Collaborative scientific document recommender

Ciprian Dorin Crăciun

22 September 2008

General definition

- extracting (useful, previously unknown) information (meta-data) from already existing data;
- generic application: we have (a high amount of) data, but:
  - we don't know what we are actually looking for;
  - we know what we want, but there is no (classic) algorithm for the task;
  - the data is improperly formatted or unstructured;
- what do we understand by information?

Applications

- (computer) security;
- digital libraries / document management;
- social and economic sciences;
- medicine (and biology):
  - patient medical records processing;
  - documentation searching and cross-referencing;

Methods

- human based — OLAP;
- hard computing (ex. decision trees);
- probabilistic approach (ex. naive Bayes);
- soft computing (ex. neural networks);

Problems

- The main problems:
  - cleaning and preprocessing;
  - (efficient) storage;
  - (in time) processing;
  - output validation;
- The main solutions:
  - distributed processing (ex. map-reduce frameworks);
  - distributed storage (ex. column stores: Google BigTable, Amazon SimpleDB);

Relations with other domains

- distributed computing;
- multi-agent systems;
- soft computing;
- probability and statistics;
- databases;

Personal interests in data mining (and connex domains)

- document and text retrieval;
- approaching data mining by using soft computing methods;
- adapting (or using) distributed (processing and storing) techniques;
- correlating (and validating) information (obtained by applied different techniques);

Current general context

- increasingly number of scientific papers;
- big databases with unstructured (or limited structured) meta-data:
  - CiteSeer, DBLP;
  - PubMed/MEDLINE;
- divergent cataloging systems:
  - OAI — Open Archive Initiative;
  - DCMI — Dublin Core Metadata Initiative;
  - DOI;
- lack of exploratory tools;
- rudimentary (free) available systems:
  - Greenstone;
  - DSpace;

Proposed solution

- solution: collaborative scientific document recommender;
- purpose: to aid and guide researchers to find more easily relevant documents for their work;
- requirements:
  - collaborative system;
  - on-line and (limited) off-line usage;
  - integration / aggregation of already existing systems;
  - open (interoperable) system;
  - visual exploration;

Why this subject?

- (relatively) new research domain (information and text retrieval);
- direct application;
- interdisciplinary relations;
- current research collective background and context;

Individual sub-tasks / research directions

- document hierarchies;
- document classification / clustering;
- user behavior;
- tuning from user feedback;
- visual exploration;

Hierarchies and documents

- multiple (customized) hierarchies;
- mapping between hierarchies;
- unsupervised hierarchy creation (with multiple categories per document):
  - fuzzy hierarchical clustering;
- (supervised, fuzzy) document classification;
- (self-contained, minimal) reference documents proposal;

User behavior and expertise

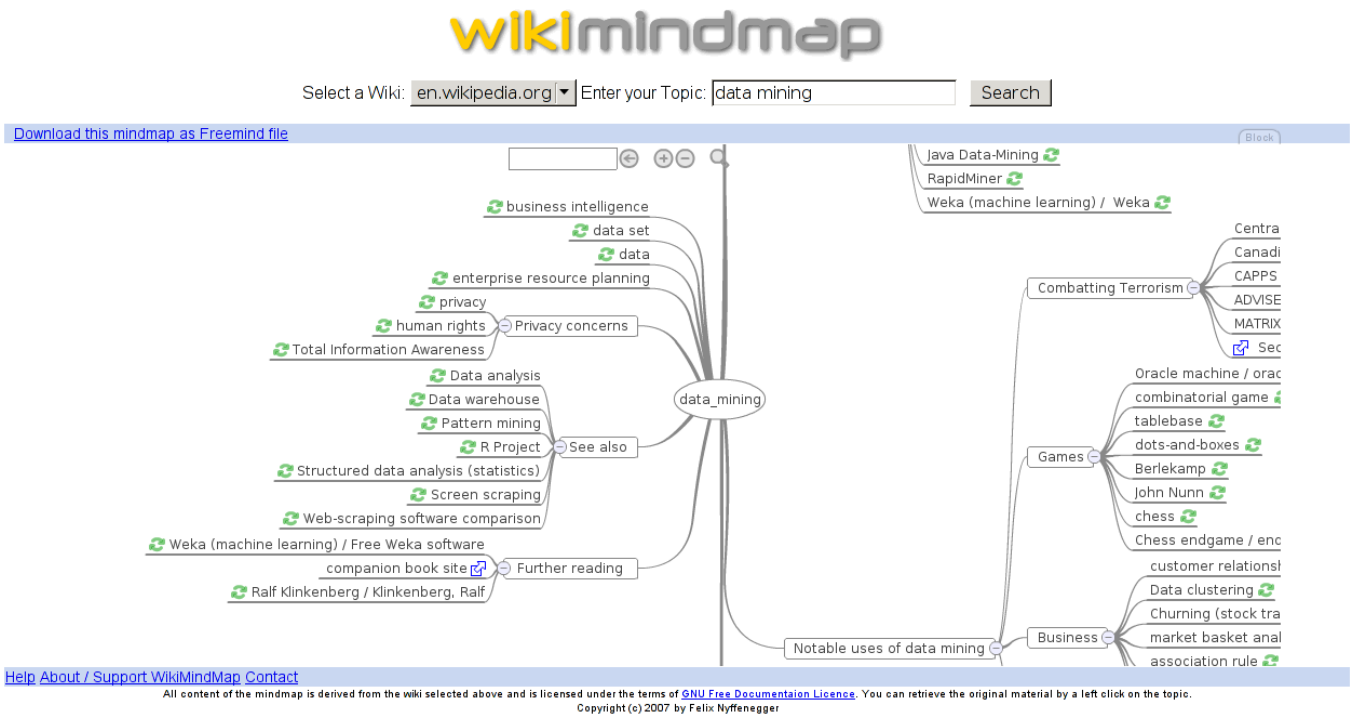
- deducing the expertise of an user for certain categories;
- integrating the user view back into the "canonical" hierarchies;
- recommending certain (group of related) documents based on user interests;

Visual exploration

- WikiMindMap — `http://wikimindmap.org/`;
- Clusterball — `http://www.chrisharrison.net/projects/clusterball`;
- WEBSOM — `http://websom.hut.fi/`;
- Musicovery — `http://musicovery.com/`;

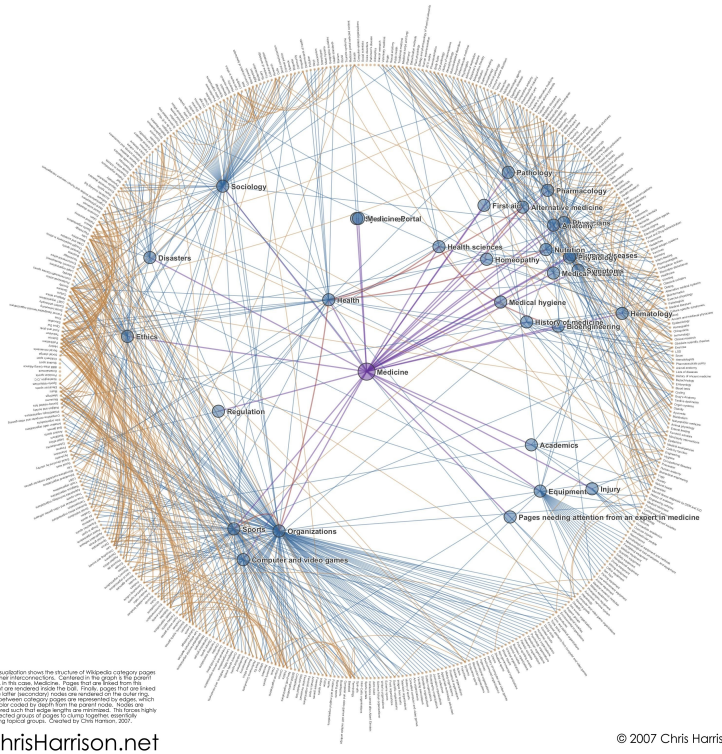
Visual exploration

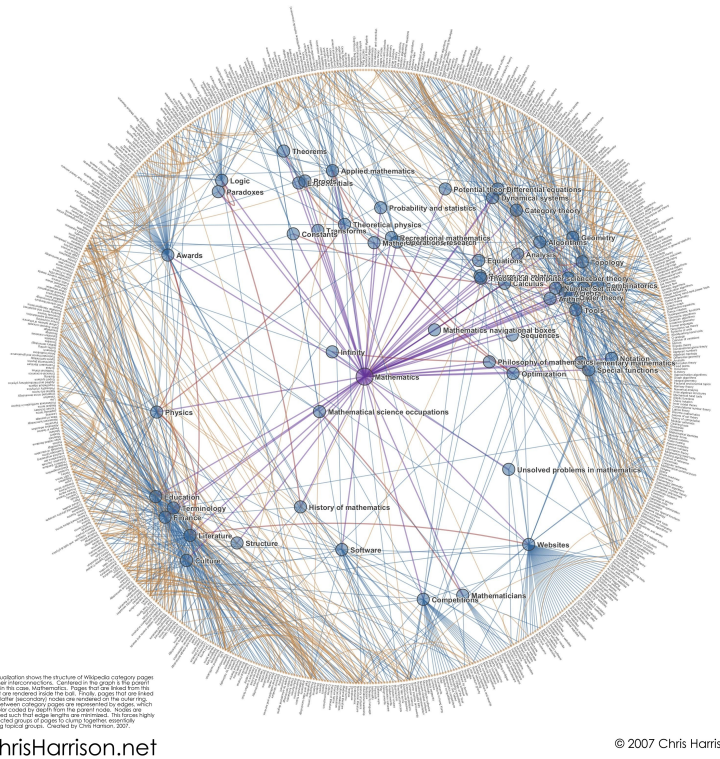
WikiMindMap



Visual exploration

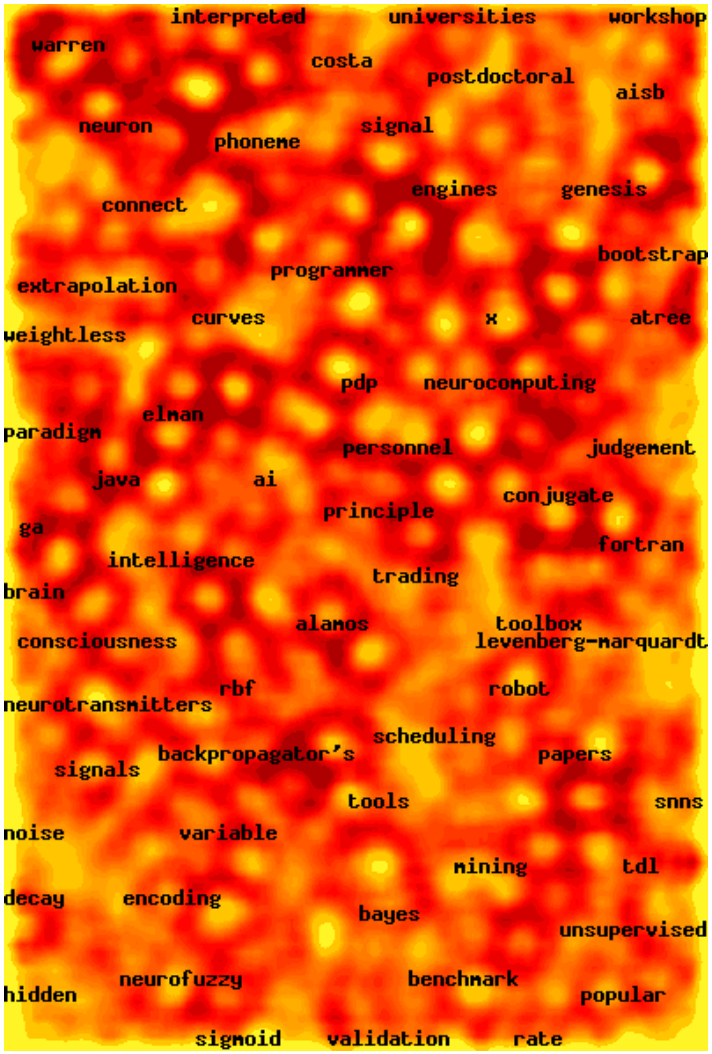
Clusterball

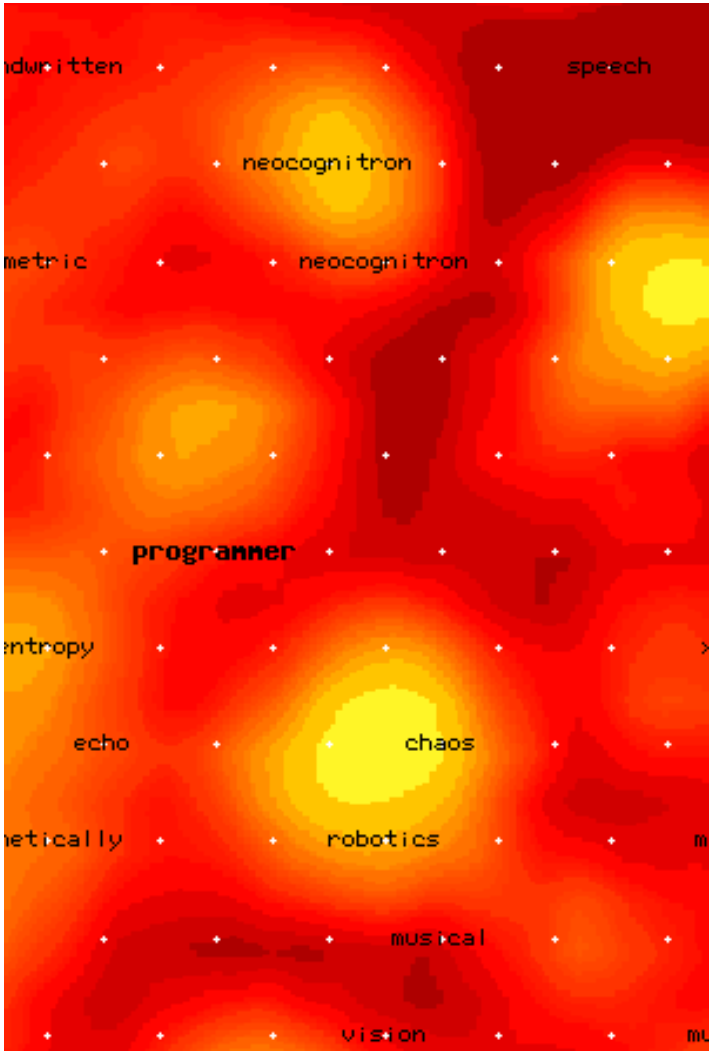




Visual exploration

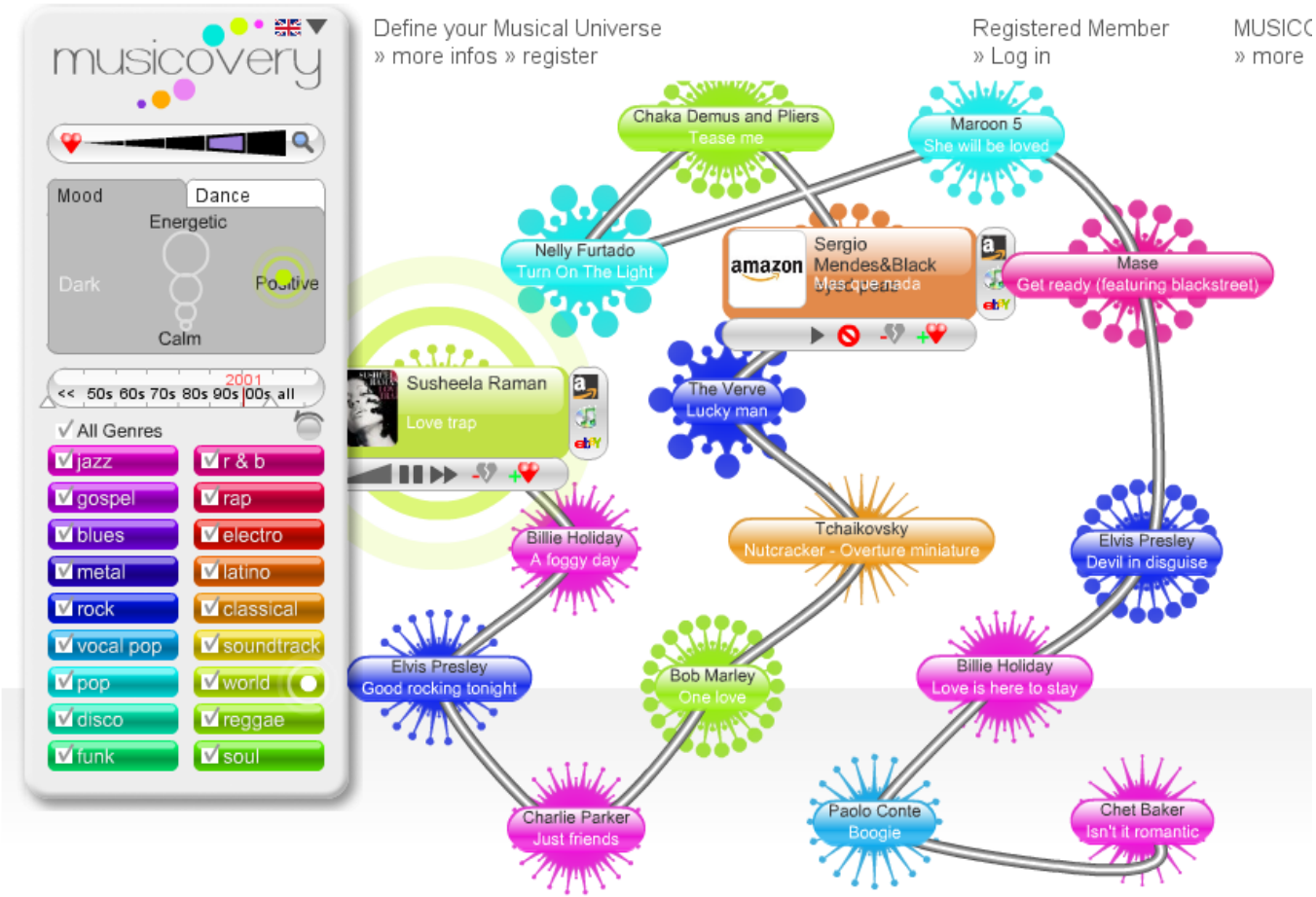
WEBSOM



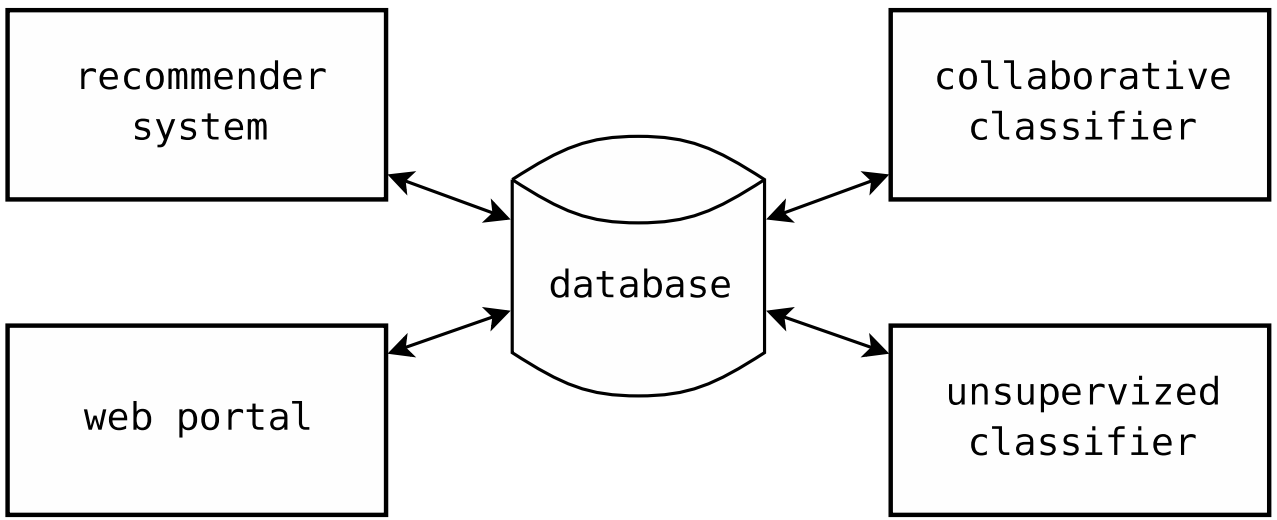


Visual exploration

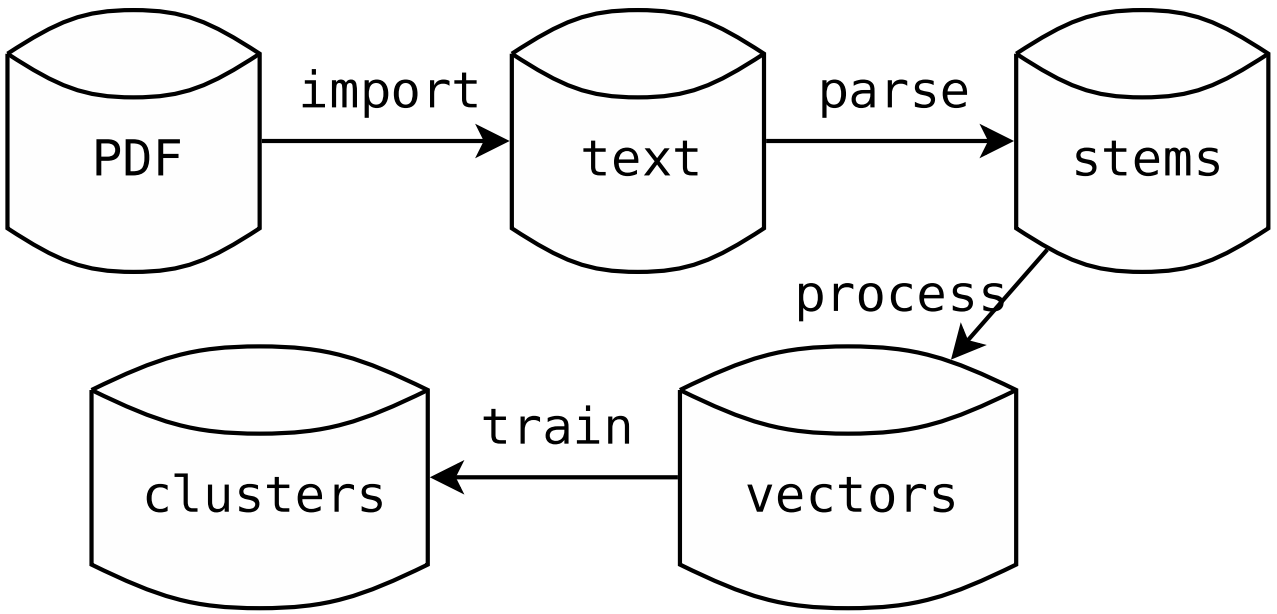
Musicovery



System architecture



Prototype workflow



Fin