The HyperBagGraph DataEdron: An Enriched Browsing Experience of Scientific Publications

Xavier Ouvrard @ UniGe & CERN supervised by:
Pr. Stéphane Marchand-Maillet @ UniGe (CH) and Dr. Jean-Marie Le Goff @ CERN

This application is a small brother of the
Collaboration Spotting project @ CERN

- Collspotting Project leader: Dr Jean-Marie Le Goff
- 1 fellowship, 3 PhD students

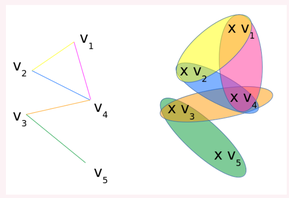
http://collspotting.web.cern.ch

Leveraging insight into your data network by viewing co-occurrences while navigating across different perspectives.

Graph:
- Set of **vertices** and set of **edges**.
- An edge links two vertices : **pairwise relationship**.

Sets:
- Regroup elements with no repetition and no order.

Hypergraphs:
- Extend graphs.
- Allow relations between multiple vertices.
- Are a family of **hyperedges** of unempty subsets of a **vertex** set.

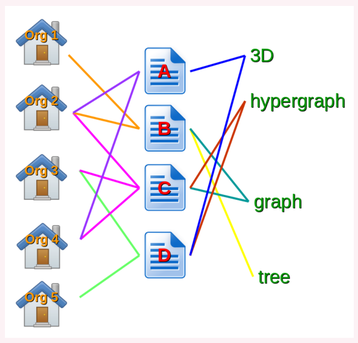


Natural multisets:
- Collection of objects with allowed repetitions.
- Constituted of a universe and of a multiplicity function on this universe
- Natural hb-graphs: the values of the multiplicity function are integer

HyperBag-graphs (Hb-graphs):
- Extend hypergraphs
- Allow duplication of elements
- Are a family of multisets - called **hb-edges** - of same universe called the **vertex** set
- Natural hb-graphs: use natural multisets

In a Scientific Publication Database:
- Metadata store information on the structure of the DataBase
- Metadata have types, that can be used either as dimension or as reference
- Data instances attached to one type of metadata can be re-grouped by using a reference => we talk about co-occurences
- **Co-occurences** are *n*-adic **relationships.**
- Co-occurences are **multisets**, often reduced to **sets**.

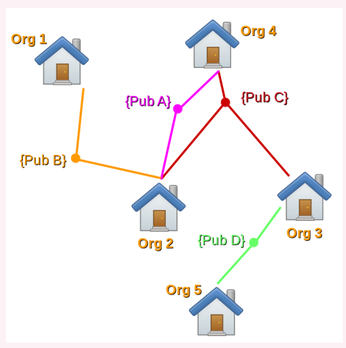
On an example:



We can retrieve in a simplified way a **hypergraph**: we use as **reference** the publication and the **co-occurences** of organisations seen as subsets of all organisations:

Pub A	Org 2, Org 4
Pub B	Org 1, Org 2
Pub C	Org 2, Org 3, Org 4
Pub D	Org 3, Org 5

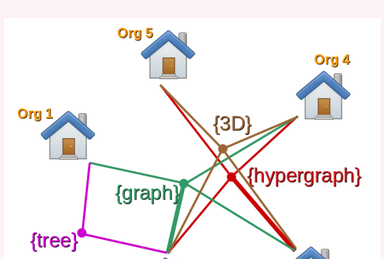
that we can visualize with an extra-node representation:



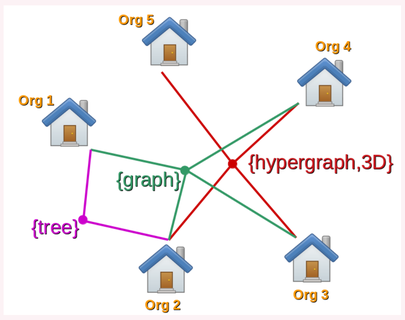
Choosing as reference keywords, we retrieve family of multisets of organisations, called a **hb-graph**

tree	$\{\{\text{Org } 1^1, \text{Org } 2^1\}\}$
graph	$\{\{\text{Org}1^1, \text{Org}2^2, \text{Org } 3^1\}\}$
hypergraph	$\{\{\text{Org } 2^1, \text{Org } 3^2, \text{Org } 4^1, \text{Org } 5^1\}\}$
3D	$\{\{\text{Org } 2^1, \text{Org } 3^1, \text{Org } 4^1, \text{Org } 5^1\}\}$

Represented by a bundled extra-node multipartite graph representation:



Which can be reduced as a hypergraph:

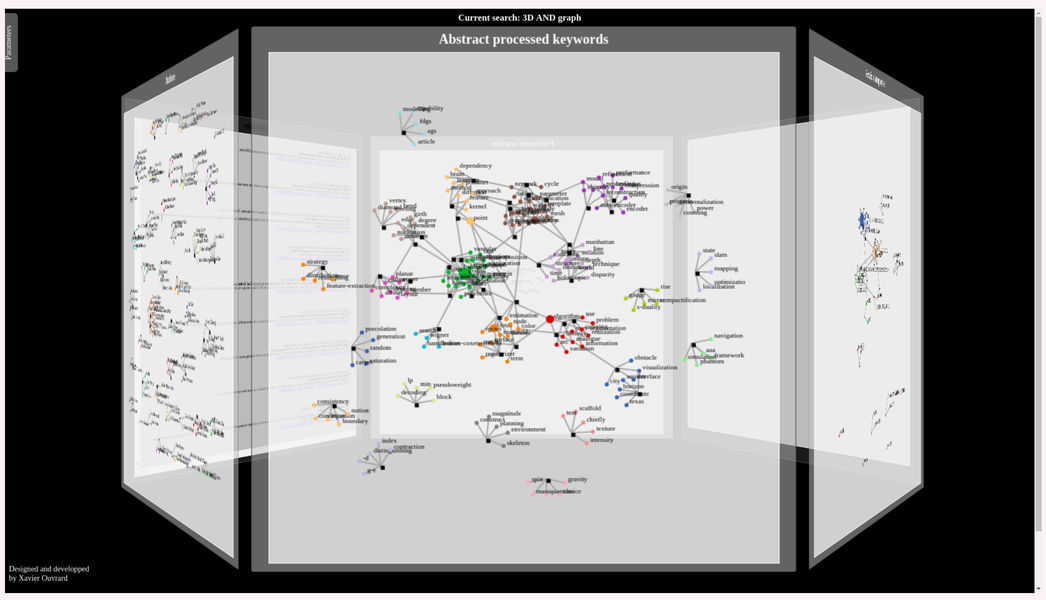


Searching on a Scientific Publication Database

With traditional verbatim browser:
- Give linear information
- To refine information: perform a new search
- Making complex query can be hazardous for most people
- Accessing the different facets of the information space require to perform different searches

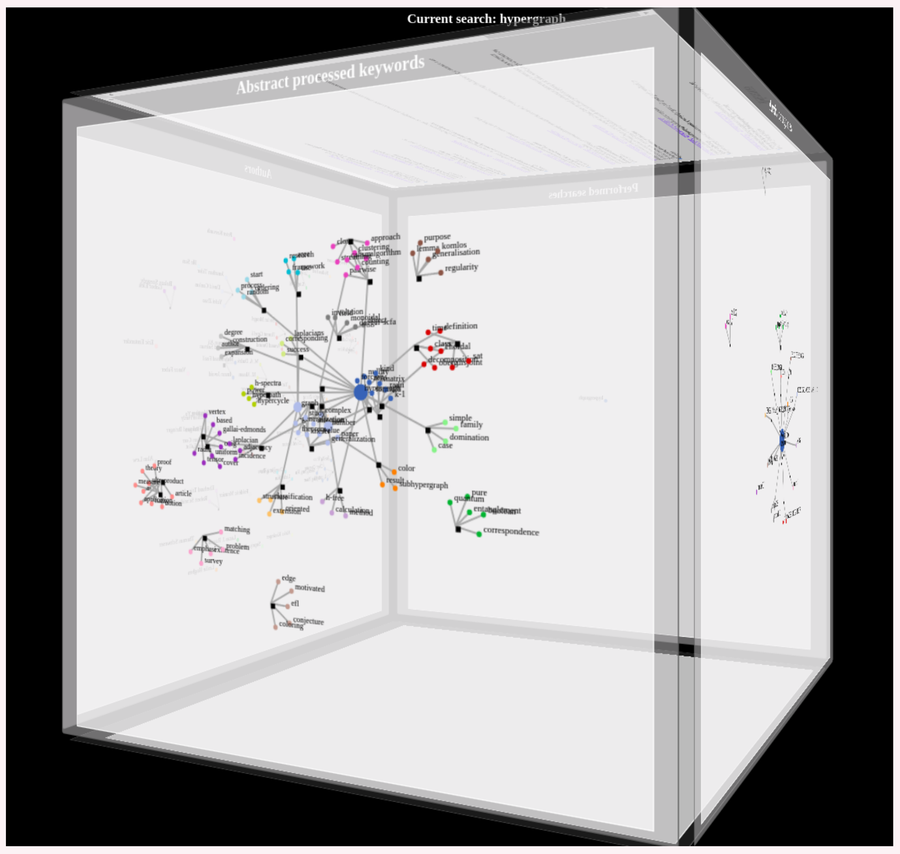
But in fact:
- A space of information is multi-facetted
- Much more information is available or can be extracted
- Use of natural language processing allow to extract keywords
- Hb-graphs highlight how the data instances are linked and allow additional information to be displayed

Switching between the different facets of the information space



Carousel view of the information space

Toggling to the DataEdron

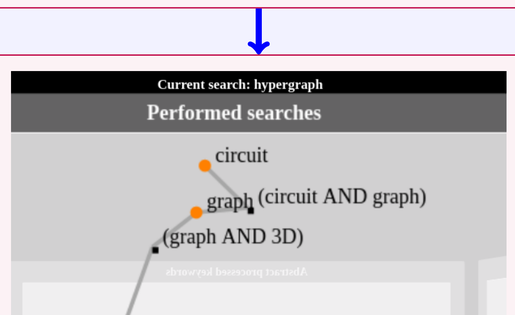


The Hb-graph DataEdron:
switching between the different facets of the information space

Additional information can be displayed:
- Dblp / Linkedin profile
- Publication abstract / full article
- Wikipedia information (for keywords)
- DuckDuckGo deambiguation and abstract...

Full interactivity of faces:
- Highlight extra-nodes through vertices on the same reference
- Highlight vertices involved in the highlighted references

How can we perform search?
- Traditional text field search
- Then queries can be built visually
- AND, OR, NOT possibilities
- The graph of search can be explored
- Possibility of merging different searches on a single graph



Recomputing facets

Which database can we search on?
- Arxiv
- Inspire

... but can be applied to any databases.

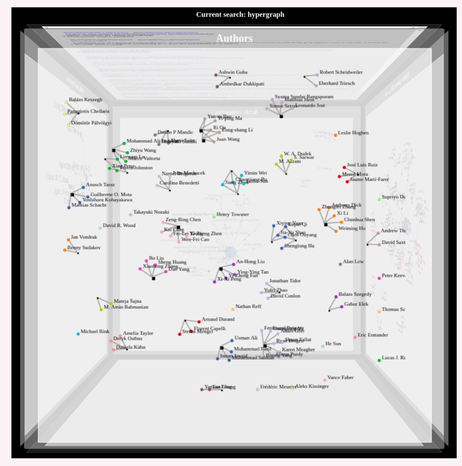
How do we proceed?
- All the queries are made online
- Everything is processed online, including keywords if they are not provided
- No intermediate storage

Which machine learning is used?
- **Natural language processing:**
  - It allows the extraction the keywords from abstract
  - Stop words are removed
  - Lemmatisation is made: only nouns are kept
  - Singularisation is made
  - Keywords are ranked using tf-idf
  - Tf: term frequency of the term in the document
  - Idf: inverse document frequency of the apparition of the term in the set of documents
- **Clustering**
  - The aim is to regroup vertices by communities
  - Vertices that are more connected than in a random graph are gathered
  - Use the modularity of Newmann
  - The Louvain community detection algorithm is fast and efficient
- **Layout embedding**
  - We use a force directed algorithm
  - It attracts vertices that are connected and repeals the ones that are disconnected
  - Works well on small graphs

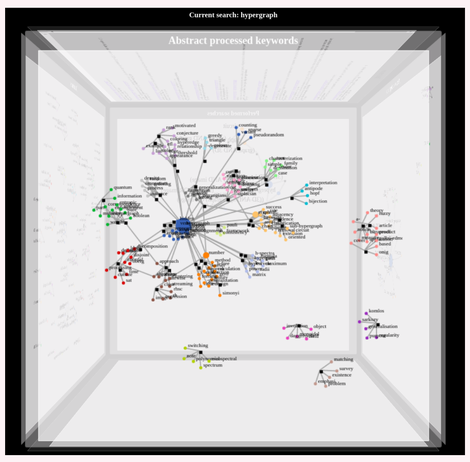
Search on the keyword: hypergraph

View the different facets

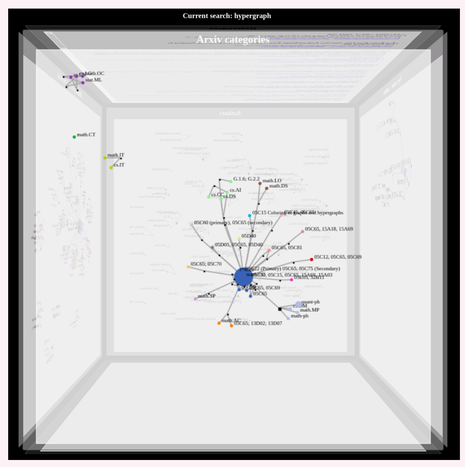
Facet: Authors



Facet: Processed keywords



Facet: Arxiv Categories



What are the research challenges solved?
- Modeling requires a strong framework
  - Particularly the switching of references is challenging
- Scaling up representations (for some applications):
  - Requires to fasten the computation
  - Only appropriated for some applications
- Finding important part of the representations:
  - A diffusion process has been proposed that allows to retrieve information on vertices and hb-edges.

What the future work is?
- Find an efficient recommandation system based on the browsing experience of the user
- Data linkage of multiple queries provenance
- Having more insights require fast extraction of information from the documents itself

Do you want more information?
- Find my articles on:

http://www.infos-informatique.net

- Contact me:

xavier.ouvrard@cern.ch