Further plans and available data sets for research in directed networks

Andras Benczur

Insitute for Computer Science and Control

Hungarian Academy of Sciences

benczur@sztaki.mta.hu

http://datamining.sztaki.hu

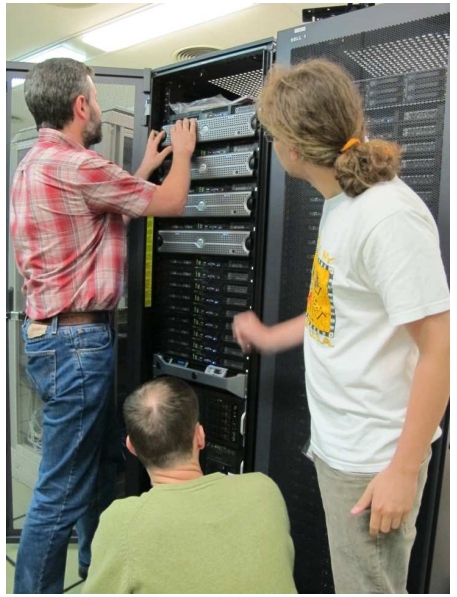
Supported by the EC FET Open project "New tools and algorithms for directed network analysis" (NADINE No 288956)

Overview

- Web classification, ClueWeb12
- Temporal ranking, learning to rank
- Metadata extraction from pdf publications
- Plagiarism Detection
- Twitter: 1TBdata available, user graph collection in progress for Andreas' data
- Distributed systems for very large problems

Hardware

- 50-node old dual core Hadoop
- 5-node new Hadoop/HBASE
- 260TB net Isilon



Automatic metadata extraction

- Careful selection of open source PDF converters
- Feature generation
  - font size, face, upper/lower case, numeric characters, symbols
  - location (centered, vertical position, spacing, page number)
  - entity list (names, institutions)
- Manual training for a Hungarian journal in Economics
- Automatic training planned by using publication DBs
- Selection of machine learning methods
  - Random forest is best, LogitBoost with trees is second best
  - Conditional random fields sound nice but not nearly as good as claimed
- Extraction depends of what we can train (manually label)
  - Author, title, institution
  - References extracted structured
  - Tables, figure captions
  - …?

Plagiarism detection

- BonFIRE Future Internet Research and Experimentation testbed
- KOPI: A plagiarism detection toolkit
  - http://kopi.sztaki.hu/
  - Translation plagiarism (English and Hungarian)
  - Now serving English Wikipedia
  - Service puts very heavy load on search index (sentence based checks, existing suboptimal code)
  - Index ported to several distributed key-value stores
  - We feed with Web data


BBC News Sport Weather Travel Future

NEWS EUROPE

Home UK Africa Asia Europe Latin America Mid-East US & Canada Business Health Sci/Environ

2 April 2012 Last updated at 14:02 GMT

Share

Hungary President Schmitt quits in plagiarism scandal

Hungary's President Pal Schmitt says he is resigning, after being stripped of his doctorate over plagiarism.

Mr Schmitt, elected in 2010, said "my personal issue divides my beloved nation rather than unites it".

"It is my duty to end my service and resign my mandate as president," he told parliament.

Last week, Budapest's Semmelweis University revoked his 1992 award after finding that much of his thesis had been copied.

Mr Schmitt, 69, won gold medals for fencing at the 1968 and 1972 Olympic Games.

Mr Schmitt was an Olympic fencing champion before his rise in politics

Crosslingual Web Classification

INTERNET MEMORY FOUNDATION

HANZO ARCHIVES

- Save resources, select quality and topic
- Legal regulation (porn, illicit content)
- Web scale data (Test: ClueWeb09 25TB – 0.5 Billion English language docs)
- We just obtained ClueWeb12

Cross-Lingual Web Spam Classification.
Garzó, Daróczy, Kiss, Siklósi, Benczúr.
WebQuality 2013 (@WWW)
The classication power of Web features.
Erdelyi, Benczur, Daroczy, Garzo, Kiss, Siklosi
Internet Mathematics, under revision

Julien Masanes

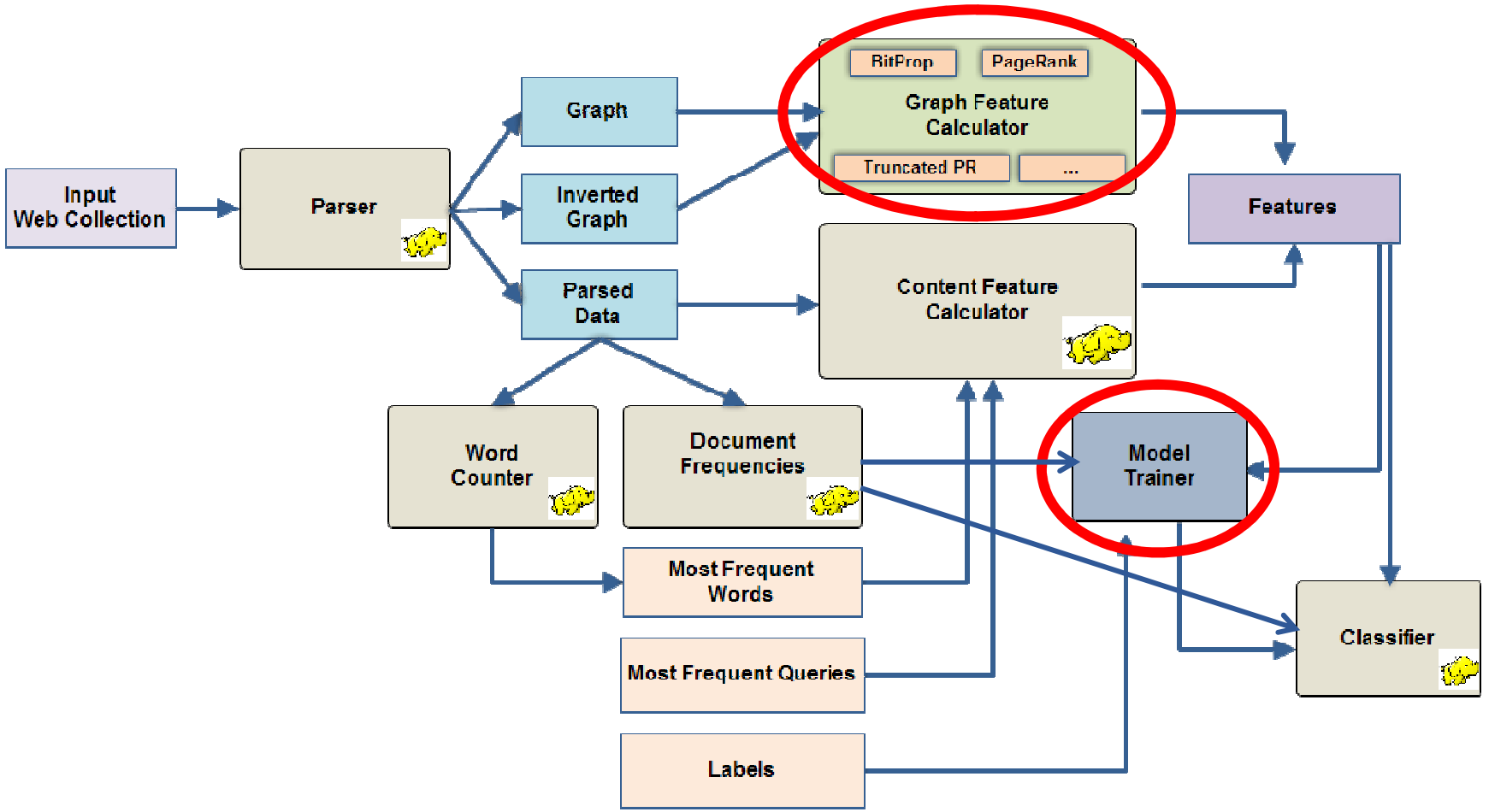
Philippe Rigaux

Internet Memory Paris

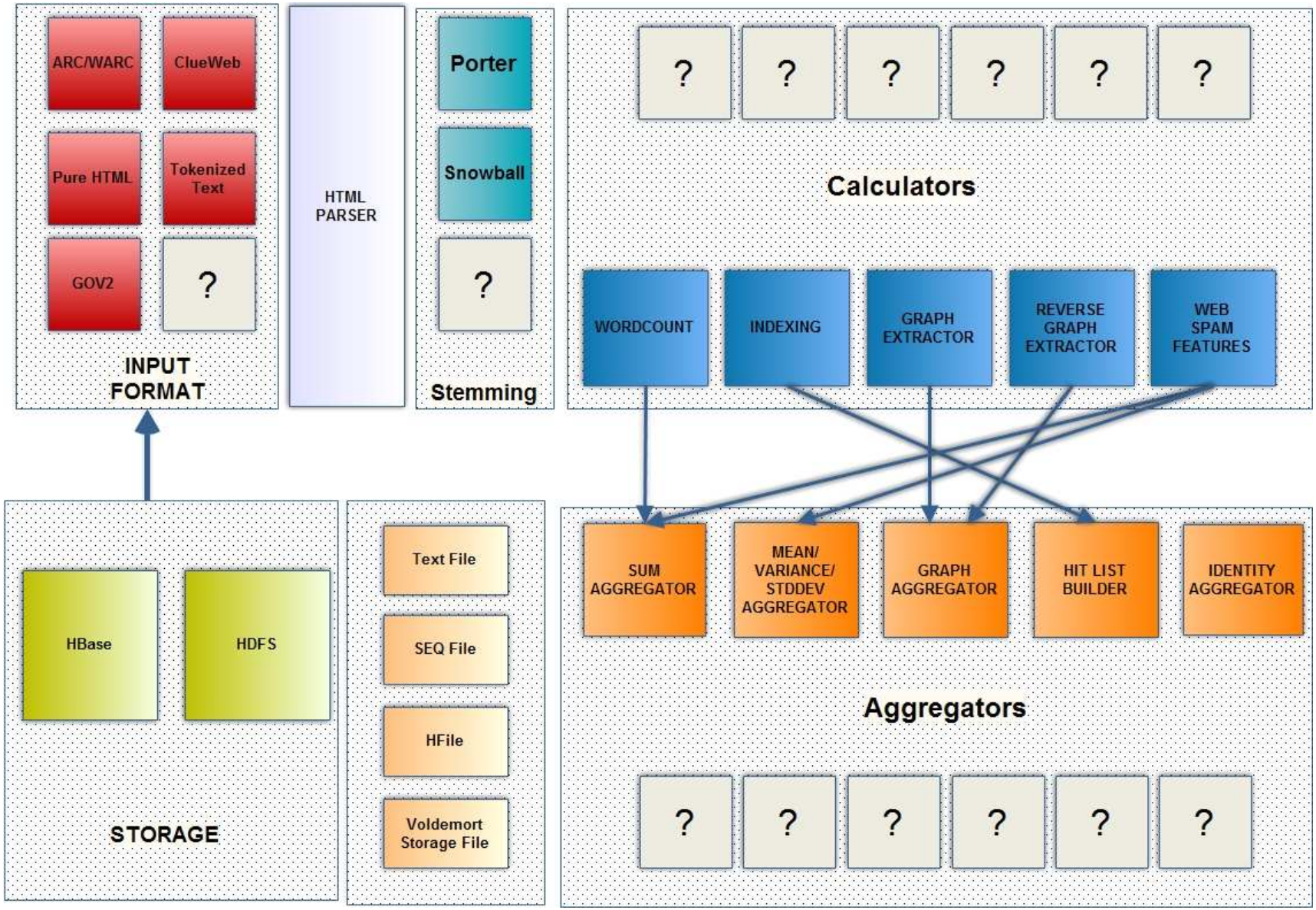
Large set of features

- Term frequency
  - tf.idf or BM25 scores for frequent terms
- Content
  - DOM, HTML, HTTP elements
  - Appearance of popular terms
  - Term, n-gram statistics, compressibility
- Linkage
  - PageRank (truncated variants; ratios)
  - Neighborhood (only approximate counting is possible)
  - TrustRank

Workflow (MapRed jobs indicated)

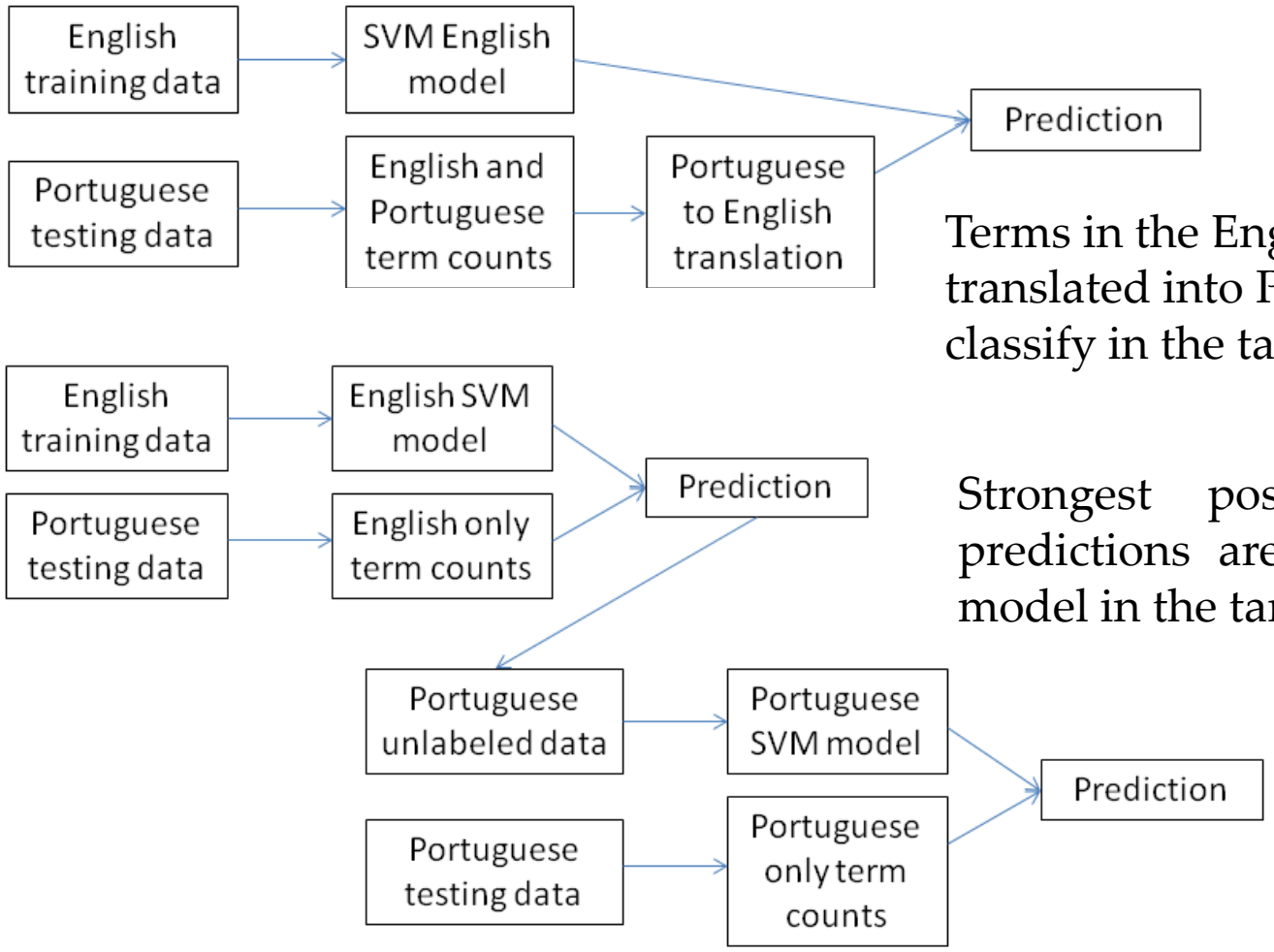


SZTAKI Web Processing Framework



Crosslingual Web Classification

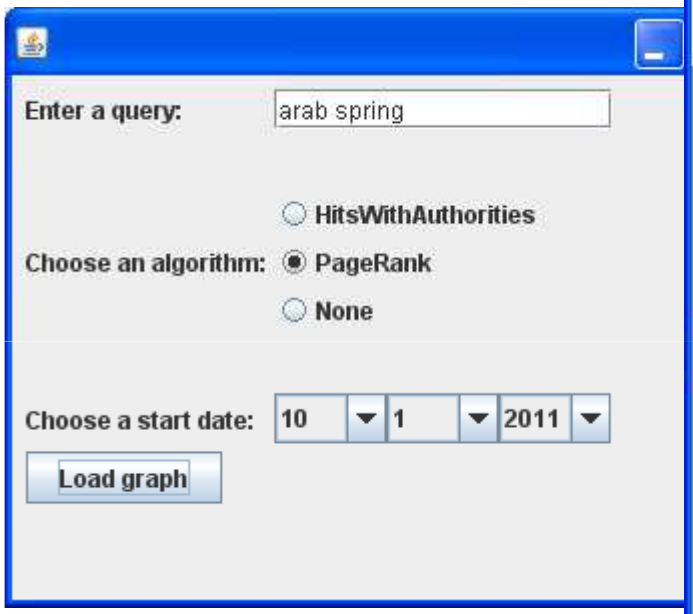
- Expensive human labeling task language by language?
- How can models be "translated"?

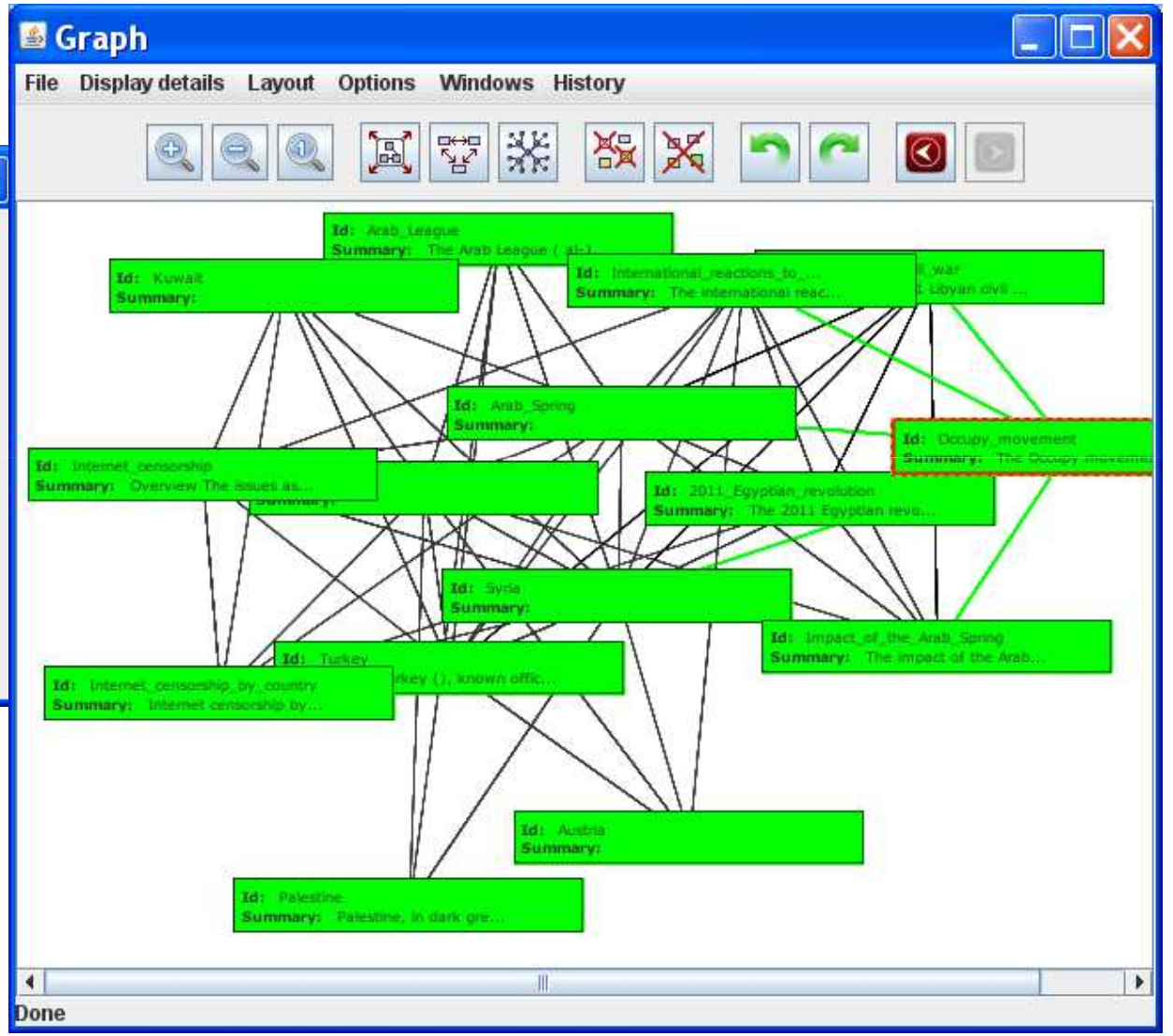


Terms in the English model translated into Portuguese to classify in the target language.

Strongest positive and negative predictions are used for training a model in the target language.

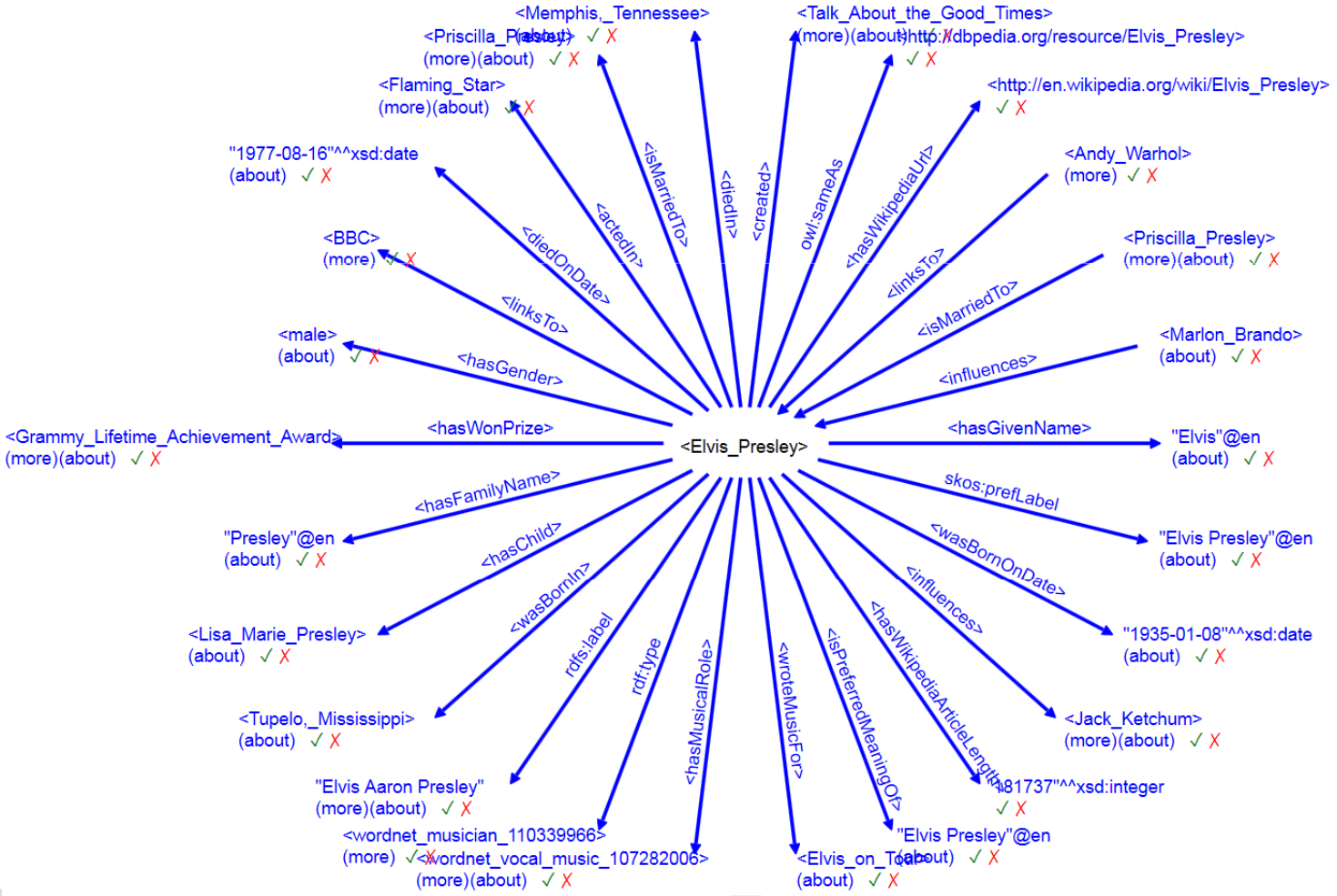
Temporal Wikipedia Search (Julianna)





Yago: Yet Another General Onthology

- By MPII Saarbrücken derived from Wikipedia WordNet and GeoNames
- 10+ million entities (persons, organizations, cities), 120+ million facts
- We are developing similar visualization as Wikipedia (prev slide)



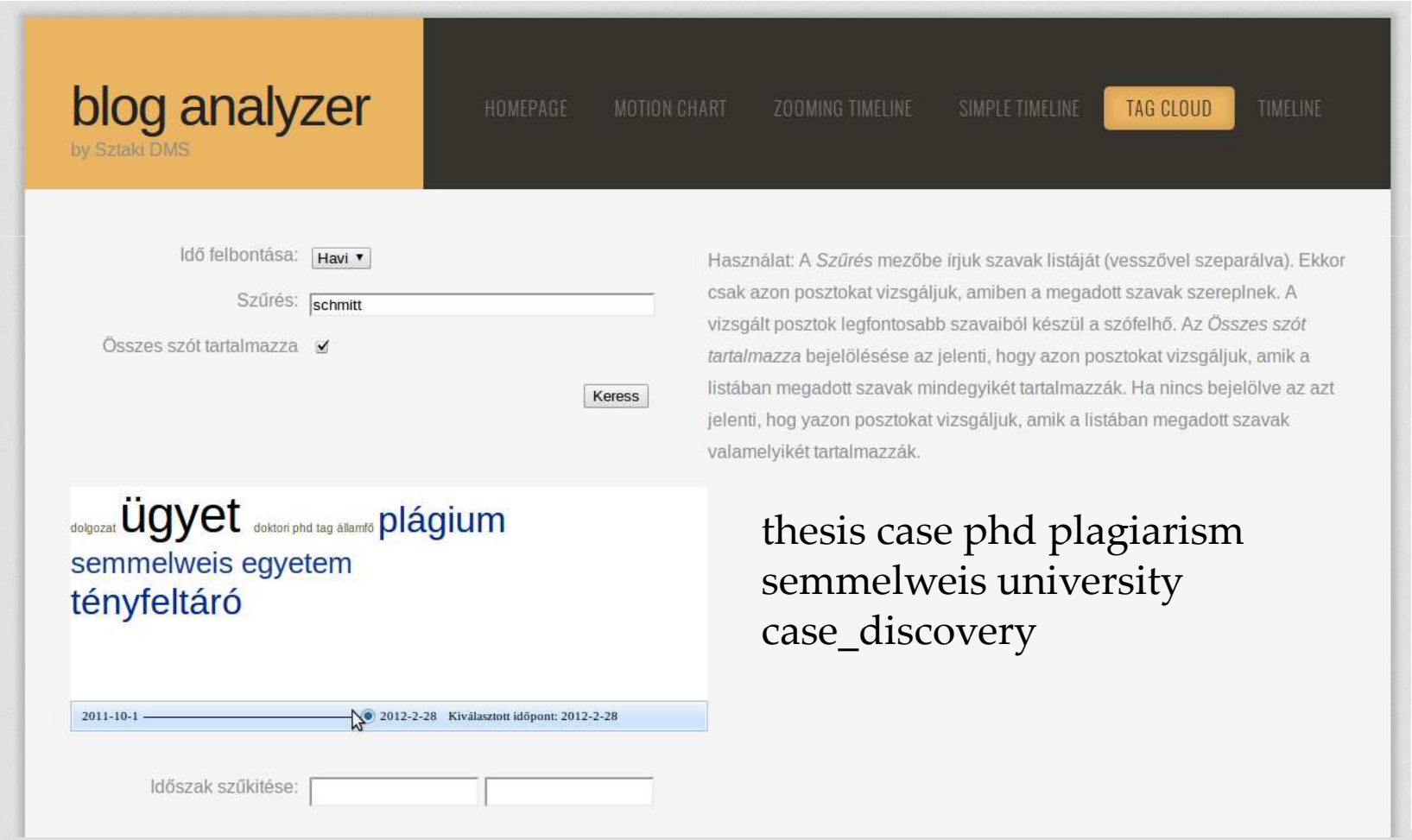
From http://yago-knowledge.org
Give us feedback:
✓ = fact is correct
✗ = fact is false

Temporal trends in blog data

Liberation_war economic
promise those engine
this_year in_effect fulfill

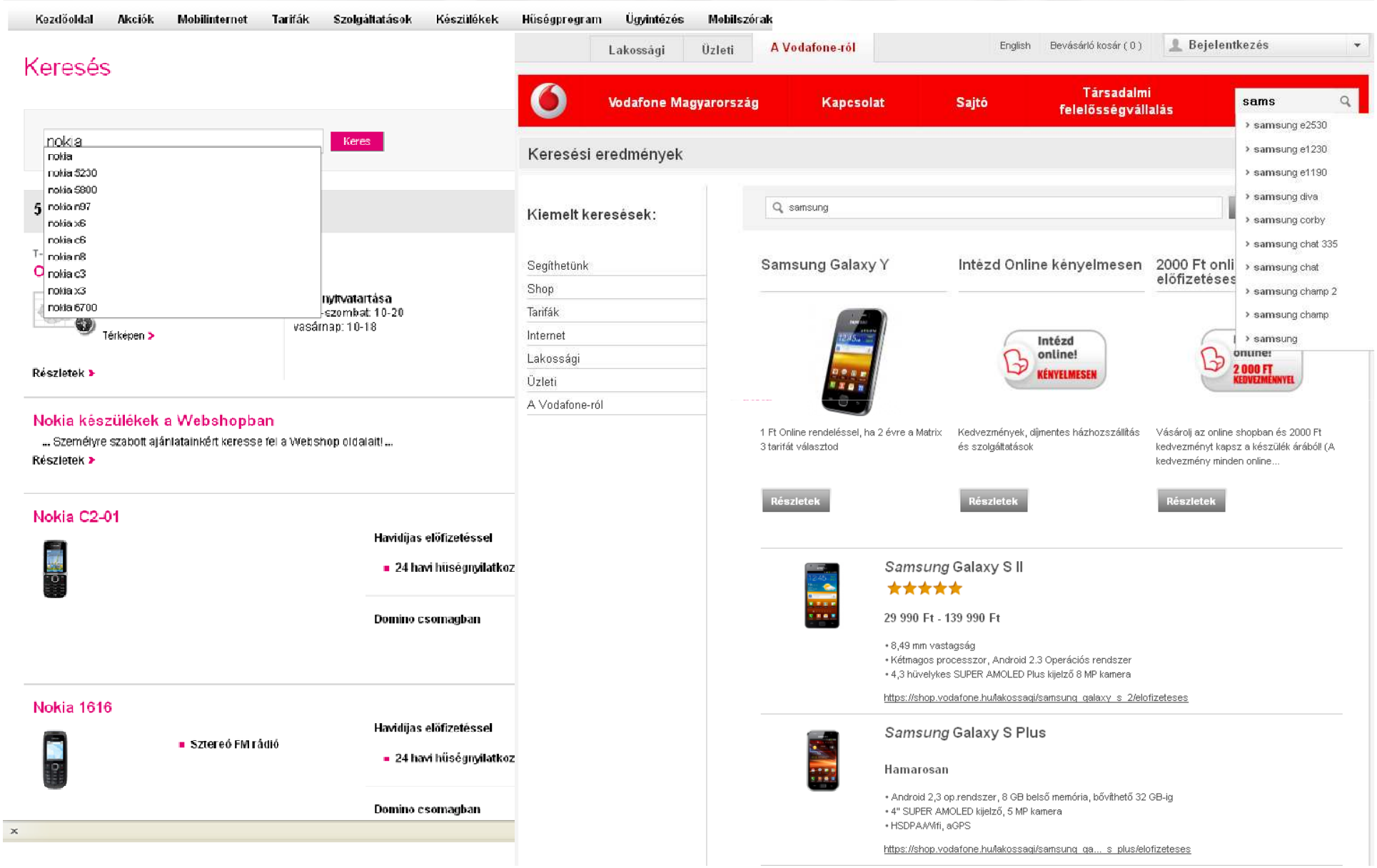
Temporal trends in blog data

- Temporal Text Mining: probabilistic models, language models
- Still in progress, challenging algorithmic issues

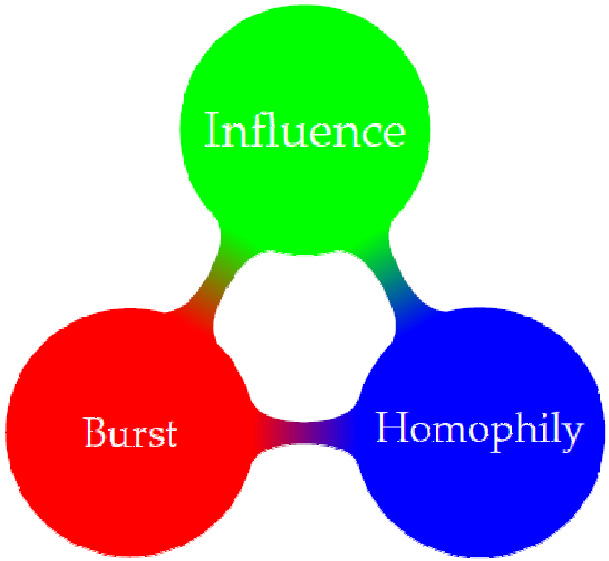


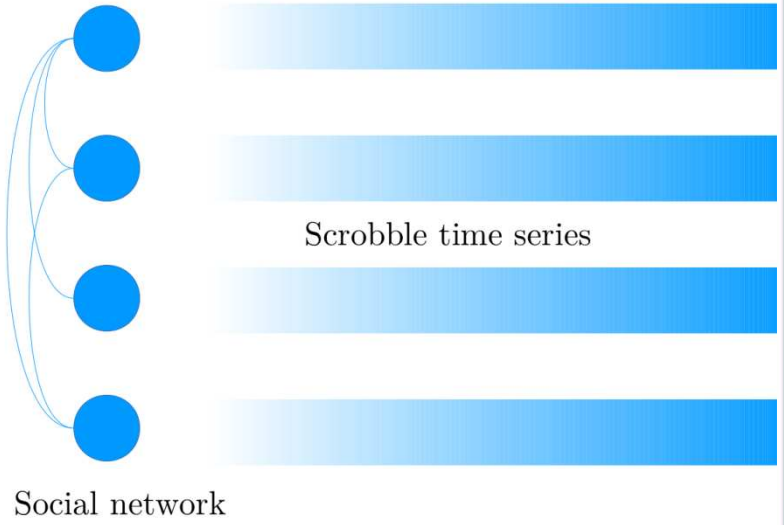
thesis case phd plagiarism semmelweis university case_discovery

SZTAKI Full Text Search Technology

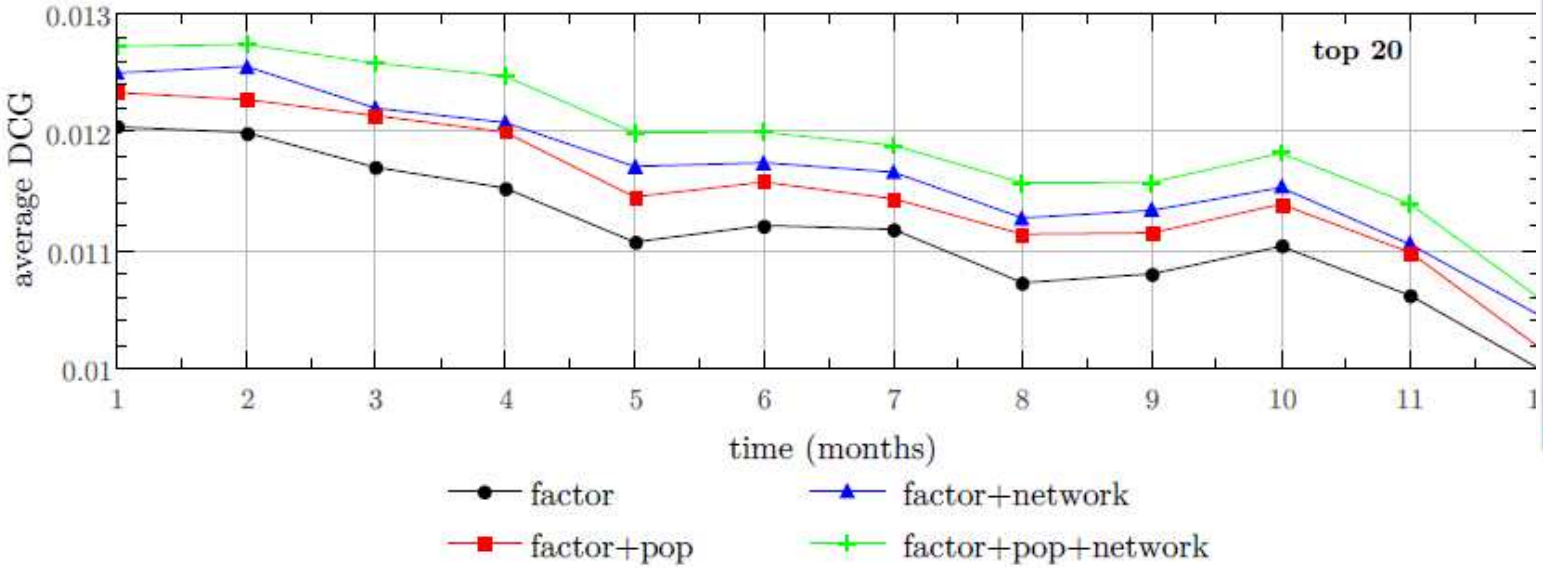


Network Influence in Recommenders





last.fm



Apply for Twitter: retweets

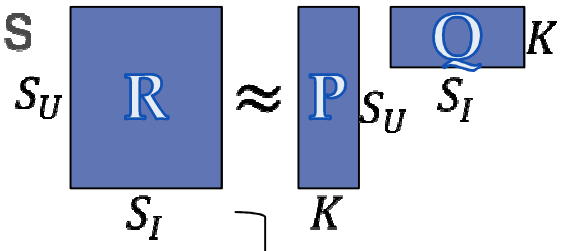
- Twitter data:
  - topics (~bursts: occupy wall street ....)
  - Andreas has 4 topics ("10o","occupy","20n","yosoy132").
- For all topics we have a set of tweets (can be a retweet)
- In numbers:
  - Follower network: 10^6 users
  - Tweets: ~ 10^5 - 10^6 per topic
- Social network (who follows who) is missing
- Needed since we only know the ROOT of a retweet sequence
- Robert is collecting the network

The Matrix Factorization recommender

- Model
  - How we approximate user preferences
  - $\hat{r}_{u,i} = {p_u}^T q_i$
- Objective function (error function)
  - What we want to minimize or optimize?
  - E.g. optimize for RMSE with regularization

$$L = \sum_{(u,i)\in Train} \left(\hat{r}_{u,i} - r_{u,i}\right)^2 + \lambda_U \sum_{u=1}^{S_U} \|P_u\|^2 + \lambda_I \sum_{i=1}^{S_I} \|Q_i\|^2$$

- Learning method
  - How we improve the objective function?
  - E.g. stochastic gradient descent (SGD)



Learning

Source of next slides:
Domonkos Tikk, CEO, Gravity

BRISMF model

- Biased Regularized Incremental Simultaneous Matrix Factorization
- Apply regularization to prevent overfitting
- To further decrease RMSE using bias values
- Model:

$$\hat{r}_{ui} = \vec{p}_u \vec{q}_i + b_u + c_i = \sum_{k=1}^{K} p_{uk} q_{ki} + b_u + c_i$$

BRISMF Learning

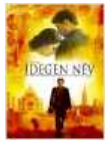
- Loss function

$$\sum_{(u,i)\in R_{train}} \left(r_{ui} - \sum_{k=1}^{K} p_{uk} q_{ki} - b_u - c_i \right)^2 + \lambda \sum_{(u,k)} p_{uk}^2 + \lambda \sum_{(i,k)} q_{ki}^2 + \lambda \sum_{u} b_u^2 + \lambda \sum_{i} c_i^2$$

- SGD update rules

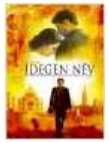
$$\Delta p_{uk} = \eta \left(e_{ui} q_{ki} - \lambda p_{uk} \right) \qquad \Delta q_{ki} = \eta \left(e_{ui} p_{uk} - \lambda q_{ki} \right)$$

$$\Delta b_u = \eta \left(e_{ui} - \lambda b_u \right) \qquad \Delta c_i = \eta \left(e_{ui} - \lambda c_i \right)$$

R					
	1	4		3	
			4	4	
	4		2		4

P	
1,2	-0,4
1,2	0,9
0,5	-0,3

Q				
1,5	0,9	-1,3	-0,1	0.6
-0,2	0,5	-0,4	1,6	0,3

R					
	1	4	3.3	3	2.4
	-0.5	3.5	4	4	1.5
	4	4.9	2	1.1	4

P	
1,4	1,1
0,9	1,9
2,5	-0,3

Q				
1,5	2,1	1,0	0.7	1.6
-1,0	0,8	1,6	1,8	0,0

Influence Learning by Gradient Descent

- Present influence recommender:
  - heuristic weighted network learning
  - no artist based learning part
- Heuristic combination of the influence and factor models
  - Is it likely that user v influences user u on artist a?
  - Can user a be influenced at all in case of artist a?
- Use SGD method to learn user and artist factors

$$\hat{r}_{uat} = \sum_{v} \Gamma(\Delta t)(\vec{p}_v \vec{q}_a + b_v + c_i)$$

Distributed learning?

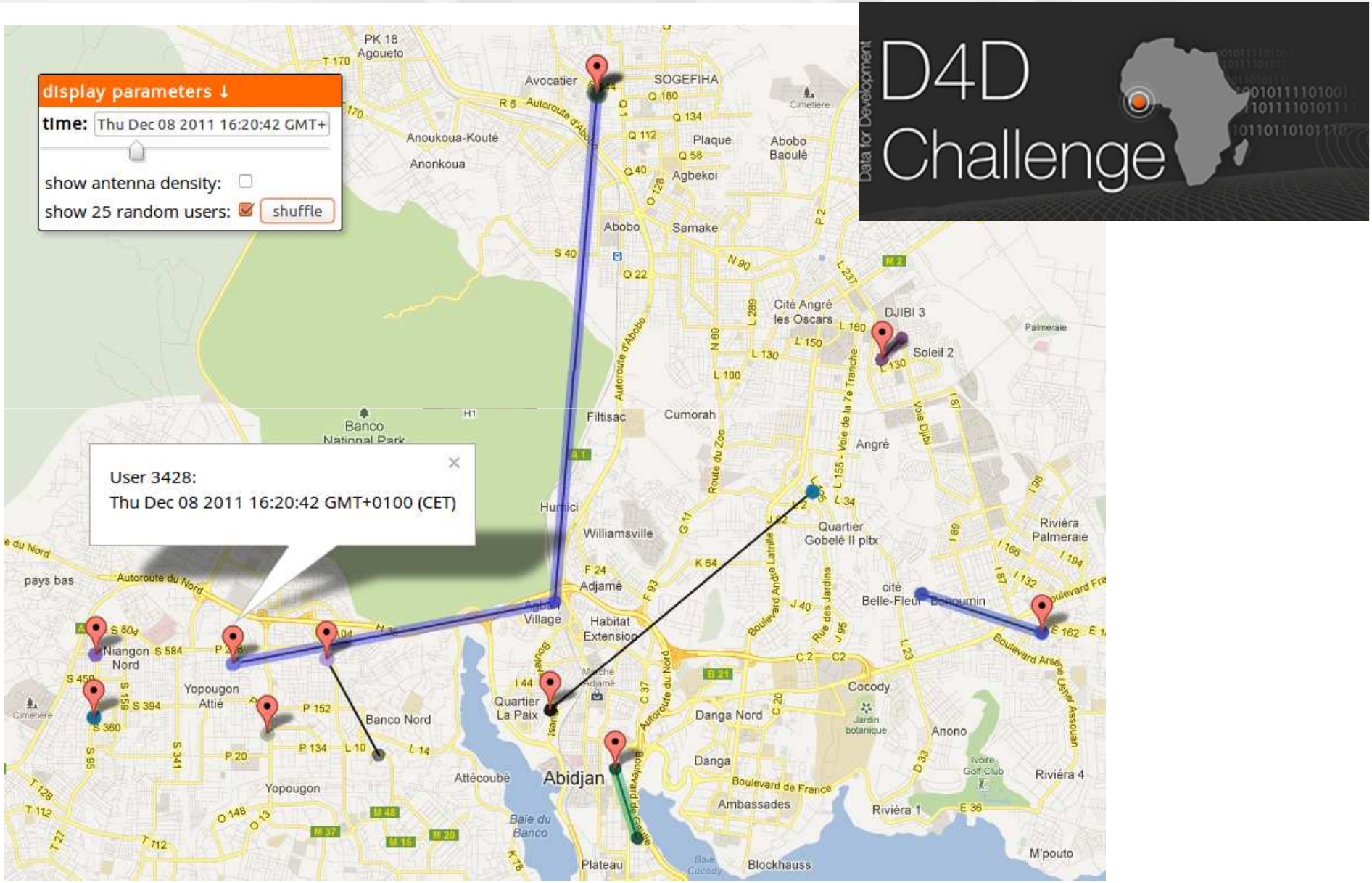
- Hadoop gathered bad reputation recently
  - Wants to be too robust, keep writing all temporal data several times to disk
  - Fails after a given number of servers
  - The learning and graph problems do more computation on less data compared to building a Google search index
- My personal choice of frameworks
  - GraphLab (Danny Bickson, HUJI)
    - Nearly as efficient as possible C++ codes
    - But very hard to write them
    - We work with them on implementing learning-to-rank methods
  - Stratosphere (Volker Markl, Kostas Tzoumas, TU Berlin)
    - Developments coordinated by TU Berlin with lots of partners incl. us
    - Promises to simplify complex workflows like the spam filter
- Yet what many applications need would be
  - Streaming (read data only once, no batch computations)
  - Fully distributed: no Facebook, Google, Netflix knowing each and every online action ever in our life – have P2P learning

A distributed systems comparison slide

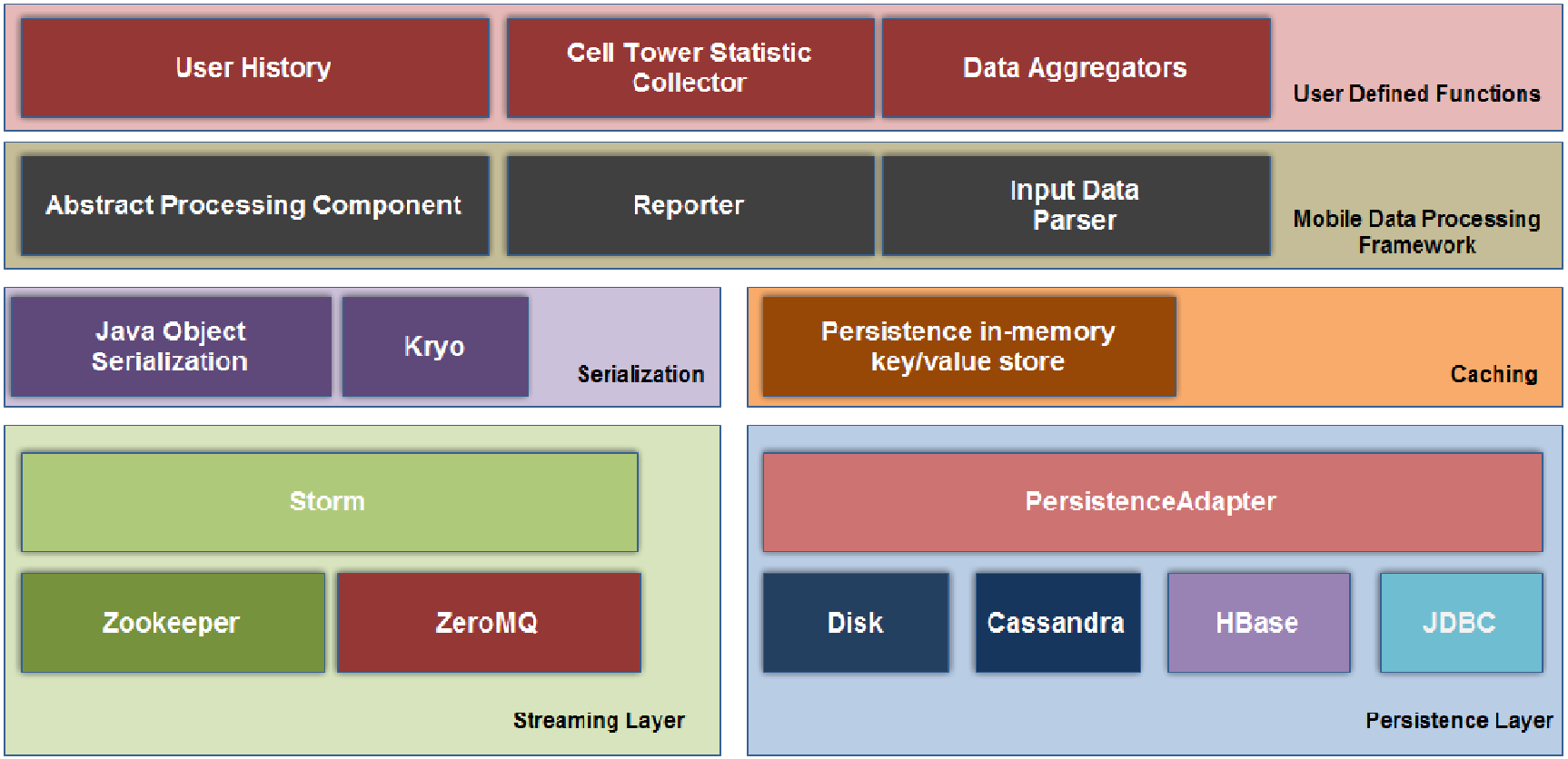
	MapReduce	Pregel	Stratosphere/Naiad	GraphLab
Programming Model	Fixed Functions – Map and Reduce	Supersteps over a data graph with messages passed	Iterative dataflow with operators and UDFs	Data graph with shared data table and update functions
Parallelism	Concurrent execution of tasks within map and reduce phases	Concurrent execution of user functions over vertices within a superstep	Concurrent execution of operators during a stage	Concurrent execution of non-overlapping scopes, defined by consistency model
Data Handling	Distributed file system	Distributed file system	Flexible data channels: Memory, Files, DFS etc.	Undefined – Graphs can be in memory or on disk
Task Scheduling	Fixed Phases – HDFS Locality based map task assignment	Partitioned Graph and Inputs assigned by assignment functions	Job and Stage Managers assign operators to available daemons/tasks	Pluggable schedulers to schedule update functions
Fault Tolerance	DFS replication + Task reassignment / Speculative execution of Tasks	Checkpointing and superstep re-execution	Operators/Task failure recovery	Synchronous and asychronous snapshots
Developed by	Google	Google	TU Berlin / Microsoft	Carnegie Mellon

"Scalable Machine Learning for Big Data" tutorial at ICDE 2012

Mobility Data Stream processing (Orange D4D)



Stream Processing Architecture Overview



Goal is to hide Storm details from user

- Streaming infrastructure pluggable (could combine with Stratosphere)
- Persistence layer pluggable

Conclusions

- Web classification plans to integrate with BUbiNG, use SZTAKI cluster to test the crawler
- Analyze ClueWeb12 and maybe a NADINE crawl?
- Temporal ranking in Wikipedia – other temporal collections?
- Use metadata extraction from online publications to infer topics and rich information that is available in full text only (beyond the usual DBLP graph analysis)
- Network analysis in the plagiarism detection tool?
- Twitter
  - Understand the 1TBdata
  - Find influences in the user graph that we collect for Andreas' data
- Distributed machine learning and graph algorithms

Questions?

András Benczúr
Head,
Informatics Laboratory
and
"Big Data" lab

http://datamining.sztaki.hu/

benczur@sztaki.mta.hu