VU VRIJE UNIVERSITEIT AMSTERDAM

LOTUS: Adaptive Text Search for Big Linked Data

F. Ilievski | W. Beek | M. van Erp | L. Rietveld | S. Schlobach

INTRODUCTION

A wealth of information is potentially available in Linked Open Data sources.

This information could be exploited by researchers and developers for tools and evaluations on a LOD-scale.

But, accesing Big Linked Data is not trivial

INTRODUCTION

A wealth of information is potentially available in Linked Open Data sources.

This information could be exploited by researchers and developers for tools and evaluations on a LOD-scale.

But, accesing Big Linked Data is not trivial

No centralized query service for Linked Data

INTRODUCTION

A wealth of information is potentially available in Linked Open Data sources.

This information could be exploited by researchers and developers for tools and evaluations on a LOD-scale.

But, accesing Big Linked Data is not trivial

No centralized query service for Linked Data

Limited natural language access to Linked Data

INTRODUCTION

A wealth of information is potentially available in Linked Open Data sources.

This information could be exploited by researchers and developers for tools and evaluations on a LOD-scale.

But, accesing Big Linked Data is not trivial

No centralized query service for Linked Data

Limited natural language access to Linked Data

Text-based retrieval is not customizable

“The lack of a global entry point to resources through a flexible text index is a serious obstacle for linked data consumption.”

“The lack of a global entry point to resources through a flexible text index is a serious obstacle for linked data consumption.”

by researchers and developers

Requirements!

REQUIREMENTS

1. Text-based queries
2. Resilience (of text search)
3. Findability (of authoritative and non-authoritative statements)
4. Availability
5. Scalability
6. Serviceability (for both machines and humans)
7. Customizability

THE LOTUS APPROACH



LOD Laundromat

A centralized Linked Data
cleaning and publishing architecture
Allows access to a big subset of the LOD Cloud

38 billion statements

LOTUS!

lotus.lodlaundromat.org

LOTUS

Linguistic entry point

LOTUS is a linguistic entry point to the LOD Laundromat data collection.

Approximate matching

Allows statements to be findable based on approximate string matching on associated literals.

Adaptive Framework

LOTUS allows the resource retrieval to be tailored to fit various use cases.

Customizability of retrieval

Retrieval options

Matching options

- Phrase matching
- Disjunctive token matching
- Conjunctive token matching
- Conjunctive token matching with character edit distance

Retrieval options

Matching options

- Phrase matching
- Disjunctive token matching
- Conjunctive token matching
- Conjunctive token matching with character edit distance

Ranking algorithms

- Length normalization
- Practical scoring function
- Phrase proximity

Content-based

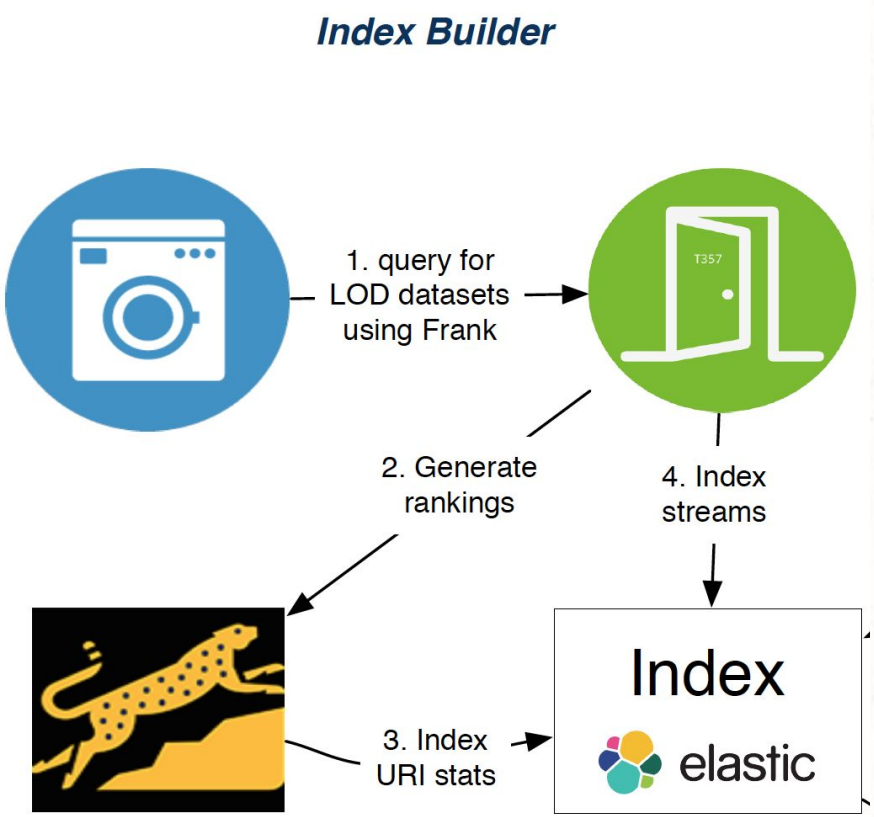
- Terminological richness
- Semantic richness
- Recency

Document-based

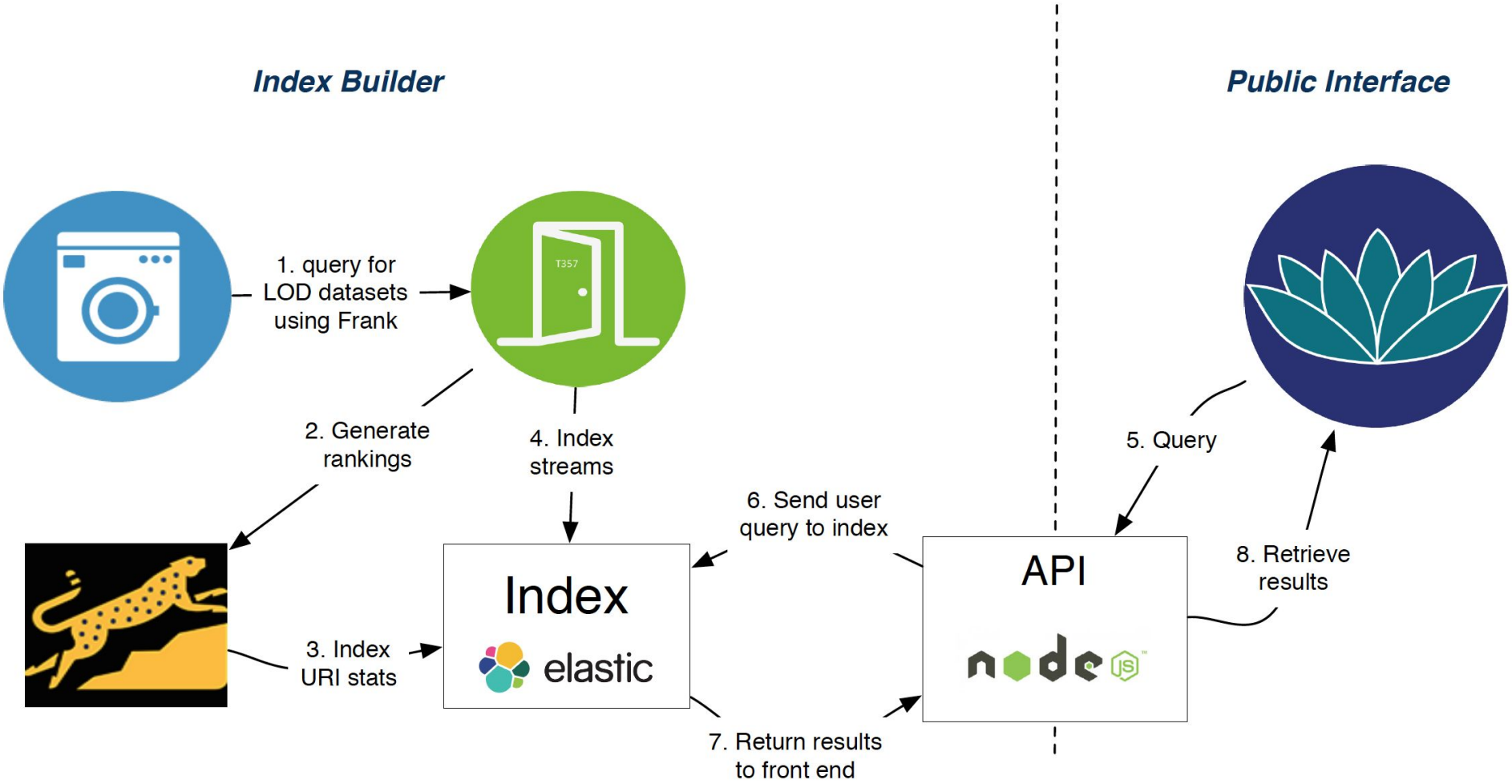
- Degree popularity
- Appearance popularity

Resource-based

Implementation



Implementation



Web Interface + API

lotus.lodlaundromat.org

Distributed architecture

4,334,672,073 Indexed literals

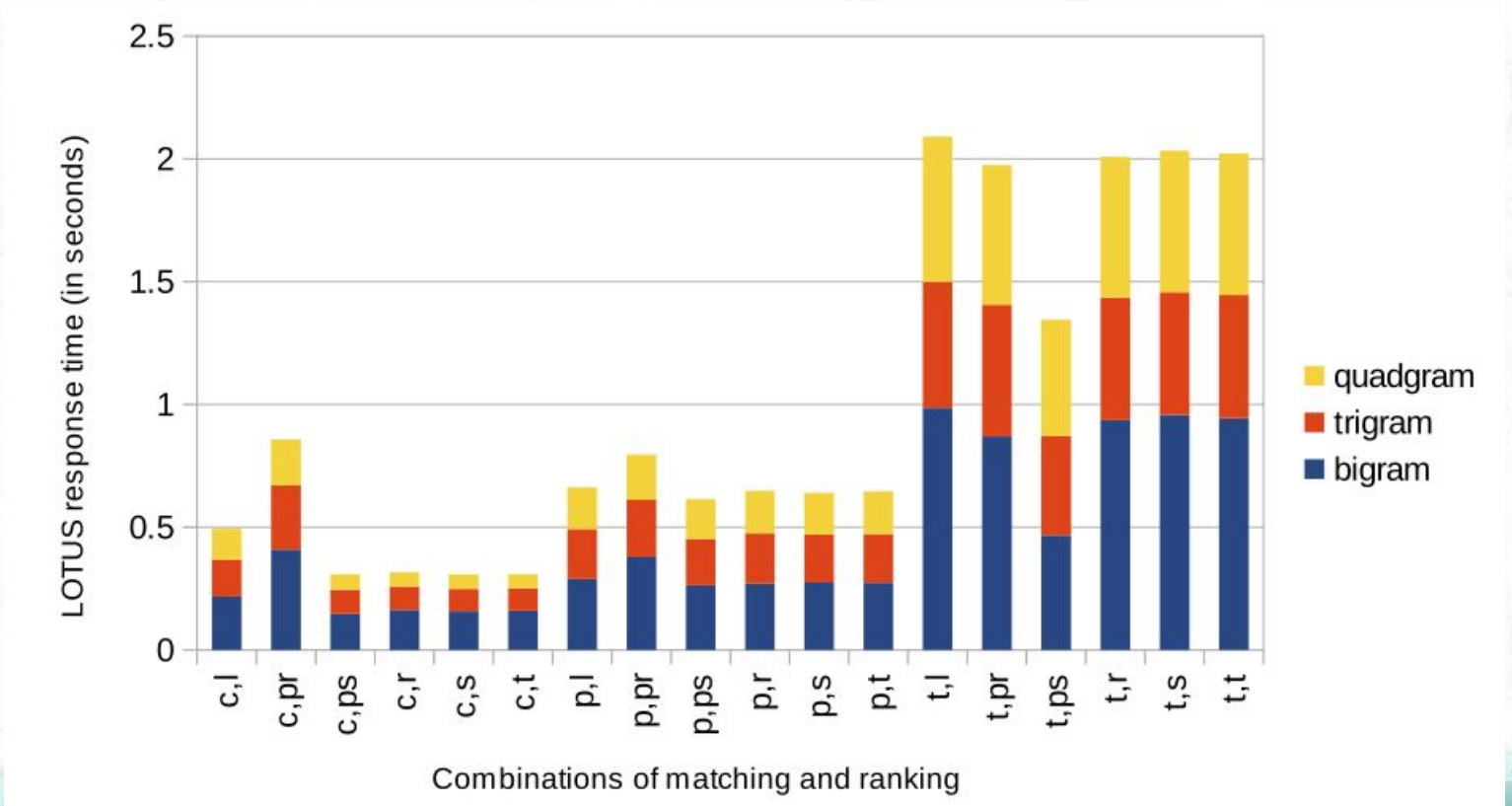
Scaled horizontally over 5 servers

Data replication to ensure high runtime availability of LOTUS

Performance!

Scaling and performance

We used 18k queries to benchmark 18 retrieval combinations of LOTUS



Conclusions

LOTUS is

A Centralized linguistic entry to Big Linked Data

LOTUS indexes over 4 billion literals from the LOD Laundromat

An Adaptive Retrieval framework

LOTUS allows its retrieval to be customized to fit users' needs by offering 32 matching+ranking options.

"connecting the dots"

LOTUS relies heavily on 2 existing systems (LOD Laundromat & ES), but fills the gap by offering a much needed tool for scientific evaluation.

Vision: Scaling applications and evaluations at LOD scale with LOD Lab

Future work

The Precision and recall

of LOTUS should be evaluated on concrete applications, such as Entity Linking and Network Analysis.

Future work

The Precision and recall

of LOTUS should be evaluated on concrete applications, such as Entity Linking and Network Analysis.

Context-dependent ranking

could be added in the future to take the query context into account in order to improve the ranking accuracy.

Thanks!

lotus.lodlaundromat.org

You can find me at
@earthling91 / f.ilievski@vu.nl