The HyperBagGraph DataEdron: An Enriched Browsing Experience of Datasets

Track: Foundation of Data Science and Engineering

Xavier Ouvrard[1,2(✉)], Jean-Marie Le Goff[1], and Stéphane Marchand-Maillet[2]

[1] CERN, 1 Esplanade des Particules, Meyrin, Switzerland
xavier.ouvrard@cern.ch
[2] University of Geneva, Carouge, Switzerland

Abstract. Traditional verbatim browsers give back information linearly according to a ranking performed by a search engine that may not be optimal for the surfer. The latter may need to assess the pertinence of the information retrieved, particularly when s·he wants to explore other facets of a multi-facetted information space. Simultaneous facet visualisation can help to gain insights into the information retrieved and call for further refined searches. Facets are potentially heterogeneous co-occurrence networks, built choosing at least one reference type, and modeled by HyperBag-Graphs—families of multisets on a given universe. References allow to navigate inside the dataset and perform visual queries. The approach is illustrated on Arxiv scientific pre-prints searches.

Keywords: Hyper-Bag-Graphs · Knowledge discovery · Visual queries · Information retrieval

1 Introduction

When browsing a textual database, traditional verbatim browsers give back linear information in the form of ranked list of short reference description. To increase the pertinence of this information, the surfer has often to perform additional searches either by refining the original search terms s·he used or by using other pertinent queries that can help her·him to refine the retrieved information.

In an information space, meaningful information can be regrouped by hierarchical classification or—non exclusive—by semantically cohesive categories that

This work is part of the PhD of X. Ouvrard, done at UniGe and funded by a doctoral position at CERN, co-supervised by Pr. S. Marchand-Maillet and Dr J.M. Le Goff. The authors are really thankful to Tullio Basaglia (CERN Library).

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-38919-2_30) contains supplementary material, which is available to authorized users.

A. Chatzigeorgiou et al. (Eds.): SOFSEM 2020, LNCS 12011, pp. 362–374, 2020.
https://doi.org/10.1007/978-3-030-38919-2_30

are combined to express concepts, called facets [1]. Those facets are linked by the physical entities contained in the search output. Choosing a type of reference linked to these entities enables the construction of a co-occurrence network per facet and enhance navigation in the information space. For instance, in scientific publications different information are linked in an article: the article reference, the authors, the main keywords... All this metadata can potentially give insights into the information space and can be chosen as reference to build co-occurrences. Choosing as reference for instance the article id, facets depict co-occurrence networks, either of homogeneous type, such as co-authors or co-keywords, or of heterogeneous types, i.e. combining multiple types together. Co-occurrences can potentially contain repetitions or require an individual weighting: modeling it requires multisets instead of sets.

We propose in this article a new way to explore an information space by using hyper-bag-graphs (hb-graphs for short)—families of multisets on a universe called the vertex set—a mathematical structure we introduced in [2]. Hb-graphs are a separate mathematical category from the one of hypergraphs. This is an important difference as hb-graphs store extra-information that can not be kept with hypergraphs and have different algebra operations. Moreover, we have shown in [3] that hb-graphs enhance exchange-based diffusion over co-occurrence networks, providing a fine vertex and hb-edge ranking.

We propose four extensions of the hypergraph framework of [4]. First, the visualisation part is extended to support hb-graphs: it is an important mathematical generalization that supports redundancy and hb-edge based weighting of vertices that requires multiset families (hb-graphs) instead of subset families (hypergraphs). Second, the new framework supports navigation of heterogeneous co-occurrence networks; in the former framework only homogeneous co-occurrences where allowed. Third, multi-references for building co-occurrences is tackled. Fourth, an application is given with Arxiv search, by the implementation of a 2.5D interface to perform visual queries and visualize the Arxiv information space.

Section 2 lists the related work and the mathematical background. Section 3 presents the hb-graph framework. Section 4 gives results and Sect. 5 concludes.

2 Related Work and Mathematical Background

2.1 Information Space Discovery

Discovering knowledge in an information space requires to gather meaningful information, either hierarchically or semantically. Semantics provide support to the definition of facets within an information space [1].

Navigation and visualisation of the information space facets have been achieved previously in many different ways. [5] uses a pivot to stroll between three facets; the approach, based on a tripartite graph, is limited to the visualisation of a small amount of pivots at the same time. In [6], an interactive exploration of implicit and explicit relations in faceted datasets is proposed. The space of visualisation is shared between different metadata with cross findings between metadata, partitioning the space in categories. [7] proposes a visual analytics

graph-based framework for exploring an information space. The labeled graph representing the dataset is explored by retrieving paths with same type vertices going through reference vertices. Visualisation facets are navigable graphs of pairwise collaborations.

2.2 Co-occurrence Networks

Data mining is only one step in the knowledge discovery processing chain. If numerical data allows rich statistics on the instances, non numerical data mining consists often in summarizing data as occurrences. Alternately, techniques using data instance similarities such as k-nearest neighbors can be used to link different occurrences: however in high dimensionality, they are limited by the curse of dimensionality, even if some techniques limit its effect [8]. Retrieving links through the dataset itself is another way of detecting co-occurrences.

If the dataset reflects existing links—as group of friends in social networks—the job is easier since an inherent co-occurrence/collaboration network can be built through the data instances. Nonetheless, links are often neither direct nor tangible: thus co-occurrences need to be built or processed from the dataset.

A dataset can be a set of physical references, stored as rows in traditional relational databases. Each physical reference has metadata instances attached to it. Metadata instance types can be either interesting for visualisation or processing additional information. The set of physical references and metadata instances used for visualisation provide the types of the network, each type being seen either as a reference or a facet of the information space. This allows—as it will be explained in the next section—the retrieval of co-occurrences in one facet, based on one reference type—which can differ from the physical reference.

2.3 Multisets and Hb-Graphs

Co-occurrences seen as collaborations are m-adic relationships of occurrences, often modeled as hypergraphs, i.e. families of subsets of a given vertex set. But hypergraphs, as they are subsets, do not support neither hyperedge-based repetition nor hyperedge-based weighting of vertices. Hb-graphs—introduced newly in [2]—as multiset families naturally allow them.

Multisets—also known as bags or msets—have been used for a long time in many domains such as text representation and image. Multisets support the individual weighting of their elements by using a **multiplicity function** on a set called the **universe**. The elements that have non-zero multiplicity value belong to the **support** of the multiset. A **natural multiset** occurs when the multiplicity function has its range in the non-negative integers[1].

More information on hypergraphs and multisets, with additional references, can be found in [3].

[1] We denote $\mathfrak{A}_m = \{x_i^{m_i} : i \in [\![n]\!]\}$ where $m_i = m(x_i)$ a mset $\mathfrak{A}_m = (A, m)$ of universe $A = \{x_i : i \in [\![n]\!]\}$, of multiplicity function m and of support $\mathfrak{A}_m^\star = \{x_i : m_i \neq 0\}$.

Table 1. Synthesis of the framework

METADATA	Schema hypergraph ↓	Related to database structure	$\mathcal{H}_{\text{Sch}} = (V_{\text{Sch}}, E_{\text{Sch}})$
	Extended schema hypergraph ↓	Store possible additional processings	$\overline{\mathcal{H}_{\text{Sch}}} = (\overline{V_{\text{Sch}}}, \overline{E_{\text{Sch}}})$
	Extracted extended schema hypergraph ↓	U: set of metadata of interest (visualisation and reference)	$\mathcal{H}_X = (V_X, E_X)$ where $V_X = U$, $E_X = \{e \cap U : e \in \overline{E_{\text{Sch}}}\}$
	Reachability hypergraph ↙↓↘	Hyperedges are connected components E_{cc} ($\subset V_X$) of $\mathcal{H}_X$	$\mathcal{H}_R = (V_R, E_R)$ $V_R = V_X$ $E_R = \{E_{\text{cc}} : E_{\text{cc}}$ c.c. of $H_X\}$
	Navigation hypergraph ↓	Choose: $e_r \in E_R$ references $R_{\text{ref}} \subset e_r$	$\mathcal{H}_N = (V_N, E_N)$ $V_N = V_R \backslash R_{\text{ref}}$ $E_N = \{e_r \backslash R : R \subseteq R_{\text{ref}} \wedge R \neq \emptyset\}$
DATA	Facet visualisation hb-graphs	Co-occurrence networks as hb-graphs	

Following [2], a **hb-graph** $\mathfrak{H} = (V, \mathfrak{E})$[2] is a family of multisets called hb-edges $\mathfrak{E} = (\mathfrak{e}_i)_{i \in [\![p]\!]}$ having the same universe $V = \{v_1, ..., v_n\}$ called the vertex set. Each hb-edge $\mathfrak{e}_i \in \mathfrak{E}$ has its own multiplicity function: $m_{\mathfrak{e}_i} : V \to \mathbb{W}$ where $\mathbb{W} \subset \mathbb{R}^+$. A hb-edge can be seen as a dependent weighted system of vertices. A hb-graph with only natural multisets as hb-edges is said **natural**. A **hypergraph** appears as a particular case of natural hb-graph with a binary value—0 or 1—for each hb-edge multiplicity.

The **support hypergraph** $\overline{\mathfrak{H}} = (V, E)$ of a hb-graph $\mathfrak{H} = (V, \mathfrak{E})$ is the hypergraph of same vertex set V and of hyperedges $E = (\mathfrak{e}_i^\star)_{i \in [\![p]\!]}$. The support hypergraph is unique for a given hb-graph. But reconstructing the hb-graph from a support hypergraph generates an infinite number of hb-graphs, showing that the information contained in a hb-graph is denser than in a hypergraph.

Hb-graph unnormalized extra-node representation is obtained by adding an extra-node per hb-edge linked to each hb-edge support vertex with a link thickness proportional to the vertex multiplicity. Figure 2.b(ii) shows an example.

3 Hb-Graph Framework

3.1 Enhancing Navigation

For the sake of clarity, we briefly summarize in Table 1 the enhancement of navigation of [4], achieved by defining different hypergraphs at the metadata level. We take as thumbnail an example based on a publication dataset. Possible metadata types are: *publication id*, title, abstract, *authors*, affiliations, addresses, *author keywords*, *publication categories*, *countries*, *organizations*, and eventually some

[2] We use fraktur font for multisets and hb-graphs: $\mathfrak{A} : A, \mathfrak{e} : e, \mathfrak{E} : E, \mathfrak{H} : H$.

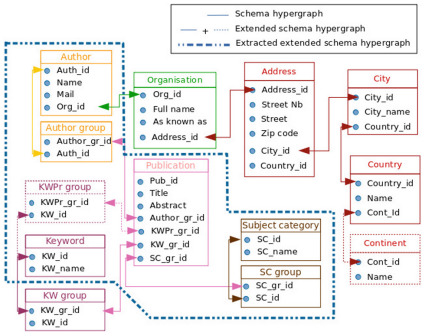


Fig. 1. Schema hypergraph, extended schema hypergraph, Extracted extended schema hypergraph: exploded view shown on an example of publication dataset

processed metadata types such as *processed keywords*, *continent, ...*[3] Enhancing navigation supposes first to define the **schema hypergraph** reflecting the relationships between the database metadata instances. We give the possibility to extend it into **an extended schema hypergraph** to store potential additional processings. Out of the latter an **extracted extended schema hypergraph** $\mathcal{H}_X = (V_X, E_X)$ is enhanced that keeps metadata instances of interest to build the co-occurrences and to be visualized; it might require some intermediate hyperedge bundling. Figure 1 shows the different hypergraphs.

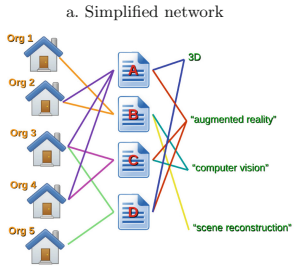
The **reachability hypergraph** $\mathcal{H}_R = (V_R, E_R)$ reflects the connected components of $\mathcal{H}_X$, with $V_R = V_X$: its hb-edges do not intersect. Hence, if $\mathcal{H}_R$ has only one hyperedge, the whole dataset is navigable. We assume that in each hyperedge of the reachability hypergraph, there is at least one metadata type or a combination of metadata types that can be chosen as the **physical reference**. The data instance related to this reference is supposed to be unique. For instance, in a publication dataset the physical reference is the publication id of the publication itself. In the example, the extracted hypergraph has only one component {publication id, authors, processed keywords, subject categories}.

Each hyperedge $e_r \in E_R$ of $\mathcal{H}_R$ leads to one new **navigation hypergraph** $\mathcal{H}_N = (V_N, E_N)$ by choosing a non-empty subset R_{ref} of e_r of possible reference types of interest. The choice of a subset R of R_{ref} allows to consider the remaining vertices of $e_r \backslash R$ as visualisation vertex types, that will be used to generate the facet visualisation hb-graphs and are called the visualisation types. Hence: $E_N = \{e_r \backslash R : R \subseteq R_{\text{ref}} \wedge R \neq \emptyset\}$. When there is only one reference of interest selected at a time in R_{ref} we denote $E_{N/1}$ for E_N. In the publication database example, many navigation hyperedges are possible; the navigation hyperedge choosing as reference publication ids is {authors, publication categories, processed keywords} while using processed keywords as reference is: {authors, publication category, publication ids}.

[3] Metadata of interest for visualisation or referencing are in italic.

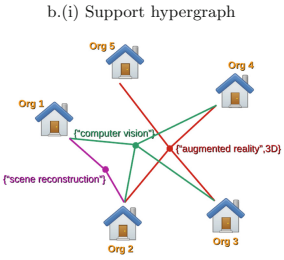
3.2 Facet Visualisation Hb-Graphs

In [4], we use sets to store co-occurrences. Nonetheless in many cases, it is worth storing additional information by joining a multiplicity—with nonnegative integer or real values—to the elements of co-occurrences. A small example emphasizes the interest of moving towards multisets: we consider the publication network of Fig. 2. In this example, building co-occurrences accounting the occurrence multiplicity induces not only a refined visualisation, with distinguishable hb-edges in between some of the vertices (augmented reality and 3D) but also yields to refined rankings of both vertices and hb-edges, as mentioned in [3]. As some parts relies on a mathematic description they have been put in Appendice. The reader can always refer to Fig. 2 for an illustration of the concepts where we choose the keywords as reference.

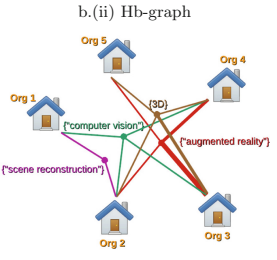


b. Co-occurrences of organizations with keyword as reference

scene reconstruction	$\{\{\text{Org } 1^1, \text{Org } 2^1\}\}$
computer vision	$\{\{\text{Org}1^1, \text{Org}2^1, \text{Org } 3^1, \text{Org } 4^1\}\}$
augmented reality	$\{\{\text{Org } 2^1, \text{Org } 3^3, \text{Org } 4^2, \text{Org } 5^1\}\}$
3D	$\{\{\text{Org } 2^1, \text{Org } 3^2, \text{Org } 4^1, \text{Org } 5^1\}\}$



Vertex weighted degree ranking:
1. Org 2;
2. Org 3; Org 4;
4: Org 5; Org 1.
Hyperedge weighted cardinality ranking:
1. CV, AR, 3D; 2. SR



Vertex weighted m-degree ranking:
1. Org 3;
2. Org 2; Org 4;
4: Org 5; Org 1.
Hb-edge weighted m-cardinality ranking:
1. AR; 2: 3D; 3: CV; 4: SR

Fig. 2. A simplified publication network with publication id, organizations, keywords.

Each physical entity d of a dataset $\mathcal{D}$ corresponds to a unique physical reference r. d is described by a set of data instances of different types that are in $\alpha \in \overline{V_{\text{Sch}}}$. We write $\mathcal{I}$ the set of data instances in $\mathcal{D}$, and t the type application that gives the type of an instance.

Hb-graphs requires a common universe taken as vertex set. We consider for each type α, its instance set $U_\alpha = \{i : i \in \mathcal{I} \wedge t(i) = \alpha\}$ of instances of $\mathcal{D}$ of type α. The common universe for the visualisation hb-graph depends on the search.

We write $\mathfrak{A}_{\alpha,r} = (U_\alpha, m_{\alpha,r})$ the **multiset** of universe U_α, of the values of type α—possibly none—that are attached to d, the physical entity of reference r. The support of $\mathfrak{A}_{\alpha,r}$ is $\mathfrak{A}^\star_{\alpha,r} = \{a_{i_1}, ..., a_{i_{k_r}}\}$. Hence, we abusively write: $\mathfrak{A}_{\alpha,r} = \left\{ a_{i_1}^{m_{\alpha,r}(a_{i_1})}, ..., a_{i_{k_r}}^{m_{\alpha,r}(a_{i_{k_r}})} \right\}$ omitting the elements of U_α that have a zero multiplicity in $\mathfrak{A}_{\alpha,r}$.

d is entirely described by its reference r and the family of multisets, corresponding to homogeneous co-occurrences of the different types α in $\overline{V_{\text{Sch}}}$ linked to the physical reference, i.e. $\left(r, (\mathfrak{A}_{\alpha,r})_{\alpha \in \overline{V_{\text{Sch}}}}\right)$.

In Fig. 2, the publication id is the physical reference. Taking as reference the publication id, the co-occurences for the Publication A of organisations are: $\left\{\text{Org } 2^1, \text{Org } 3^1, \text{Org } 4^1\right\}$ and of keywords are: $\left\{3\text{D}^1, \text{augmented reality}^1\right\}$. The example in Fig. 2.b shows a reference that is not the physical reference.

Type heterogeneity in co-occurrences can enable simultaneous view of different types in a single facet. To allow type heterogeneity in co-occurrences, we consider a partition Γ of the different types in $\overline{V_{\text{Sch}}}$. Each type belonging to an element ν of the partition Γ will be visualized simultaneously in a co-occurrence: it enriches the navigation process, allowing heterogeneous co-occurrences. An interesting case is when ν has a semantic meaning and elements of ν appear as an "is a" relationship. For instance in a publication database organizations regroups "institute" and "company". Also, we consider $\mathfrak{A}_{\nu,r} \triangleq (U_\nu, m_{\nu,r})$, where $U_\nu \triangleq \bigcup_{\alpha \in \nu} U_\alpha$, of support $\mathfrak{A}^\star_{\nu,r} \triangleq \bigcup_{\alpha \in \nu} \mathfrak{A}^\star_{\alpha,r}$ such that $m_{\nu,r}(a) \triangleq \begin{cases} m_{t(a),r}(a) & \text{if } a \in \mathfrak{A}^\star_{\nu,r} \\ 0 & \text{otherwise} \end{cases}$.

d is entirely described in the case of heterogeneous co-occurrences by $\left(r, (\mathfrak{A}_{\nu,r})_{\nu \in \Gamma}\right)$. The homogeneous co-occurrences are retrieved when all $\nu \in \Gamma$ are singletons.

Performing a search on the dataset retrieves a set $\mathcal{S}$ of physical references r. In the single-reference-restricted navigation hypergraph, each hyperedge $e_N \in E_{N/1}$ describes accessible facets relatively to a chosen reference type $\rho \in V_N \backslash e_N$. Given a partition $\gamma \in \Gamma_N$, where $\Gamma_N \triangleq \{\nu \cap e_N : \nu \in \Gamma\}$ is the induced partition of e_N related to the partition Γ of $\overline{V_{\text{Sch}}}$, the associated facet shows the visualisation hb-graph $\mathfrak{H}_{\gamma/\rho,\mathcal{S}}$ where the hb-edges are the heterogeneous co-occurrences of types in γ relatively to reference instances of type ρ (γ/ρ as short) retrieved from the different references in $\mathcal{S}$.

We then build the co-occurrences γ/ρ by considering the set of all values of type ρ attached to all the references $r \in \mathcal{S}$: $\Sigma_\rho \triangleq \bigcup_{r \in \mathcal{S}} \mathfrak{A}^\star_{\rho,r}$. Each element s of Σ_ρ is mapped to a set of physical references $R_s \triangleq \left\{r : s \in \mathfrak{A}^\star_{\rho,r}\right\} \in \mathcal{P}(\mathcal{S})$ in which they appear: we write r_ρ the mapping. The multiset of values $\mathfrak{e}_{\gamma,s}$ of types $\alpha \in \gamma$ relatively to the reference instance s is $\mathfrak{e}_{\gamma,s} \triangleq \biguplus_{r \in R_s} \mathfrak{A}_{\gamma,r}$.

The **raw visualisation hb-graph** for the facet of heterogeneous co-occurrences γ/ρ attached to the search $\mathcal{S}$ is defined as: $\mathfrak{H}_{\gamma/\rho,\mathcal{S}} \triangleq \left(\bigcup_{r \in \mathcal{S}} \mathfrak{A}^\star_{\gamma,r}, (\mathfrak{e}_{\gamma,s})_{s \in \Sigma_\rho}\right)$. Fig. 2.b(ii) gives an example of such a raw visualisation hb-graph.

Since some hb-edges can possibly point to the same sub-mset of vertices, we build a reduced visualisation weighted hb-graph from the raw visualisation

hb-graph. To achieve it we define: $g_\gamma : s \mapsto \mathfrak{e}_{\gamma,s}$ and $\mathcal{R}$ the equivalence relation such that: $\forall s_1 \in \Sigma_\rho, \forall s_2 \in \Sigma_\rho : s_1 \mathcal{R} s_2 \Leftrightarrow g_\gamma(s_1) = g_\gamma(s_2)$.

Considering a quotient class $\overline{s} \in \Sigma_\rho/\mathcal{R}$[4], we write $\overline{\mathfrak{e}_{\gamma,\overline{s}}} \triangleq g_\alpha(s_0)$ where $s_0 \in \overline{s}$.

$\overline{E_\gamma} \triangleq \left\{\overline{\mathfrak{e}_{\gamma,\overline{s}}} : \overline{s} \in \Sigma_\rho/\mathcal{R}\right\}$ is the support set of the multiset $\{\{\mathfrak{e}_{\gamma,s} : s \in \Sigma_\rho\}\}$: $\overline{\mathfrak{e}_{\gamma,\overline{s}}} \in \overline{E_\gamma}$ is of multiplicity $w_\gamma\left(\overline{\mathfrak{e}_{\gamma,\overline{s}}}\right) = |\overline{s}|$ in this multiset.

It yields: $\{\{\mathfrak{e}_{\gamma,s} : s \in \Sigma_\rho\}\} = \left\{\overline{\mathfrak{e}_{\gamma,\overline{s}}}^{w_\gamma\left(\overline{\mathfrak{e}_{\gamma,\overline{s}}}\right)} : \overline{s} \in \mathcal{S}_\rho/\mathcal{R}\right\}$.

Let $\tilde{g_\gamma} : \overline{s} \in \Sigma_\rho/\mathcal{R} \mapsto \mathfrak{e} \in \overline{E_\gamma}$, then $\tilde{g_\gamma}$ is bijective. $\tilde{g_\gamma}^{-1}$ allows to retrieve the class associated to a given hb-edge; hence the associated values of Σ_ρ to this class—which will be important for navigation. The references associated to $\mathfrak{e} \in \overline{E_\gamma}$ are $\bigcup_{s \in \tilde{g_\gamma}^{-1}(\mathfrak{e})} r_\rho(s)$. The **reduced visualisation weighted hb-graph** for the search $\mathcal{S}$ is defined as $\mathfrak{H}_{\gamma/\rho,w_\gamma,\mathcal{S}} \triangleq \left(\bigcup_{r \in \mathcal{S}} \mathfrak{A}^\star_{\gamma,r}, \overline{E_\gamma}, w_\gamma\right)$.

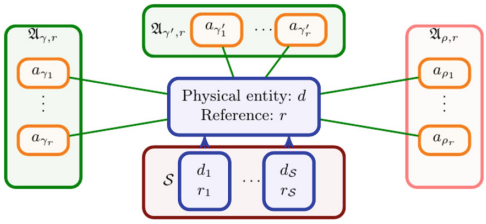


Fig. 3. Navigating between facets of the information space

Using the support hypergraph of the visualisation hb-graphs retrieves the results given in the case of homogeneous co-occurrences in [4]: hence [4] appears as a particular case of the new hb-graph framework.

3.3 Navigability Through Facets

As for a given search $\mathcal{S}$ and a given reference ρ, the sets Σ_ρ and $R_s, s \in \Sigma_\rho$ are fixed, the navigability can be ensured between the different facets. We consider a group of types γ, its visualisation hb-graph $\mathfrak{H}_{\gamma/\rho,w_\gamma}$ and a subset A of the vertex set of $\mathfrak{H}_{\gamma/\rho,w_\gamma}$. We target another group of types γ' of heterogeneous co-occurrences referring to ρ for visualisation. Figure 3 illustrates the navigation.

We suppose that the user selects elements of A as vertices of interest from which s·he wants to switch facet. Hb-edges of $\overline{E_\gamma}$ which contain at least one element of A are gathered in $\overline{E_\gamma}\big|_A \triangleq \left\{\mathfrak{e} : \mathfrak{e} \in \overline{E_\gamma} \wedge (\exists x \in \mathfrak{e} : x \in A)\right\}$. Using the application $\tilde{g_\gamma}^{-1}$ we retrieve the corresponding class of references of type ρ associated to the elements of $\overline{E_\gamma}\big|_A$, to build the set of references $\overline{V}\big|_A$ of type ρ involved in the building of co-occurrences of type γ'. Each of the classes in $\overline{V}\big|_A$

[4] $\Sigma_\rho/\mathcal{R}$ is the quotient set of Σ_ρ by $\mathcal{R}$.

contains instances of type ρ that are gathered in a set $\mathcal{V}_{\rho,A}$. Each element of $\mathcal{V}_{\rho,A}$ is linked to a set of physical references by r_ρ. Hence we obtain the physical reference set involving elements of A: $\mathcal{S}_A \triangleq \bigcup_{s\in\nu_{\rho,A}} R_s$.

The raw visualisation hb-graph $\mathfrak{H}_{\gamma'/\rho}\big|_A \triangleq \left(\bigcup_{r\in\mathcal{S}_A} \mathfrak{A}^\star_{\gamma',r}, (\mathfrak{e}_{\gamma',s})_{s\in\mathcal{V}_{\rho,A}}\right)$ in the targeted facet is now enhanced using $\mathcal{S}_A$ as search set instead of set $\mathcal{S}$. To obtain the reduced weighted version we use the same approach as above. The multiset of co-occurrences retrieved includes all occurrences that have co-occurred with the references attached to one of the elements of A selected in the first facet. Of course if $A = A_{\gamma,S}$ the reduced visualisation hb-graph contains all the instances of type γ' attached to physical entities of the search $\mathcal{S}$.

In Fig. 2.b(ii), with $A = \{\text{Org1}\}$, allows to retrieve two hb-edges: computer vision—attached to PubB and PubC—and scene reconstruction—PubB. Hence: $S_A = \{\text{PubB}, \text{PubC}\}$. Switching to the Publication facet and keeping as reference keywords, two hb-edges $\{\text{PubB}^1, \text{PubC}^1\}$ and $\{\text{PubB}^1\}$ are retrieved. The same with $A = \{\text{Org1}, \text{Org2}\}$ retrieves all the co-occurences of Publications with reference to keywords.

The reference type can always be shown in one of the faces as a visualisation hb-graph where all the hb-edges are constituted of the reference itself with multiplicity the number of time the reference occurs in the hb-graph.

Ultimately, by building a multi-dimensional network organized around groups of types, one can retrieve very valuable information from combined data sources. This process can be extended to any number of data sources as long as they share one or more types. Otherwise the reachability hypergraph is not connected and only separated navigations are possible.

3.4 The Case of Multiple References

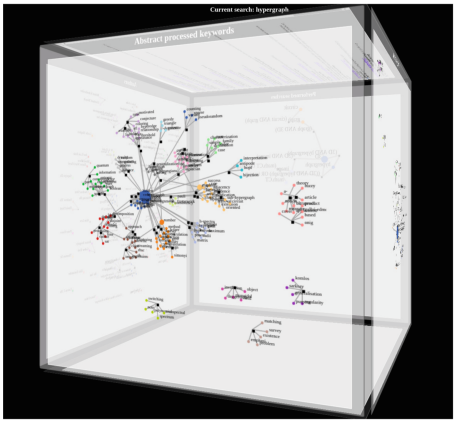
Extending co-occurrences to multiple references chosen in $e_R \in E_R$ is not straightforward. There are two ways of doing so: a disjunctive and a conjunctive way. We consider the set $R \subset e_R$ of references and $e_N = e_R \backslash R$ the visualisation types.

In the disjunctive way, each co-occurrence is built using the same approach than before considering successively each type $\rho \in R$. This is particularly adapted for types that are partitioning the physical references. It is the case for instance in the aggregation of two databases on two different kind of physical data, such as publication and patent, and the co-occurrence of the chosen navigation type is built referring in this case either to a publication or (non-exclusive) to a patent. The hb-graphs obtained are built by extending the family of hb-edges.

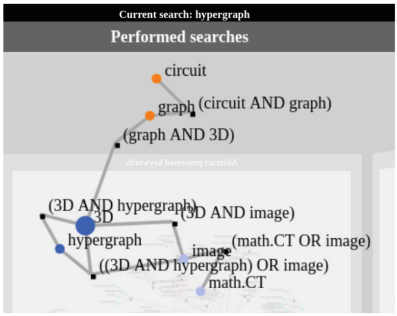
In the conjunctive approach, we start by building the cross product of instances of the references and retrieve co-occurrences of elements for which the data d is attached to the corresponding values of cross-reference instances. Hence co-occurrences are restricted to the simultaneous presence of reference instances attached to the physical entity.

3.5 The DataHbEdron[5]

The DataHbEdron provides soft navigation between the different facets of the information space. Each facet of the information space corresponding to a visualisation type includes a visualisation hb-graph viewed in its 2D extra-node representation with a normalised thickness on hb-edges [2]. The different facets are embedded in a 2.5D representation called the DataHbEdron. The DataHbEdron can be toggled between a cube with six faces—Figure 4—and a carousel shape with n faces—not shown here due to the lack of space—to ease navigation between facets. The reference face shows a traditional verbatim list of references corresponding to the search output.



(a) Cube shape



(b) Performed search

Fig. 4. DataHbEdron: cube shape.

Individual faces of the DataHbEdron show different facets of the information space: the underlying visualisation hb-graphs support the navigability through facets. Hb-edges can be selected interactively between the different facets; since each hb-edge is linked to a subset of the references, the corresponding references can be used to highlight information in the different facets as well as in the face containing the reference visualisation hb-graph.

4 Results, Evaluation and Conclusion

4.1 Use Case

We applied this framework to perform searches and visual queries on the Arxiv database allowing simultaneous visualisation of the different facets of the information space constituted by authors, extracted keywords and subject categories.

[5] A video demo is available on: https://www.infos-informatique.net.

The tool developed is now part of the Collaboration Spotting family[6]. When performing a search, the standard Arxiv API[7] is used to query the Arxiv database. The queries can be formulated either by a text entry or done interactively directly using the visualisation: queries include single words or multiple words, with possible Boolean query operators—AND, OR and NOT—and parenthesis groupings. The querying history is stored and presented as an interactive hb-graph to visualize the construction of complex queries including refinement of the queries already performed. Each time a new query is formulated, the corresponding metadata is retrieved by the Arxiv API.

When performing a search on Arxiv, the query is transformed into a vector of words. Arxiv relies on Lucene's built-in Vector Space Model of information retrieval and the Boolean model. The most relevant documents are retrieved based on a similarity measure between the query vector and the word vectors associated to individual documents. The API returns the top n highest scored document metadata associated to the document. Metadata, filled by authors during their submission of a preprint, contains different information such as authors, Arxiv categories and abstract.

The facets are shown on the DataHbEdron with additional faces: the first face shows the Arxiv reference visualisation hb-graph with a layout similar to classical textual search engines. The second face corresponds to the visualisation hb-graph of co-authors. The third face depicts the visualisation hb-graph of co-keywords extracted from the abstracts using classical natural language processing and TF-IDF that is used as keyword multiplicity. The fourth face shows the hb-graph of Arxiv categories. The fifth face shows past or reloaded queries of the session.

Any node on any face is interactive to highlight information from one face to another showing the hb-edges that are mapped through the references. Queries can be built using the vertices of the hb-graph, either isolated or in combination with the current search using AND, OR and NOT. The first query is the only one required to be typed in. Merging queries of different users is immediate as they correspond to hb-edges of a hb-graph. Queries are evolving, gathered, stored and re-executable months later. The surfer has the possibility to display additional contextual information related to authors using DBLP, to keywords using DuckDuckGo for disambiguation and Wikipedia.

4.2 Evaluation

The validity of our framework is asserted by the mathematical construction completeness and robustness: we have achieved the possibility to navigate inside the dataset by showing co-occurrences in a sufficient refined way to support all the information extracted. As this model has been instantiated through a user interface in the use case of Arxiv, but, also, as mentioned previously, on some other sample data using csv files, its versatility is ensured. We have gathered in Table 2 some of the non-exhaustive features that allows to compare our solution

[6] http://collspotting.web.cern.ch/.
[7] https://arxiv.org/help/api/index.

Table 2. Elements of comparison (see text for details)

	Verbatim browser	PivotPath [5]	PivotSlice [6]	CS core [7]	DataEdron cube [4]	DataHbEdron
output	linear	tripartite graph	graph	graph	linear & hypergraph	linear & hb-graph
#facets	1	3	many	many	4	many
view per facet	no	no	no	yes	yes	yes
simultaneous facet views	no	yes	yes	no	yes	yes
heterogeneous co-occurrences	x	no	no	yes	no	yes
multiple references	x	no	no	disjunctive	no	conjunctive, disjunctive
zoom in data	new query	no	yes	yes	no	yes
filter data	new query	no	yes	yes	no	by visual queries
visual query	no	no	yes, restricted to current search	yes, restricted to current search	no	yes, even with new search
redundancy in co-occurrences	x	no		no	no	yes
information extraction	limited	pivot change	elaborated questions	elaborated questions	elaborated questions	elaborated questions
combination of facets	no	no	yes	yes	yes	yes
type of ranking	binary cosine similarity	no		number of references per vertex	hyperedges and vertices	hb-edges and vertices

with others. The user interface uses a 2.5D approach, but it is out of the scope of this article to make any claim on the quality of the interactions a user can have with such an interface.

5 Future Work and Conclusion

The framework supports dataset visual queries, possibly contextual, that either result from searches on related subjects or refine the current search: it enables full navigability of the information space. It provides powerful insights into datasets using simultaneous facet visualisation of the information space constructed from the query results. This framework is versatile enough to enhance user insight into many other datasets, particularly textual and multimedia ones.

References

1. Ranganathan, S.R.: Elements of Library Classification. Asia Publishing House, Mumbai (1962)
2. Ouvrard, X., Le Goff, J.-M., Marchand-Maillet, S.: Adjacency and tensor representation in general hypergraphs. part 2: multisets, hb-graphs and related e-adjacency tensors. arXiv preprint arXiv:1805.11952 (2018)

3. Ouvrard, X., Le Goff, J.-M., Marchand-Maillet, S.: Diffusion by exchanges in hb-graphs: highlighting complex relationships extended version. arXiv:1809.00190v2 (2019)
4. Ouvrard, X., Le Goff, J., Marchand-Maillet, S.: Hypergraph modeling and visualisation of complex co-occurence networks. Electron. Notes Discrete Math. **70**, 65–70 (2018)
5. Dörk, M., Riche, N.H., Ramos, G., Dumais, S.: PivotPaths: strolling through faceted information spaces. IEEE Trans. Vis. Comput. Graph. **18**(12), 2709–2718 (2012)
6. Zhao, J., Collins, C., Chevalier, F., Balakrishnan, R.: Interactive exploration of implicit and explicit relations in faceted datasets. IEEE Trans. Vis. Comput. Graph. **19**(12), 2080–2089 (2013)
7. Agocs, A., Dardanis, D., Le Goff, J.-M., Proios, D.: Interactive graph query language for multidimensional data in collaboration spotting visual analytics framework. ArXiv e-prints, December 2017
8. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 604–613. ACM (1998)