

M Ű E G Y E T E M 1 7 8 2

Feedback Prediction for Blogs

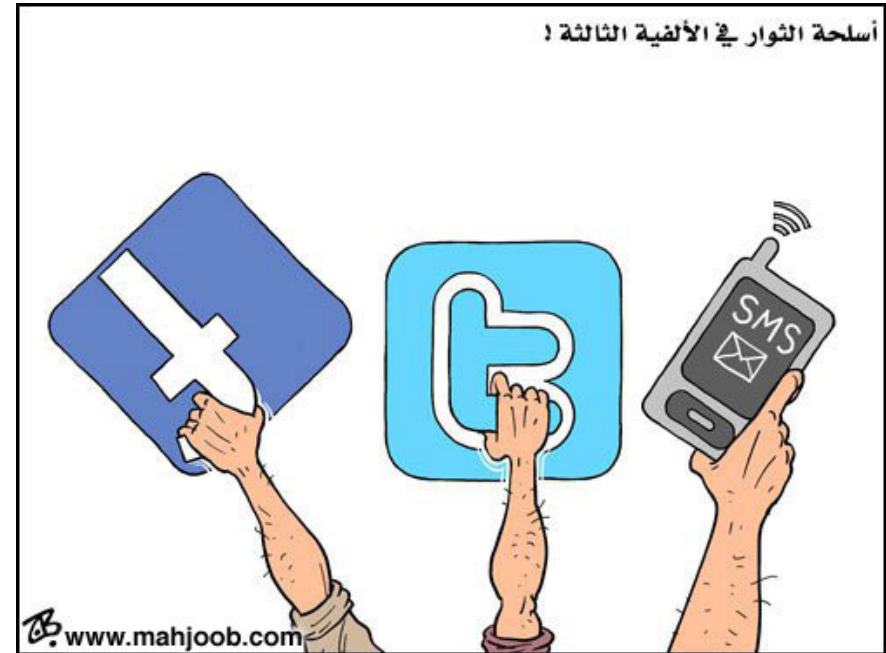
Krisztián Buza

Department of Computer Science and Information Theory
Budapest University of Technology and Economics

buza@cs.bme.hu

Introduction

- Scope
 - data mining in social media
- Goal
 - prediction of relevance of recently-appeared social media entries in the near future (like weather forecasts)
- Major results
 - We developed and tested a proof-of-concept prototype
 - Publication of the collected data



Domain-specific concepts

- *Source*: generates documents
- *Document*
 - *Main text (or: text)*
(text may change over time → potentially several versions of document texts)
 - *Feedbacks*
 - *Links*
 - **Temporal aspects** are relevant for all the above components of a document

Domain-specific Concepts



Document

Source of the document:
torokgaborelemez.blog.hu

Main text of the document

Links to other documents
(Trackbacks)

Feedbacks

Domain-specific concepts

The image shows a Facebook post by Henrikas Dapkus. The post features a photograph of several decorated Easter eggs on a table, with markers and other items scattered around. The text of the post is in Lithuanian: "Dapkus dazo kiaušinius. Gražių šv. Vėlykų, mieli draugai!". Below the photo, there are interaction options like "Gefällt mir", "Kommentieren", and "Teilen". There are also four comments from other users: Rūta Minčiūnaitė, Gitana Kurmytė, Eimantas Buu, and Auguste Mikšyte. The post is highlighted with a blue border.

Document

Source of the document:
Henrikas Dapkus

Main text of
the document

Feedbacks

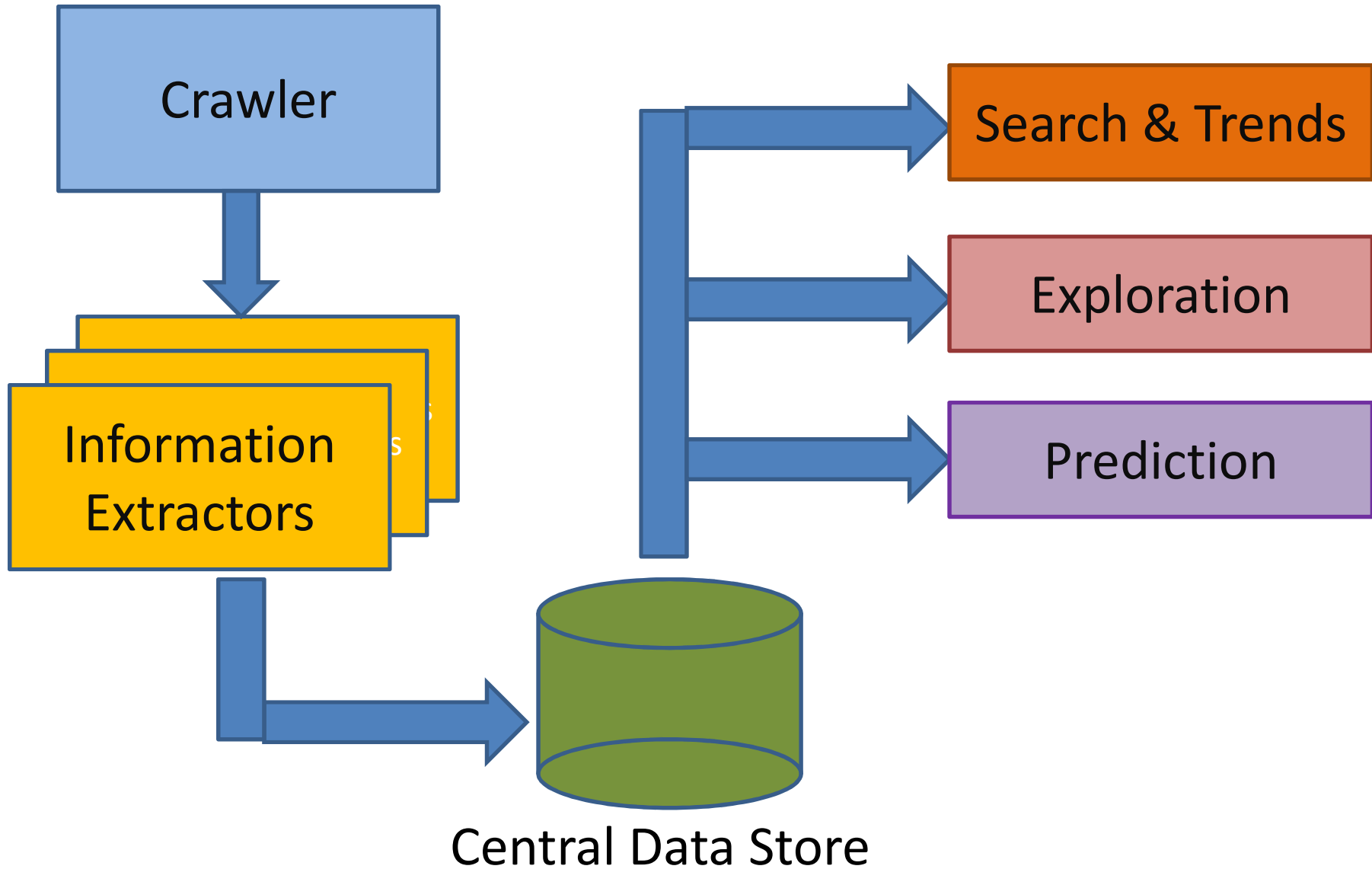
Thousands of blogs,
tweets,... appeared
about our company
in the last days.
Which ones should
we reply to?



Problem Formulation

For the documents that
appeared in the last 72 hours,
predict the number
of new feedbacks, i.e., the
number of feedbacks in
the next 24 hours.

System schema



Crawler

The screenshot shows the 'Crawler' configuration window of the Social Web Miner application. The window has a title bar with the text 'Social Web Miner' and standard Windows window controls. Below the title bar is a tabbed interface with the following tabs: 'Crawler', 'Extractors', 'Extractor Assignments', 'Search & Trends', 'Data Exploration', and 'Prediction'. The 'Crawler' tab is active, and a sub-tab 'Configure Information Extractors' is selected. The configuration is organized into several sections:

- Domain:** A text field containing 'blog.hu'.
- Seeds:** A list box containing five URLs: 'http://torokgaborelemez.blog.hu/', 'http://konspiralo.blog.hu/', 'http://telefonkozpont.blog.hu/', 'http://kerekagy.blog.hu/', and 'http://autozz.blog.hu/'. Below the list box is an input field with the placeholder '(type new seed here)'. To the right of the list box are three buttons: 'Delete', 'Clear', and 'Invert'. Below the input field is an 'Add' button.
- Special Parameters:** A text field containing 'blog.hu,?fullcommentlist=1#comments'. Below it is an input field with the placeholder '(type special parameter setting here)'. To the right are three buttons: 'Delete', 'Clear', and 'Invert'. Below the input field is an 'Add' button.
- Max. number of pages to crawl:** A text field with the value '50000'.
- Max. crawling depth:** A text field with the value '12'.
- Delay (ms):** Two text fields, the first with '200' and the second with '100'.
- Checkboxes:** Two checked checkboxes: 'Save crawled pages' and 'Log crawling process'.
- Buttons:** At the bottom of the window are two buttons: 'Save' and 'Reset'.

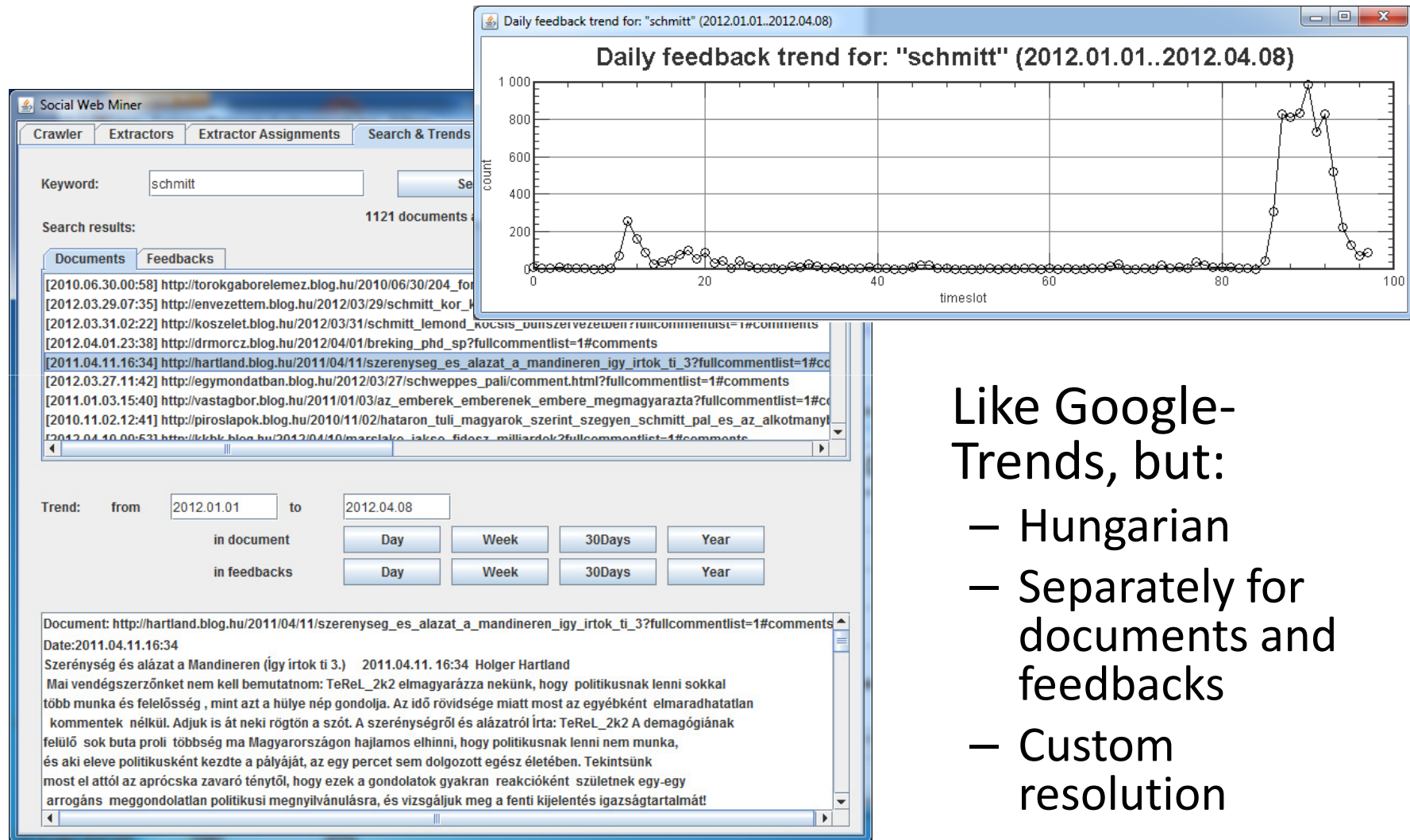
Information Extractors

The screenshot shows the 'Social Web Miner' application window with the 'Extractors' tab selected. The interface is divided into several sections:

- Available Extractors:** A list of extractors including BlogURLExtractor1, BlogURLExtractorKonsprialo, BlogURLExtractorNapizeje, TrackbackExtractor, FeedbackExtractor (highlighted), and TextExtractor. Below this list is an input field containing 'FeedbackExtractor' and an 'Add new' button.
- Buttons:** 'Delete' and 'Clear' buttons are located to the right of the available extractors list.
- Tags and pieces of information to extract:** A table with columns for 'Open Tag' and 'Closing Tag'.

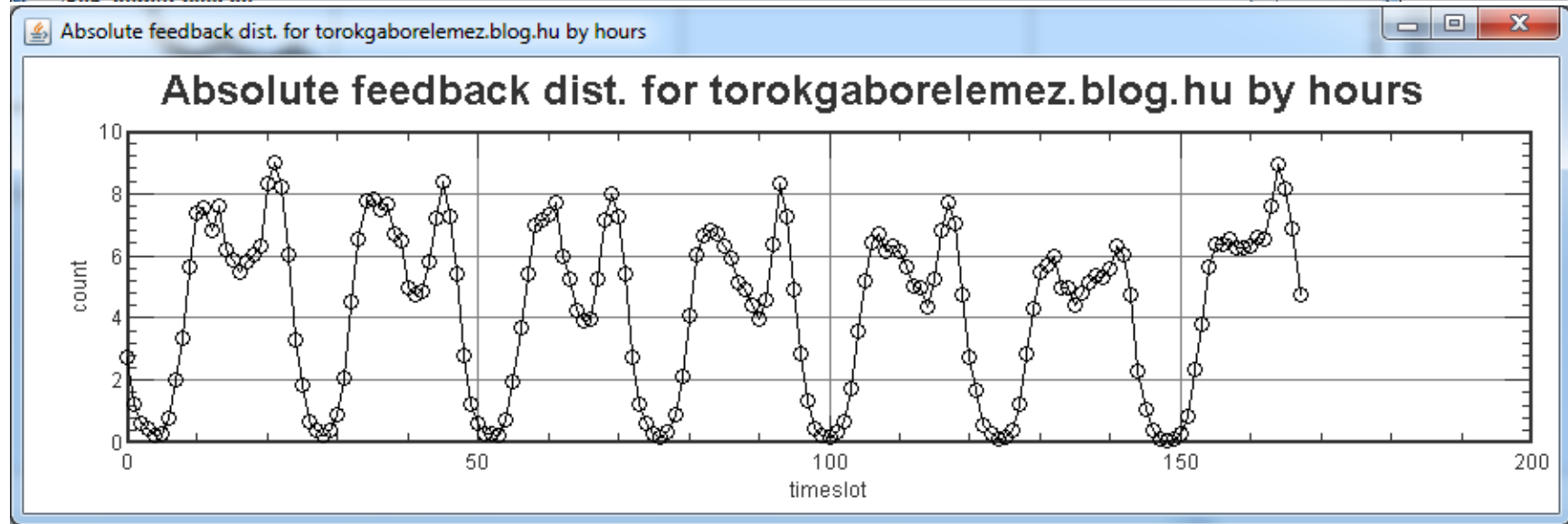
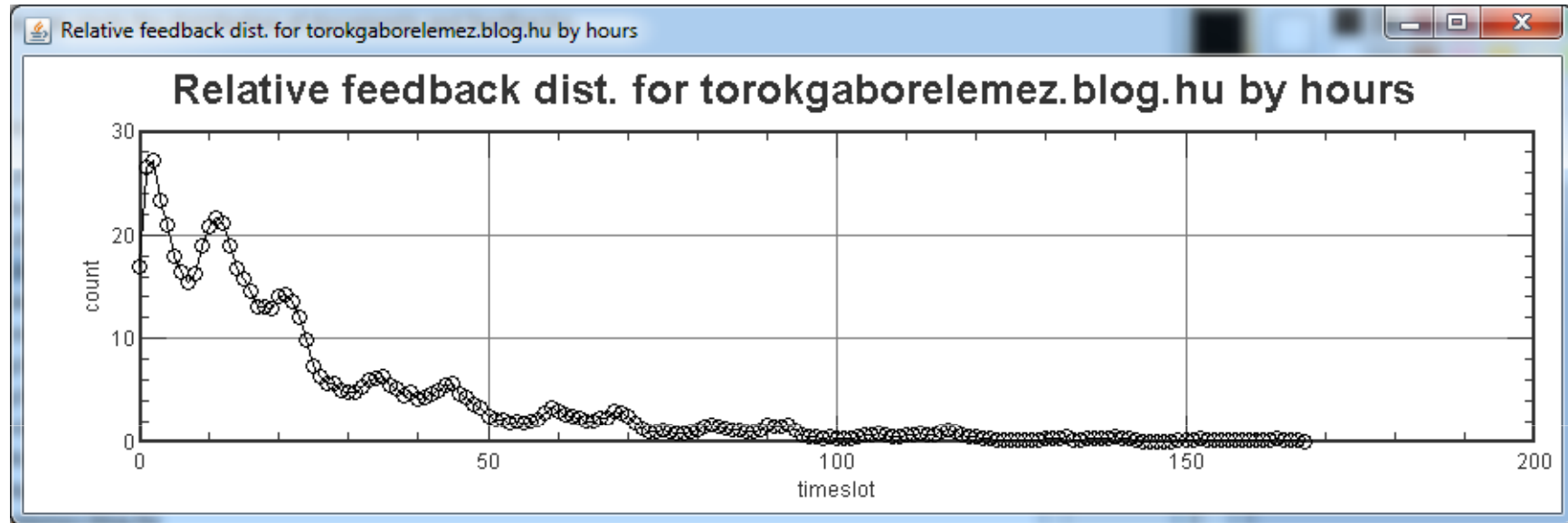
	Open Tag	Closing Tag
List tags	<input type="text" value='	<input type="text"/>
Entity tags	<input type="text" value='<div class="comment"'/>	<input type="text"/>
<input checked="" type="checkbox"/> Text	<input type="text" value='<div class="commenttext"'/>	<input type="text" value="</div>"/>
<input type="checkbox"/> URL	<input type="text"/>	<input type="text"/>
<input checked="" type="checkbox"/> Date/Time	<input type="text" value='	<input type="text" value=""/>
- Extraction constraints:** A list of checkboxes: 'Link before Text', 'Link before Time', 'Text before Link', 'Text before Time', 'Time before Link', and 'Time before Text' (checked).
- Date/Time extraction:** Includes a checkbox for 'Extract date/time from URL' with an adjacent 'URL time tag:' input field, and a checked checkbox for 'Set date/time as current if no date/time can be extracted'.
- Text extraction:** Includes a checkbox for 'Extracted text should contain HTML-tags'.
- Update Selected:** A button at the bottom right, preceded by a '=====>' separator.

Search & Trends



- Like Google-Trends, but:
- Hungarian
 - Separately for documents and feedbacks
 - Custom resolution

Data Exploration



Prediction

The screenshot shows the 'Prediction' tab of the Social Web Miner application. The interface is divided into several sections: Feature extraction, Prediction model, and Results.

Feature extraction:

- Train data from: 2010.01.01.00:00 to: 2011.12.15.00:00
- Prediction (test) data from: 2012.02.01.00:00 to: 2012.02.28.00:00
- Forecast period: 24 hours, Lookback period: 72 hours
- Step: 24 hours
- File name prefix: extractedFeatures

Additional features:

- Weekday indicators
- Parent features
- Text features (TF)
- Number of TF: 100
- Min.support: 20

Prediction model:

- Multilayer perceptron Structure: 5,2 Epochs: 1000 Learning Rate: 0.05 Momentum: 0.01
- k-NN k: 5
- RBF Network Number of clusters: 100
- REP-tree M5P-tree Linear Regression
- Bagging Number of elementary models: 100

Results:

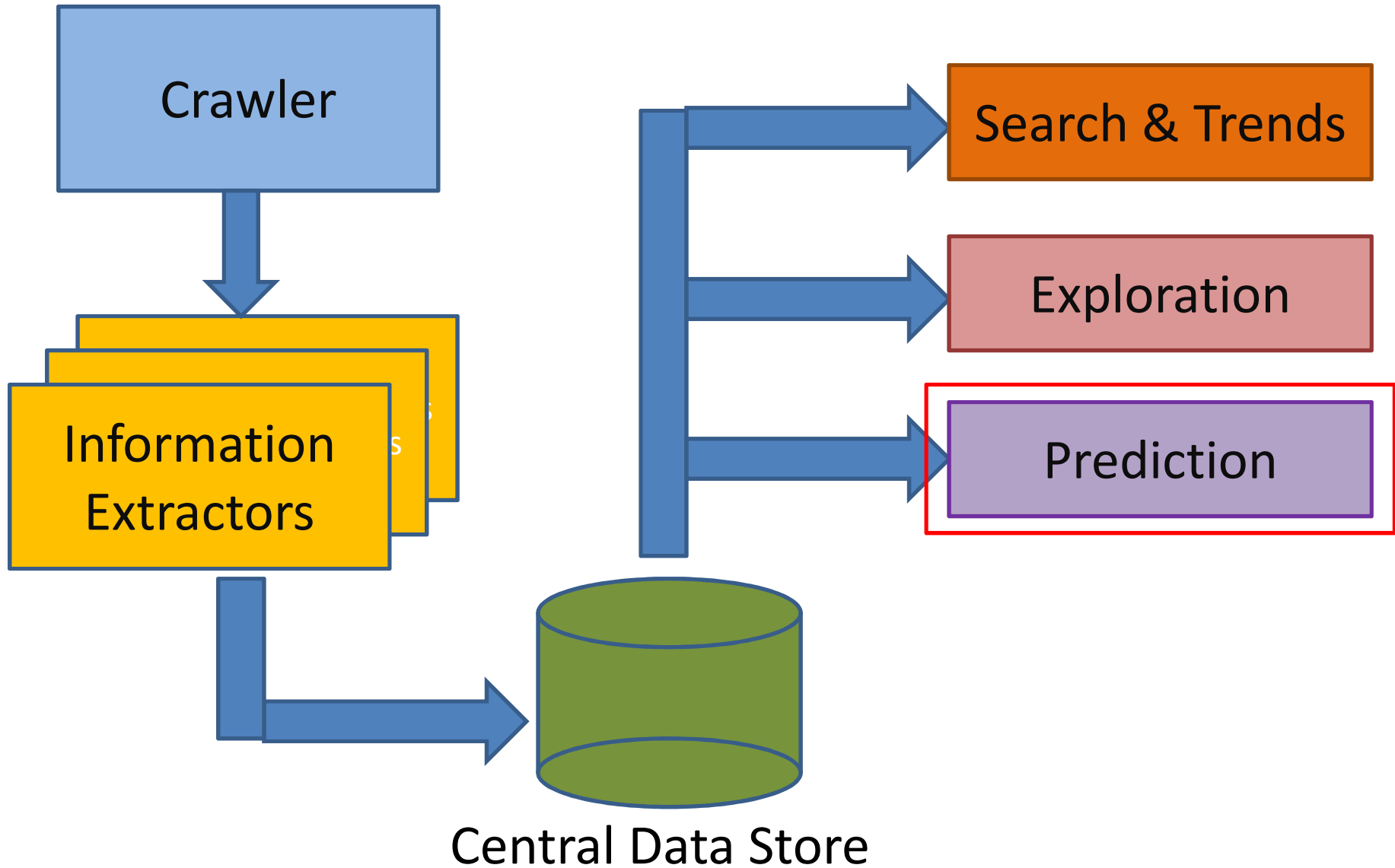
1.2778759	1.0	http://boldogokasajtkeszitok.blog.hu/2012/02/26/ecrasez_l_inf_me?fullco
0.22766016	0.0	http://kkbk.blog.hu/2012/02/27/megvedte_a_fidesz_a_kommunizmus_ugynokei
0.22663249	0.0	http://envezettem.blog.hu/2012/02/25/a_resti_fontosabb_mint_a_nyugdij?F
0.103179865	0.0	http://telefonkozpont.blog.hu/2012/02/26/masok_irtak_a_facebook_a_lancl

Summary:

(Please note that evaluation is meaningless if predictions are made for future.)

Hits @ 10	6.178571	1.6701926
Hits @ 20	12.821428	2.3153784
AUC @ 10	0.8944222	0.084635146
AUC @ 20	0.8652075	0.060107005

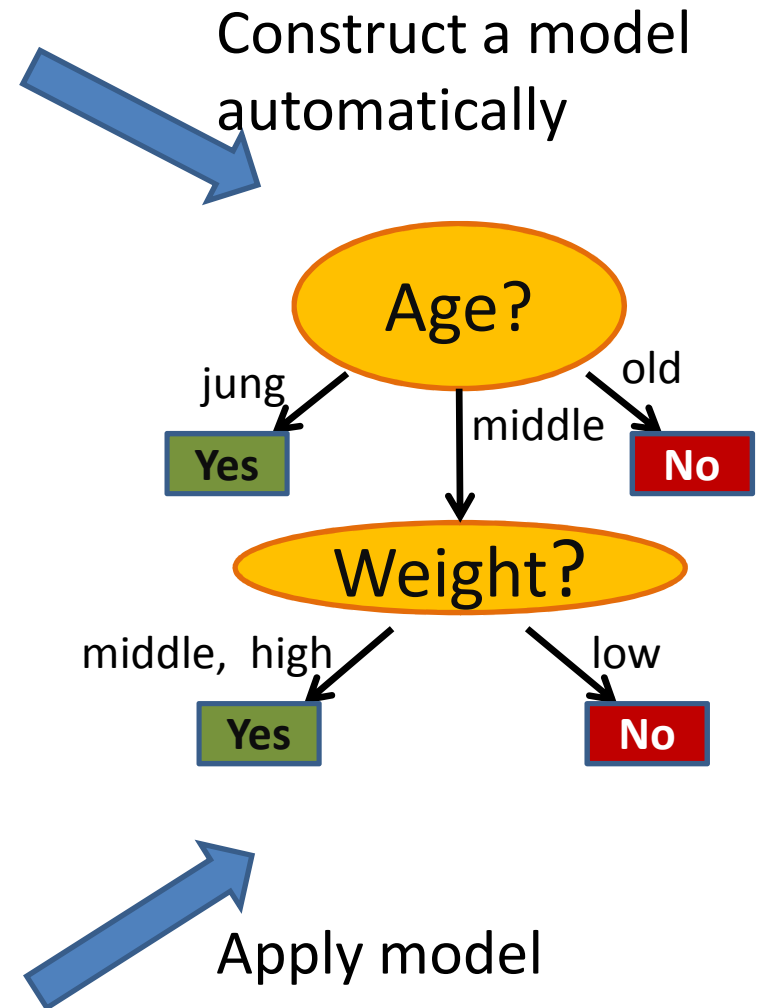
System schema



Machine Learning

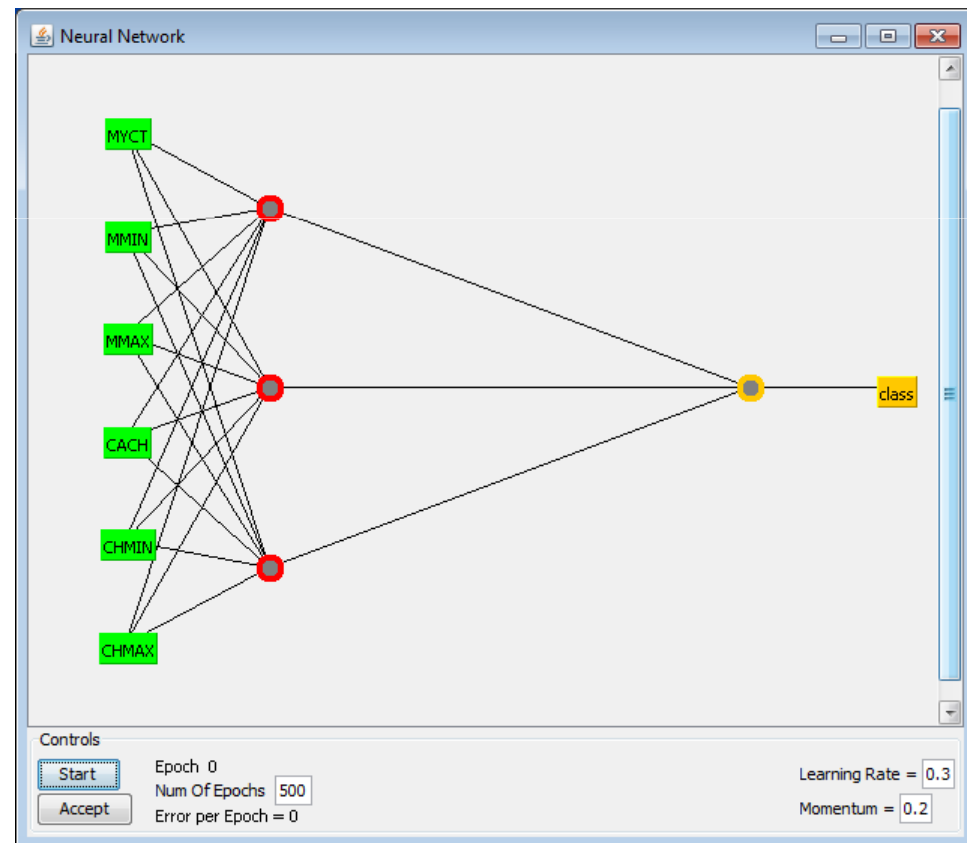
ID	Age	Weight	Sport	Purchase chocolate cake
1	Jung	Low	Yes	Yes
2	Old	Middle	No	No
3	Middle	Hi	No	Yes
4	Old	Middle	Yes	No
5	Jung	Hi	No	Yes
...

ID	Age	Weight	Sport	Purchase...
101	Middle	Low	No	?
102	Old	Low	No	?
103	Jung	Middle	No	?
...



Machine Learning

- Models we used:
 - Regression trees:
M5P, REPTree
 - Neural networks
 - RBF Networks
 - K-NN
 - (Linear) Regression
 - Ensemble Models:
bagging, stacking



Feature Extraction

- In total, we extract up to several hundreds of features, for example:
 - Basic Features
 - Number of links/feedbacks in the last 24 hours
 - How the number of feedbacks/links increase
 - Aggregation of such features by source
 - Textual Features
 - Most significant bag of words features (language specific preprocessing)
 - Weekday Features
 - Parent Features

Evaluation

- Data:
 - 37 279 documents collected from Hungarian blogs
 - 6,17 GB (plain HTML, without images, sounds, etc.)
- Temporal train and test split
 - Train data: Year 2010 and 2011
 - Test data: February and March 2012
- We tried various models and feature sets
 - In total: several months of computational time

Evaluation Procedure

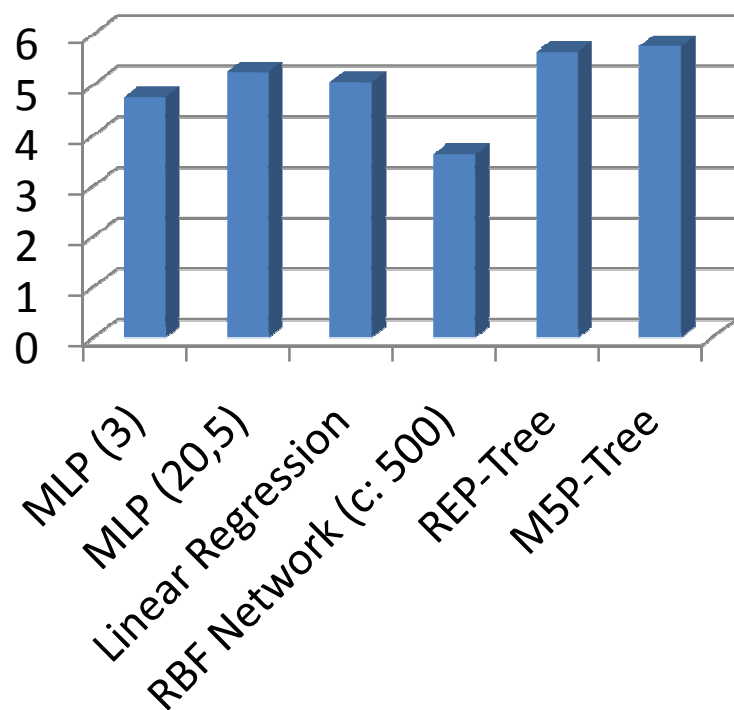
- Select a base date/time
 - e.g. 2012.03.01.12:00
- Simulate that the current time is the selected base date/time, and make predictions according to that time
 - e.g. we predict the number of feedbacks in the time interval between 2012.03.01.12:00 and 2012.03.02.11:59
- Compare the predictions with what happened in the next 24 hours relative to the base date/time
- Various base dates/times – average results

Evaluation Metrics

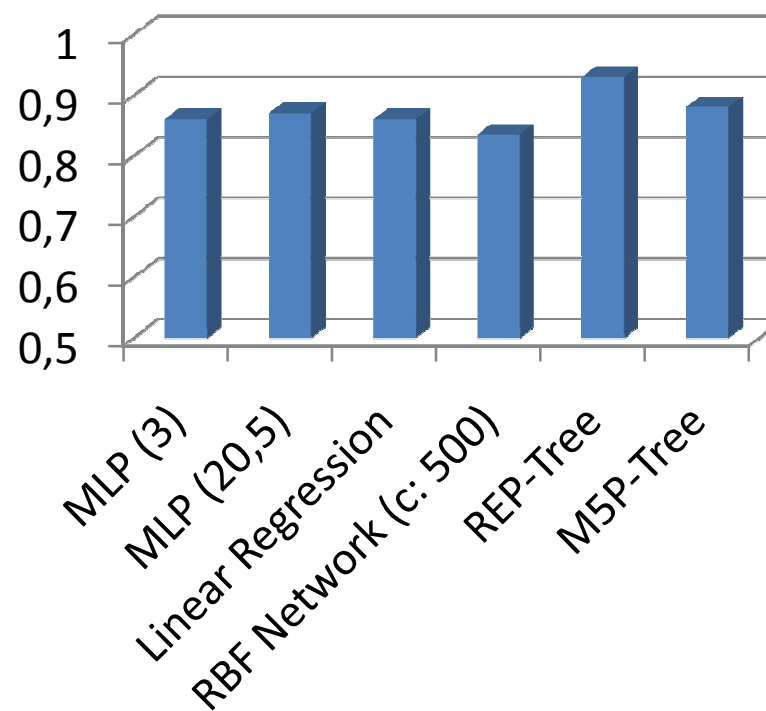
- Average of Hit@10
 - out of the 10 documents predicted to be the most relevant, how many belong to the most relevant 10 documents
- AUC@10
 - consider the 10 most relevant documents according to the ground truth
 - let these 10 documents belong to the positive class, other documents belong to the negative class
 - calculate AUC of the predictions

Performance of the examined models

Hits@10



AUC@10



All Features

(Basic features + Textual Features (200) + Weekday Features + Parent Features)

Effect of the Feature Set

Model	Basic	Basic + Weekday	Basic + Parent	Basic + Textual
MLP (3)	5,533 ± 1,384 0,886 ± 0,084	5,550 ± 1,384 0,884 ± 0,071	5,612 ± 1,380 0,894 ± 0,062	4,617 ± 1,474 0,846 ± 0,084
MLP (20,5)	5,450 ± 1,322 0,900 ± 0,080	5,483 ± 1,323 0,910 ± 0,056	5,383 ± 1,292 0,914 ± 0,056	5,333 ± 1,386 0,896 ± 0,069
k-NN (k: 20)	5,433 ± 1,160 0,913 ± 0,051	5,083 ± 1,345 0,897 ± 0,061	5,400 ± 1,172 0,911 ± 0,052	3,933 ± 1,223 0,850 ± 0,060
RBF Net (clusters: 500)	4,750 ± 1,456 0,876 ± 0,067	4,667 ± 1,300 0,871 ± 0,062	4,517 ± 1,284 0,877 ± 0,061	3,567 ± 1,359 0,824 ± 0,066
Linear Regression	5,283 ± 1,392 0,876 ± 0,088	5,217 ± 1,343 0,869 ± 0,097	5,283 ± 1,392 0,875 ± 0,091	5,083 ± 1,215 0,864 ± 0,096
REP Tree	5,767 ± 1,359 0,936 ± 0,038	5,583 ± 1,531 0,931 ± 0,042	5,683 ± 1,420 0,932 ± 0,043	5,783 ± 1,507 0,902 ± 0,086
M5P Tree	6,133 ± 1,322 0,914 ± 0,073	6,200 ± 1,301 0,907 ± 0,084	6,000 ± 1,342 0,913 ± 0,081	6,067 ± 1,289 0,914 ± 0,068
		☹	☹	☹

Effect of Bagging

Model	Basic	Basic + Bagging (100)
MLP (3)	5,533 ± 1,384 0,886 ± 0,084	5,467 ± 1,310 0,890 ± 0,080
MLP (20,5)	5,450 ± 1,322 0,900 ± 0,080	5,633 ± 1,316 0,903 ± 0,069
k-NN (k: 20)	5,433 ± 1,160 0,913 ± 0,051	5,450 ± 1,102 0,915 ± 0,051
RBF Net (clusters: 20)	4,117 ± 1,253 0,854 ± 0,063	4,333 ± 1,135 0,867 ± 0,054
Linear Regression	5,283 ± 1,392 0,876 ± 0,088	5,150 ± 1,327 0,881 ± 0,082
REP Tree	5,767 ± 1,359 0,936 ± 0,038	5,850 ± 1,302 0,934 ± 0,039
M5P Tree	6,133 ± 1,322 0,914 ± 0,073	5,783 ± 1,305 0,926 ± 0,048
		😊

Experimental Results – Lessons Learned

- Hit@10: around 5-6
 - Much better prediction than naïve models (e.g. averaging by source or random)
- M5P tree and REPTree seem to work best
- Neural networks work fine
- SVM: unacceptable training time
- Ensembles:
 - do not really improve (bagging, stacking)
- Basic features are the most relevant ones



Source: <http://www.sterlingtimes.org>

Can YOU do it better?

- Show it!
- Download the data from <http://www.cs.bme.hu/~buza/blogdata.zip>

Possible future work

- Advanced search
 - logic operations between keywords, ontologies, synonyms, inferencing, LSA, ranking of results...
- Enhanced prediction
 - higher accuracy, more detailed prediction: predict positive / negative feedbacks separately, personalized prediction: who comments what?, methods: matrix factorization, graph-based techniques, enhanced ensembles, enhanced classifiers (more options)
 - Concept drift, transfer learning techniques
- Clustering of documents (e.g. by topic)
- Topic tracking, and topic evolution
- Advanced visualization: standard deviation in plots, etc.
- Further domains (not only Hungarian blogs)
- Scaling: develop new, specialized index structures?
- Technology: use database server? Save trained prediction model?
- Non-textual entries (image, audio, video, etc.)

Conclusion

- Unbelievable growth of the importance of social media: US president elections, Revolutions in the Islamic world...
- Industrial proof-of-concept application for data mining in social media
 - Focus: feedback prediction for blogs
- Publication of the collected data
<http://www.cs.bme.hu/~buza/blogdata.zip>