ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph

Philippe Lemoisson, Guillaume Surroca, Clement Jonquet, Stefano A. Cerri

HAL Id: lirmm-01605411

https://hal-lirmm.ccsd.cnrs.fr/lirmm-01605411

Submitted on 11 May 2018

ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph

Philippe Lemoisson[1,2]**, Guillaume Surroca**[3]**, Clement Jonquet**[3] **and Stefano A. Cerri**[3]

[1] Cirad, TETIS, Montpellier, France
[2] TETIS, APT, Cirad, CNRS, Irstea, Univ. Montpellier, Montpellier, France
[3] Laboratory of Informatics Robotics Microelectronics of Montpellier (LIRMM), University of Montpellier & CNRS, France

philippe.lemoisson@cirad.fr ; {surroca, jonquet, cerri}@lirmm.fr

Abstract: Formal data is supported by means of specific languages from which the syntax and semantics have to be mastered, which represents an obstacle for collective intelligence. In contrast, informal knowledge relies on weak/ambiguous contributions e.g., *I like*. Reconciling the two forms of knowledge is a big challenge. We propose a brain-inspired knowledge representation approach called ViewpointS where formal data and informal contributions are merged into an adaptive knowledge graph which is then topologically, rather than logically, explored and assessed. We firstly illustrate within a mock-up simulation, where the hypothesis of knowledge emerging from preference dissemination is positively tested. Then we use a real-life web dataset (MovieLens) that mixes formal data about movies with user ratings. Our results show that ViewpointS is a relevant, generic and powerful innovative approach to capture and reconcile formal and informal knowledge and enable collective intelligence.

Keywords: knowledge representation, knowledge graph, semantic Web, social Web, collective intelligence, nature-inspired computational model, formal/informal knowledge, serendipitous learning.

1. Introduction

Formal data is supported by means of specific languages, e.g., the chemical description of an organic compound or the conceptual model of a relational database. Mastering the syntax and semantics of such a language is therefore a strong pre-condition for turning the data into knowledge. This happens to be a serious obstacle on the path towards shared understanding and collective intelligence. The phenomenon is well known in computer science: despite significant investment in modelling languages such as Unified Modeling Language, the dialog between final users and software analysts remains painful and sometimes sterile (Anthony J H Simons & Graham 1998). The phenomenon is even more accurate in multidisciplinary contexts. For example, when tight collaboration and mutual understanding is required between senior chemists and computer scientists building an electronic encyclopaedia (Krief et al. 2008; Lemoisson & Cerri 2005). Cross connecting ontologies in order to obtain machine-readable data (the so called semantic Web) is a sticking point, due to the fact that each ontology's evolution is domain-dependent (Karapiperis & Apostolou 2006). Even when the ontologies belong to the same domain (e.g., biomedicine), finding the right language (e.g., an ontology) for annotating a specific dataset is in itself a challenge as illustrated in (Jonquet et al. 2010). Shared understanding of formal data, i.e., yielding to the semantic Web (Berners-Lee et al. 2001) is a key issue, even more when societal problems require a multidisciplinary treatment.

In contrast, informal knowledge relies on weak/ambiguous contributions such as "I recommend this paper" or "you should try with this keyword" or "on this topic, you should

meet this researcher". Such contributions can never be assessed correct or fully reliable, nevertheless they do play a key role in our daily lives, even in our scientific lives, for instance when a colleague recommends a paper, or a subject of investigation, or another colleague. They express most of the time spontaneous answers to the basic need of socializing (O'Reilly 2009); it also happens that they are collected with a scientific purpose, e.g., when stakeholders are asked to draw in turn, on a shared map, their personal view about the local context (Saquali et al. 2009). Within the social Web, folksonomies and recommendations emerge on top of countless subjective personal views; these structures are unfortunately not actionable in the logical paradigm, because of their lack of semantics, and absence of centrally controlled coherence (Mikroyannidis 2007; Lux & Dosinger 2007).

Reconciling the two forms of knowledge is a big challenge and significant efforts have been made to bridge the gap between the so-called semantic and the social Web (Gruber 2008). It is for instance a major issue for the Web science community today (Hendler et al. 2008). Facing this issue, Breslin et al. 2011 envisage two options which might be called respectively "post-formalizing" and "pre-formalizing" the informal content:

- The first option (Bojārs et al. 2008; Cimiano & Staab 2004) consists in translating social data into the semantic Web. Close to this first option, some authors (Aberer et al. 2004) promote the bottom-up evolution of ontologies by synthetizing the results of user interactions, e.g., building a lexical resource by aligning the vocabularies of patients and health professionals (Donald et al. 2015). However, the parallel and independent evolutions of social Web contents and ontologies make this kind of bridging delicate to maintain (Uren et al. 2006), (Mika 2007; Mikroyannidis 2007).
- The second option consists in leveraging Web 2.0 services to feed the semantic Web, e.g., integrating structured data for describing movies when rating them (Ankolekar & Krötzsch 2007). However, spreading standards within the open world of Web 2.0 may seem challenging, if not questionable.

To reconcile the two ecosystems however is the key of a *collective knowledge system* in the sense coined by Gruber in (Gruber 2008): a collective knowledge system provides tailored advice on top of collected knowledge, structured data and high level automated expertise able to bring in new levels of understanding.

In this paper we propose a brain-inspired knowledge representation approach where formal data and informal contributions are merged into an adaptive knowledge graph which is then topologically, rather than logically, explored and assessed. Our approach called ViewpointS aims at yielding such a system by: (i) merging informal and formal knowledge into a unified knowledge graph following the brain metaphor, (ii) empowering feedback and learning processes and (iii) providing means for assessing/qualifying the emerging *collective knowledge*. ViewpointS relies on three assumptions:

1. A viewpoint is a subjective connection between two objective knowledge resources; the aggregation of these connections between two given resources can be viewed as a synapse between two neurons;
2. The knowledge graph or "associative memory" formed by all the viewpoints (formal versus informal, proactive versus reactive) is a selectionist system evolving continuously according to user's interactions in the metaphor of a collective brain: each interaction is equivalent to the tuning of a synapse;
3. The topological structures which appear when adopting a user's perspective actualize assessable knowledge.

Our conceptual framework is inspired from the social Web: it builds upon connections established by "agents" about "things" (resources or persons); moreover it records the provenance of each connection as a mean to ensure trustworthiness. Due to the associative

nature of the memory built upon the connections, it continuous evolves along the interactions of the contributors. Rather than undertaking logical assessment (incompatible with informal content), we define *perspectives* (sets of quantification rules tuned to the interpretation of the user's context, which apply on the viewpoints) and topologically explore the knowledge map resulting from such an interpreted knowledge graph. In this paper, we propose and define the ViewpointS formalism supporting the storage, exploitation and assessment of collective knowledge. We firstly illustrate within a mock-up simulation, where the hypothesis of knowledge emerging from preference dissemination is positively tested. This proof of concept is then checked against a real-life web dataset (MovieLens) that mixes formal data about movies with user ratings. Our results show that ViewpointS is a relevant, generic and powerful innovative approach to capture and reconcile formal and informal knowledge as well as a step forward in assessing and exploiting collective intelligence.

The rest of the paper is organised as follows. Section 2 presents our sources of inspiration within the state of the art. Section 3 gives a detailed presentation of the formalism supporting the storage, exploitation and assessment of knowledge. In section 4, a first proof of concept is given: we assess *collective knowledge* emerging from serendipitous learning through a mock-up simulation called "The three princes of Serendip". In section 5, we validate our hypothesis by processing the MovieLens dataset about movies descriptions and ratings. We firstly exhibit the topological structure reflecting the formal knowledge about movies, then input informal knowledge issued from user ratings and study the plasticity of the topological structure under distinct perspectives. Finally, section 6 summarizes the key novel aspects and value added and discusses directions for future work.

2. Related work and inspirations

Starting from the question of 'knowledge', a major source of inspiration has been the Theory of Neuronal Group Selection (Edelman & Reeke 1982; Edelman n.d.; Edelman & Tononi 2000) which recently re-appeared in the front scene under the name of 'connectome' (Seung 2012). According to this theory, the human brain is not a store of fixed or coded attributes to be called up and assembled as in a computer; instead, it results from a process of continual re-categorisation within a network (the cortex) of about 30 billion neurons and 1 million billion synapses. One central and striking assumption in this theory is that most of the brain global/macro capacities rely on a single local/micro mechanism: the variation of the synapses' strengths as a feedback of individual value-systems to experience. In ViewpointS, the key idea is twofold: (i) the unit of knowledge is a connection (we call it *viewpoint*) between two *knowledge resources* (as defined in 3.1); and (ii) the wiring harness of *viewpoints* between a given pair of *knowledge resources* plays the role of 'synapse interconnecting two neurons'; we therefore call it *synapse*. We de facto abandon conceptual models such as (Chein & Mugnier 2008) and rule-based approaches such as (Noh et al. 2010). Instead we allow the emergence of *collective knowledge* in a selection process analogue to that described in (Edelman n.d.), by letting each individual reflect his/her own understanding and/or feedback through the emission of new *viewpoints* updating the *synapses*. Such selection processes exist in various contexts of the social Web: learning character recognition through Recaptchas (von Ahn et al. 2008), evolving folksonomies (Krause & Aras 2009), building powerful semantics on top of lexical data (Lafourcade 2007), recommending (Adomavicius & Tuzhilin 2005). In this paper, we propose to implement a generic, context-free, selection process accepting both formal data and informal contributions.

When going into the syntactic aspects, our starting point was from the recent shift in the Web paradigm bringing the contributor back in the loop (O'Reilly 2009). Bringing back the contributor can be done at a global/macro level through the notions of 'provenance'

(https://www.w3.org/TR/prov-o) and 'named graphs' (Carroll et al. 2005). Or at the local/micro level by extending the traditional bipartite model of ontologies with a social dimension. For instance, (Mika 2007) presents a tripartite model actor-concept-instance based on the commitment of a user in terms of classifying an instance as belonging to a certain concept. Quite emblematic of this local/micro level storage of the provenance is the social tagging system Flickr built upon three dimensional units representing agent/user contributions in terms of tags associated to resources/images: agent-resource-tag. It is interesting to note that contributor-centred views based on triples made their way simultaneously in the semantic and social Web, but with distinct goals and formalisms (Weinberger 2011). Whereas the triples of (Mika 2007) secured the possibility of structure and hierarchy by preserving the central notions of concept and instance, the triples of social tagging systems are deprived from any directly specified parent-child relationship and lead to folksonomies characterised by the absence of hierarchy and lack of semantics (Mikroyannidis 2007) (Lux & Dosinger 2007).

In ViewpointS, we started from a tripartite model agent-resource-tag generalizing Flickr by expressing 'the belief of an agent that the tag is appropriate with respect to the resource'. It should be noted that the classical denotation of viewpoints as beliefs, such as (Self 1990; Attardi & Simi 1995) was deliberately oversimplified in order to yield a topological, rather than logical, exploitation of the "wisdom of the crowd" in a manner similar to (Mika 2007; Markines et al. 2009; Specia & Motta 2007). We then made one decisive step by replacing the triple agent-resource-tag by the more abstract triple agent-resource-resource in which agents may themselves be resources.

Based on these complementary influences, we have set the conditions of a selection process operating on top of the agent-resource-resource triples by aggregating all the triples connecting two given *knowledge resources* into higher level binary links called *synapses*; those build *knowledge maps* where the notion of 'semantic proximity' can be defined. The notions of 'semantic similarity' or 'semantic proximity' are central in the contexts of annotation, disambiguation, concept alignment or information retrieval (Harispe et al. 2013; Lee et al. 2008), (Pedersen et al. 2007). As surveyed in (Markines et al. 2009), various similarities can be defined within triples, sometimes recursively, e.g., in (Quattrone et al. n.d.): "two tags are similar in that they link similar resources and two resources are similar in that they are linked by similar tags". We made the hypothesis that any user's context could be translated into a set of quantification rules for the *viewpoints* and called such a set a *perspective*; this is in line with (Kim et al. 2008) where it is written that "ontologies should be evaluated with respect to a particular goal, application or scenario rather than merely for the sake of an evaluation". Once a *perspective* is adopted, the initial heterogeneous semantics carried by the *viewpoints* are transformed into proximities within a *knowledge map*. As a consequence, the agents can use the *viewpoints* of others (transformed into proximities) when browsing *knowledge maps* and reversely update the *knowledge graph* through *viewpoints* expressing their feedback. Along these exploitation/feedback cycles, the shared knowledge is continuously elicited against the beliefs of the agents in a selection process supported by the evolving strength of *synapses*, as it happens in the brain (Edelman & Tononi 2000). We have previously demonstrated (Lemoisson et al. 2013) the learning ability of a knowledge base implementing this conceptual framework and showed how it allows the search and discovery of knowledge through a search engine prototype for scientific publications (Surroca et al. 2014). We have also assessed the diffusion of preference systems resulting from this selection process (Surroca et al. 2015): we simulated 'proximity-guided serendipity' in the context of several scenarii parametrized by the degree of acceptance of resources outside one's own preference system.

3. Conceptual framework, formalism and knowledge assessment

In this section, we first present a conceptual framework, then a formalism where the formal and the informal merge into a unified and adaptive memory, and we finally propose means exploiting the topology of the memory in order to assess *collective knowledge*. This formalism has been implemented in Java; it is publicly available at:
https://github.com/sifrproject/viewpoints_kernel

3.1. Conceptual framework

The ViewpointS approach elaborates upon the connections, established by the members of a community, between pairs of objects of interest called "knowledge resources". The high-level ontology shown in Figure 1 introduces the *knowledge resources* and the roles they play.
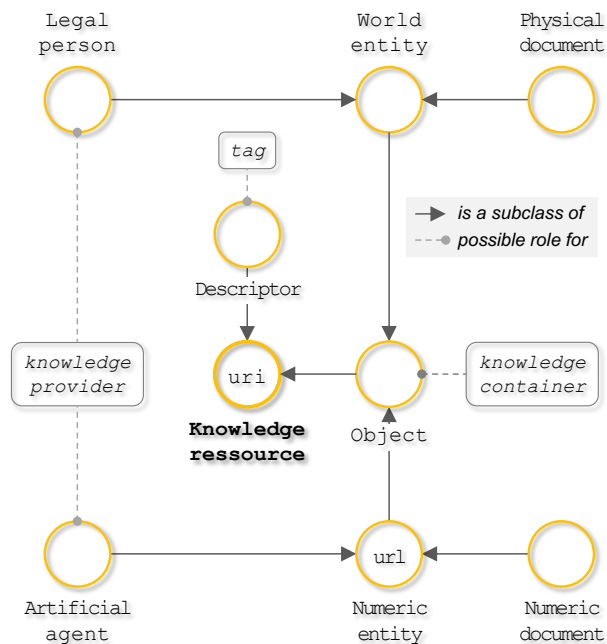

Figure 1: Knowledge resources

A knowledge resource can be:

- either an `object`, i.e., either a `world entity` such as a person, a company, a spatial entity, a monument … or a `numeric entity` hosted by a computer, such as a Web page or an algorithm, i.e., an entity belonging to the 'numeric matter' as coined by (Vial 2013);

- or a `descriptor`, i.e., a meaningful expression within a language such as a keyword, an item in an ontology, a topic. The distinction between `objects` and `descriptors` implicitly refers to the position adopted by J. Searle in (Searle 1969)[1].

Besides, a *knowledge resource* may participate to knowledge construction through three possible roles: i) "tag": this role is played by `descriptors` qualifying `objects`, ii) "knowledge container": this role is played by `objects` embedding some knowledge or iii) "knowledge provider": this role is played by `objects` delivering knowledge.

---

[1] According to J. Searle: i) there is a real world that exists independently of us, independently of our experiences, our thoughts, our language; ii) we have direct perceptual access to that world through our senses, especially touch and vision; iii) words in our language have reasonably clear meanings and can be used to refer to or talk about real objects of the world.

All *knowledge resources* are identified by an URI; `numeric entities` have URLs.[2] All *knowledge resources* have a "preview": in the case of a `numeric entity` a preview service will be provided by the browser accessing the URL; in the case of `world entities`, at least one numeric preview is requested; in the case of `descriptors`, an simple explicit identifier stands for preview, e.g., "Fujiyama", "majestuous."

In the use cases we have handled so far, *knowledge resources* belong to five mutually exclusive classes:

- `Legal person`: `world entity` "performing acts and undertaking obligations" (Carroll et al. 2005); it can play two roles: "knowledge provider" when creating connections or "knowledge container" as a connectable `object`.

- `Artificial agent`: numeric entity creating connections in the role of "knowledge provider" or playing the role of "knowledge container" as a connectable `object`.

- `Physical document`: `world entity` playing the role of "knowledge container", such as a real book on a real shelf.

- `Numeric document`: `numeric entity` playing the role of "knowledge container", such as a tweet or a Web page.

- `Descriptor`: meaningful linguistic expression playing the role of "tag".

In the following, we shall write `agent` to designate indifferently a `legal person` or an `artificial agent`, i.e., potential knowledge providers.

The ViewpointS approach elaborates upon the connections established by `agents` between couples of *knowledge resources*. For instance, when an internaut tags an image with a keyword in Flick'r, we consider that an `agent` (the internaut) connects a `numeric document` (the image) to a `descriptor` (the keyword). In a similar way, when I upload in an open archive the reference of a book authored by myself, we consider that an `agent` (myself) connects a `physical document` (the book) to an `agent` (myself); in that latter case, the same `agent` plays the two roles of "knowledge provider" and "knowledge container". Each of these connections reflects the beliefs of a given agent/subject, i.e., it is subjective by nature; we reify/objective it and call it *viewpoint*. As illustrated in Figure 2, a *viewpoint* keeps track of the provenance of the connection (the `agent` playing the role of "knowledge provider"); it has a type 'θ' ('index', 'author', 'like' …) and a time stamp 'τ'.
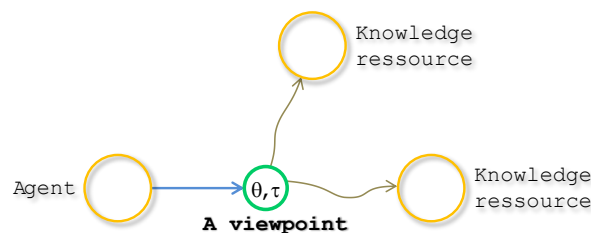

Figure 2: The subjective connection called "viewpoint"; the blue arrow gives the provenance, 'θ' gives the semantics, 'τ' gives the time stamp

3.2. Formalism

Let R be a set of *knowledge resources,* let a_1 be an `agent` taken in R, r_2 and r_3 two arbitrary elements of R, let τ be a standard time stamp,

---

[2] In the semantic web vision, a Uniform Resource Locator (URL) is a specific type of Uniform Resource Identifier (URI) to identify web addresses.

let θ be formed of two parts: i) a polarity indicating whether the proximity between the knowledge resources should be reinforced (positive polarity) or weakened (negative polarity) and ii) semantics, such as standard unambiguous semantic Web properties e.g., `foaf:knows`, `cv:jobType`. In this paper, we consider only positive polarities and shall therefore omit them e.g., we denote 'matches' instead of '(+, matches)'; θ is called the `type` of the *viewpoint*,

---

The *viewpoint* (a_1, {r_2, r_3}, θ, τ) stands for: the `agent` a_1 believes at time τ that r_2 and r_3 are related according to the semantics carried by the `type` θ.

`a₁=EmitterOf(w); {r₂,r₃}=ConnectedBy(w).`

We denote `W` the set of all *viewpoints*.

Let `EDGES` be the set of all links woven between the *viewpoints* and the *knowledge resources*:

`EDGES={(a,w)| w∈W; a=EmitterOf(w)} ∪ {(w,r)|w∈W; r∈ConnectedBy(w)}.`

The *knowledge graph* is the bipartite graph built by taking each *viewpoint* and linking it to its emitter and connected resources:

$$KG_{R,W} = (R∪W, \text{EDGES})$$

For instance, let G_3, B_7 and G_5 (respectively a big green circle, a big blue square and a small green circle) be `world entities`, let `peon-size` be an `agent` detecting sizes, let `princeBlue` be an `agent` fond of blue, let 'big' and 'small' be `descriptors`, the four viewpoints illustrated in Figure 3 express the following knowledge:

`peon-size` believes that G_3 matches 'big'

`peon-size` believes that B_7 matches 'big'

`peon-size` believes that G_5 matches 'small'

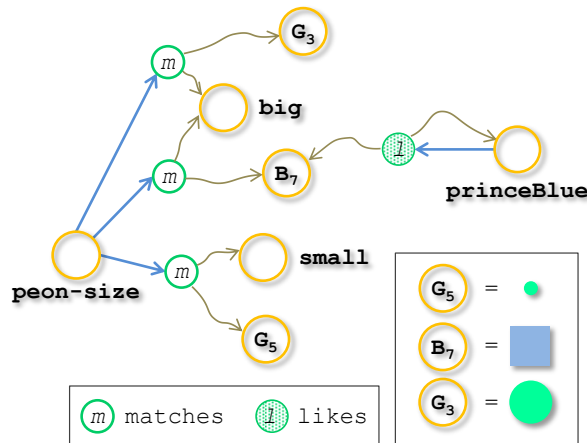`princeBlue` believes that `princeBlue` likes B_7

Figure 3: A knowledge graph $KG_{R,W}$

In order to exploit the knowledge, we build *perspectives* defining rules for quantifying the *viewpoints*. It may be default rules adopted by a group of users in a recurrent context e.g., information retrieval, or specific rules filtering `KG_{R,W}` according to preferences such as: ignoring the *viewpoints* anterior to a given date, privileging the *viewpoints* emitted by some agents or privileging *viewpoints* of a given `type`.

The preliminary step in building a *knowledge map* consists in grouping all the *viewpoints* connecting any given pair of *knowledge resources* into a higher level link called a *synapse*.

We denote `S` the set of all *synapses*. We also consider the `synapse` function:

`. synapse: R²→2^W`

`(r₁,r₂)→{w∈W|Connected(w)={r₁,r₂}}`

Setting a perspective U consists in defining rules for valuating each *synapse*. We consider two functions ^{U}map and ^{U}reduce and combine them into ^{U}synapse.

. ^{U}map evaluates each *viewpoint* w of the *synapse* individually, as a function of 'θ', 'τ' and 'EmitterOf(w)'.

. ^{U}reduce aggregates all the ^{U}map results into a single positive or null value; it may be a simple sum or a sophisticated function with a threshold or attenuation.

^{U}synapse is finally the composition of 3 functions:

. ^{U}synapse: $R^2 \rightarrow \Re+$

IF synapse(r_1,r_2)==∅ THEN ^{U}synapse (r_1,r_2)=0 ELSE ^{U}synapse(r_1,r_2)= ^{U}reduce(^{U}map(synapse (r_1,r_2))).

We call *knowledge map* and denote ^{U}KM$_{R,S}$ the undirected labelled graph interpreting KG$_{R,W}$ through the perspective U. The vertices are the *knowledge resources*. An edge is built for each ^{U}synapse (r_1,r_2)>0.

Figure 4 illustrates this interpretation process with a perspective U⁰ where the aggregation function computing the value of *synapses* is the sum.

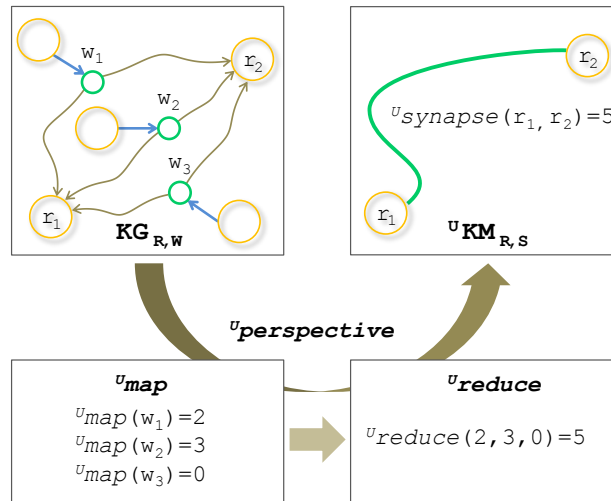*Figure 4: Building a knowledge map ^{U}KM$_{R,S}$*

One single KG$_{R,W}$ will therefore be interpreted into several distinct ^{U}KM$_{R,S}$ depending on the *perspective* adopted. In order to exhibit knowledge in ^{U}KM$_{R,S}$, we exploit its topology with standard graph algorithms taking into account the *perspective* U.

We firstly consider a metric semantic distance called SBPD (for "shortest bounded path distance"). SBPD scans all the paths between two given resources, using Dijkstra's algorithm for exploration, filters all paths exceeding a threshold and computes the shortest path. Therefore, the worst case complexity of the algorithm is $O(|W|^2|R|^2)$; however this complexity is not reached because the spreading is cut as soon as the distance exceeds the bound.

Alternative distances can be defined, such as MBPD (for "multi bounded paths distance"). MBPD scans all the paths between two given resources, filters those exceeding a threshold, computes the *synapse* equivalent to the maximal flow and then computes the distance according to this *synapse*.

We use the following definitions:

Given a perspective U and a distance ^{U}Ψ defined on R^2, given a *knowledge resource* 'r', given a parameter 'm', we call neighbourhood of 'r' and we denote ^{U,m}Ψ-neighbours(r) the set of all *knowledge resources* such as: ^{U}Ψ(x, r)≤ m.

'm' is called the radius of the neighbourhood.

3.3. Assessment of the collective knowledge

Assessing knowledge may seem a hazardous enterprise in a context characterized by weak semantics. Whereas standard measurements exist in the case of explicit information retrieval (e.g., precision at position 'n', mean average precision, and normalized discount cumulative gain, as listed in (Qin et al. 2010)), it is hopeless to try and assess the relevance of a knowledge resource when it is encountered serendipitously. Similarly, whereas knowledge acquisition can be measured in face of explicit pedagogical objectives, these measures do not apply to informal contexts. Assessing informal learning remains an open issue, despite a great deal of research work (Sefton-Green 2004).

Our objective in this section is to provide a generic measure for assessing knowledge in the absence of any explicit conceptual schema, i.e., relying upon the topology only.

Let us suppose a use case where internauts are browsing through a movie dataset, guided only by the proximities between *knowledge resources*. If the immediate neighbours of each currently displayed `object`/movie are statistically of same `descriptor`/genre, internauts will tend to visit movies 'of same genre', i.e., they will learn 'genres' despite they do not use the concept of 'genre' in their browsing. This phenomenon is typical of recommender systems (Adomavicius & Tuzhilin 2005), (Yamaba et al. 2013). We may therefore consider that according to the settings of this use case, the `objects`/movies are "sorted by" `descriptors`/genres.

In a more general way, we study how a collection 'O' of *knowledge resources* (e.g., a collection of `objects`) can be sorted according to another collection 'K' (e.g., a collection of `descriptors`). The idea is the following: if close elements of 'O' are locally homogeneous with respect to 'K', i.e., have similar elements of 'K' in their respective neighbourhoods, then 'the collection O is sorted according to the collection K'. We take the following definition:

> Let `U` be a perspective, let $^U\Psi$ be a distance defined on R^2, let 'm' be the `radius` of neighbourhoods, let $O=\{o_i\}_{i \in |O|}$ and $K = \{k_j\}_{j \in |K|}$ be two collections of *knowledge resources*, we denote $K(o_j)$ the vector $\{k^j(o_i)\}_{j \in |K|}$ of dimension $|K|$ such that $k^j(o_i)$=number of occurrences of 'k_j' in $^{U,m}\Psi$-neighbours(o_i).
>
> ^{U,m,K}localHomogeneity(O) is the average value of cosine similarity $(K(o_1), K(o_2))$ computed upon all the pairs (o_1, o_2) verifying $^U\Psi(o_1, o_2) \leq m$.

If ^{U,m,K}localHomogeneity$(O)=1$, we have an information: when two elements of O are $^{U,m}\Psi$-neighbours, they are also $^{U,m}\Psi$-neighbours of the same elements of K, i.e. they are "described identically by K"; if ^{U,m,K} localHomogeneity$(O)=0$, we have no information about the description of O by K. In this definition, the `radius` 'm' stands for the precision of the *knowledge map*; our hypothesis is that the `radius` is a key factor in the knowledge assessment. We firstly test this hypothesis on a mock-up in section 4, and then on a real Web dataset in section 5.

4. Learning colours: "The three Princes of Serendip" simulation

4.1. Objective of the simulation

The objective of the simulation is three-fold:

a) illustrate how the ViewpointS formalism captures formal data and informal contributions into an adaptive knowledge graph;

b) topologically assess knowledge emerging from serendipitous preference dissemination;

c) study the importance of the `radius` in the knowledge assessment.

"The three Princes of Serendip" is a simulation based on a *knowledge graph* populated with coloured geometrical objects (denoted 'objects' in the following), `descriptors` and `agents`.

The formal knowledge is represented by connecting the `objects` to `descriptors` about shapes and sizes ('square', 'triangle', 'circle', 'big', 'medium', 'small'). The colours are not formally represented. Serendip is a world where the <u>notion</u> of colour is absent: the `descriptors` 'red', 'blue', 'green' <u>do not</u> exist in the knowledge graph. However three `agents` (the princes), each with the ability to discriminate one specific colour, disseminate their respective preferences via *viewpoints*; these serendipitous connections stand for the informal knowledge.

In this section, we firstly illustrate the formalism, and then track the progressive apparition of the notion of colour, as a consequence of the dissemination of the princes' preferences.

In the Serendip simulation, we adopt the default perspective U^0: i) each *viewpoint* is weighted `1` and ii) the aggregation function computing the value of *synapses* is the `sum` function.

4.2. Description of the simulation

The simulation involves two kinds of `agents`: the peons and the princes.

The formal knowledge about shapes and sizes is represented by *viewpoints* of `type` `vps:matches` emitted by `peon-size` and `peon-shape`. Doing so, they weave at τ_0 an initial graph where each `object` is connected to two `descriptors` qualifying its size and its shape. The lower part of Figure 5 illustrates six *viewpoints* expressing this formal knowledge:

```
(peon-size,{G₅,small}, vps:matches,τ₀)
(peon-size,{B₇,big}, vps:matches,τ₀)
(peon-size,{G₃,big}, vps:matches,τ₀)
(peon-shape,{G₅,circle}, vps:matches,τ₀)
(peon-shape,{B₇,square}, vps:matches,τ₀)
(peon-shape,{G₃,circle}, vps:matches,τ₀)
```

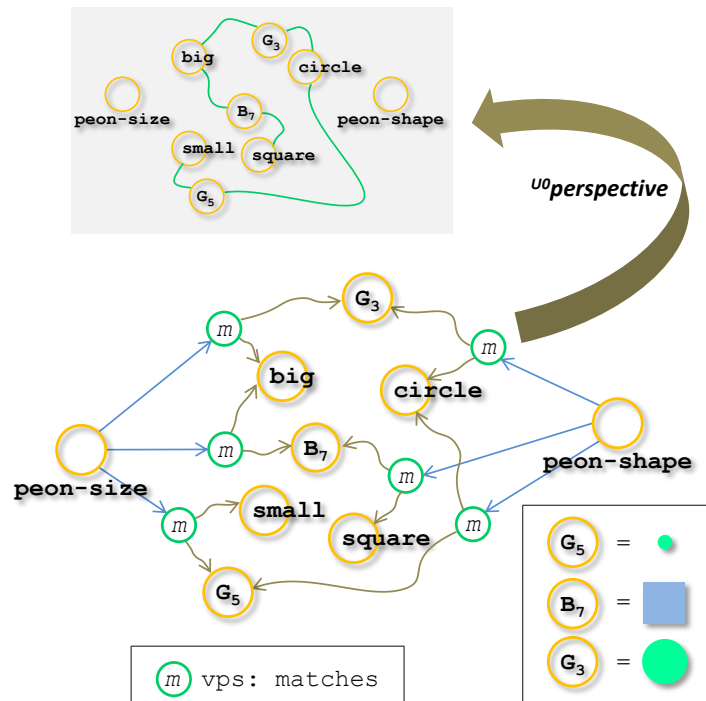The upper part of Figure 5 illustrates the *knowledge map* under the perspective U^0.

Figure 5: The peons emit formal knowledge

The informal knowledge about colours is represented by *viewpoints* of `type` `vps:likes` emitted by the princes of Serendip, namely `princeRed`, `princeGreen` and `princeBlue`. Each prince has the ability to discriminate one specific colour and will react only to this colour during his serendipitous explorations. This is simulated by encoding the colour in the names

of the geometrical `objects`: `R` (for "Red"), `G` (for "Green"), and `B` (for "Blue"). The lower part of Figure 6 illustrates an extension of the previous *knowledge graph* resulting from the input of three *viewpoints* by the princes:

```
(princeGreen,{princeBlue,G₅}, vps:likes,τ₁)
(princeGreen,{princeBlue,G₃}, vps:likes,τ₂)
(princeBlue,{princeBlue,B₇}, vps:likes,τ₃)
```
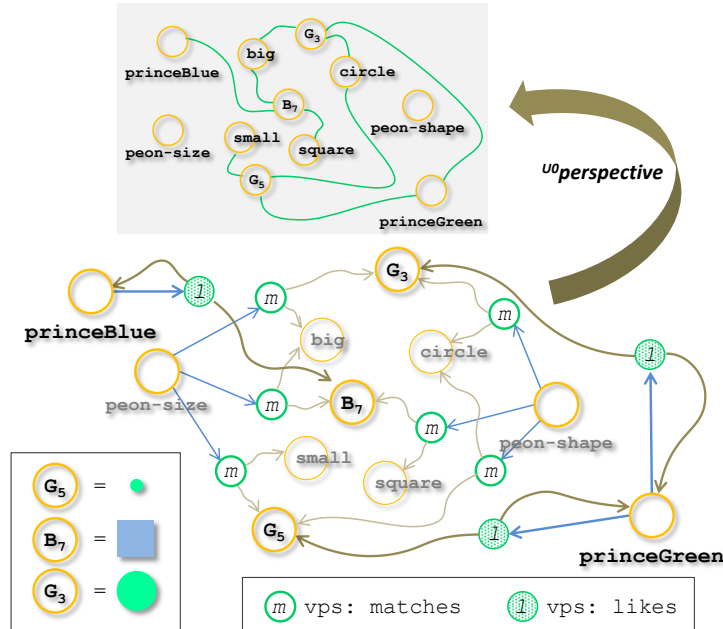

Figure 6: The princes emit informal knowledge

The upper part of Figure 6 illustrates the updated *knowledge map* under the perspective `U⁰`. We compare in Figure 7 the neighbourhood of `G₅` before (left side) and after (right side) the interactions of the princes.

Figure 7: Impact of the informal knowledge

SBPD and MBPD are successively used for computing the distances. Under `U⁰`, all the *synapses*, consisting of a single *viewpoint* weighted `1`, are valued `1`.

Using SBPD, the new path added on the right side of Figure 7 by the new neighbour 'princeGreen' is not shorter that the initial path: the distance between 'G₃' and 'G₅' is '2' on both sides; as a consequence the SBPD distance remains unchanged.

Using MBPD, we have $MBPD(G_3,G_5)=2$ on the left whereas $MBPD(G_3,G_5)= 1/((2*2)/(2+2))=1$ on the right; as a consequence the MBPD distance decreases when the informal knowledge input by princeGreen appears.

We adopt in the next section a protocol simulating a continuous flow of informal knowledge emitted by the princes in order to study this aspect in more details and test our hypothesis concerning the importance of the precision.

4.3. Protocol and measurements

We consider a set of 270 geometrical objects so that each of the 27 combinations of size, shape and colour is represented 10 times. 'small', 'medium', 'big' are encoded in the object names by 's', 'm', 'b'; 'square', 'triangle', 'circle' by 'q', 't', 'c' and 'red', 'green', 'blue' by 'R', 'G', 'B', e.g., 'bBq205' is a big blue square.

We firstly initialise the *knowledge graph* by the formal knowledge of shapes and sizes provided extensively by peon-size and peon-shape.

Then the three princes deliver the informal knowledge along N^{cycles} "cycles" of interactions simulating serendipitous trips. A "cycle" is defined as follows:

- a cycle consists in N^{trips} "trips"

- a trip consists in: i) choosing a prince randomly with the probabilities p^{Red} for princeRed, p^{Green} for princeGreen, p^{Blue} for princeBlue and ii) let the chosen prince choose an object at random and then go through $N^{interactions}$ "interactions"

- an interaction of one prince consists in: i) evaluating the current object and ii) choosing a new object within the $^{U,m}\Psi$-neighbours of the current object with a distribution of probabilities corresponding to their respective proximities; this corresponds to a serendipitous exploration in breadth.

- evaluating an object consists in emitting a viewpoint of type vps:likes in case its coded-colour corresponds to the preference of the prince.

In the whole simulation the chosen perspective is U^0 described above, the distance function is SBPD described above.

The measurements are applied at the end of each cycle. Let us denote O the population of all objects. Let us denote 'Shapes' the collection of descriptors {'square', 'triangle', 'circle'}. In order to assess formal knowledge, we measure:

$$^{U0,m,Shapes} localHomogeneity(O)$$

In order to assess informal knowledge, we assimilate the princes to the colours they prefer and denote 'Colours' the collection {'princeRed', 'princeGreen', 'princeBlue'}. We measure:

$$^{U0,m,Colours} localHomogeneity(O)$$

Two series of runs of the simulation, designated by $RUN_{1,m}$ an $RUN_{2,m}$, are presented hereunder, where 'm' is the radius.

$RUN_{1,m}$ corresponds to a balanced activity of the three princes (pRed= pGreen= pBlue= 33,33%) i.e., the three colours are equally diffused.

$RUN_{2,m}$ corresponds to the exclusive activity of princeRed (pRed= 100%) i.e., only the implicit knowledge of the red colour is diffused.

KG
Objects : 270 geometrical objects + 2 peons + 3 princes
Descriptors: 6 (3 sizes and 3 shapes)
formal knowledge: 270*2 *viewpoints*
informal knowledge: max=7*50*10 *viewpoints*

Ncycles=12; NTrips=50; NInteractions =10

perspective U_0
formal knowledge & informal knowledge: weight(*viewpoint*)='1'

$RUN_{1,m}$	$RUN_{2,m}$
$p^{Red} = 33,33\%$	$p^{Red} = 100\%$
$p^{Green} = 33,33\%$	$p^{Green} = 0\ \%$
$p^{Blue} = 33,33\%$	$p^{Blue} = 0\%$

'm' takes the values: '1.5', '2', '2.5', '3', '3.5' and '4'.

4.4. Results and discussion

Figure 8 illustrates the two series of parallel runs. Each $RUN_{i,m}$ is illustrated by two curves respectively measuring $^{U_0,m,Shapes}localHomogeneity(O)$, and $^{U_0,m,Colours}localHomogeneity(O)$ along 12 cycles corresponding to successive flows of contributions from the princes expressing their preferences (informal knowledge). In the discussion, we simply denote $^{Shapes}lH(O)$ and $^{Colours}lH(O)$.

A first observation is that the minimum for both measures is 0.33 whatever the radius 'm', except in the $RUN_{2,m}$ with m>2. This is a direct consequence of the symmetry of this toy example: 0.33 is the probability of having an object with same shape, or same colour, in the neighbourhood of a given object when all the knowledge about shapes and colours has been symmetrically expressed.

A second observation is that $^{Shapes}lH(O)$ always decreases along the cycles. We interpret it as a competition between the sorting by shapes and the sorting by colours.

A third observation is that $^{Colours}lH(O)$ is always higher in $RUN_{1,m}$ compared to $RUN_{2,m}$. This can be interpreted as: "the graph learns/sorts colours better when 3 teachers are promoting 3 colours than with 1 teacher is favouring 1 colour".

Going into to the discussion about the key values of the radius 'm' in the U_0 perspective：

- as long as m<2, the formal knowledge is undetectable (the *synapses* linking two objects of same shape through peon-shape are bounded by '1'; peon-shape never qualifies twice a given object, thus SBPD distances are ≥2, see Figure 7) . Only the informal knowledge is detected (each prince is allowed to qualify a given object several times, which yields shorter distances).

- when 2≤m<4, the formal knowledge is detected ($^{Shapes}lH(O)$ takes high values at cycle₀); the competition starts between the two sortings.

- when m>2, the informal knowledge becomes messy because of the great number of close neighbours brought in by the reiterate evaluations of the princes.

- when m≥4, the minimum 0.33 is reached for both measures. Even the formal knowledge becomes messy because objects of different shapes become neighbours by transitivity: the local homogeneity is lost. For instance, consider the path bBq5-square-bBq2-big-bGt7 linking a 'big blue square' to a 'big green triangle' through another 'big blue square', firstly because of the shape, then because of the size, SPBD (bBq5, bGt7)=4; as a consequence, a blue square and a green triangle are neighbours.

We can summarize this part of the discussion by saying that in the U_0 perspective, $1 \le m < 2$ is the relevant `radius` range for sorting the objects by colours, whereas $2 \le m < 4$ is relevant for sorting the objects by shapes.

Recapitulating the answers to our initial questions:

- the serendipitous preference dissemination by the princes yields a (informal) knowledge of colours that comes in competition with the (formal) knowledge of shapes;
- this phenomenon can be observed only when adopting the relevant "precision" (radius).

The next section is aimed at checking our hypotheses against real data.
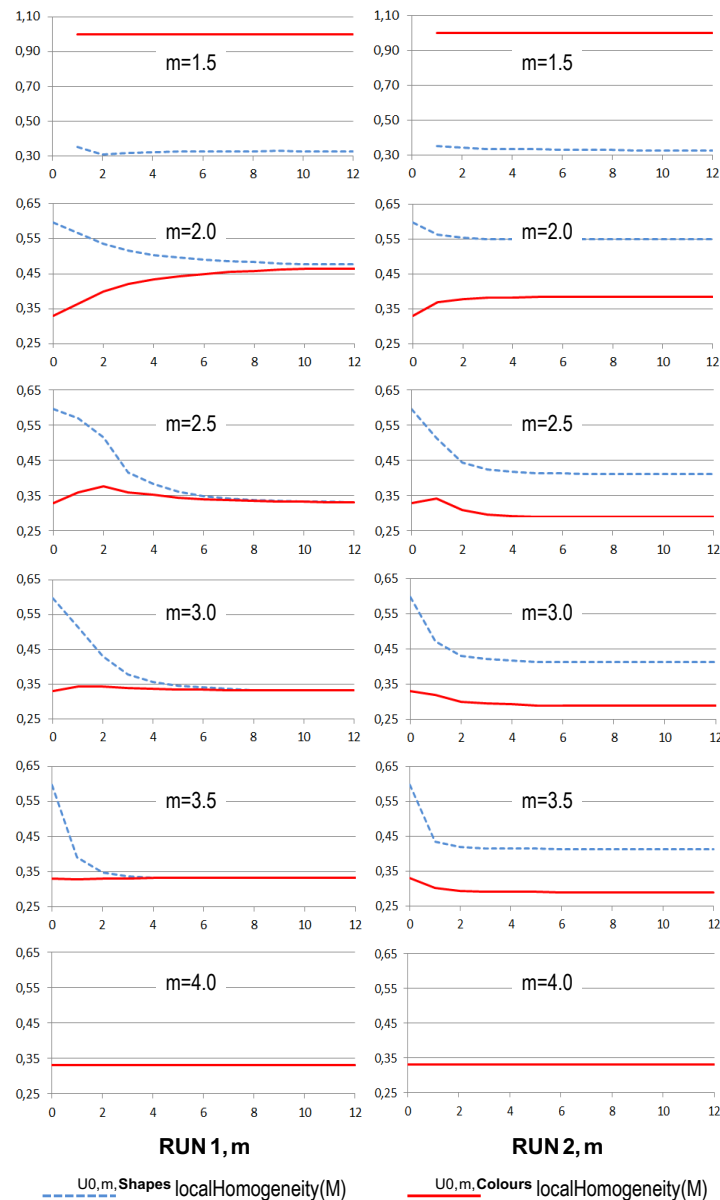

Figure 8: Compared evolutions of formal knowledge (shapes) versus informal knowledge (colours)

5. The MovieLens experiment

5.1. Objective of the experiment

In this section, we take a Web dataset where formal knowledge issued from explicit semantics is mixed with informal knowledge issued from social contributions: MovieLens

(http://datahub.io/dataset/movielens). The complete dataset consists of 100.000 movies and 1.000.000 ratings which have been collected by the GroupLens Research Project at the University of Minnesota. Their data have played the role of experimental matter for many authors such as (Peralta 2007; O'Connor & Herlocker 1999; Harpale & Yang 2008), mostly in the context of studies about collaborative filtering methods for recommendation. A corpus of sophisticated measurements and analyses is available in different research works such as (Jung 2012). In this section, we experiment with three objectives:

a) to illustrate how a Web dataset can be transcribed into a *knowledge graph*;
b) to illustrate the role of *perspectives* in the topological sorting of objects;
c) to assess concurrence or complementarity between a classification of movies by 'genres' and a classification through 'ratings'.

5.2. Modelling the MovieLens data

Figure 9 has been borrowed from (Peralta 2007) to describe the conceptual schema of the Web dataset. The red frames show the matching with Figure 10.
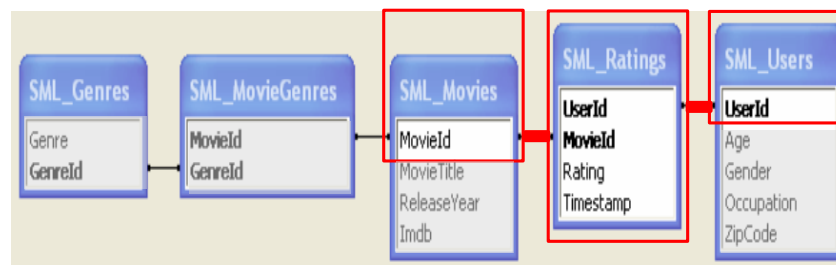

Figure 9: The MovieLens conceptual schema

To illustrate the transcription of this schema in ViewpointS, we take three examples:

- (formal) movie description: the *viewpoint* `(AA-MovieLens,{movieX, genreY}, mo:genre,τ)` transcribes the fact: the `artificial agent` 'AA-MovieLens' connects the `object` 'movieX' to the `descriptor` 'genreY'
- (formal) commenter profile: the *viewpoint* `(userX,{userX,occupationY},cv:jobType,τ)` transcribes the fact: the `legal person` 'userX' connects himself to the `descriptor` 'occupationY'
- (informal) rating: the *viewpoint* `(userX, {movieY, ***}, vpml:rating,τ)` transcribes the fact: the `legal person` 'userX' connects the `object` 'movieY' to the `descriptor` '***'

Figure 10 illustrates the ViewpointS model; the ratings link together two subgraphs representing respectively the formal knowledge of movies' descriptions and users' profiles. The red frame shows the matching with Figure 9.

The systematic use of standard unambiguous semantic Web properties in the transcription process, e.g., `mo:genre`, `foaf:gender`, automatically provides a shared language, than can be exploited for building specific perspectives. For instance, the prefix 'mo:' identifies the Movie Ontology (www.movieontology.org) which unambiguously defines the meaning of the resources and properties used within this dataset. By systematically using such tags, we enable ViewpointS to embrace the semantic Web standards and import/export RDF representations.
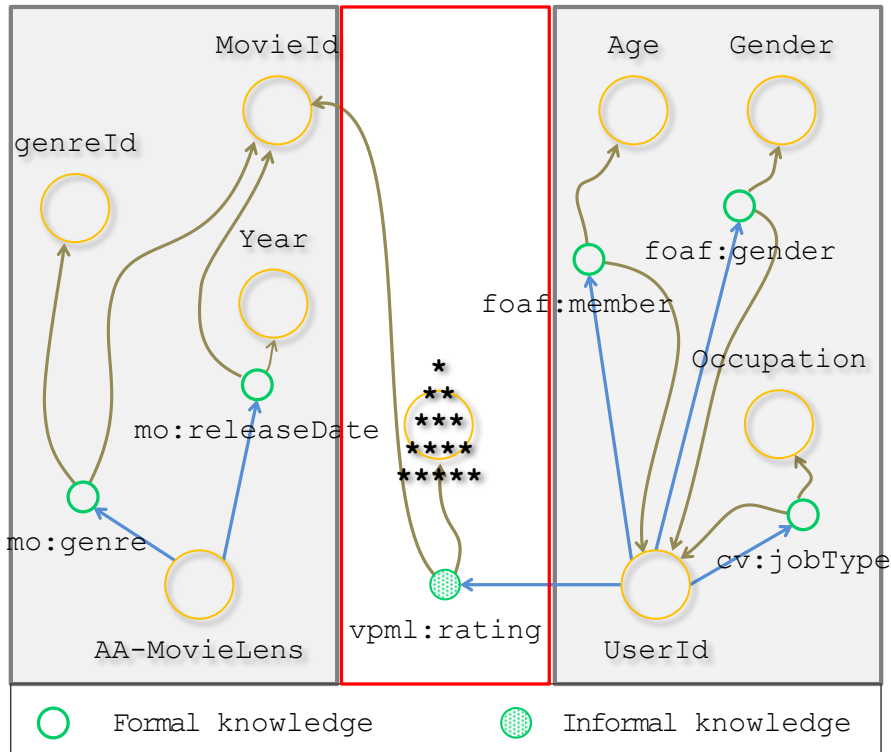
Figure 10: The ViewpointS model for MovieLens

5.3 Protocol and measurements

We take an extract of the original dataset: 1682 movies rated by 943 commenters providing 5.000 ratings.

We firstly initialise the *knowledge graph* by transcribing the formal knowledge. There are 1682 movies and 943 users. Each movie is linked to one or more of the 18 genres (2 genres is the average) and to one year; each user is linked to one age, one gender, and one occupation.

Measures (cycle$_0$, m=1) and *Measures* (cycle$_0$, m=2) are operated immediately after this initialization, where 'm' denotes the radius chosen for computing neighbourhoods.

We then arbitrary split the ratings into 5 subsets of 1000 ratings to be input as informal knowledge. Each run of the experiment will consist in transcribing these subsets one by one along five successive cycles.

Measures (cycle$_i$, m=1) and *Measures* (cycle$_i$, m=2) are operated at the end of each cycle.

Measures (cycle$_i$, m) are specified as follows:

Let us denote M the population of the 1682 movies,

let us denote Genres the collection of 18 genres provided by MovieLens,

let us denote Ratings the collection $\{*, **, ***, ****, *****\}$, we compute:

- $^{Ui,m,\textbf{Genres}}$ localHomogeneity(M)
- $^{Ui,m,\textbf{Ratings}}$ localHomogeneity(M)

In this experiment, the distance function is SBPD;

'm' denotes the radius chosen for computing neighbourhoods. Three series of RUNs designated by RUN$_{1,m}$, RUN$_{2,m}$ and RUN$_{3,m}$ are presented hereunder.

Each RUN$_{i,m}$ corresponds to a perspective U$_i$. Whatever U$_i$, the ^{U}reduce function is the sum; the perspectives differ only by the ^{U}map functions.

U$_1$ corresponds to priority given to formal knowledge, i.e., vpml:rating are weighted 1 by ^{U}map and all other *viewpoints* are weighted 3.

U_2 corresponds to balanced importance of formal versus informal knowledge, i.e., all *viewpoints* are weighted 1.

U_3 corresponds to priority given to informal knowledge, i.e., `vpml:rating` are weighted 3 and all other *viewpoints* are weighted 1.

KG
Objects : 1682 'MovieId' + 943 'UserId' Descriptors: 127 formal knowledge: 7471 *viewpoints* among which 2893 *viewpoints* of type mo:genre informal knowledge: 5000 *viewpoints*

	$N^{cycles}=5$ 'm'=radius
$RUN_{1,m}$	perspective U_1 formal knowledge: $^Umap(viewpoint)=$'3' informal knowledge: $^Umap(viewpoint)=$'1'
$RUN_{2,m}$	perspective U_2 formal & informal knowledge: $^Umap(viewpoint)=$'1'
$RUN_{3,m}$	perspective U_3 formal knowledge: $^Umap(viewpoint)=$'1' informal knowledge: $^Umap(viewpoint)=$'3'

5.4 Results and discussion

The upper part of Figure 11 presents the 6 runs $RUN_{1,1}$ to $RUN_{3,2}$; the lower part of the figure presents $RUN_{1,1}$, $RUN_{2,2}$ and $RUN_{3,2}$ in more details.

$^{Ui,m,\textbf{Genres}}$localHomogeneity(M) and $^{Ui,m,\textbf{Ratings}}$localHomogeneity(M) are drawn for each run.

RUN 1, m   **RUN 2, m**   **RUN 3, m**

- - - $^{Ui,m,\textbf{Genres}}$localHomogeneity(M)   ——— $^{Ui,m,\textbf{Ratings}}$localHomogeneity(M)
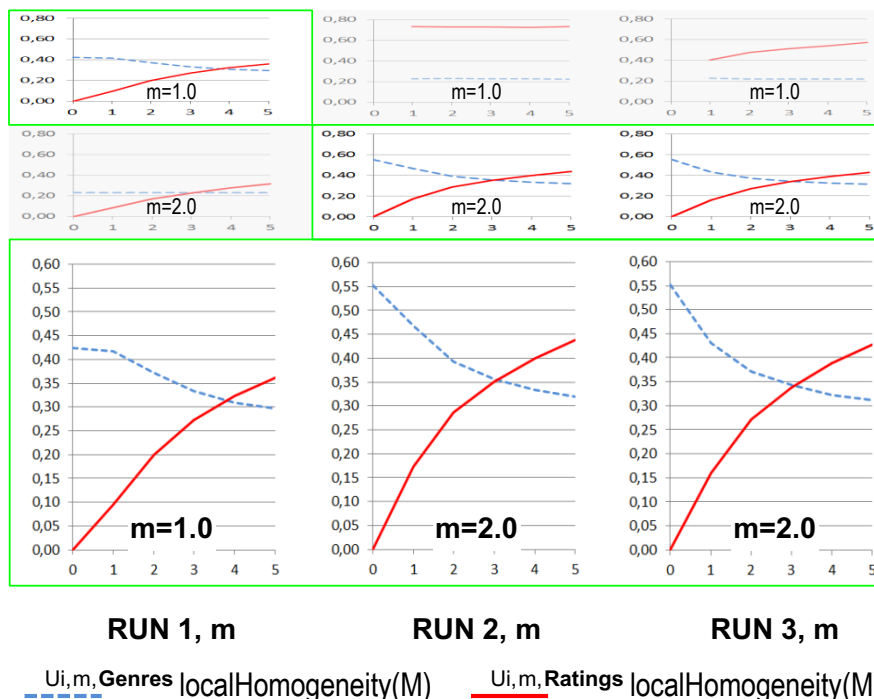
Figure 11: Compared evolutions of formal knowledge (genres) versus informal knowledge (ratings)

We first compare the values for cycle$_0$, when ratings are absent of the *knowledge graph*. In the *perspective* U_1, $^Umap=$'3' for *viewpoints* of type mo:genre yielding strong *synapses* between movies and genres. In consequence, the suitable radius for observing $^{Genres}lH(M)$ is 'm=1', not 'm=2'. When we use the radius 'm=2' in the computation of SBPD, distinct genres

become close by transitivity through movies, so that movies become in turn close by transitivity through genres; movies' neighbourhoods become heterogeneous in genres i.e., $\text{Genres}lH(M)$ becomes low.

In the *perspectives* U_2 and U_3, Umap='1' for *viewpoints* of type `mo:genre`. The suitable `radius` for observing $\text{Genres}lH(M)$ is 'm=2', not 'm=1'. When we use the `radius` 'm=1' in SBPD, movies stay far from genres, thus movies of same genre are not brought into in the same neighbourhood by transitivity; movies' neighbourhoods stay heterogeneous in genres i.e., $\text{Genres}lH(M)$ stays low.

Looking at cycle_1 to cycle_5, we observe that both 'm=1' and 'm=2' are suitable, although $\text{Ratings}lH(M)$ is better observed when 'm=1' .

According to this preliminary observations, only $\text{RUN}_{1,1}$, $\text{RUN}_{2,2}$ and $\text{RUN}_{3,2}$ (lower part of Figure 11) can be used for comparing the simultaneous evolutions of $\text{Genres}lH(M)$ and $\text{Ratings}lH(M)$.

We observe that $\text{Genres}lH(M)$ always decreases while $\text{Ratings}lH(M)$ always increases. The curves (approximatively) intersect at:

0.32 for $\text{RUN}_{1,1}$	0.35 for $\text{RUN}_{2,2}$	0.34 for $\text{RUN}_{2,3}$

This intersection varies little with the perspective; the highest value '0.35' is nevertheless obtained in the balanced perspective U_2. This is quite similar to what has been observed in section 4: the graph better exhibits "knowledge" in balanced conditions (note: what is balanced here in the *knowledge map* resulting from the *perspective* chosen, whereas in the Serendip simulation, it was the *knowledge graph* itself resulting from the princes' interactions)

This experiment on a non-symmetrical dataset shows concurrence and complementarity between a classification of movies by 'genres' and a classification through 'ratings'. This is analog to the competition between shapes and colours in the Serendip simulation. It can be interpreted as "genres and ratings are independent sorting dimensions". Moreover, the intersection of curves occurs between cycle_3 and cycle_4, i.e., after approximatively 5000*3/5=3000 *viewpoints* of type `vpml:ratings`, to be compared with 2893 *viewpoints* of type `mo:genre`. This can be interpreted as: "ratings are as powerful as genres for sorting movies".

Whereas all the curves of the Serendip simulations corresponded to various values of the `radius` within a single *perspective*, both the *perspective* and the `radius` vary in the MovieLens experiment. A second important result is a refinement of our previous hypothesis: given a *perspective*, given a question (may a collection O be sorted according to a collection K?), the `radius` is a key element for answering. *Knowledge maps* and their topological exploitations via `localHomogeneity` prove to be useful tools for assessing *collective knowledge*. Moreover by observing the evolution of a given *knowledge map* along the interaction of users, we may envisage assessing collective learning.

6. Conclusion and perspectives

We have presented a model and formalism where both the explicit semantics of the linked data and the informal contributions of Web 2.0 users can be expressed as fine-grained subjective units of knowledge called *viewpoints*. The set of all *viewpoints* form an adaptive knowledge graph, which can be put in perspective and yield knowledge maps. We have defined two distances based on the viewpoints' evaluation and aggregation: SBPD and MBPD, and proposed 'localHomogeneity' as a topology-based measure for assessing emergent knowledge.

When simulating "The three Princes of Serendip", we have proved the potential of our measurements to capture both the explicitly reified knowledge and the implicit knowledge

hidden behind subjective contributions; we have also proved informal learning by watching the evolutions of a knowledge map.

We then have tested the approach against real Web data by experimenting with the MovieLens dataset. We have shown the ability of our model to capture the semantics of the data, and compared the impact of the subjective contributions (the ratings) on the formal knowledge under different perspectives. Doing so, we have stepped forward in assessing and exploiting collective intelligence in contexts accepting both formal and informal data.

We are currently working on the design of an API offering intuitive browsing of the knowledge and one-click feedback exploiting the context. We are also in search of a benchmark such as those offered on LETOR (Qin et al. 2010) to evaluate the efficiency of our distance in terms of information retrieval. Finally, in order to automatically elicit the best radius with respect to given question, we are experimenting with genetic algorithms.

The next step in our agenda is to prove the concept in real life scenarios, i.e., to invite users to elicit knowledge collectively by using the ViewpointS approach. Evaluation and quality of the knowledge (Lichtenstein & Parker 2009) will be crucial metrics in those scenarios. Two use cases are planned, both oriented towards cross-disciplinary discoveries: one in the agronomics domain, the other in the biomedical domain, both in the context of ongoing research projects.

Acknowledgements

References

Aberer, K. et al., 2004. Emergent Semantics Principles and Issues. In D. Lee, YoonJoon and Li, Jianzhong and Whang, Kyu-Young and Lee, ed. *Database Systems for Advanced Applications*. Springer Berlin Heidelber, pp. 25–38.

Adomavicius, G. & Tuzhilin, A., 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), pp.734–749.

von Ahn, L. et al., 2008. reCAPTCHA: human-based character recognition via Web security measures. *Science (New York, N.Y.)*, 321(5895), pp.1465–8.

Ankolekar, A. & Krötzsch, M., 2007. The two cultures: Mashing up Web 2.0 and the Semantic Web. In *16th international conference on World Wide Web*. pp. 825–834.

Anthony J H Simons & Graham, I., 1998. *37 Things that Don't Work in Object-Oriented Modelling with UML*,

Attardi, G. & Simi, M., 1995. A Formalization Of Viewpoints. *Fundamenta Informaticae*, 23(2,3,4), pp.149–173.

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web. *Scientific American*.

Bojārs, U. et al., 2008. Using the Semantic Web for linking and reusing data across Web 2.0 communities. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1), pp.21–28.

Breslin, J.G., Passant, A. & Vrandečić, D., 2011. Social semantic Web. In S. B. Heidelberg, ed. *Handbook of Semantic Web Technologies*. pp. 467–506.

Carroll, J.J. et al., 2005. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web - WWW '05*. New York, New York, USA: ACM Press, p. 613.

Chein, M. & Mugnier, M.-L., 2008. *Graph-based Knowledge Representation: Computational Foundations of Conceptual Graphs*,

Cimiano, P. & Staab, S., 2004. Learning by googling. *ACM SIGKDD Explorations Newsletter*, 6(2), pp.24–33.

Donald, M. et al., 2015. Construction d ' un vocabulaire patient / médecin dédié au cancer du sein à partir des médias sociaux. *Ingénierie des Connaissances*.

Edelman, G.M., *Neural Darwinism: The theory of neuronal group selection.*,

Edelman, G.M. & Reeke, G.N., 1982. Selective networks capable of representative transformations, limited generalizations, and associative memory. *Proceedings of the National Academy of Sciences of the United States of America*, 79(6), pp.2091–5.

Edelman, G.M. & Tononi, G., 2000. *A Universe of Consciousness: How Matter Becomes Imagination*,

Gruber, T., 2008. Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1), pp.4–13.

Harispe, S. et al., 2013. The Semantic Measures Library and Toolkit : fast computation of semantic similarity and relatedness using biomedical ontologies. , pp.1–2.

Harpale, A.S. & Yang, Y., 2008. Personalized active learning for collaborative filtering. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*. New York, New York, USA: ACM Press, p. 91.

Hendler, J. et al., 2008. Web Science: An interdisciplinary approach to understanding the World Wide Web. *Communications of the ACM*, 51(7), pp.60–69.

Jonquet, C., Musen, M.A. & Shah, N.H., 2010. Building a Biomedical Ontology Recommender Web Service. *Biomedical Semantics*, 1(S1).

Jung, J.J., 2012. Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB. *Expert Systems with Applications*, 39(4), pp.4049–4054.

Karapiperis, S. & Apostolou, D., 2006. Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*, pp.199–216.

Kim, H.L. et al., 2008. The state of the art in tag ontologies: a semantic model for tagging and folksonomies. In Dublin Core Metadata Initiative, pp. 128–137.

Krause, M. & Aras, H., 2009. Playful Tagging — Folksonomy Generation Using Online Games. *Proceedings of the 18th international conference on World wide web*, pp.1207–1208.

Krief, A. et al., 2008. Building the EnCOrE dictionary collaboratively: strategy and practice. *Knowledge and Learning*, 4(2/3), pp.203–216.

Lafourcade, M., 2007. Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*.

Lee, W.-N. et al., 2008. Comparison of ontology-based semantic-similarity measures. *AMIA, Annual Symposium 2008*, pp.384–8.

Lemoisson, P. & Cerri, S.A., 2005. Interactive Knowledge Construction In The Collaborative Building Of An Encyclopedia. *Applied Artificial Intelligence*, 19(9–10), pp.933–966.

Lemoisson, P. et al., 2013. Viewpoints: An Alternative Approach toward Business Intelligence. In *eChallenges e-2013 Conference*, pp. 8, Dublin, Irland.

Lichtenstein, S. & Parker, C.M., 2009. Wikipedia model for collective intelligence: a review of information quality. *International Journal of Knowledge and Learning*, 5(3/4), p.254.

Lux, M. & Dosinger, G., 2007. From folksonomies to ontologies: employing wisdom of the crowds to serve learning purposes. *International Journal of Knowledge and Learning*, 3(4/5), p.515.

Markines, B. et al., 2009. Evaluating similarity measures for emergent semantics of social tagging. In *18th international conference on World wide web - WWW '09*. Madrid, Espagne: ACM Press, pp. 641–650.

Mika, P., 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1), pp.5–15.

Mikroyannidis, A., 2007. Toward a social semantic web. *Computer*.

Noh, T. et al., 2010. Learning the emergent knowledge from annotated blog postings. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4), pp.329–339.

O'Connor, M. & Herlocker, J., 1999. Clustering items for collaborative filtering. In *Proceedings of the ACM SIGIR workshop on ....*

O'Reilly, T., 2009. *What is Web 2.0*, Google Books.

Pedersen, T. et al., 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Biomedical Informatics*, 40(3), pp.288–299.

Peralta, V., 2007. Extraction and integration of movielens and imdb data.

Qin, T. et al., 2010. LETOR: A benchmark collection for research on learning to rank for information retrieval.

Information Retrieval, 13(4), pp.346–374.

Quattrone, G. et al., Measuring Similarity in Large-scale Folksonomies.

Saquali, M. et al., 2009. The PBRM (perception-based regional mapping): A spatial method to support regional development initiatives. *Applied Geography*, 29(3), pp.358–370.

Searle, J., 1969. *Speech acts: an essay in the philosophy of language*, Cambridge, UK: Cambridge University Press.

Sefton-Green, J., 2004. Literature Review in Informal Learning with Technology Outside School.

Self, J., 1990. *Computational Viewpoints*,

Seung, S., 2012. *Connectome: How the Brain's Wiring Makes Us Who We Are*, Houghton Mifflin Harcourt.

Specia, L. & Motta, E., 2007. Integrating Folksonomies with the Semantic Web. *The semantic web: research and applications*, pp.624–639.

Surroca, G. et al., 2015. Preference Dissemination by Sharing Viewpoints : Simulating Serendipity. In *7th Intertnational Conference on Knowledge Engineering and Ontology Development KEOD'15*, 2(2), pp. 402-409, Lisbon, Portugal.

Surroca, G. et al., 2015. Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique. In *25èmes Journées Francophones d'Ingénierie des Connaissances, IC'14*, pp. 12, Clermont-Ferrand, France

Uren, V. et al., 2006. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1), pp.14–28.

Vial, S., 2013. *L'être et l'écran - Comment le numérique change la perception*,

Weinberger, H., 2011. Tagging Web 2.0 content in context. *International Journal of Knowledge and Learning*, 7(3/4), p.157.

Yamaba, H., Tanoue, M. & Takatsuka, K., 2013. On a serendipity-oriented recommender system based on folksonomy. *Procedia Computer Science*, 22, pp.276–284.