Topic Tracking Framework

Bogdan Benea
Mihai Paraschiv
Florin Popescu
Alexandru Salajan
Irina Tudose

Main objective

Create a framework on top of which various news topic exploring applications can be developed.

(Project Proposal Stage)

Specifications?

Entry: Entries are collected by crawling processes from various sources. A common entry type is a news article. (feeds)

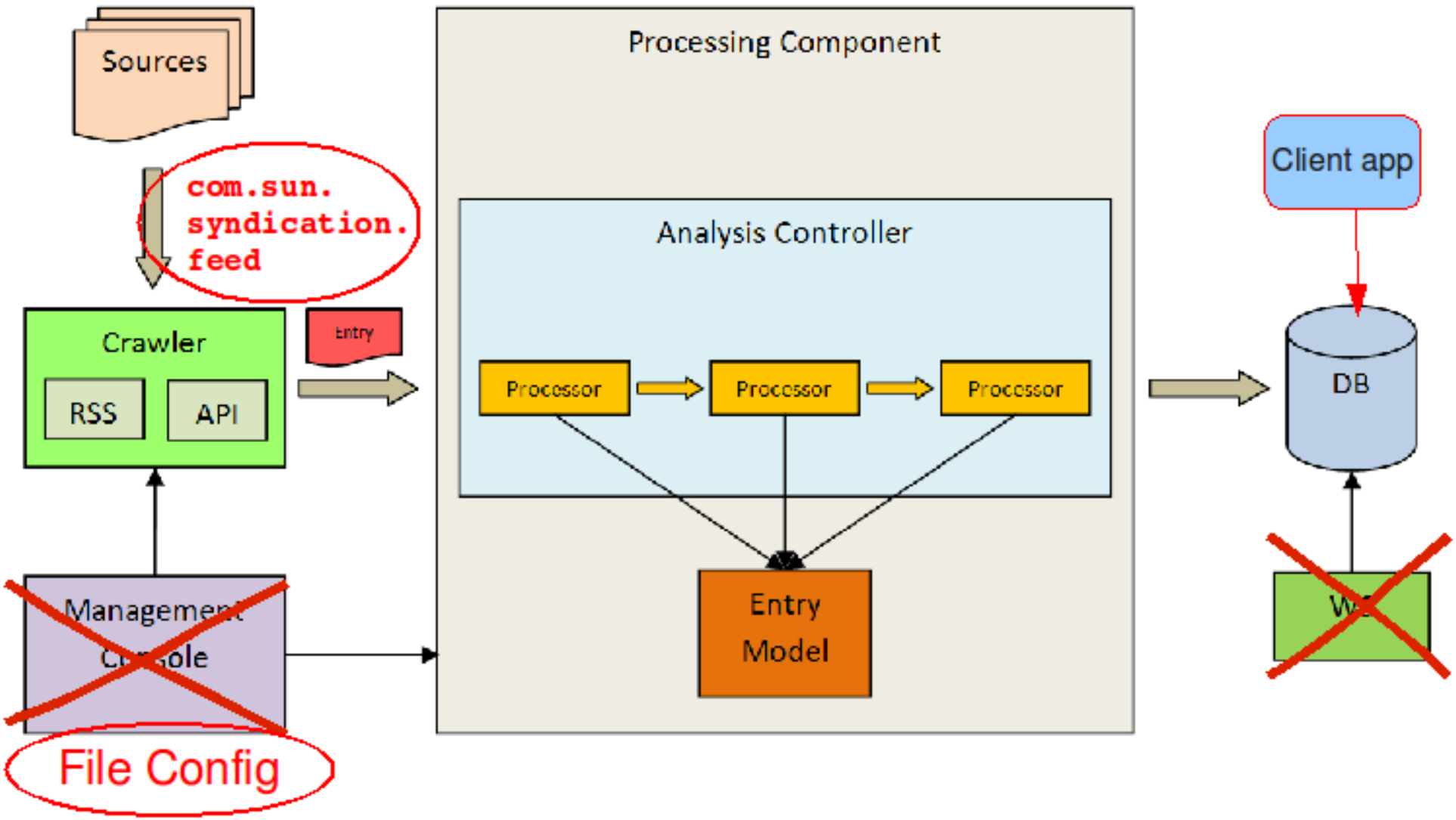
Topic: A collection of entries (articles). Topics are organized in a hierarchy. (no hierarchy)

Source: Feed or channel from where entries are collected. These can be added manually or can be detected by a crawler. (no crawler)

NamedEntity: These are extracted from the text and are good indicators of topic similarity.

(Project Requirements Stage)

Architecture?



(Project Requirements Stage)

Design - Analyzing new entries ?

Phase 1: Statistics computation

- the entry model initially contains only features identified by the crawler (e.g., title, content, tags)
- the model is enhanced with statistics - ranking of the terms (e.g., term frequency) and named entities

Phase 2: Topic detection

- For each available topic:
  - Compute the similarity ***s*** between the entry model and the topic (e.g., term frequency - inverse document frequency)
  - Compute the relevance score ***r*** of the topic for the given entry
- Select the topic with the highest ***r*** (create a new topic using this entry if ***r*** is less than the predefined threshold)

Phase 3: Topic update

- augment the cluster with the entry model

(Project Design Stage)

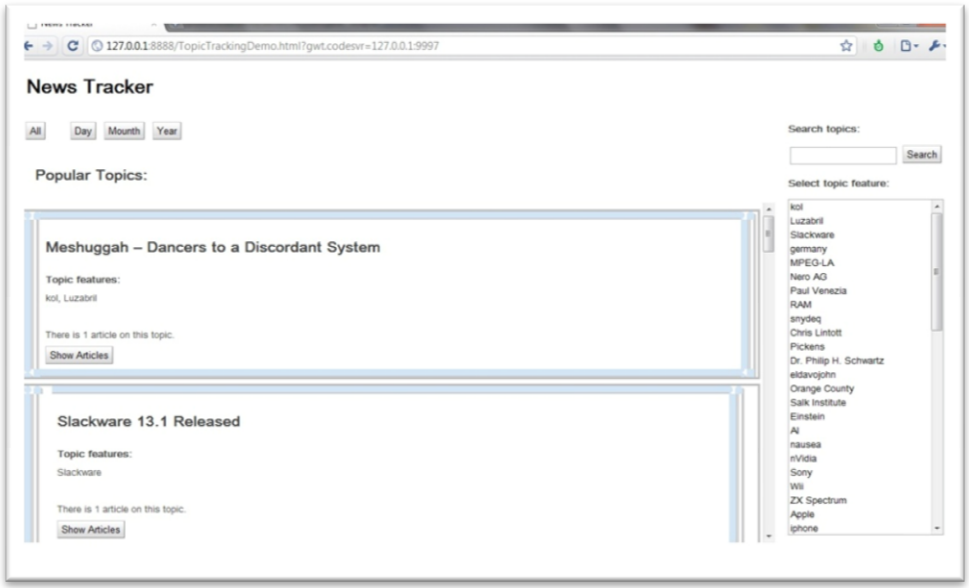
Testing?

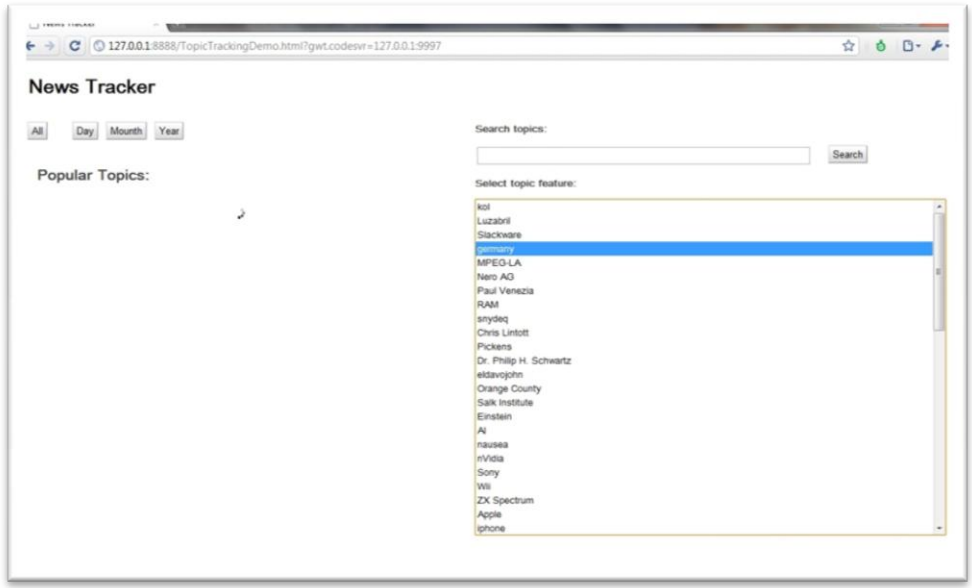
JUnit 4 testing:

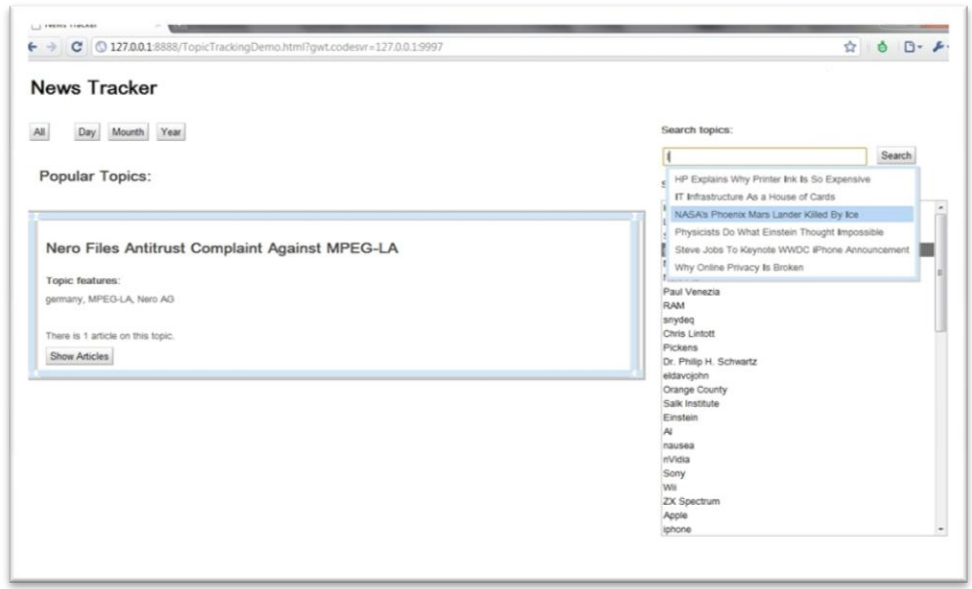
- analysis module (*ttf.test.analysis*)
- TfIdf module (*ttf.test.tfidf*) - Term frequency Inverse Document frequency
- tokenizer module (*ttf.test.tokenizer*)
- alchemyAPI (*ttf.test*)

(Development & Testing Stage)

Demo application







Tools

Java packages & APIs:

- mysql-connector-java
- apache commons ...
- apache log4J
- JUnit 4
- HTML Parser 2.0 (http://htmlparser.sourceforge.net/)
- ICU4J (http://site.icu-project.org/)
- ROME (https://rome.dev.java.net/

- http://www.alchemyapi.com/

(Development & Testing Stage)

Project Sites

Source Code:
http://code.google.com/p/topic-tracking-framework/

Docs:
https://sites.google.com/site/vvpnewstracker/