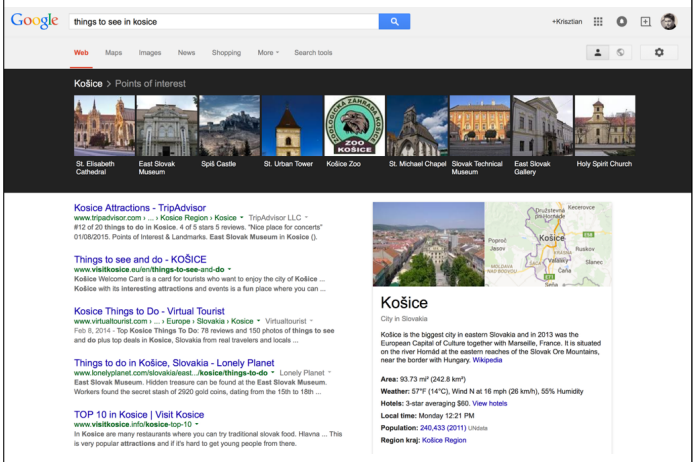
Evaluation Initiatives for Entity-oriented Search

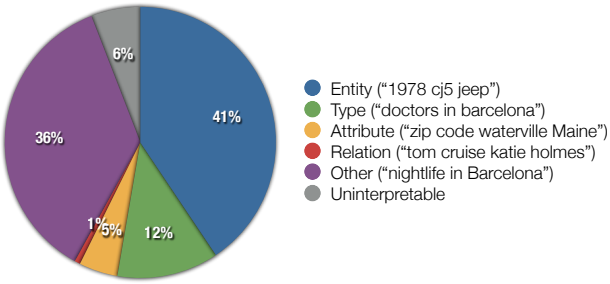
Krisztian Balog
University of Stavanger

KEYSTONE Spring WG Meeting 2015 | Kosice, Slovakia, 2015

Evaluation Initiatives for
Entity-oriented Search
Krisztian Balog
University of Stavanger

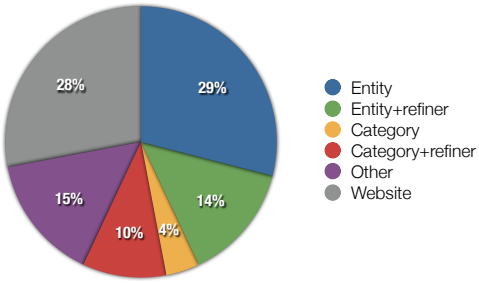


Distribution of web search queries (Pound et al., 2010)



Pound, Mika, and Zaragoza (2010). Ad-hoc object retrieval in the web of data. In WWW '10.

Distribution of web search queries (Lin et al., 2011)



Lin, Pantel, Gamon, Kannan, and Fuxman (2012). Active objects. In WWW '12.

Evaluation Initiatives for
Entity-oriented Search
Krisztian Balog
University of Stavanger

Evaluation Initiatives for
Entity-oriented Search

- Shared tasks at world-wide evaluation campaigns (TREC, CLEF)
- Provide Ingredients for evaluating a **given task**
  - Data collection
  - Gold-standard annotations
  - Evaluation metrics

In this talk

- Keyword-based search in knowledge bases
- Query understanding with the help of knowledge bases
- Test collections, metrics, evaluation campaigns

1

Entity Retrieval

Ad-hoc entity retrieval

- **Input**: keyword query
  - "telegraphic" queries (neither well-formed nor grammatically correct sentences or questions)

ben franklin

american embassy nairobi

meg ryan war

Sweden Iceland currency

worst actor century

Chernobyl

- **Output**: ranked list of entities
- **Collection**: semi-structured documents

Evaluation initiatives

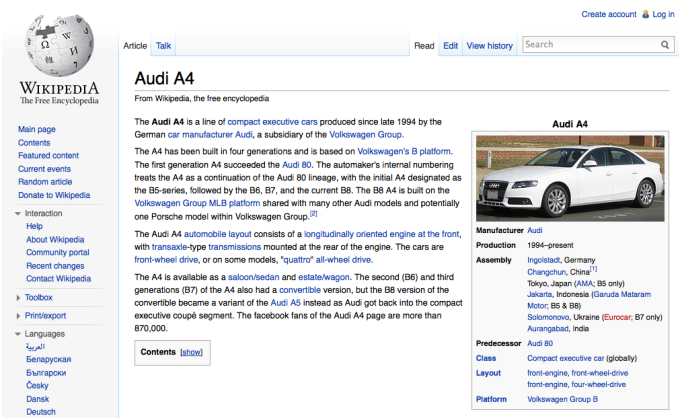
- INEX Entity Ranking track (2007-09)
  - Wikipedia
- TREC Entity track (2009-11)
  - Web crawl, including Wikipedia
- Semantic Search Challenge (2010-11)
  - BTC2009 (Semantic Web crawl, ~1 billion RDF triples)
- INEX Linked Data track (2012-13)
  - Wikipedia enriched with RDF properties from DBpedia and YAGO

Common denominator: DBpedia

bit.ly/dbpedia-entity

Query set	#q (orig)	#q (used)	avg q_len	avg #rel
INEX-XER "US presidents since 1960"	55	55	5,5	29,8
TREC Entity "Airlines that currently use Boeing 747 planes"	20	17	6,7	13,1
SemSearch ES "Ben Franklin"	142	130	2,7	8,7
SemSearch LS "Axis powers of World War II"	50	43	5,4	12,5
QALD-2 "Who is the mayor of Berlin?"	200	140	7,9	41,5
INEX-LD "England football player highest paid"	100	100	4,8	37,6
Total		485	5,3	27

Balog and Neumayer (2013). A Test Collection for Entity Retrieval in DBpedia. **In SIGIR '13.**



Semi-structured data

```
dbpedia:Audi_A4
```

```
foaf:name                          Audi A4
rdfs:label                         Audi A4
rdfs:comment                       The Audi A4 is a compact executive car
                                   produced since late 1994 by the German car
                                   manufacturer Audi, a subsidiary of the
                                   Volkswagen Group. The A4 has been built [...]
dbpprop:production                 1994
                                   2001
                                   2005
                                   2008
rdf:type                           dbpedia-owl:MeanOfTransportation
                                   dbpedia-owl:Automobile
dbpedia-owl:manufacturer           dbpedia:Audi
dbpedia-owl:class                  dbpedia:Compact_executive_car
owl:sameAs                         freebase:Audi A4
is dbpedia-owl:predecessor of      dbpedia:Audi_A5
is dbpprop:similar of              dbpedia:Cadillac_BLS
```

Baseline models

- Standard document retrieval methods applied on entity description documents
- E.g., language modeling (MLM)

$$P(e|q) \propto P(e)P(q|\theta_e) = \underbrace{P(e)}_{\text{Entity prior}} \prod_{t \in q} \underbrace{P(t|\theta_e)}_{\text{Entity language model}}{}^{n(t,q)}$$

Entity prior
Probability of the entity being relevant to *any* query

Entity language model
Multinomial probability distribution over the vocabulary of terms

Fielded models

- Fielded extensions of document retrieval methods
- E.g., Mixture of Language Models (MLM)

$$P(t|\theta_d) = \sum_{j=1}^{m} \underbrace{\mu_j}_{\text{Field weights}} P(t|\underbrace{\theta_{d_j}}_{\text{Field language model}})$$

Field weights
$\sum_{j=1}^{m} \mu_j = 1$

Field language model
Smoothed with a collection model built from all document representations of the same type in the collection

Setting field weights

- Heuristically
  - Proportional to the length of text content in that field, to the field's individual performance, etc.
- Empirically (using training queries)
- Problems
  - Number of possible fields is huge
    - It is not possible to optimise their weights directly
  - Entities are sparse w.r.t. different fields
    - Most entities have only a handful of predicates

Predicate folding

- **Idea**: reduce the number of fields by grouping them together based on, e.g., type

Name	foaf:name	Audi A4
	rdfs:label	Audi A4
Attributes	rdfs:comment	The Audi A4 is a compact executive car produced since late 1994 by the German car manufacturer Audi, a subsidiary of the Volkswagen Group. The A4 has been built [...]
	dbpprop:production	1994 2001 2005 2008
Out-relations	rdf:type	dbpedia-owl:MeanOfTransportation dbpedia-owl:Automobile
	dbpedia-owl:manufacturer	dbpedia:Audi
	dbpedia-owl:class	dbpedia:Compact_executive_car
	owl:sameAs	freebase:Audi A4
In-relations	is dbpedia-owl:predecessor of	dbpedia:Audi_A5
	is dbpprop:similar of	dbpedia:Cadillac_BLS

Probabilistic Retrieval Model for Semistructured data (PRMS)

- Extension to the Mixture of Language Models
- Find which document field each query term may be associated with

$$P(t|\theta_d) = \sum_{j=1}^{m} \mu_j P(t|\theta_{d_j})$$

Mapping probability
Estimated for each query term

$$P(t|\theta_d) = \sum_{j=1}^{m} P(d_j|t) P(t|\theta_{d_j})$$

Kim, Xue, and Croft (2009). Probabilistic Retrieval Model for Semistructured data. In ECIR'09.

Estimating the mapping probability

$$P(t|C_j) = \frac{\sum_d n(t, d_j)}{\sum_d |d_j|}$$

Term likelihood
Probability of a query term occurring in a given field type

Prior field probability
Probability of mapping the query term to this field before observing collection statistics

$$P(d_j|t) = \frac{P(t|d_j)P(d_j)}{P(t)}$$

$$P(t) = \sum_{d_k} P(t|d_k)P(d_k)$$

Example (IMDB)

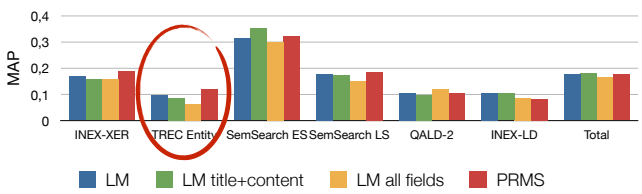
meg ryan war

d_j	$P(t\|d_j)$
cast	0,407
team	0,382
title	0,187

d_j	$P(t\|d_j)$
cast	0,601
team	0,381
title	0,017

d_j	$P(t\|d_j)$
genre	0,927
title	0,07
location	0,002

Baseline results



Related entity finding

- TREC Entity track (2009-2011)
- **Input**: natural language query
- **Output**: ranked list of entities
- **Collection**: combination of unstructured (web crawl) and structured (DBpedia)
- Queries contain an entity (E), target type (T), and a required relation (R)
  - Entity and target type are annotated in the query

Examples

airlines that currently use Boeing 747 planes
target type — input entity

Members of The Beaux Arts Trio
target type — input entity

What countries does Eurail operate in?
target type — input entity

Modeling related entity finding

- Exploiting the available annotations in a three-component model

$$p(e|E,T,R) \propto \underbrace{p(e|E)}_{\text{Co-occurrence model}} \cdot \underbrace{p(T|e)}_{\text{Type filtering}} \cdot \underbrace{p(R|E,e)}_{\text{Context model}}$$

Bron, Balog, and de Rijke (2010). Ranking Related Entities: Components and Analyses. In CIKM '10.

Examples

Can we obtain such annotations automatically?

airlines that currently use Boeing 747 planes
target type — input entity

Members of The Beaux Arts Trio
target type — input entity

What countries does Eurail operate in?
target type — input entity

2 Understanding Queries

Wikification

- Recognizing concepts in text and linking them to the corresponding entries in Wikipedia

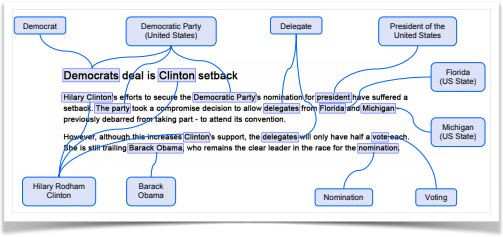


Image taken from Milne and Witten (2008b). Learning to Link with Wikipedia. In CIKM '08.

Entity linking

- Typically only named entities are annotated
- Reference KB can be different from Wikipedia
- Usages
  - Improved retrieval, enabling semantic search
  - Advanced UX/UI, help users to explore
  - Knowledge base population/acceleration

Entity linking methods

- Mention detection
  - Identifying entity mentions in text
- Candidate entity ranking
  - Generating a set of candidate entries from the KB for each mention
- Disambiguation
  - Selecting the best entity for a mention
  - Machine learning; features: commonness, relatedness, context, …

Entity linking evaluation

ground truth $\hat{A}$

Košice is the biggest city in eastern **Slovakia** and in 2013 was the **European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak Ore Mountains**, near the border with **Hungary**.

system annotation A

Košice is the biggest city in eastern **Slovakia** and in 2013 was **the European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak** Ore Mountains, near the border with **Hungary**.

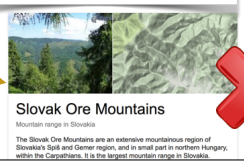
Entity linking evaluation

ground truth $\hat{A}$

Košice is the biggest city in eastern **Slovakia** and in 2013 was the **European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak Ore Mountains**, near the border with **Hungary**.

system annotation A

Košice is the biggest city in eastern **Slovakia** and in 2013 was **the European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak** Ore Mountains, near the border with **Hungary**.



Entity linking evaluation

ground truth $\hat{A}$

Košice is the biggest city in eastern **Slovakia** and in 2013 was the **European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak Ore Mountains**, near the border with **Hungary**.

system annotation A

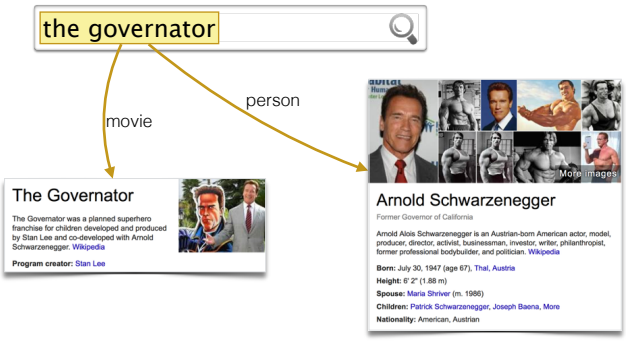
Košice is the biggest city in eastern **Slovakia** and in 2013 was **the European Capital of Culture** together with **Marseille**, **France**. It is situated on the river **Hornád** at the eastern reaches of the **Slovak** Ore Mountains, near the border with **Hungary**.

$$P = \frac{|A \bigcap \hat{A}|}{|A|} \qquad R = \frac{|A \bigcap \hat{A}|}{|\hat{A}|} \qquad F = \frac{2 \cdot P \cdot R}{P + R}$$

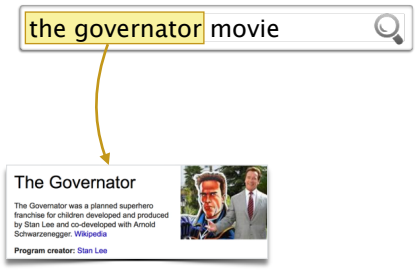
Entity linking for queries

- Challenges
  - search queries are short
  - limited context
  - lack of proper grammar, spelling
  - multiple interpretations
  - needs to be fast

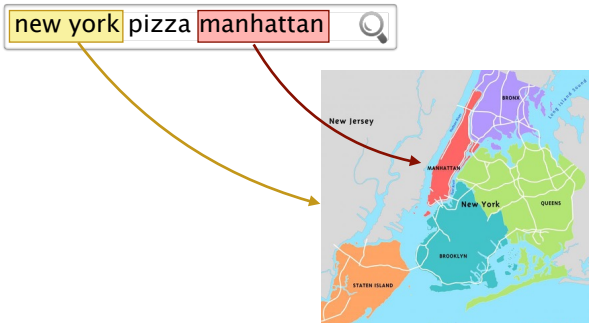
Example



Example



Example



Example

ERD'14 challenge

- **Task**: finding query interpretations
- **Input**: keyword query
- **Output**: sets of sets of entities
- **Reference KB**: Freebase
- Annotations are to be performed by a web service within a given time limit

Evaluation

new york pizza manhattan

ground truth $\hat{I}$

New York City, **Manhattan**

New York-style pizza, **Manhattan**

system annotation I

New York City, **Manhattan** ✓

New York-style pizza ✗

$$P = \frac{|I \bigcap \hat{I}|}{|I|} \qquad R = \frac{|I \bigcap \hat{I}|}{|\hat{I}|} \qquad F = \frac{2 \cdot P \cdot R}{P + R}$$

ERD'14 results

Rank	Team	F1	latency
1	SMAPH Team	0.7076	0.49
2	NTUNLP	0.6797	1.04
3	Seznam Research	0.6693	3.91

Single interpretation is returned

http://web-ngram.research.microsoft.com/erd2014/LeaderBoard.aspx

3

Living Labs for IR Evaluation

Living Labs for IR evaluation

http://living-labs.net
@livinglabsnet

Goals & focus

- **Overall goal**: make information retrieval evaluation more realistic
  - Evaluate retrieval methods in a *live setting* with *real users* in their *natural task environments*
- **Focus**: medium-sized organizations with fair amount of search volume
  - Typically lack their own R&D department, but would gain much from improved approaches

What is in it for participants?

- Access to privileged commercial data
  - (Search and click-through data)
- Opportunity to test IR systems with real, unsuspecting users in a live setting
  - (Not the same as crowdsourcing!)

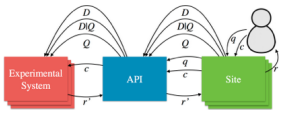
"Give us your ranking, we'll have it clicked!"

Research aims

- Understanding of online evaluation and of the generalization of retrieval techniques across different use cases
- Specific research questions
  - Are system rankings different when using historical clicks from those using online experiments?
  - Are system rankings different when using manual relevance assessments ("expert judgments") from those using online experiments?

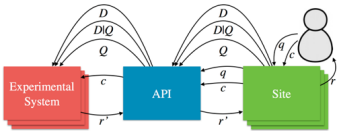
Key idea

- Focus on frequent (head) queries
  - Enough traffic on them (both real-time and historical)
  - Ranked result lists can be generated offline
- An API orchestrates all data exchange between live sites and experimental systems



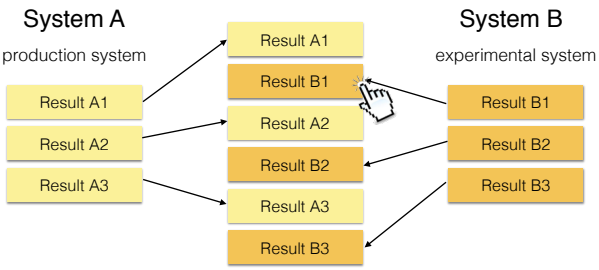
Balog, Kelly, and Schuth (2014). Head First: Living Labs for Ad-hoc Search Evaluation. In CIKM '14.

Methodology



- Queries, candidate documents, historical search and click data made available
- Rankings are generated for each query and uploaded through an API
- When any of the test queries is fired, the live site request rankings from the API and interleaves them with that of the production system
- Participants get detailed feedback on user interactions (clicks)
- Ultimate measure is the number of "wins" against the production system

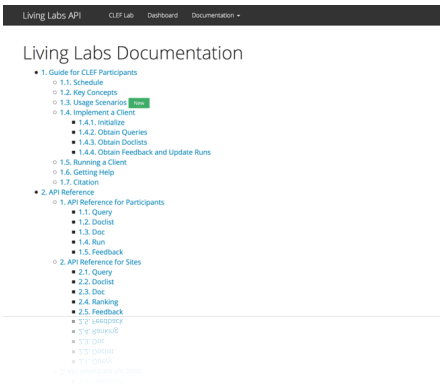
Interleaving



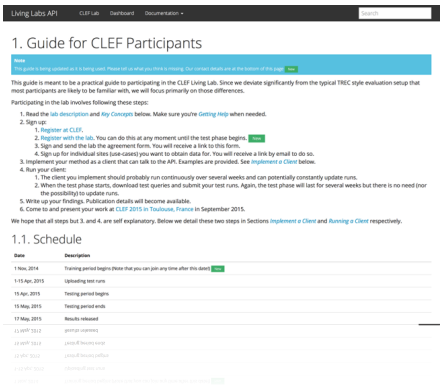
Partners & Use-cases

	Product search	Local domain search	Web search
Provider	regiojatek.hu	uva.nl	seznam.cz
Data	raw queries and (highly structured) documents	raw queries and (generally textual) documents	pre-computed document-query features
Site traffic	relatively low (~4K sessions/day)	relatively low	high
Info needs	(mostly) transactional	(mostly) navigational	vary

API doc: doc.living-labs.net



Guide for CLEF participants

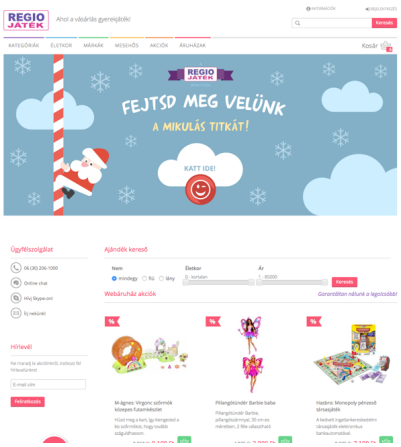


Partners & Use-cases

	Product search	Local domain search	Web search
Provider	regiojatek.hu	uva.nl	seznam.cz
Data	raw queries and (highly structured) documents	raw queries and (generally textual) documents	pre-computed document-query features
Site traffic	relatively low (~4K sessions/day)	relatively low	high
Info needs	(mostly) transactional	(mostly) navigational	vary

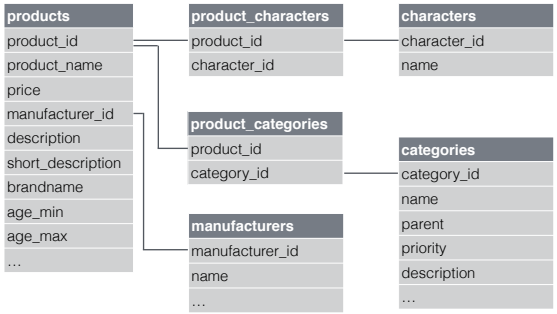
Product search

- Ad-hoc retrieval over a product catalog
- Several thousand products
- Limited amount of text, lots of structure
  - Categories, characters, brands, etc.



Homepage of the year 2014
in e-commerce category

Relational database



Logs

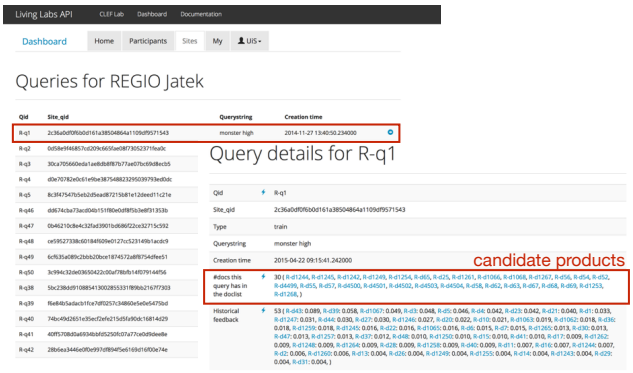
timestamp session_id event event attributes

```
1402052038;7902b47fbbd45360a672670367cccf8d;IDX;28178,28188,33533,36450,32394,34188,76395,856
1402052039;4561d3d945d8981e177eac00f6025809;QRY;kisvakond puzzle;;
    1;47702,07921,80875,36419,65736,09726,85683,09679,37294,15847,80878,00131,03994,17713
1402052040;fd593d671402dd5ea21e39d00aa30e26;PRD;16402
1402052040;fd593d671402dd5ea21e39d00aa30e26;PRC;16402;1;80875,36419,65736,09726,85683,09679,
1402052041;fd593d671402dd5ea21e39d00aa30e26;IDX;34188,34561,34543,30649,27747,28188,28178,76
1402052043;44ee8beae78426b0eb5662d25191984c;PRD;71983
1402052043;44ee8beae78426b0eb5662d25191984c;PRC;71983;1;88986,73681,49141,02215,85632,85633,
1402052043;897deef88cfd77d59cf2eebe6c9cab43;CAT;12;9;07975,07976,76715,99293,88844,64632,6452
1402052044;dc1db4bef11f027a0a8d161b8354a318;IDX;00841,34543,76395,28178,34188,33039,33533,33
1402052045;bb3a12f50e5c1a53765885a90a8d8992;IDX;00841,28188,27747,76395,30649,28178,34188,36
```

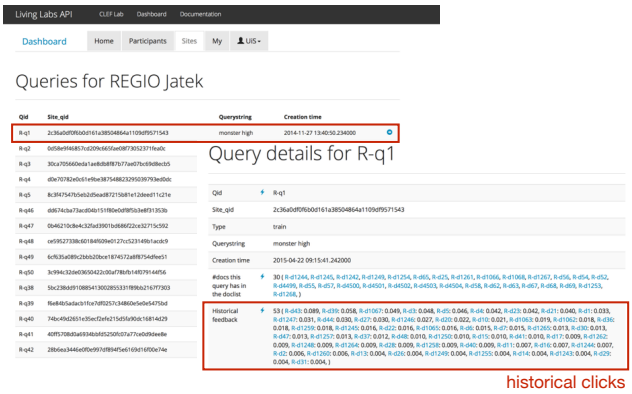
…

```
1402052039;4561d3d945d8981e177eac00f6025809;QRY;kisvakond puzzle;;
    2;00132,13867,64253,13866,64252,69406,00127,36152,69405,14792,92991,00126,15758,20313
1402052068;4561d3d945d8981e177eac00f6025809;QRY;puzzle;gender:0|category_id:0|price:1001-200
    bonus:;1;47702,09679,36152,01918,19991,19989,06633,19995,99888,51805,36150,81522,05655,6
1402052076;4561d3d945d8981e177eac00f6025809;QRY;puzzle;gender:0|category_id:0|price:1001-200
    bonus:;8;32497,28307,08246,00064,32471,27995,34018,32491,61540
```

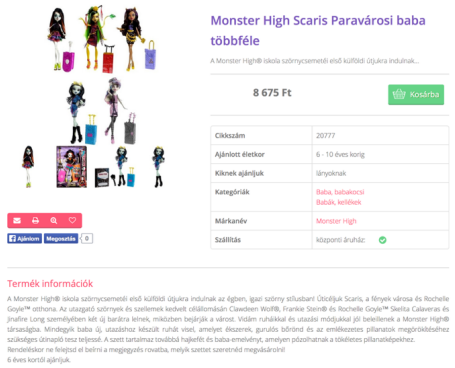
Test queries



Test queries



Product data




```
{
  "content": {
    "age_max": 10,
    "age_min": 6,
    "arrived": "2014-08-28",
    "available": 1,
    "brand": "Mattel",
    "category": "Bab\u00e1k, kell\u00e9kek",
    "category_id": "25",
    "characters": [],
    "description": "A Monster High\u00ae iskola sz\u00f6rnycsemet\u00e9i [...]",
    "gender": 2,
    "main_category": "Baba, babakocsi",
    "main_category_id": "3",
    "photos": [
      "http://regiojatek.hu/data/regio_images/normal/20777_0.jpg",
      "http://regiojatek.hu/data/regio_images/normal/20777_1.jpg",
      [...]
    ],
    "price": 8675.0,
    "product_name": "Monster High Scaris Parav\u00e1rosi baba t\u00f6bbf\u00e9le",
    "queries": {
      "clawdeen": "0.037",
      "monster": "0.222",
      "monster high": "0.741"
    },
    "short_description": "A Monster High\u00ae iskola sz\u00f6rnycsemet\u00e9i
      els\u0151 k\u00fcl\u00f6ldi \u00fatjukra indulnak..."
  },
  "creation_time": "Mon, 11 May 2015 04:52:59 -0000",
  "docid": "R-d43",
  "site_id": "R",
```


frequent queries that led to the product

Results

Method	impr. total	impr. per query	#clicks	CTR
1	1884	18.84	581	0.31
2	1407	14.07	630	0.45
3	1334	13.34	483	0.36

Join us!

living-labs.net

Living Labs for IR evaluation

- Evaluation will continue to run even after the CLEF deadline
- Additional use-cases are welcome

Summary

- Entities as the unit of retrieval
- Understanding queries with the help of entities
- From traditional test-collection based evaluations to *evaluation as a service*

Questions?

Contact | @krisztianbalog | krisztianbalog.com