eswc15

Ranking Entities in the Age of Two Webs
An Application to Semantic Snippets

Mazen Alsarem*
Pierre-Edouard Portier* Sylvie Calabretto* Harald Kosch**

* Université de Lyon, CNRS, INSA de Lyon, LIRIS, UMR5205, F-69621, France
** Universität Passau, Innstr. 43, 94032 Passau, Germany
How can we use the Web of data to improve the user experience when interacting with a Web search engine result page?
Portorož, is an Adriatic Mediterranean coastal settlement in the Municipality of Piran in southwestern Slovenia. Its modern development began in the late 19th century with appearance of first health resorts. In the early 20th century it became one of the grandest seaside resorts in Europe...
Sparse graph
- A lot of annotated entities
- Poorly connected graph
- Noise and off-topic entities
Sparse graphs with associated textual data.

Portorož, is an Adriatic Mediterranean coastal settlement in the Municipality of Piran in southwestern Slovenia. Its modern development began in the late 19th century with appearance of first health resorts. In the early 20th century it became one of the grandest seaside resorts in Europe...

Economy of Portorož is mainly based on tourism and gaming industry.

Bernardin is a tourist complex located in the western part of Portorož.

Portorož belongs to the coastal municipality of Piran, located in Slovenia.

Abstract from DBpedia

From Web of Data

Windows of text centered on the surface forms associated with the entity

From Web of Documents
Mazen Alsarem: Ranking Entities in the Age of Two Webs

Web Information retrieval system

User information need

Documents ranked list

Web of Data

Web page

Algorithm

RDF Graph

Automatic annotation

- Rank the entities
- Filter the noise

- A lot of annotated entities
- Poorly connected graph
- Noise and off-topic entities

Context

Web page’s text
Objective

So, what we need is:

"A query-biased entities ranking algorithm"

- Adapted to the context of automatically annotated web pages
- Making good use of contextual knowledge obtained from the user's query and the Web page's textual content
Related works are almost always based on modifications of the PageRank algorithm.

At the core of the PageRank algorithm we find the power iteration method that propagates the importance of the nodes through the graph until the steady vector of a corresponding Markov chain is attained.

For this method to converge, the stochastic matrix that represents the structure of the graph is combined with a rank-1 stochastic matrix corresponding to a probability distribution on the nodes and representing some prior knowledge about the importance of the nodes.
Modification Of PageRank

With the classical PR, we have no prior knowledge about the importance of the nodes, therefore we use a uniform probability distribution (Random surfer).

All PageRank modifications are based on using a peculiar source of prior knowledge about the importance of the nodes.
We have three contributions:

- **LDSVD**, a new query-biased SVD-based ranking algorithm returning a prior-distribution on the nodes of a RDF graph by using implicit relationships obtained from the textual data associated with the nodes.

- **LDRANK**, a modified PageRank algorithm able to combine prior-distributions obtained from different sources of prior-knowledge into a consensual prior-distribution.

- **ENsEN**, a semantic-snippet generation system based on the LDRANK algorithm.
Compute a prior-distribution about the importance of the entities given associated textual data and a user's query.

First Contribution

LDSVD
Linked Data *Singular Value Decomposition*
LDSVD Approach

1. Build $R \leftarrow \text{the sparse entities-terms matrix}$, that represents the entities in the terms' space.

2. Build $\text{Info}_\text{need} \leftarrow \text{entities detected in the text of the query}$, that represents the user information need.

3. In the original terms' space, we increase the importance of the “$\text{Info}_\text{need}$” entities.

4. Forcing them to move away from the origin in the reduced space obtained from $\text{SVD}$.

5. We then observe how much other entities were pushed away from the origin (in the SVD reduced space).

6. Entities with a more pronounced movement away from the origin of the reduced space are potentially related to the user's information need.

The Scoring

The score of an entity is proportional to the magnitude of its movement away from the origin of the SVD-reduced space.
A Query-biased Ranking Algorithm for the entities of an RDF graph
Ranking entities on the Web of data: Existing approaches

- **Modification of PageRank**
  - Types of the links between ontologies
    - Swoogle, Ontology Rank, Ding et al 2012.
  - Types of predicates between entities
    - PopRank, Nie et al 2005.
  - Topical relationships between entities
    - RareRank, Wei et al 2011.

- **Ranking of the Web pages within which the entities were detected**

- **Based on learning to rank**

So, what do we propose?
We propose an algorithm that modifies PR using three prior knowledge sources.

- **The Random surfer**
  - Equidistrib
    - Classical PageRank

- **The results from SE**
  - Hitdistrib
    - Fafalios et al., 2014

- **Textual implicit relationships**
  - Svddistrib
    - Our approach LDSVD

**LDRANK: Algorithm description**
**LDRANK: Algorithm description**

Aggregation of three prior knowledge sources

- **Equidistrib**
- **Hitdistrib**
- **Svddistrib**

- Classical PageRank
- Fafalios et al., 2014
- Our approach LDSVD

- We apply a **consensual linear opinion pool algorithm**.
- Iterative algorithm
- At each step, **expert i** re-evaluates its distribution as a **linear combination** of the distributions of all the experts.
- The weight associated by **expert A** to the distribution of **expert B** is proportional to the **distance** between the two distributions.
- The authors define this distance such that the process **converges towards a consensus**.

---

* Fafalios, P., Tzitzikas, Y.: Post-analysis of keyword-based search results using entity mining, linked data, and link analysis at query time (2014).
No evaluation dataset suited to our context
• Query-biased ranking of entities in sparse graphs with text associated to the nodes.

Crowdsourcing approach for making our evaluation dataset
1) Data collection
• 30 queries: selected randomly from "Yahoo! Search Query Tiny Sample"
• For each query:
  • Keep top 5 results returned by Google search
• For each Web page:
  • Extract main raw textual content*
  • Annotation by DBpedia Spotlight ==> ~80 entities per Web page.

2) The job
• CrowdFlower platform
• Evaluate the annotations with respect to the user information need.

* by applying the algorithm proposed by Kohlschtter, Fankhauser, and Nejdl, 2010
3) **Microtask**

- Scoring the relevance of the annotations of a single sentence.
- For each annotated entity the worker is asked to evaluate the relevance of the annotation with respect to the query.
  - irrelevant (0),
  - marginally relevant (1),
  - fairly relevant (2),
  - and highly relevant (3).
- **10 workers** per job == 10 judgments.

---

Our system thinks that "MLB" is referring to the concept (Major League Baseball), is it?

- Irrelevant
- Marginally relevant
- Fairly relevant
- Highly relevant

**Major League Baseball:**

Major League Baseball (MLB) is a professional baseball organization that constitutes one of the four major professional sports leagues in North America. It is the oldest league of the four. Teams play ...

* We used the ordinal scale proposed by Jarvelin and Kekalainen when they introduced the DCG graded relevance
4) Quality Control

- Measure the agreement with the Krippendorff’s alpha coefficient:
  - We obtained an alpha of 0.22 which can be considered a fair agreement*.
  - But comparing to existing works (crowdsourcing, information retrieval), we cannot be satisfied with an alpha of 0.22.
- To improve the quality, we removed the workers that often disagreed with the majority.
  - =>
    - alpha of 0.46
    - 96.5% jobs still have at least 3 workers
    - only 0.7% jobs done by only 1 worker.

*According to Landis and Koch’s scale
**Comparison of:**

1. **Unmodified PageRank**
2. **HitPageRank** (Fafalios et al 2014)
3. **SVDPPageRank**: PageRank modified with our LDSVD ranking strategy
4. **LDRANK**: Consensual mixture of the three previous prior probability vectors.

We used the NDCG (Normalized Discounted Cumulative Gain) metric.
Comparison of the execution time for the four different strategies
Enhanced Search Engine

A software system that enhances a SERP with semantic snippets

Third Contribution
ENsEN: Introduction

Snippets Vs. Semantic snippets

A snippet is an excerpt from a Web page, determined at query-time, it tries to help the user in making a decision about the relevance of the corresponding web page to his information need.

A semantic snippet aims to improve this process of decision making and the whole exploration by making explicit the relationship between the user information need and the most relevant entities in a web page.
ENsEN: Workflow

Pre-processing

Building RDF graph (& Annotation)

Ranking Entities (LDRANK)

Graph extension

Sentence Selection (Machine learning)

Snippet UI generation

Query

Text (for each result)

RDF Graph

Extended Graph

Sentence for each entity

Ranked list of entities

User interaction

Data flow

Web service API call

SERP: Search Engine Results Page

Search engine

DBpedia

SPARQL endpoint

SERP

Text (for each result)

RDF Graph

Ranked list of entities

Sentence for each entity

Extended Graph

Graph extension

Sentence Selection (Machine learning)

Snippet UI generation

ENsEN

Pre-processing
ENsEN: Interface

Mazen Alsarem: Ranking Entities in the Age of Two Webs
Erasmus Darwin (12 December 1731 – 18 April 1802) was an English physician. One of the key thinkers of the Midlands Enlightenment, he was also a natural philosopher, physiologist, slave-trade abolitionist, inventor and poet.

Grandfather of Charles Darwin and a philosopher, botanist, and naturalist in his own right.

Excerpt from the document:

One of the key thinkers of the Midlands Enlightenment, he was also a natural philosopher, physiologist, slave-trade abolitionist, inventor and poet.

Annotated concepts:

- Abolitionism in the
- Anna Seward
- Anthropomorphism
- Ashbourne Derbyshire
- Associationism
- Birmingham
- Boarding school
- Botany
- Bourgeoisie
- Breadsall
- Brookfield Community
- Carl Linnaeus
- Charles Darwin
- Chasetown F.C.
- Chemistry

Excerpt from the document:

Erasmus Darwin arrived at his conclusions through an "integrative" approach: he used
Abstract

Evolution is the change in the inherited characteristics of biological populations over successive generations. Evolutionary processes give rise to diversity at every level of biological organisation, including species, individual organisms and molecules such as DNA and proteins. All life on Earth is descended from a last universal ancestor that lived approximately 3.8-3.5 billion years ago. Repeated speciation and the divergence of life can be inferred from shared sets of biochemical and morphol...

In the document

Although some of his ideas on how evolution might occur are quite close to those of Lamarck, Erasmus Darwin also talked about how competition and sexual selection could cause changes in species: "The final course of this contest among males seems to be, that the strongest and most active animal should propagate the species which should thus be improved".

As a naturalist, he formulated one of the first formal theories on evolution in Zoonomia, or, The Laws of Organic Life (1794-1796).
ENsEN: Crowdsourcing-based User Evaluation

- **10 tasks** from the “Yahoo!Answers Query To Questions” dataset.
- Each task was made of **three questions** on a **common topic** (we have the answers).
- We collected **20 judgments** for each task.
- **Half** of the workers were asked to use our system, and the other half used Google.
ENsEN : Crowdsourcing-based User Evaluation

A better correctness rate...

... with less time spent!
Conclusion & perspectives

Conclusion:
• Proposal of an algorithm (**LDRANK**) for ranking the entities of a sparse RDF graph, given the knowledge of a user’s information need and using both the Web of data and the Web of documents.
• Integration of this algorithm in an efficient **semantic snippets generation system**

Perspectives:
• Evaluation of the potentials of this approach for **exploratory search**.
Live Demo:
http://liris-qir.insa-lyon.fr:8080/ENsEN/

Project page:
http://liris.cnrs.fr/drim/projects/ensen/

Google:
search « LIRIS ENsEN »

YouTube:
https://www.youtube.com/user/malsarem
or search « Alsarem ENsEN »

Github:
https://github.com/ALSAREM/ENsEN

Thank You