ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph
Philippe Lemoisson, Guillaume Surroca, Clement Jonquet, Stefano A. Cerri

To cite this version:

HAL Id: lirmm-01605411
https://hal-lirmm.ccsd.cnrs.fr/lirmm-01605411
Submitted on 11 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ViewpointS: capturing formal data and informal contributions into an adaptive knowledge graph

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

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 (Bojärs 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 synthesizing 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 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.

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. 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 ‘q’ (‘index’, ‘author’, ‘like’ …) and a time stamp ‘t’.

![Figure 2: The subjective connection called “viewpoint”; the blue arrow gives the provenance, ‘0’ gives the semantics, ‘t’ gives the time stamp](image)

**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 $\tau$ 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 \( \theta \) 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)’; \( \theta \) is called the type of the viewpoint.

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

We denote \(W\) the set of all viewpoints.

Let \(\text{EDGES}\) be the set of all links woven between the viewpoints and the knowledge resources:

\[
\text{EDGES} = \{(a, w) | w \in W; a = \text{EmitterOf}(w)\} \cup \{(w, r) | w \in W; r \in \text{ConnectedBy}(w)\}.
\]

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

\[
\text{KG}_{R,W} = (R \cup W, \text{EDGES}).
\]

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 \(\text{KG}_{R,W}\)](image)

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 \(\text{KG}_{R,W}\) 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:

\[
\text{synapse}: R^{2} \rightarrow 2^{W} \\
(r_1, r_2) \mapsto \{w \in W | \text{Connected}(w) = \{r_1, r_2\}\}.
\]
Setting a perspective \( U \) consists in defining rules for valuating each synapse. We consider two functions \( \map \) and \( \reduce \) and combine them into \( \synapse \).

- \( \map \) evaluates each viewpoint \( w \) of the synapse individually, as a function of \( '0', 't' \) and \( 'EmitterOf(w)' \).
- \( \reduce \) aggregates all the \( \map \) results into a single positive or null value; it may be a simple sum or a sophisticated function with a threshold or attenuation.

\( \synapse \) is finally the composition of 3 functions:

\[
\synapse: \mathbb{R}^2 \rightarrow \mathbb{R}^+ \\
\text{IF } \synapse(r_1, r_2) = \emptyset \text{ THEN } \synapse(r_1, r_2) = 0 \text{ ELSE } \synapse(r_1, r_2) = \reduce(\map(\synapse(r_1, r_2)))
\]

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

Figure 4 illustrates this interpretation process with a perspective \( U^0 \) where the aggregation function computing the value of synapses is the sum.

![Figure 4: Building a knowledge map \( \km_{R,S} \)](image)

One single \( KG_{R,W} \) will therefore be interpreted into several distinct \( \km_{R,S} \) depending on the perspective adopted. In order to exhibit knowledge in \( \km_{R,S} \), 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 \( \Psi \) defined on \( \mathbb{R}^2 \), given a knowledge resource \( 'r' \), given a parameter \( 'm' \), we call neighbourhood of \( 'r' \) and we denote \( \Psi_{\text{\text{-}}n\text{-neighbours}(r)} \) the set of all knowledge resources such as: \( \Psi(x, r) \leq m \).
- \( 'm' \) 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 \( \Psi \) be a distance defined on \( \mathbb{R}^2 \), 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 \( K(o_i) \) the vector \( \{k_j(o_i)\}_{j \in |K|} \) of dimension \( |K| \) such that \( k_j(o_i) = \text{number of occurrences of } k_j \text{ in } U \cdot \Psi \text{-neighbours } (o_i) \).
| \( \Psi \) localHomogeneity \((O)\) is the average value of cosine similarity \((K(o_i), K(o_j))\) computed upon all the pairs \((o_i, o_j)\) verifying \( \Psi(o_i, o_j) \leq m \).

If \( \Psi(\cdot,\cdot) \) localHomogeneity \((O)\) = 1, we have an information: when two elements of O are \( U \cdot \Psi \)-neighbours, they are also \( U \cdot \Psi \)-neighbours of the same elements of K, i.e. they are “described identically by K”; if \( \Psi(\cdot,\cdot) \) 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 notion of colour is absent: the descriptors ‘red’, ‘blue’, ‘green’ do not 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 $t_0$ 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:

- $(\text{peon-size}, \{G_5, \text{small}\}, vps$ matches, $t_0)$
- $(\text{peon-size}, \{B_7, \text{big}\}, vps$ matches, $t_0)$
- $(\text{peon-size}, \{G_3, \text{big}\}, vps$ matches, $t_0)$
- $(\text{peon-shape}, \{G_5, \text{circle}\}, vps$ matches, $t_0)$
- $(\text{peon-shape}, \{B_7, \text{square}\}, vps$ matches, $t_0)$
- $(\text{peon-shape}, \{G_3, \text{circle}\}, vps$ matches, $t_0)$

The upper part of Figure 5 illustrates the knowledge map under the perspective $U_0$.

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

Figure 5: The peons emit formal knowledge
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:

- $(\text{princeGreen}, \{\text{princeBlue}, G_5\}, \text{vps:likes}, t_1)$
- $(\text{princeGreen}, \{\text{princeBlue}, G_3\}, \text{vps:likes}, t_2)$
- $(\text{princeBlue}, \{\text{princeBlue}, B_7\}, \text{vps:likes}, t_3)$

The upper part of Figure 6 illustrates the updated knowledge map under the perspective $U_0$. We compare in Figure 7 the neighbourhood of $G_5$ before (left side) and after (right side) the interaction of the princes.

![Diagram](image)

**Figure 6:** The princes emit informal knowledge

SBPD and MBPD are successively used for computing the distances. Under $U_0$, 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 than the initial path: the distance between 'G3' and 'G3' is '2' on both sides; as a consequence the SBPD distance remains unchanged.

Using MBPD, we have MBPD(G1, G2)=2 on the left whereas MBPD(G3, G3)= 1/(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 Ncycles “cycles” of interactions simulating serendipitous trips. A “cycle” is defined as follows:
- a cycle consists in Ntrips “trips”
- a trip consists in: i) choosing a prince randomly with the probabilities pRed for princeRed, pGreen for princeGreen, pBlue for princeBlue and ii) let the chosen prince choose an object at random and then go through Ninteractions “interactions”
- an interaction of one prince consists in: i) evaluating the current object and ii) choosing a new object within the s.mΨ-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 Uφ 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:

$U_{φ, r}.Shapes.localHomogeneity(O)$

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:

$U_{φ, c}.Colours.localHomogeneity(O)$

Two series of runs of the simulation, designated by RUN_i,n an RUN_i,m are presented hereunder, where ‘m’ is the radius.

RUN_i,n corresponds to a balanced activity of the three princes (pRed= pGreen= pBlue= 33.33%) i.e., the three colours are equally diffused.

RUN_i,m corresponds to the exclusive activity of princeRed (pRed= 100%) i.e., only the implicit knowledge of the red colour is diffused.
4.4. Results and discussion

Figure 8 illustrates the two series of parallel runs. Each \text{RUN}_i,m \ is illustrated by two curves respectively measuring $\text{O}_0,m,\text{Shapes} \ local \ homogeneity \ (O),$ and $\text{O}_0,m,\text{Colours} \ local \ homogeneity \ (O)$ along 12 cycles corresponding to successive flows of contributions from the princes expressing their preferences (informal knowledge). In the discussion, we simply denote $\text{Shape}^1H(O)$ and $\text{Colour}^1H(O)$.

A first observation is that the minimum for both measures is 0.33 whatever the radius ‘m’, except in the \text{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 $\text{Shape}^1H(O)$ 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 $\text{Colour}^1H(O)$ is always higher in \text{RUN}_1,m compared to \text{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 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 $\leq 2$, 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 $2\leq m<4$, the formal knowledge is detected ($\text{Shape}^1H(O)$ takes high values at cycle $0$); 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 $m\geq 4$, 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 $\text{bBq5-square-bBq2-big-bGt7}$ 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, $1 \leq m < 2$ is the relevant radius range for sorting the objects by colours, whereas $2 \leq m < 4$ 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)](image)

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
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.

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

- (formal) movie description: the viewpoint \(\text{(AA-MovieLens, } \{\text{movieX, genreY}\}, \text{ mo:genre, t})\) transcribes the fact: the artificial agent ‘AA-MovieLens’ connects the object ‘movieX’ to the descriptor ‘genreY’

- (formal) commenter profile: the viewpoint \(\text{(userX, } \{\text{userX, occupationY}\}, \text{ cv:jobType, t})\) transcribes the fact: the legal person ‘userX’ connects himself to the descriptor ‘occupationY’

- (informal) rating: the viewpoint \(\text{(userX, } \{\text{movieY, ***}\}, \text{ vpml:rating, t})\) 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., \(\text{mo:genre, foaf:gender}\), automatically provides a shared language, than can be exploited for building specific perspectives. For instance, the prefix ‘\(\text{mo:}\)’ 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.
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 (cycle0, m=1) and Measures (cycle0, 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 (cyclei, m=1) and Measures (cyclei, m=2) are operated at the end of each cycle.

Measures (cyclei, 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:

- $U_{i,m,Genres} \text{localHomogeneity}(M)$
- $U_{i,m,Ratings} \text{localHomogeneity}(M)$

In this experiment, the distance function is SBPD;

'm' denotes the radius chosen for computing neighbourhoods. Three series of RUNs designated by RUN1,m, RUN2,m and RUN3,m are presented hereunder.

Each RUNi,m corresponds to a perspective $U_i$. Whatever $U_i$, the $U\text{reduce}$ function is the sum; the perspectives differ only by the $U\text{map}$ functions.

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

<table>
<thead>
<tr>
<th>KG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects : 1682 'MovieId' + 943 'UserId'</td>
</tr>
<tr>
<td>Descriptors: 127</td>
</tr>
<tr>
<td>formal knowledge: 7471 viewpoints among which 2893 viewpoints of type mo:genre</td>
</tr>
<tr>
<td>informal knowledge: 5000 viewpoints</td>
</tr>
</tbody>
</table>

\[N^\text{view}=5\]
\[m'=\text{radius}\]

- **RUN$_1$,m** perspective $U_1$
  - formal knowledge: 'map(viewpoint)='3''
  - informal knowledge: 'map(viewpoint)='1''

- **RUN$_2$,m** perspective $U_2$
  - formal & informal knowledge: 'map(viewpoint)='1''

- **RUN$_3$,m** perspective $U_3$
  - formal knowledge: 'map(viewpoint)='1''
  - informal knowledge: 'map(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$_{3,1}$ and RUN$_{3,2}$ in more details.

$U_{i,m},\text{Genres} \quad \text{localHomogeneity}(M)$ and $U_{i,m},\text{Ratings} \quad \text{localHomogeneity}(M)$ are drawn for each run.

![Figure 11: Compared evolutions of formal knowledge (genres) versus informal knowledge (ratings)](image)

We first compare the values for cycle0, when ratings are absent of the knowledge graph. In the perspective $U_i$, 'map='3' for viewpoints of type mo:genre yielding strong synapses between movies and genres. In consequence, the suitable radius for observing $\text{Genres}_{iH}(M)$ 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., \( \text{Genres}^1H(M) \) becomes low.

In the perspectives \( U_2 \) and \( U_3 \), ‘map=1’ for viewpoints of type mo:genre. The suitable radius for observing \( \text{Genres}^1H(M) \) 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., \( \text{Genres}^1H(M) \) stays low.

Looking at cycle_1 to cycle_5, we observe that both ‘m=1’ and ‘m=2’ are suitable, although \( \text{Ratings}^1H(M) \) 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 \( \text{Genres}^1H(M) \) and \( \text{Ratings}^1H(M) \).

We observe that \( \text{Genres}^1H(M) \) always decreases while \( \text{Ratings}^1H(M) \) always increases. The curves (approximately) intersect at:

\[
\begin{align*}
0.32 & \text{ for } \text{RUN}_{1,1} \\
0.35 & \text{ for } \text{RUN}_{2,2} \\
0.34 & \text{ for } \text{RUN}_{3,2}
\end{align*}
\]

This intersection varies little with the perspective; the highest value ‘0.35’ is nevertheless obtained in the balanced perspective \( U_0 \). 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 approximately \( 5000 \times 3/5 = 3000 \) 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

This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University of Montpellier, CNRS and Cirad.

References


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


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


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


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


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