kasneci Fabian M. Suchannek Georgiana Irfim Maya Ramanath Gerhard Weikum

Languange-model-based Rnaking 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?

Which politicians are also scientists

Search

Advanced Search
Preferences

Web Results **1** - **10** of about **562,000** for **Which politicians are also scientists**.

DFG - Press Release No. 8, 2008 - Politics has taken Notice of Science

14 Feb 2008 **...** He emphasised that he was very pleased about how clearly the voice of the scientific community had been heeded by the **politicians**. He **also** **...**
www.dfg.de/en/news/press_releases/2008/press_release_2008_08.html - 18k - Cached - Similar pages

Political science - Wikipedia, the free encyclopedia

Political **scientists** may serve as advisers to specific **politicians**, **...** The antecedents of Western politics can **also** trace their roots back even earlier **...**
en.wikipedia.org/wiki/Political_science - 52k - Cached - Similar pages

Top **Scientists** Want Research Free From Politics - CommonDreams.org

Scientists have rarely if ever been involved in implementing policy related to their discoveries. That is the domain of the **politicians**, who are bought and **...**
www.commondreams.org/archive/2008/02/15/7085/ - 68k - Cached - Similar pages

The Political Graveyard: Christian **Scientist Politicians**

Christian **Scientist**. Still living as of 2003. See **also**: congressional biography. **....** for a specific individual, try the alphabetical index of **politicians**. **...**
politicalgraveyard.com/group/christian-**scientist**.html - 32k - Cached - Similar pages

Climate Science & Politics

As a **scientist**, you should teach **politicians** that in science "consensus" is not **...** I am ready for this debate and most of the other climate skeptics **also**. **...**
climatepatrol.wordpress.com/ - 47k - Cached - Similar pages

Science vs. politics gets down and dirty - USATODAY.com

Scientists and **politicians** have disagreed throughout history, of course, **...** we must **also** be willing when necessary to reject the wrong ways," Bush said **...**
www.usatoday.com/tech/science/2007-08-05-science-politics_N.htm - 59k - Cached - Similar pages

Religion vs. science vs. politics - Cosmic Log - msnbc.com

Druyan **also** has been working on a totally new TV series that would serve as a **.....** If **Scientists** and Theologians cannot reconcile the two, **Politicians** **...**
cosmiclog.msnbc.msn.com/archive/2007/12/20/528690.aspx - 84k - Cached - Similar pages

Mayan gods Greek gods common Search Advanced Search Preferences

Web Results **1** - **10** of about **459,000** for **Mayan gods Greek gods common**.

Godchecker.com - Your Guide To The **Gods**. Mythology with a twist!
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.**god**checker.com/ - 27k - Cached - Similar pages

Mayan Mythology : **Gods**, Goddesses, Spirits, Legends of the **Maya**
Maya 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.**god**checker.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, **...**
www.encyclocentral.com/20336-Related-**Mayan_Gods**_Deity_Depicting_Kings.html - 59k - Cached - Similar pages

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

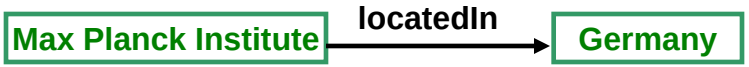
- Framework
  - Data model
  - Query language
  - Ranking model
- Evaluation
  - Setting
  - Metrics
  - Results

Framework (Data model)

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

Framework (Data model)

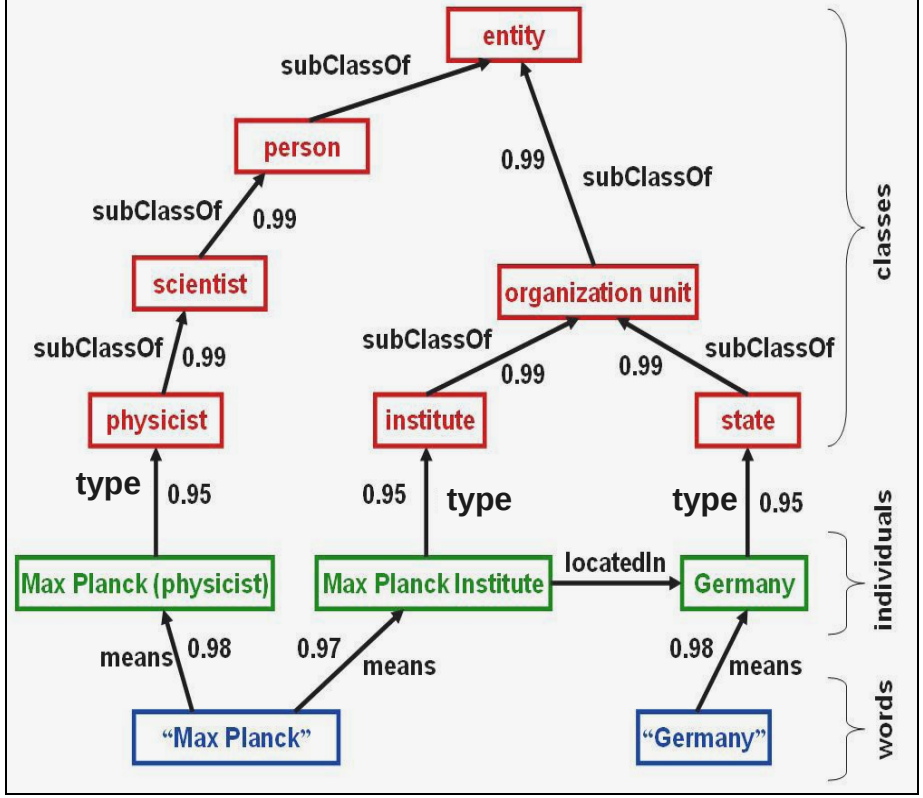
- Entity-relationship (ER) graph
  - Node label : entity
  - Edge label : relation
  - Edge weight : relation "strength"
- Fact
  - Represented by an edge



- Evidence pages for a fact *f*
  - Web pages from which *f* was derived
- Computation of fact confidence (i.e. edge weights) :

$$conf(f) = \frac{1}{n_f} \sum_{i=1}^{n_f} ExtrConf(f, P_i) \cdot Trust(P_i)$$

Excerpt from YAGO: Suchanek et al. WWW 2007



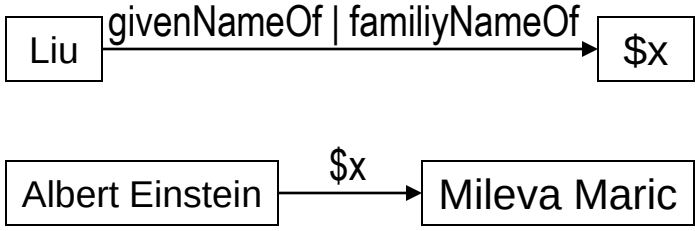
Framework (Query language)

- *R* : set of relationship labels
- *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 $r \in RegEx(R) \cup V$.

Examples:



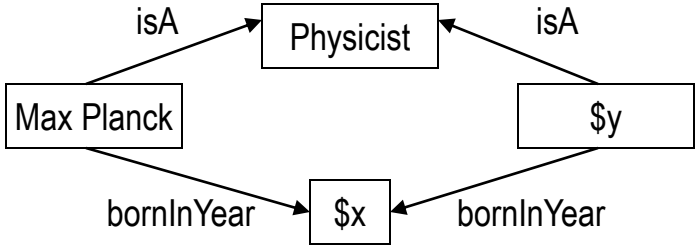
Framework (Query language)

- **Definition (NAGA query)**

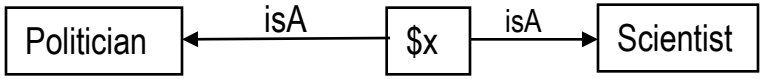
  A *NAGA query* is a connected directed graph in which each edge represents a fact template.

- **Examples**

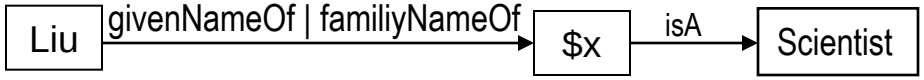
1) Which physicist was born in the same year as Max Planck?



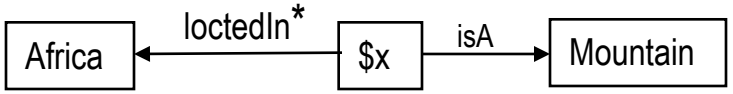
2) Which politician is also a scientist?



3) Which scientist are called Liu?



4) Which mountain is located in Africa?



5) What connects Einstein and Bohr?

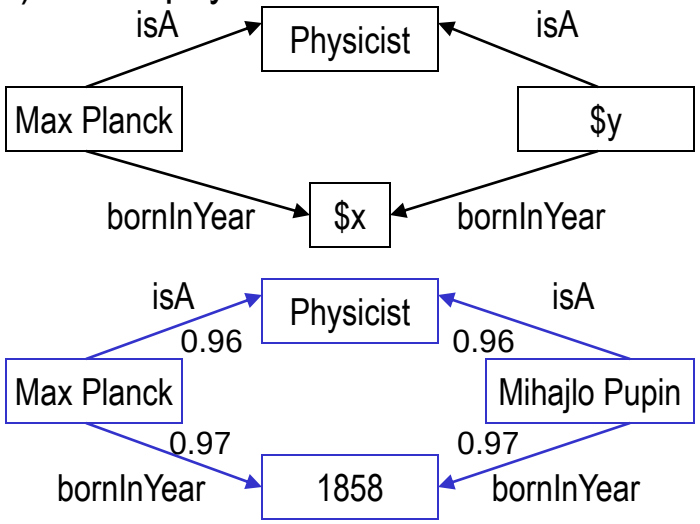
Framework (Query language)

- **Definition (NAGA answer)**

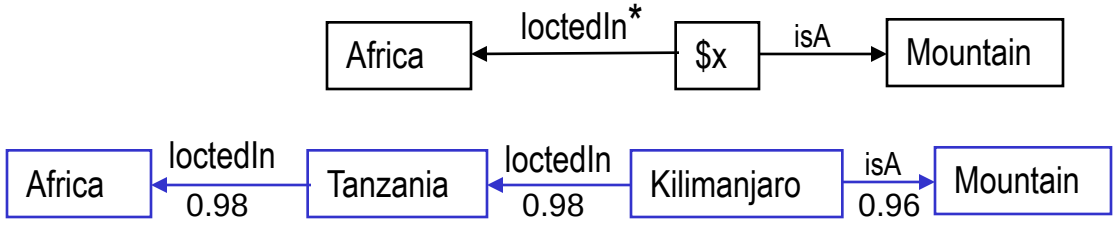
  A *NAGA answer* is a subgraph of the underlying ER graph that matches the query graph.

- **Examples**

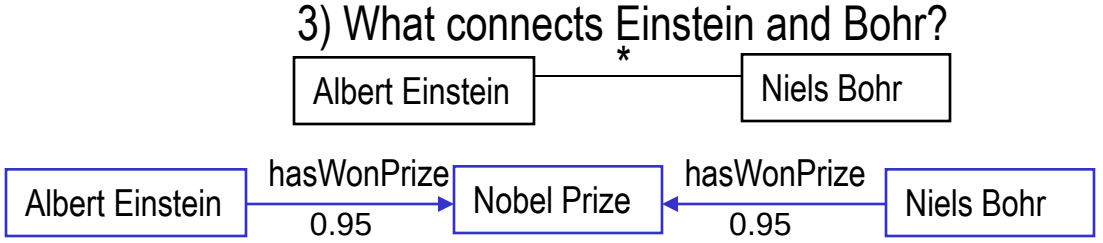
1) Which physicist was born in the same year as Max Planck?



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

"**Max Planck born in Kiel**"

bornIn (Max_Planck, Kiel) (Source: Wikipedia)

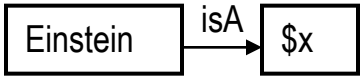
"**They believe Elvis hides on Mars**"

livesIn (Elvis_Presley, Mars) (Source: The One and Only King's Blog)

Informativeness

prominent results preferred

- Frequency of facts



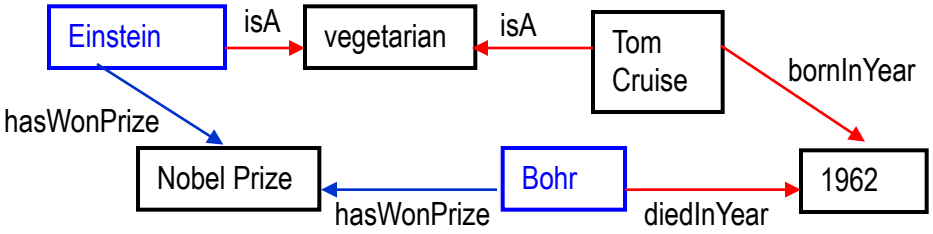
Einstein isa scientist

Einstein isa vegetarian

Compactness

Prefer "tightly" connected answers

- Size of the answer graph



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

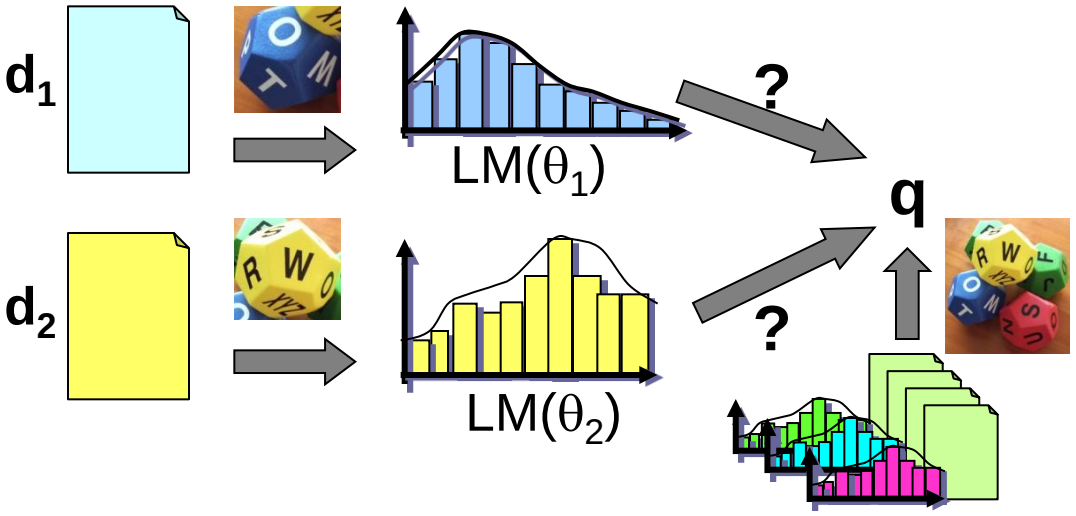
- Size of the answer graph

NAGA exploits ***language models*** for ranking

Framework (Ranking model)

Statistical Language Models for Document IR

[Maron/Kuhns 1960, Ponte/Croft 1998, Lafferty/Zhai 2001]



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

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

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 $q_1 q_2 \ldots q_n$, e.g. Albert Einstein —isA→ $x

  Given g with facts $g_1 g_2 \ldots g_n$, e.g. Albert Einstein —isA→ Physicist

  We use *generative mixture models* to compute $P[q \mid g]$

using generative mixture model

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

estimated using knowledge base graph structure

background model

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

based on IE accuracy and authority analysis

$$P(Physicist \mid Albert\ \ Einstein, isA) = \frac{P(Albert\ \ Einstein, isA, Physicist)}{P(Albert\ \ Einstein, isA)} = \frac{P(Albert\ \ Einstein, isA, Physicist)}{\sum_{*} P(Albert\ \ Einstein, isA, *)}$$

estimated by correlation statistics

Framework (Ranking model)

- **Estimating Confidence**

- The maximum likelihood estimator for confidence is:

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

- where P(f 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.

Provided by the extraction mechanism

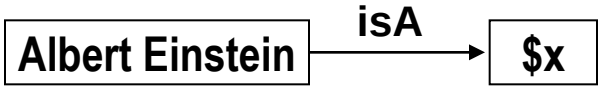
confidence value $c(f) = \frac{1}{n}\sum_{i=1}^{n} acc(f, p_i) \bullet tr(p_i)$

similar to PageRank

Framework (Ranking model)

- **Estimating Informativeness**

Consider



Possible results

NAGA Ranking (Informativeness)

$$P(Physicist \mid Albert\ \ Einstein,\ isA) = \frac{P(Albert\ \ Einstein,\ isA,\ Physicist)}{\sum_{*} P(Albert\ \ Einstein,\ isA,\ *)}$$

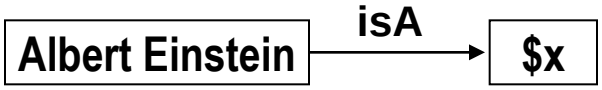
$$\approx \frac{\#GoogleHits(Albert\ \ Einstein,\ Physicist)}{\#GoogleHits(Albert\ \ Einstein)} > \frac{\#GoogleHits(Albert\ \ Einstein,\ Vegetarian)}{\#GoogleHits(Albert\ \ Einstein)}$$

$$\approx P(Vegetarian \mid Albert\ \ Einstein,\ isA) = \frac{P(Albert\ \ Einstein,\ isA,\ Vegetarian)}{\sum_{*} P(Albert\ \ Einstein,\ isA,\ *)}$$

Framework (Ranking model)

- **Estimating Informativeness**

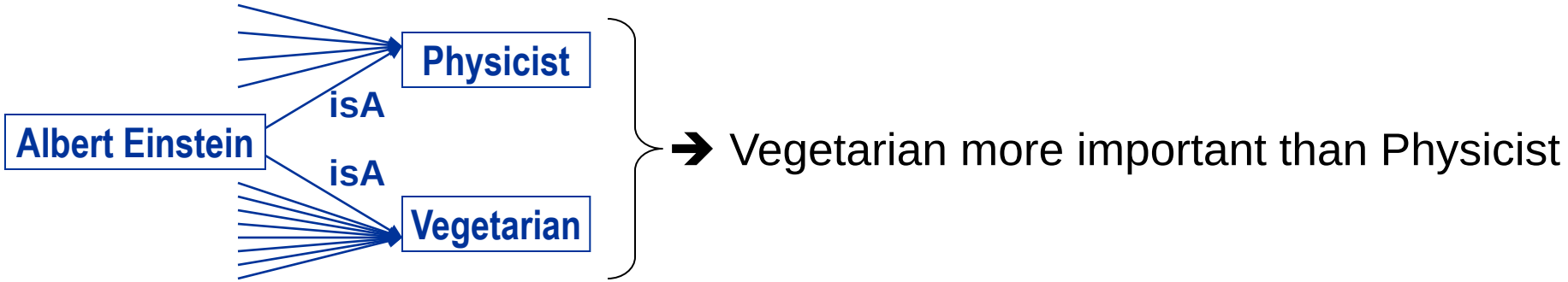
Consider



Possible results

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

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



Framework (Ranking model)

- **Background model**

- P(qi), 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 q1 = $y bornIn Ulm and q2 = $y isA scientist.*
- *If there are many people born in Ulm, but there are only few scientists overall, this suggests giving a higher weight to q2.*

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

Benchmark	Question with NAGA translation
TREC	When was Shakespeare born? Shakespeare bornInYear $x In what country is Luxor? Luxor locatedIn $x $x isA country
SphereSearch	In which movies did a governor act? $y isA governor $y actedIn $z $z isA movie What was discovered in the 20th century? $x discoveredInYear $y $y after 1900 $y before 2000
OWN	Who produced or directed the movie "Around the World in 80 Days"? $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

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

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 Rnaking for Queries on RDF-Graph

Framework in Brief(1)

- Knowledge Graph
  - G = <V, A, l_V, l_A, L>
    - L: a set of lables
    - V: a set of nodes
    - $A \subseteq V \times V \times L$ a set of labled arcs
    - l_V : V → L, an injective function that returns the label of a node
    - l_A : A → L, an injective function that returns the label of an arc such that lA ((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, …, $t_{|A|}$}

Framework in Brief(2)

- A key – augmented knowledge graph G
  - Derived by enriching the knowledge graph with a function KG: A → 2KEY 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 importantce
- 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.

Woody_Allen produced ?x .
Woody_Allen directed ?x

- Keywo[illegible]ructured Queries
  - Allows keyword to be associtaed
  - E.g. Woody_Allen produced ?x{*murder lover*}
Woody_Allen directed ?x
- Relax[illegible]
  - Allows for approximation matching of queries
  - -- alleviates the problem of “too few results”
  - E.g. Woody_Allen produced ?x{*murder*}
Woody_Allen ?v ?x

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

Purely structured queries with relaxation				
Dataset	**OWN**	**WOR**	**BANKS**	**NAGA**
IMDB	**0.880**	0.751	0.777	0.798
LT	**0.876**	0.787	0.721	0.869
Keyword-augmented queries with relaxation				
Dataset	**OWN**	**WOR**	**BANKS**	**NAGA**
IMDB	**0.884**	0.722	0.782	0.776
LT	**0.853**	0.835	0.690	0.782

Table 12: Avg. NDCG for all evaluation queries