Viewpoints: An Alternative Approach toward Business Intelligence

Philippe Lemoisson, Guillaume Surroca, Stefano A. Cerri

eChallenges e-2013 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2013
ISBN: 978-1-905824-40-3

Viewpoints: An Alternative Approach toward Business Intelligence

Philippe LEMOISSON[1], Guillaume SURROCA[2], Stefano CERRI[2]
[1]*Cirad, Campus International de Baillarguet, Montpellier, 34398, France*
Tel: +33 4 67 59 37 42, Fax: + 33 4 67 59 38 27, Email:philippe.lemoisson@cirad.fr
[2] *Lirmm, Laboratory of Informatics, Robotics, Microelectronics of Montpellier University of Montpellier & CNRS, 161 rue Ada, Montpellier, 34095, France*
Tel: +33 467 41 85 00, Email:guillaume.surroca@gmail.com; Email:cerri@lirmm.fr

Abstract: Business intelligence is crucial for synergy and competitiveness in the world of business; the challenge is to capitalize upon the experiences, relationships and knowledge of all members over time. In this paper, we address business intelligence and more generally collective intelligence through a social approach based on viewpoints. The collective knowledge is stored within a bi-partite graph populated by agents, documents and topics on one side and by viewpoints on the other side. Viewpoints consist in labelled triples (agent, document, topic). We define a semantic distance on this structure. We then engage a selectionist process: information retrieval is based on existing viewpoints while feedbacks yield new viewpoints. The implemented framework and algorithms are tested through a simulation.

1. Introduction

Business intelligence (BI) is crucial for synergy and competitiveness in the world of business. In a 1958 article [1], IBM researcher Hans Peter Luhn used the term; in that paper he defined intelligence as: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal." Although the collective dimension is absent from this initial definition, BI can clearly be considered as part of a broader phenomenon called collective intelligence (CI) in our highly interconnected societies. The challenge is twofold: i) motivate people to participate to a collective endeavour based on knowledge sharing, ii) fully exploit a distributed set of resources (documents, people and computers) for producing new information and help elaborate new knowledge.

In this paper we consider the second aspect of the challenge through three basic questions:

- how to help each individual find trustful documents (texts, datasets, images …) on a particular topic?
- how to help her/him find the right people to exchange, argue and capitalize on a particular subject?
- how to empower the emergence of new knowledge within the community?

BI mostly results of the compilation of semi-structured or unstructured data (texts but also images), as stated in [2]; this considerably disadvantages approaches needing a rationalization of terminology. In BI or CI, the question of meaning arises in first place, especially the question of the reference to "knowledge supports owning an URI, i.e. documents".

Our "attitude" facing this question is that of H. Halprin. In [3], Halprin firstly recalls a controversy between two philosophical positions: i) Kripke/ Berners-Lee asserting that "the URI means what the owner says or thinks it does", so that the URI unifies access and reference and ii) Russell/Haynes emphasizing that "reference is absolutely unconstrained

except by formal semantics", so that "the relationship between access and reference is essentially arbitrary". The author then refers to Wittgenstein, depicts the historical path linking the understatement "meaning is use" to the discipline of information retrieval, and proves the benefit of "relevance feedback" [4] by experimenting on a set of 200 queries. He concludes that "finding and giving meaning to URIS on the Semantic Web can be built out of the social semantics implicitly given by the searching behaviour of ordinary users".

Among many arguments in favour of a thesis involving the humans in information retrieval is the following: if the process was to be achieved on the basis of formal semantics only, we should be facing what is defined in [5], [6] as an "AI-complete" problem. To escape this difficulty, some kind of human arbitration during the retrieving process is necessary; an illustration can be found in [7] where experts supervise the reading and collecting of web data by an artificial agent.

We therefore consider a combination of AI (artificial intelligence) techniques such as classifiers, machine learning and DHC (distributed human computation). An emblematic use of DHC is discussed in [8]: humans' micro efforts to recognize characters are integrated by an artificial agent in charge of book digitization. Among many illustrations of this new paradigm, we can cite two: DHC helps and solve the co-reference resolution problem in [9], it is also adopted in [10] where semantic matching is done with the help of folksonomies. To describe this new collaboration paradigm, the metaphor of a global brain has been used by many authors, such as Bernstein, Klein and Malone: "As the scale, scope, and connectivity of these human-computer networks increase, we believe it will become increasingly useful to view all the people and computers on our planet as constituting a kind of global brain" [11]. It must be noted however that a full implementation of this metaphor would require the emergence of new knowledge, and even of "new levels of understanding", which according to [12] has not been assessed yet.

Our paper addresses the emergence of new knowledge within a community sharing information. We propose a conceptual framework intended to help and answer the three questions above, and simultaneously open the way for an graphical assessment of collective learning. Our approach is based on three ideas:

1. We consider a "knowledge space" where agents, documents and topics are symmetrically considered as "objects". Within this structure, we consider triples ($agent_x$, $object_1$, $object_2$) labelled by +1 or -1. We call these labelled triples "viewpoints" and interpret them in the following way : +1 means "$agent_x$ believes that $object_1$ and $object_2$ are "semantically close" ; -1 means "$agent_x$ believes that $object_1$ and $object_2$ are "semantically distant".
2. We take advantage of this "semantic relatedness" for information retrieval.
3. We engage a selectionist process. The coevolution of viewpoints is ruled by the agents' feedbacks to information retrieval: retrieval is based on existing viewpoints while feedbacks yield new viewpoints.

According to this vision, users are agents located within a knowledge ecosystem: a network of documents, topics and other agents. The foundational chunk of information is not a binary link between a document and a topic but a ternary relation including the agent that emits the viewpoint. The building of knowledge is cumulative in a fashion similar to what occurs in the Open Source movement: when capitalizing on someone else's viewpoints previously expressed, we engage in offering our own viewpoints for the benefit of future users. The reputation of agents, the relevance of documents with respect to particular topics, the proximity or importance of topics will continuously evolve according to the interplay of viewpoints. The knowledge space reflects in real time the evolution of both the outside world and of the community knowledge.

2. Objectives

In this paper, we have two objectives:

- to enter in the details of the conceptual framework supporting our approach;
- to test the use of the "semantic distance" in information retrieval by simulating the co-evolution of viewpoints in the context a community of artificial agents.

3. Methodology

Our representational basis for the knowledge space is "purely topological": it consists in a graph linking people, documents and topics through viewpoints. This representation does not involve formal semantics, nor relies on vector spaces (see [13] for an extended analysis of the semantical exploitation of high dimensional vector spaces), nor necessitates (but it can benefit from it) a previously built lexical database such as WordNet. Therefore the semantic distances such as those analysed in [14] do not apply; instead we define a socially constructed ψ-distance playing the role of "semantic distance". A decisive improvement in information retrieval (IR) is expected from the topology itself and the use of the ψ-distance.

3.1 – The knowledge space

The knowledge space is a bipartite graph KG populated with two classes of nodes: class "O" and class "V".

There are three subclasses of "O":

- the objects of subclass A are interpreted as "agents"; those can be human or artificial (e.g. topic miners or embedded logics). The concept of agent unifies all knowledge providers.
- the objects of subclass D are interpreted as "documents". The concept of document unifies all knowledge supports (texts, maps, videos…).
- the objects of subclass T are interpreted as "topics". The concept of topic unifies all knowledge descriptors, i.e. all means of describing agents, documents, or subjects of enquiry.

The objects of class V are viewpoints; they express beliefs. They are formalized below:

Definition 1: We call viewpoint a pair $\mathbf{v} = (\mathbf{u}, \boldsymbol{\alpha})$

i. $\mathbf{u}$ is a couple consisting in on agent (the emitter of the viewpoint) and a pair of objects of O; $\mathbf{u} = (a_1, \{o_2, o_3\})$ means: "agent a_1 has something to say about the pair $\{o_2, o_3\}$". It is denoted $u = a_1 \rightarrow \{o_2, o_3\}$.

ii. $\boldsymbol{\alpha} \in \{-1, +1\}$ is the evaluation given by agent a_1 about the semantic proximity between the objects o_2 and o_3; $\alpha = +1$ means "semantically close" ; $\alpha = -1$ means "semantically distant ".

Definition 2: The knowledge space is a directed graph built upon O and V. It is denoted KG:

- the elements of $O \cup V$ provide the vertices of KG
- the elements of V are labelled by (α)
- the edges of OG are the directed links built from the elements of V: each $v = a_1 \rightarrow \{o_2, o_3\}$ provides three directed edges: $a_1 \rightarrow v$, $v \rightarrow o_2$ and $v \rightarrow o_3$.

3.2 – The ψ-distance

The basic concept in the definition of the ψ-distance is the notion of "jump". A "jump" is a triple (o_1, v, o_2) where o_1 is adjacent to v and o_2 is adjacent to v in KG; o_1 and o_2 are said to be "ψ-adjacent" in O. A path in O is a sequence of "ψ-adjacent" objects.

This leads to the following definition[1]:

Definition 3:

i. ψ-distance (o_i, o_k) = the smallest path-distance between o_i and o_k along all sequences of ψ-adjacent objects.
ii. ^{m}Neighbourhood (o_i) = {o_j such that ψ-distance (o_i, o_j) ≤ m}. Neighbourhood (o_i) contains at least 'o_i'.

ψ-distance (o_i, o_k) will be called the "semantic distance[2]" between o_i and o_k in O.

3.3 – A protocol for simulating the coevolution of viewpoints

In order to simulate the use of the "semantic distance" in information retrieval, we implement the co-evolution of viewpoints in the context a community of artificial agents.

We consider fixed sets of artificial agents (A), documents (D) and topics (T); O=A∪D∪T.

We initialize the community knowledge by building an initial set V_0 of viewpoints of the type (a_i→{o_j, o_k},+1), where (a_i, o_j, o_k) is randomly chosen in A×O×O.

Let KG be the bipartite graph built upon O and V_0.

Let β be a "permeability" parameter characterizing the average ratio of acceptance of new viewpoints by an agent; β belongs to [0, 1][3].

Let us call a "run" the following sequence:

1. an agent a_i is chosen at random
2. an object o_q is chosen at random, o_q stands for the query made by a_i
3. the answer to the query is computed by using the semantic distance (see section 5); it consists in all the objects of O situated at a distance smaller than 'm' from 'o_q'. Let R= ^{m}Neighbourhood(o_q) be the answer.
4. we call ^{known}R the part of R consisting in the objects 'o_k' about which a_i had already emitted viewpoints of the type (a_i→{o_q, o_k},+1) or (a_i→{o_q, o_k},-1)
5. let ^{new}R = R - ^{known}R the part of the answer which is new for a_i
6. a_i emits[4] a viewpoint for each object o_k in ^{new}R; this viewpoint is (a_i→{o_q, o_k},+1) with the probability β and (a_i→{o_q, o_k},-1) with the probability (1-β)
7. we call "satisfaction[5]" the proportion of accepted objects within ^{new}R
8. we call "relevance[6]" the proportion of objects (either in ^{known}R or in ^{new}R) associated to a positive viewpoint within R

[1] in order to keep this part concise, we skip the progressive steps of the formalization.

[2] if we consider a class "Emotion" as a subclass of "Topic" then the distance also deals with the "proximity with emotions".

[3] β=1 would mean that a_i accepts all the proposed objects, and is therefore very « permeable » to the others'viewpoints; on the contrary β=0 would mean that a_i refuses all that she/he does not already know.

[4] in the implemented algorithm, the viewpoints are stored in a buffer and emitted at the end of each go.

[5] we consider that the satisfaction of the agent only depends of the relative relevance of new objects

One "go" consists in X "runs" potentially involving as many different agents; we compute the average "satisfaction" and "relevance" at the end of each "go". A "simulation" consists in Y "gos".

4. Technology

The implementation has been done using Java, the JUNG graph display API and JOpenChart.

5. Developments

The ^{m}Neighbourhood algorithm presented below is inspired from Dijkstra's algorithm. We label the objects of KG with a computed semantic distance which plays the role of the weights of Dijkstra's algorithm. The worst case complexity of the algorithm is $O(|V|^2|O|^2)$ with |V| = number of Viewpoints and |O| = number of objects. However this complexity is never reached as the spreading is limited to the exact and sufficient amount of expands to compute the ^{m}Neighbourhood.

```
mNeighbourhood ( KnowledgeGraph kg, Object searchedObject)
Queue todo ← { searchedObject } ;
List neighbourhood ← { ∅ };
Init() ; // sets all nodes distances to infinity ; all lists to "empty" ; all variables to "null" or 0
While todo ≠ { ∅ }
        Object oi ← pop( todo ) ;

        For each Viewpoint vj ∈ oi.neighbours

        If vj ∉ oi.referentViewpoints

                vj.pred ← oi ;

                For each Object ok ∈ vj.neighbours

                        If ok ≠ oi

                                ok.referentViewpoints ∪ { vj } ;

                                sum ← ∑( { Viewpoint vl | vl.pred = oi } ) ;
                                If sum > 0
                                        jumpDist ← 1 / sum ;
                                        newDist ← oi.dist + jumpDist ;
                                        If( newDist < ok.dist OR ok.dist < 0 OR
                                           Vj.pertinence = -1 )
                                                ok.dist ← newDist ;
```

[6] the relevance is computed on the basis of the whole answer to the query; an answer can be relevant even if nothing new is proposed

todo ∪ { o_k } ;

If o_k.dist ≤ m

neighbourhood ∪ { o_k }

Figure 1 illustrates the ^{m}Neighbourhood algorithm applied to the agent 'a_3'. Each node is labelled as follows: <node name> | d = <distance>; d < 0 is equivalent to "disconnected".

According to the distances labeling the nodes:

- ^{0,5} Neighbourhood (a_3) ={a_3, d_0}
- ^{1}Neighbourhood (a_3) ={a_3, d_0, t_2, d_3, a_0, d_4, a_2}
- ^{2}Neighbourhood (a_3) ={a_3, d_0, t_2, d_3, a_0, d_4, a_2, a_4, t_0, a_1, t_1, t_3}

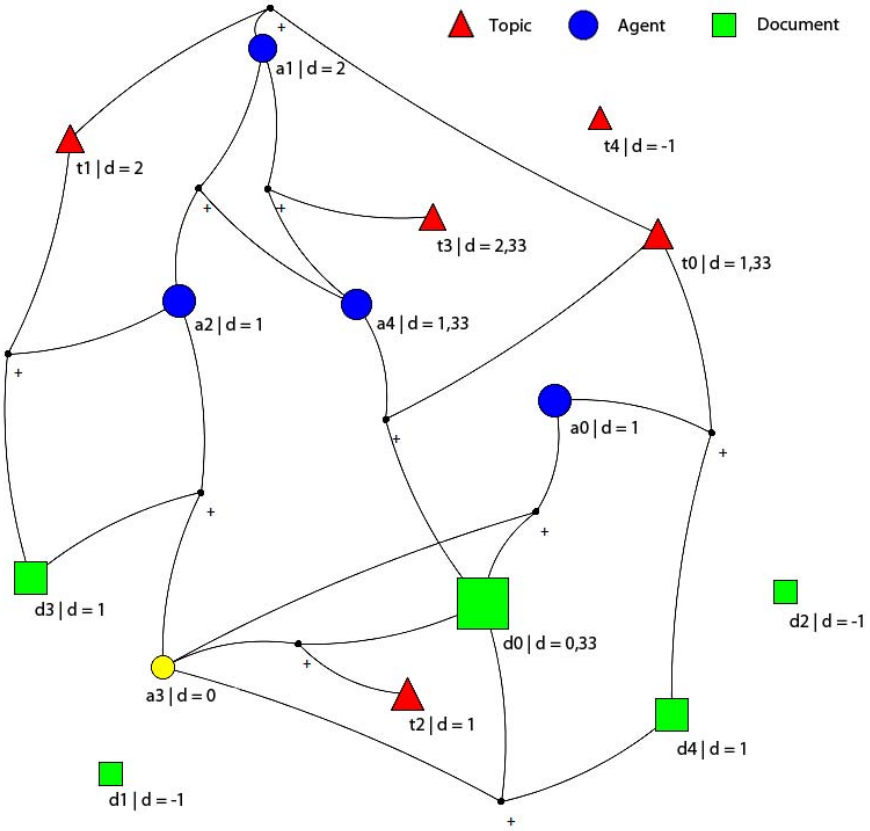


Figure 1: the semantic distance in O, resulting from a computation in KG

6. Results

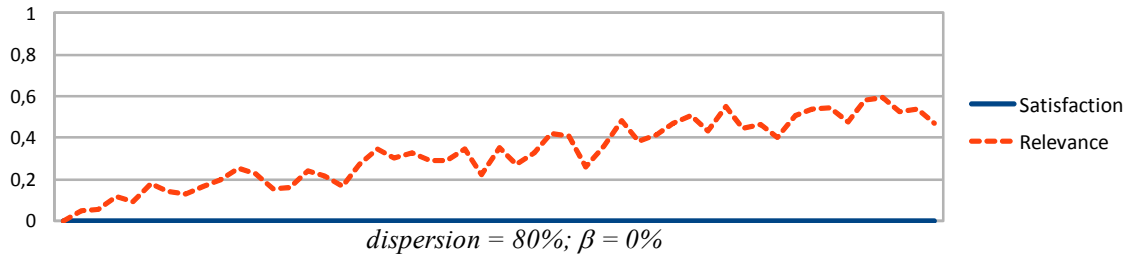
The KG structure and IR algorithm depicted above have been used in 3 successive simulations, using the parameters detailed in Table 1; only the β parameter differs from one simulation to the other:

Table 1: parameters for the simulations 1,2 and 3

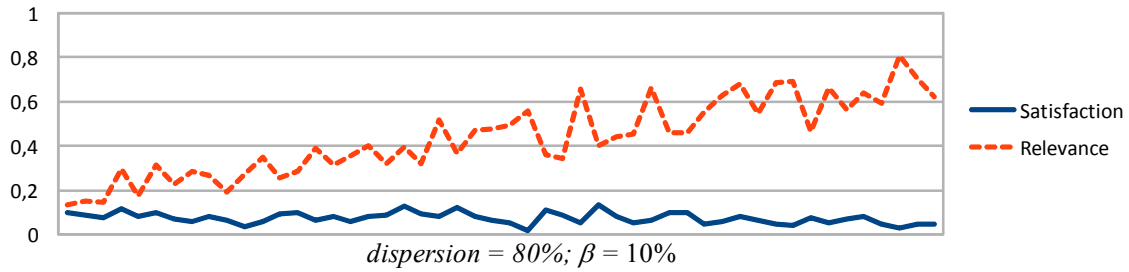
Card(A) = 20	Fixed number of agents
Card(D)=10	Fixed number of documents
Card(T)=20	Fixed number of topics
Card(V_0)=10	Initial number of viewpoints
m=1	Parameter of the ^{m}Neighbourhood
X=10	Number of "runs"
Y=50	Number of "gos"
β1= 0%	"permeability" parameter for simulation 1
β2= 10%	"permeability" parameter for simulation 2
β3= 30%	"permeability" parameter for simulation 3

This set of parameters leads to reasonable computational times (less than 300 seconds with 3 Ghz quadcore). The following curves provide three main results:

1. the average satisfaction of the agents is equal to their permeability. This was strongly expected since both correspond to the acceptance of new viewpoints; the only interest of the "satisfaction" curves is to show the dispersion of the random tries.
2. when the permeability is null (β=0%), the relevance grows from 0 to 0.5; this growth is approxatively linear. It can be explained by the fact that all the unknown retrieved objects (those generating negative viewpoints in KG) become more distant; the algorithm therefore provides less and less answers and this leads to an increased proportion of relevant ones.
3. when the agents accept to learn from the community the increase in relevance grows in proportion with their permeability; this was somewhat expected. The maximum relevance reaches 0.6 (=0.5+0.1) when β=10%, it reaches 0.8 (=0.5+0.3) when β=30%.



dispersion = 80%; β = 0%



dispersion = 80%; β = 10%

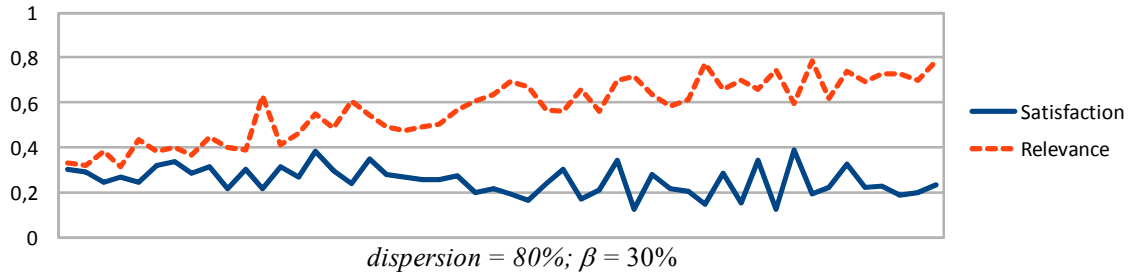


Figure 2: simulation results with various values for permeability β

The above results demonstrate the operationality and confirm the expected behaviour of the ψ-distance. They also demonstrates however that simulations based on probabilistic behaviours can provide no more than variations around a probabilistic equilibrium ; they do not open to the study of the self-structuring ability of the graph.

7. Business Benefits

In business, considerable economic interests depend on the capacity of companies to engage their employees into a collective retrieval of salient information among tons of public data on market, clients and competitors. Face of this situation, the need for collective intelligence is crucial, especially the ability to provide quick and relevant answers to the questions listed in the introduction. Moreover the opportunity of intergenerational transmission of knowledge is priceless, especially at a turning moment when baby-boomers retire.

Our approach deals with all these questions in a unified way. We have demonstrated simple requests about a single object, the approach applies as well (by intersecting neighbourhoods in KG) to complex requests involving several objects such as "I remember my colleague 'a' recommending some documents in complement of document 'd' about the topic 't', please help me find these documents."

Once a smart interface for emitting viewpoints is designed and implemented, the approach will hopefully appear as little demanding and much rewarding; it is therefore expected to yield a structured knowledge space well suited for Business Intelligence.

8. Conclusions

In this paper, we have presented a conceptual framework aimed at empowering Business Intelligence. We have formalized the notion of "viewpoint" in a knowledge space populated by agents, documents and topics; we have defined a semantic distance within this knowledge space; we have implemented an IR algorithm based on the semantic distance.

We have tested the algorithm computing the distance by simulating a cycle of information retrieval and feedbacks in a society of virtual agents. This simulation demonstrates a positive correlation between the "permeability" of the agents and the average subjective relevance.

Our current objectives are to study of the self-structuring ability of the graph in the context of a real-life scenario and then to take advantage of sophisticated analytics such as those defined in [15] in order to assess collective learning.

References

[1] Luhn, H P (1958). "A Business Intelligence System". IBM Journal 2 (4): 314. doi:10.1147/rd.24.0314.
[2] Inmon, B. & A. Nesavich, "Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence", Prentice Hall 2007, pp. 1-13
[3] Halpin, H. (2009). Social meaning on the web: from Wittgenstein to search engines.

[4] Callan, J. P., Croft, W. B., & Harding, S. M. (1992, January). The INQUERY retrieval system. In Database and Expert Systems Applications (pp. 78-83). Springer Vienna.
[5] Shapiro, S. C. (1992). Encyclopedia Of Artificial Intelligence Second Edition. New Jersey: A Wiley Interscience Publication.
[6] Yampolskiy, R. V. AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System, ISRN Artificial Intelligence, vol. 2012, Article ID 271878, 6 pages, 2012. doi:10.5402/2012/271878
[7] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., & Mitchell, T. M. (2010, July). Toward an architecture for never-ending language learning. In Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010) (Vol. 2, No. 4, pp. 3-3).
[8] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. Science, 321(5895), 1465-1468.
[9] Yang, Y., Singh, P., Yao, J., Au Yeung, C. M., Zareian, A., Wang, X., ... & Shadbolt, N. (2011). Distributed human computation framework for linked data co-reference resolution. The Semantic Web: Research and Applications, 32-46.
[10] Togia, T., & McNeill, F. (2011, July). It makes sense... and reference: the role of folksonomy in formal ontology matching. In Workshop on discovering meaning on the go in large heterogeneous data (p. 31).
[11] Abraham Bernstein, Mark Klein, and Thomas W. Malone. 2012. Programming the global brain. Commun. ACM 55, 5 (May 2012), 41-43.
[12] 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), 4-13.
[13] Sahlgren, M. (2006). The Word-space model (Doctoral dissertation, Stockholm University).
[14] Budanitsky, A., & Hirst, G. (2001, June). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources (Vol. 2).
[15] Halpin, H., Robu, V., & Shepherd, H. (2007, May). The complex dynamics of collaborative tagging. In Proceedings of the 16th international conference on World Wide Web (pp. 211-220).