Exemplar Queries: Unleashing the Power of Knowledge Graphs

Davide Mottin
Matteo Lissandrini  Themis Palpanas  Yannis Velegrakis

University of Trento
dbTrento - Database and Information Management Group

May 2, 2014
The search statistics

Percentage of people who

use search engines 93%
find useful info 91%
are satisfied 73%
get conflicting info 41%
get partial info 34%

Only the 70% of users knows how to search correctly
An example - Traditional Queries

Google search results for "google and youtube acquisition"

About 25,100,000 results (0.31 seconds)

List of mergers and acquisitions by Google - Wikipedia, the ...
en.wikipedia.org/.../List_of_mergers_and_acquisitions_by_Go...
Google has acquired over 100 companies, with its largest acquisition being the ... service company Next New Networks, which became YouTube Next Lab and ...

History of YouTube - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/History_of_YouTube
Jump to Purchase by Google (2006) - The deal to acquire YouTube closed on November 13, and was, at the time, Google's second largest acquisition.

Google To Acquire YouTube for $1.65 Billion in Stock ...
googlepress.blogspot.com/.../google-to-acquire-youtube-for-165_09.HT...
MOUNTAIN VIEW, Calif., October 9, 2006 – Google Inc. (NASDAQ: GOOG) announced today that it has agreed to acquire YouTube, the consumer media ...

Google's Best And Worst Acquisitions - ABC News
abcnews.go.com › Money › ABC News
Jan 15, 2014 - Of all the companies Google has acquired, which have been the most ... AdMob and YouTube—"although at YouTube the revenue and profit ...

Google in Talks To Buy YouTube For $1.6 Billion - WSJ.com
online.wsj.com/news/.../SB11601481385788491... The Wall Street Journal
Google is in talks to acquire YouTube for about $1.6 billion. A deal would catapult
The user knows that they are looking for news about that specific acquisition.
The user knows that she is looking for news about that specific acquisition. But, if she wants to know about IT companies acquisitions?
First search - describe the information

Google

**IT companies acquisitions**

About 176,000,000 results (0.30 seconds)

**List of mergers and acquisitions by Google - Wikipedia, the ...**

Google has acquired over 100 **companies**, with its largest **acquisition** being the purchase of Motorola Mobility, a mobile device manufacturing company, ...

**List of mergers and acquisitions by Microsoft - Wikipedia, the ...**
[en.wikipedia.org/.../List_of_mergers_and_acquisitions_by_Mi...](en.wikipedia.org/.../List_of_mergers_and_acquisitions_by_Mi...)

Since Microsoft's first **acquisition** in 1987, it has purchased an average of six **companies** a year. The company purchased more than ten **companies** a year ...

**List of mergers and acquisitions by Yahoo! - Wikipedia, the ...**
[en.wikipedia.org/.../List_of_mergers_and_acquisitions_by_Ya...](en.wikipedia.org/.../List_of_mergers_and_acquisitions_by_Ya...)

As of April 2008, the company's largest **acquisition** is the purchase of Broadcast.com, an Internet radio company, for $5.7 billion, making Broadcast.com ...

Xobni - Rockmelt - Snip.it

**News for IT companies acquisitions**

- **Corporate Acquisitions Of Startups -- Why Do They Fail?**
  - [Forbes - 6 days ago](Forbes - 6 days ago)
Problems

- Not always possible to describe the intended results
- Many acquisitions missing
Second search - Find a similar case

Google

Acquisition like Google and Youtube

About 4,120,000 results (0.35 seconds)

Is Facebook deal like Google buying YouTube? - MarketWatch
www.marketwatch.com/.../is-facebook-deal-like-google-bu...
Feb 20, 2014 - Facebook's $19 billion acquisition of WhatsApp is similar to Google buying YouTube, analysts say.

Boutiques.com Traffic Drops 94% – Did Google Give Up On ...
www.signature9.com/.../boutiques-com-traffic-drops-94-did-google-give...
As big as their acquisitions can be, Google's track record on non-search ... For every YouTube, Google likely has 2 Dodgeballs; and if their drop in traffic is any ... They're going to more precisely defined sites like ShopStyle (product search) ...

Google in Talks To Buy YouTube For $1.6 Billion - WSJ.com
online.wsj.com/news/.../SB11601481385788491...
Google is in talks to acquire YouTube for about $1.6 billion. ... Like Web browsers and search engines before them, YouTube and social-networking sites are ...

Was YouTube a good acquisition for Google? - Quora
www.quora.com/Google-YouTube-Acquisition.../Was-YouTube-a...
Google-YouTube Acquisition (October 2006): Was YouTube a good ... From a distance, it looks like Google basically left the product (and team) alone while ...

History of YouTube - Wikipedia, the free encyclopedia
wikipedia.org/wiki/History_of_Youtube
Second search - Find a similar case

Google search for "Acquisition like Google and Youtube"

About 4,120,000 results (0.35 seconds)

Is Facebook deal like Google buying YouTube? - MarketWatch
www.marketwatch.com/.../is-facebook-deal-like-google-buying-youtube
Feb 20, 2014 - Facebook's $19 billion acquisition of WhatsApp is similar to Google buying YouTube, analysts say.

Boutiques.com Traffic Drops 94% – Did Google Give Up On ...
www.signature9.com/.../boutiques-com-traffic-drops-94-did-google-give... As big as their acquisitions can be, Google's track record on non-search ... For every YouTube, Google likely has 2 Dodgeballs; and if their drop in traffic is any ... They're going to more precisely defined sites like ShopStyle (product search) ...

Google in Talks To Buy YouTube For $1.6 Billion - WSJ.com
online.wsj.com/news/.../SB11601481385788491... The Wall Street Journal
Google is in talks to acquire YouTube for about $1.6 billion. ... Like Web browsers and search engines before them, YouTube and social-networking sites are ...

Was YouTube a good acquisition for Google? - Quora
www.quora.com/Google-YouTube-Acquisition-Was-YouTube-a... Quora
Google-YouTube Acquisition (October 2006): Was YouTube a good ... From a distance, it looks like Google basically left the product (and team) alone while ...
Yahoo!-del.icio.us or Microsoft-Skype are not present as existing acquisitions.
A new perspective: Exemplar Queries

Exemplar: 

**Acquisition like Google and Youtube**

About 4,120,000 results (0.35 seconds)

**Google** in Talks To Buy **YouTube** For $1.6 Billion - WSJ.com
online.wsj.com/news/.../SB11601481385786491...  The Wall Street Journal
Google is in talks to acquire YouTube for about $1.6 billion. ... Like Web browsers and search engines before them, YouTube and social-networking sites are ...

**Yahoo.icio.us? – Yahoo Acquires Del.icio.us | TechCrunch**
techcrunch.com/2005/12/09/yahoo-acquires-delicious/  TechCrunch
Dec 8, 2005 - ... instant message conversation with Joshua Schachter, the founder of Del.icio.us. I was asking him for any comment on the Yahoo acquisition.

**Microsoft’s Hefty Phone Bill for Skype - WSJ.com**
online.wsj.com/.../SB100014240527487...  The Wall Street Journal
by Nick Wingfield - in 236 Google+ circles
Microsoft racked up a whopping $8.5 billion phone bill to buy Skype even ... Whether Microsoft can make a Skype acquisition work—especially at such a rich ...

**Facebook buys WhatsApp for $19 billion - Feb. 19, 2014**
money.cnn.com/2014/02/19/.../facebook-whatsapp/  CNNMoney
by Adrian Covert - in 1,019 Google+ circles
Feb 19, 2014 - WhatsApp, the largest messaging service in the world, will now be under the ... and the Web, Facebook has acquired WhatsApp for $19 billion.
Google-Youtube are treated as examples of the needed resource.

Results are similar acquisitions (e.g. Yahoo!-del.icio.us)
Outline

1. Scenario
2. Problem
   - Graph model
   - Problem definition
3. Proposed solution
   - Challenges
   - Exact solutions
   - Approximate solution
4. Ranking and top-k
5. Experimental evaluation
6. Related Work
7. Conclusions
Data Model: Knowledge Graphs

- Knowledge graph
- Directed labeled multigraph (multiple edges between the same pair of nodes)
- Nodes are concepts
- Edges are relationships
- Nodes and edges have labels
Knowledge graph

Directed labeled multigraph (multiple edges between the same pair of nodes) where

- Nodes are concepts
- Edges are relationships
- Nodes and edges have labels
A graph based approach

Exemplar Queries

May 2, 2014

D. Mottin (Unitn)
A graph based approach

An answer to an exemplar query $Q$ is a graph edge-isomorphic to $Q$
A graph based approach

An answer to an exemplar query $Q$ is a graph edge-isomorphic to $Q$
Definition

The evaluation of an exemplar query $Q_e$ on a database $D$, denoted as $xmpEval(Q_e)$, is the set \{a | \exists s \in eval(Q_e) \land a \approx s\}, where $a$ and $s$ are structures in $D$ and the symbol $\approx$ indicates a similarity function.

Problem

Given a knowledge graph $D$, an exemplar query $Q$, we want to find the ranked list of isomorphic answers $Q$ in $D$, after having found the sample $S$ in $D$. 
Research Challenges

1. Formally Define Exemplar Queries
2. Understand which data model better supports Exemplar Queries
3. Define the appropriate similarity function \(\approx\)
4. Rank the answers of an Exemplar Query
5. Compute the answers in real time on a semantically reach database
Solution overview

Exemplar Query

Exemplar Query Answering

Answers

I:

II:
Exemplar Queries

1. Traditional Evaluation
2. Find Similar
3. Refine
4. Ranking & Top-k

Exemplar Query

Answers

I:
II:
Solution overview

Exemplar Query

Exemplar Query Answering

Answers

I: 
II: 

Samples

(1) Traditional Evaluation

(2) Find Similar

Solutions

(3) Refine

(4) Ranking & Top-k

Refinements
The XQ Algorithm

**Input:** Database $D: \langle N, E \rangle$

**Input:** User Query $Q$

**Output:** Set of relevant answers $\mathcal{Q}$

1. $\mathcal{Q} \leftarrow \emptyset$
2. $S \leftarrow \text{eval}(Q)$
3. $D' \leftarrow \text{PruneAndRestrict}(D)$
4. $n_s \leftarrow \text{selectARandomNode}(S)$
5. for each $n \in E$ do
6.   $A \leftarrow \text{FindIsomorphicSubg}(S, n_s, D', n)$
7. if $A \neq \emptyset$ then
8.   $\mathcal{Q} \leftarrow \mathcal{Q} \cup A$
9. $\text{Rank}(\mathcal{Q})$
10. return $\mathcal{Q}$
The XQ Algorithm

**Input:** Database \( D : \langle N, E \rangle \)

**Input:** User Query \( Q \)

**Output:** Set of relevant answers \( \mathcal{Q} \)

1. \( \mathcal{Q} \leftarrow \emptyset \)
2. \( S \leftarrow \text{eval}(Q) \)
3. \( D' \leftarrow \text{PruneAndRestrict}(D) \)
4. \( n_s \leftarrow \text{selectARandomNode}(S) \)
5. **for each** \( n \in E \) **do**
6. \( A \leftarrow \text{FindIsomorphicSubg}(S, n_s, D', n) \)
7. **if** \( A \neq \emptyset \) **then**
8. \( \mathcal{Q} \leftarrow \mathcal{Q} \cup A \)
9. \( \text{Rank}(\mathcal{Q}) \)
10. **return** \( \mathcal{Q} \)

**Issues**

- (sub)-graph isomorphism is NP-complete [2]
- efficiency decreases with the number of query nodes
- many solutions can be discarded in advance
The XQ Algorithm

Input: Database $D$: $\langle N, E \rangle$
Input: User Query $Q$
Output: Set of relevant answers $\mathcal{Q}$

1. $\mathcal{Q} \leftarrow \emptyset$
2. $S \leftarrow \text{eval}(Q)$
3. $D' \leftarrow \text{PruneAndRestrict}(D)$
4. $n_s \leftarrow \text{selectARandomNode}(S)$
5. for each $n \in E$ do
6.   $A \leftarrow \text{FindIsomorphicSubg}(S, n_s, D', n)$
7.   if $A \neq \emptyset$ then
8.     $\mathcal{Q} \leftarrow \mathcal{Q} \cup A$
9. end if
10. $\text{Rank}(\mathcal{Q})$
11. return $\mathcal{Q}$

Algorithms
1. IterativePruning: reject non matching nodes
2. RelevantNeighborhood: restrict the search space to “near” nodes
Exemplar Queries Answering: An efficient solution

Query

```
q1
  ↓
  a
  ↓
  q2
  ↓
b  c
  ↓
  q3  q4
```

Graph

```
1
  ↑
a  a
  → 3
  ↓
b  a
  → 6
  ↓
c
  → 5
  ↓
c
  → 4
  ↓
```

Idea: represent nodes as vectors of neighbors labels
Idea: represent nodes as vectors of neighbors labels
Exemplar Queries Answering: An efficient solution

Query node q1

Graph node 1

Idea: represent nodes as vectors of neighbors labels
Pruning Theorem

**Definition (d-neighbor)**

Let $n \in N$ be a node of a database $D = \langle N, E \rangle$. The node $n_i \in N$ is a $d$-neighbor of $n$ if there exists a shortest path from $n$ to $n_i$ of length at most $d$. The $d$-neighborhood of $n$, denoted as $\mathcal{N}_d(n)$, is the set of $d$-neighbors of $n$.

**Theorem**

Given a database $D = \langle N, E \rangle$ and a user sample $S$, let $\mathcal{N}_d$ and $\mathcal{N}_d^S$ be the $d$-neighborhood of $D$ and $S$ respectively. If there exists a subgraph-isomorphism $\mu : E_S \rightarrow E$, then

$$\forall n_s \in E_S, \mathcal{N}_d^S(n_s) \subseteq \mathcal{N}_d(n), n \in E, n \in \mu(n_s)$$
Query node q1

Graph node 1

Difference vector
Exemplar Queries Answering: An efficient solution

Query node q1

Node 1 matches, map q1 with 1
Exemplar Queries Answering: An efficient solution

Query node q1

Graph node 2

Difference vector
Exemplar Queries Answering: An efficient solution

Query node q1

Node 2 does not match
Start from a query node $q$
The IterativePruning Algorithm

1. Start from a query node $q$
2. Match $q$ with the graph nodes
The Iterative Pruning Algorithm

1. Start from a query node $q$
2. Match $q$ with the graph nodes
3. For each adjacent node of $q$
4. Find nodes in the graph from candidate map of $q$ matching the edge
Start from a query node $q$

2. Match $q$ with the graph nodes

3. For each adjacent node of $q$

4. Find nodes in the graph from candidate map of $q$ matching the edge

5. Repeat 2. with an adjacent node of $q$ until all nodes have been visited
The Iterative Pruning Algorithm

1. Start from a query node $q$
2. Match $q$ with the graph nodes
3. For each adjacent node of $q$
4. Find nodes in the graph from candidate map of $q$ matching the edge
5. Repeat 2. with an adjacent node of $q$ until all nodes have been visited

Theorem

*Given a user sample $S$, if Algorithm Iterative Pruning terminates with a complete exploration of the nodes in $S$, then there exists in $\mu$ a simulation $R$ of the user sample $S$.***
Finding Relevant nodes

Idea

1. Not all the nodes are equally relevant.
2. Nodes "far" from the query are less related.
Finding Relevant nodes

Idea

1. Not all the nodes are equally relevant.
2. Nodes “far” from the query are less related.
**Finding Relevant nodes**

**Idea**

1. Not all the nodes are equally relevant.
2. Nodes “far” from the query are less related.
How closely related are two nodes in a graph?
Relevance: Personalized PageRank

Personalized PageRank provides the relevance score [9]
Relevance: Personalized PageRank

Personalized PageRank provides the relevance score [9]
Personalized PageRank provides the relevance score [9]
Relevance: Personalized PageRank

Personalized PageRank provides the relevance score [9]

Diagram showing relationships between keywords like Google, YouTube, Yahoo!, Menlo Park, S. Clara, and their relevance scores.
Proximity Score as the probability of visiting the node
Relevance: Personalized PageRank

Proximity Score as the probability of visiting the node

- **Business**
  - **Auto company**
    - **G. M. (1/120)**
    - **Opel**
  - **IT company**
    - **del.icio.us**
  - **Search Engine**
    - **YouTube**
      - **Menlo Park**
    - **Yahoo! (7/40)**
      - **S. Clara**
    - **Google (1)***

*Exemplar Queries*
\[ \mathbf{v} = (1 - c) \mathbf{A} \mathbf{v} + c \mathbf{p} \]
Random Walk with Restart Computation

\[ \mathbf{v} = (1 - c)\mathbf{Av} + c\mathbf{p} \]

- \( \mathbf{v} \): column vector
- \( \mathbf{v}[u] \): probability to be at node \( u \)
\[ \mathbf{v} = (1 - c) \mathbf{A} \mathbf{v} + c \mathbf{p} \]

- \( \mathbf{v} \): column vector
- \( \mathbf{v}[u] \): probability to be at node \( u \)
- \( \mathbf{p} \): column vector of zeros
  - apart for the starting node \( \mathbf{p}_i \)
    - set to 1, i.e. \( \mathbf{p}[i] = 1 \)
\[ \mathbf{v} = (1 - c) \mathbf{A} \mathbf{v} + c \mathbf{p} \]

- \( \mathbf{v} \): column vector
- \( \mathbf{v}[u] \): probability to be at node \( u \)
- \( \mathbf{p} \): column vector of zeros apart for the starting node \( p_i \) set to 1, i.e. \( p[i] = 1 \)
- \( \mathbf{A} \): column normalized adjacency matrix of the graph.
- \( \mathbf{A}[u, v] \): transition probability to \( u \) given that the current state is node \( v \).
Random Walk with Restart Computation

\[ \mathbf{v} = (1 - c) \mathbf{A} \mathbf{v} + c \mathbf{p} \]

- \( \mathbf{v} \): column vector
- \( \mathbf{v}[u] \): probability to be at node \( u \)
- \( \mathbf{p} \): column vector of zeros apart for the starting node \( \mathbf{p}_i \) set to 1, i.e. \( \mathbf{p}[i] = 1 \)
- \( \mathbf{A} \): column normalized adjacency matrix of the graph.
- \( \mathbf{A}[u, v] \): transition probability to \( u \) given that the current state is node \( v \).
- \( c \): restart probability
Edge probability computed as inverse frequency

\[ I(e_{ij}^{\ell}) = I(\ell) = \log \frac{1}{P(\ell)} = -\log P(\ell) \]

\[ P(\ell) = \frac{|\mathcal{R}^\ell|}{|\mathcal{R}|}. \]

Edge labels in the query assigned a higher weight

Real-time simulation of PageRank (Weighted Particle Filtering [4])
Combination of two factors:

1. Structural: similarity of two nodes in terms neighbor relationships
2. Distance-based: the PageRank already computed
Structural similarity
Distance-based similarity
Ranking function

Linear combination of structural and distance-based similarity

\[ \rho(n_s, n) = \lambda S(n_s, n) + (1 - \lambda)v[n] \]

\( S(n_s, n) \): structural similarity (cosine similarity between the vectorial representation of each node, where the vector contains the labels in the neighborhood of the node)

\( v[n] \): Personalized PageRank vector (distance-based similarity)
Experimental Setup

- **Dataset**
  - Freebase: 52M nodes, 232M edges (we are the first using the entire Freebase!)
  - Freebase Internet Domain: 2M nodes, 6M edges
  - Synthetic (number of nodes): 0.5M, 1M, 5M, 10M, 20M nodes, 1K queries
  - Synthetic (number of queries): 0.5K, 1K, 2K, 5K, 10K queries, 10M nodes

- **Testset**: 90 queries manually mapped from AOL query logs

- **Baseline**: NeMa (Neighborhood Matching) Algorithm [6]: approximate answers on graphs

**Measures**

- **Query time**: sum of times to run the different algorithms
- **Quality and usefulness of the answers** (from the user study)
Scalability results (10M nodes)

RelevantNeighborhood is stable with respect to the number of answers
Scalability results (10M nodes)

RelevantNeighborhood is stable with respect to the number of answers

< 150ms to get the answers
Scalability results (10M nodes)

RelevantNeighborhood is stable with respect to the number of answers

< 150ms to get the answers

Constant time with increasing number of answers
Quality results

Q1: Google - YouTube - Menlo Park

- Google - YouTube - Menlo Park
- Yahoo! - LAUNCH Media - Stanford U
- Yahoo! - Musicmatch - Stanford U
- Yahoo! - Right Media - Stanford U
- Yahoo! - Inktomi Corporation - Stanford U

Q2: Condom - Sex - HIV infection

- Water purification - Fecal-oral - Cholera
- Smoking cessation - Vector - Diabetes
- Oral Transm. - Cerebral palsy - Cytomegal.
- Water purification - Fecal-oral - Cholera

- Safe sex - Sex - HIV infection
- Sexual abstinence - Sex - HIV infection
- Safe sex - Vertical transmission - Hepatitis B
- Safe sex - Vertical transmission - Syphilis
- Hand washing - Droplet Contact - Cold
Baseline Comparison

- Modified NeMa to include label matching
- Tested on Freebase on restricted graphs produced by RelevantNeighborhood
- Query time becomes more than 13 hours, w.r.t. 0.1s of our solution
- Output is a set of approximate answers and not related queries
Usefulness

- **Quality**
  - 92% people say that Exemplar Queries are useful
  - 62% already had the need for such a service
Usefulness

Quality

- 92% people say that Exemplar Queries are useful
- 62% already had the need for such a service

Comparison

Users asked to choose results using two previous methods and ours.

- 64% people choose Exemplar Queries
- 30% people choose other approaches
Related work

- **Query modification:**
  - Query refinement: modify the original query to restrict or enlarge the answer set [8, 7]
  - Query diversification: include results that are somehow diverse and new with respect to the one in the original query answer set [3, 10]

- Knowledge-base mapping: match plain text to entities in a knowledge base [1]

- Querying graphs: find similar or isomorphhic structures in a graph [6, 5]
Conclusions and Future Work

We introduce

- the novel problem of finding exemplar queries
- a new model to discover the exemplar queries in a knowledge graph
- exact and Approximate solutions for this problem and a novel ranking function
- the exemplar queries problem that tries to fill the gap in the information search with a novel and interesting paradigm

We show

- the proposed service is missing and people perceived it as important
- proposed solutions are scalable and fast
- our methods work on one of the biggest knowledge graphs (Freebase) producing results in real time
time for questions
References I

A survey on ontology mapping.

The complexity of theorem-proving procedures.

Finding dimensions for queries.
In *CIKM*, 2011.

Topic-sensitive pagerank.
In *WWW*, 2002.

Neighborhood based fast graph search in large networks.
In *SIGMOD*, 2011.
Nema: Fast graph search with label similarity.
In *PVLDB*, 2013.

Relaxing join and selection queries.
In *PVLDB*, 2006.

Interactive query refinement.
In *EDBT*, 2009.

Fast random walk with restart and its applications.
In *ICDM*, 2006.

Inferring the most important types of a query: a semantic approach.