Information Evolution And The Future Of The Web

Massimo Marchiori

Abstract: The Web is an incredible informative media, but its same success has turned it into a two-sided phenomenon. On the one side, it has got plenty of information, a richness of contributions and data that is unprecedented in the history of humanity. On the other side, the sheer size of the information present in the Web make it hard to access and to use at all levels, humans and computer-based. In this paper we present a high level perspective on this two-sided nature of web information, by introducing the general concept of Trust Scenario and see its implications by defining major axes of information, and the corresponding directions they can lead to in the future of the Web.

Index Terms: World Wide Web, Semantic Web, Information management, Trust, Deception, Modal Logics.

1 Introduction

THE growth of the World Wide Web in these years has been phenomenal, and as a result it has become a major informative channels. The same success of the informative Web has brought several side-effects, leading to usability problems, both by humans and also by automatic processors. So nowadays the following questions on the informative Web have become fundamental: how do we treat this information, how do we give some order, and possibly help its intelligent reuse? So, how can we face these challenges? One view that has been taken is the so-called Semantic Web (cf. [1],[2],[3]), that tries to tackle this formidable problem by providing better semantic information. The basic idea is to extend the classic informative layer via further information layers that provide additional semantic context and logic. Using these extra information layer allows for a much more powerful handling of Web information, enabling automated reasoning and therefore helping a better interaction between people and information. But the Semantic Web is just one ingredient in the overall recipe: in order to understand the full problem of the current informative Web, we need to stay at a more general level, understand how it works, what are the possibilities into action, and then shape possible scenarios. The paper is organized as follows. First, we start with some basic definitions about the World Wide Web and its composing structure. Then, we proceed by introducing the notion of informative axes, using the informative correspondence with another media field, journalism. Then, we introduce the other side of the coin, the negative components that make up the Web as a whole. In that section, the key notions of Trust Scenario and Deception are introduced. Next, we introduce the golden ratio of the Web, and proceed by showing its application in the various evolution that the web can represent. The information axes can be combined into several ways, and each choice correspond to a possible evolutionary scenario for information handling on the web. Finally, we show examples of the other "bad" components of the web, pertaining to the trust problem: and introduce the Web skews, together with a first classification.

- *Massimo Marchiori, Department of Mathematics, University of Padua,Via Trieste 63, 35121 Padova, Italy E-mail: Massimo@math.unipd.it*

2 Preliminaries

We consider the World Wide Web in its approximation of "universal information space" where there are certain resources that are retrieved by dereferencing a certain URL (cf. [4]. In other words, more technically, we just consider the Web under the assumption that the HTTP GET method is the only one to be used (in fact, this approximation gathers, at least architecturally, a good part of the WWW, as GET is architecturally a "universal operator", in the sense of category theory, for most of the HTTP methods that collect information). So, we can view the WWW as a "dereference map" δ from URLs to byte streams, with the intended meaning that $\delta(u)=s$ if and only if, in the real WWW, there is a machine such that retrieving (GET) the URI u gives as a result the byte stream s. When we later add semantics and meaning (depending on the particular application we use), we are essentially using an interpretation (let's say Ψ) of such web objects that can give us more knowledge. That is the one that can allow, in trust scenarios, to lower deception. Most of the times the World Wide Web Consortium (W3C) sets up a standard (for example, for the Semantic Web), Ψ is refined.

3 The Information Axes

We start by better clarifying the same concept of information. In order to do this, we relate the Web to a somehow similar field: journalism. Journalism has had the same problem since its inception: you have to report and classify a bit of information, but here "information" is as wide as the information we have nowadays in the WWW, potentially treating every subject. So, what's the way out? How can we tame such variety of information? One way out, which proved to be quite successful, is to use the so-called five W's, which are five axes that somehow identify the information event. These are the well known in the journalism field and self-explanatory:

1. WHAT
2. WHERE
3. WHO
4. WHEN
5. WHY

So, can we view the information on the Web using similar axes? What about reusing these five W's for the information present in the Web? Historically, the five W's have already been used explicitly inside some web application, for instance in XML dialects (cf. [5] for one of the earliest historical uses), but what about reasoning about them at the most abstract

level? Can they help, for instance in building a better, or different, Semantic Web?

4 The Other Side of the Coin

The Web and its information axes are just one side of the coin. The other side of the coin is a totally different world, a world that takes into account human behaviour in all its aspects, positive and negative. So far the information view we have seen corresponds to the positive sides: we want to state information, and in the best possible ways. But we also have the negative sides, and they must be properly factored in, in order to get the overall complexity of the Web system, and be able to proceed in the best way. The major problem that lies in this "dark part" of the Web is trust. The problem of trust is generally a fundamental one in computer science, and even more in a distributed and decentralized environment like the WWW. In order to talk about trust, we can try to abstract and give a more or less formal definition that we can later reuse to define some terminology. So, in general we can define:

Definition 1 (Trust Scenario). A trust scenario is a quintuple (T ,R, U, S, τ) so defined:

1. A "trust property" T (that can be computationally intractable).
2. A "test property", τ (that is usually computationally tractable).
3. A "universe" U of entities (e.g., software agents, persons, etc.).
4. A number R (R ϵ [0, 1]), indicating the "real" probability that τ implies T .
5. A mapping S from U to [0, 1], indicating for every entity e ϵ U the "Subjective" probability S(e) that τ implies T.

Note that a trust scenario usually is not fixed but depends on an environment E, which can contain the information on how to compute the probabilities, and that can be itself dependant on a number of factors, like time for instance. In the following, when talking individually about entities and test properties, we shall always mean them within an understood environment and trust scenario (an "E-TRUSτ"). Having defined what a trust scenario is, we can now use it to somehow formally define when problems with trust occur, i.e., when we have deception:

Definition 2 (Deception). Deception occurs for an entity e when R << S(e) So, in general, we can say that in a trust scenario deception occurs when there is an entity such that deception occurs for it. The severity of a deception could of course be quantified in various degrees, both locally for an entity e (e.g. by using the gap measure S(e)-R), and globally by measuring its diffusion in the universe U (for instance, in case of a finite universe, by averaging the local gap measure, or by fixing a threshold and measuring how much of the universe has a deception higher than that).

5 The Golden Ratio

The aspect of Trust is not the only one that we need to factor in order to obtain the full picture of information handling on the Web. In the WWW, resources do not come for free, but there is a cost for creation and modification. Every solution for the WWW may bring some benefits, but usually also implies new creation/modification of information, and this cost must be taken into account, because that could be a big obstacle to the widespread adoption of such solution. Therefore, a key parameter to take into consideration for success is the cost/benefit ratio, the true golden ratio for Web evolution. The cost/benefit (for instance, to diminish deception of some trust scenario) is not just a static measure but a dynamic one: its variations thru time impact adoption of new technologies, and subtly turn the Web into a complex system. At first-order approximation, the cost/benefit ratio must be sufficiently low for users to adopt the solution and to build critical mass, so to create a possible network effect. An initial ratio that is too high can just prevent adoption, even if the long-term evolution of the system would bring the situation to a very low level of the ratio itself. In other words, the cost/benefit ratio can generate situation of local maxima that act like bareers to the development of a technology or line of evolution.

6 The Light Basic Axes

It is of utmost importance to minimize the cost of representing additional information in the WWW. This means that building up from the five W's axes should not be too expensive. In other worlds, like in journalism, the right compromise should be found between efficiency and effectiveness. So, we should strive to obtain the information given by the five W's in the most economical possible way, almost "zero-cost" if possible. Is there such a way? The answer is yes, at least for four or the five axes:

1. zero-cost WHAT == the resource (at least the message-body)
2. zero-cost WHERE = yes, the URI of the resource (Content-Location or Request-URI)
3. zero-cost WHO = yes, the URI authority (Host)
4. zero-cost WHEN = yes, the time when the resource was transmitted (Date)
5. zero-cost WHY = no.

Note that these definitions depend on our original assumption of the Web based on the dereferencing map δ (cf. Section 2). But similar lines of reasoning can be applied to different contexts, Web or Internet based. In the following, when applicable, zero-cost W's are understood.

7 Evolutions

What does it mean for a standard or for an application to be "Web"? In many cases, such standard/application doesn't take into account the mapping δ, but just takes into consideration the message-body (cf. [5]) of the image of δ, in some cases integrated with the information about the MIME type. Simply speaking, this is tantamount to considering "web pages". Restated, such standards/applications are posing the WHAT axis equal to such web pages. This is the starting point, and we can therefore define a first kind of World Wide Web:

W1 = WHAT

The current classic architecture of the Semantic Web stays in the W1 (where WHAT = message-body). The problem is that, to build a reasonably effective Semantic Web (or in any case, to increase the semantic content, therefore diminishing deception) can have a very high cost. The problems that nowadays Semantic Web mass adoption is experiencing mostly stem from this aspect: the initial local minimum of the cost/benefit ratio. Granted that W1, the status quo, is a rather unsatisfactory situation, we can look in general at what

additional information axes we have at our disposal, and consider other possible lines of evolution obtained for instance by extending the W1 using the information provided by the other W axes. Therefore, Ψ (the W1) can be increasingly integrated with the zero-cost WHERE, WHO and WHEN, giving various possible scenarios for evolution:

1. three flavors of W2 ((WHAT, WHERE), (WHAT, WHO), (WHAT, WHEN)),
2. two flavors of W3 ((WHAT, WHERE, WHO), (WHAT, WHERE, WHEN), (WHAT, WHO, WHEN)),
3. and one W4 (WHAT, WHERE, WHO, WHEN).

Therefore All of these combinations form a wide array of possible choices: the more information we add, the more power we get, but at the potential expenses of more computational resources (space and time).

8 Modal Logics

The extra information provided by the W layers can be profitably reused in inference mechanisms, just like W1 has been used in the Semantic Web. The various W's give a kind of temporal modal logic (see [8],[9],[10],[11]): WHERE == world , WHO == world, WHEN == time. As common to modal logics, statements expressed in the same world can usually combine seamlessly, using the operators that the interpretation I provides; as WHERE specializes WHO, this means that choosing a W2 or W3 with a WHO (and without a WHERE) will generally allow many more inferences than choosing a W2 or W3 with a WHERE. On the other hand, the WHEN component is troublesome, as it represents a time instant, and so in general composition becomes practically impossible. Therefore, in order to allow a more useful use of WHEN, we can relax the composition rules, which is equivalent to change our interpretation of the timed logic. For instance, one possible choice could be to employ some assumption of local time consistency (cf. [7]), therefore assuming that web resources stay somehow stable within some time intervals. This changes the interpretation of WHEN from a single instant to a time interval, allowing more inferences to take place. The price is that the approximation given in the choice of the stable time interval will likely make the deception increase, so there is a trade-off. However, this trade-off can be mitigated by using appropriate probability distributions of the "local stability" of a resource (therefore, passing to fuzzy/probabilistic reasoning). Another possible choice that doesn't necessarily use the local time consistency assumption is to change the definition of WHEN, which is now rather simplistic (Date), and add for example the information about cacheability of the resource, and the expiration date: this gives right away a timed interval structure, which can be quite useful. The price to pay is that appropriate cache information can have a cost. However, the benefits are quite high, because this information not only can help produce many more useful inferences in a W2, W3 or W4, but help in general the performance of the WWW (the primary reason in fact why cache information is present...). So, this approach might be worth exploiting, making a semantic/information layer (W1) interact with an operative layer that seems far, but that in fact can provide useful semantic information and grant extra inference power. Finally, of course, more sophisticated approaches are possible, where some or all of the information in the WHERE/WHO/WHEN/WHY axes is refined by integration with the information in Ψ. These intermediate solutions are a tempting way to overcome the limitations of the simplest W2, W3 and W4 solutions, while still keeping reasonably low the cost/benefit ratio.

9 Information-flow Skews

The approach that we have seen so far is based on principles, but it has to be noted that other complementary views must be taken into consideration when analysing for instance trust scenario. Problems may occur, coming from malicious attempts to increase deception over time. In such cases, it is not uncommon to use all possible means: many trust problems on the Web usually occur because of so-called information-flow skews. A skew occurs when there is a treatment of the information flow in the WWW that departs from the high-level standard architecture of the Web, and that the user cannot see. There are at least three main skews that we can categorize:

1. The Visual Skew
2. The Navigation Skew
3. The Protocol Skew

The Visual Skew occurs when not all the data flow goes back to the user, and can be synthesized with the slogan

"What is you see is not what you get".

In practice, this skew exploits the possibility that how a resource is rendered on the screen/medium (and so, what the user perceives) can be much different from what is actually in the resource. One of the classic cases where Visual Skew shows its appearance is the so-called search engine persuasion (sep) (cf. [12]), also sometimes known (improperly) as search engine spam or with the trendier wording of search engine optimisation (seo). Sep is the phenomenon of artificially "pumping up" the ranking of a resource in search engines, so to get a higher position (with all that means in terms of visibility and advertisement). Most of the techniques used in sep just profit in various ways of the visual skew (see [12],[13]), so to apparently present to the user a certain resource which is in fact quite different under the surface. The Navigation Skew occurs when not all the WWW navigation is explicitly presented to the user. For instance, if we click on a link (i.e., request a resource on the WWW), we expect that we are just fetching the corresponding page. But this is not true: for example, frames and images are automatically loaded for us. This apparent facility, however, leaves the door open for the navigation skew, as it means essentially that the authors of a resource can make us click on the page they want (!). Well-known examples of use of the navigation skew are banner ads and pop-up windows, all employing this skew in its various flavors. But even worst, the navigation skew makes possible applications that are potentially quite dangerous for users, like tracking systems (a la DoubleClick and Engage). Typically, such privacy-risky applications might employ a combination of skews (for instance, using so-called "web bugs", images that use the navigation skew to send data, and the visual skew to hide, therefore resulting invisible). The Protocol Skew occurs when the WWW protocols (e.g. HTTP) are abused (for instance, turning a stateless connection into a connection with state). For instance, the HTTP information flow in some cases should be from server to user (i.e., if we request a page, its only the server that gives us information). But this architectural principle is not always followed in reality, as for example many

sites tend to collect so-called “clickstream” information (what you requested, when you did it, what is your computer internet address, etc). Again, this skew allows to collect information “under the rug”, and can therefore become quite a problem for the user’s privacy. Such problem can be worsened a lot when abuse of this skew is performed via aggregation: for instance, use of dynamic links (URIs that are generated on the fly) together with appropriate use of other clickstream information can make such tracking easily work not just for a single click, but for an entire session.

10 Conclusion

Current attempts to formally introduce extra layers of reasoning (like the classic views of the Semantic Web) have insofar stayed within a rather limited setting. Once we broaden our information view, we can identify a much broader range of information that constitutes, at a first-order approximation, the backbone of the Informative Web itself. The positive side can be modeled by using the informative axes system proper of journalism. The negative side can be modeled by using the general definitions of Trust Scenario and the corresponding instances of Deception. All these aspects, that balance pro’s and con’s, are eventually ruled by the golden ratio, the ratio of cost versus benefit of a specific solution w.r.t. a corresponding user or system. This brings to the natural tendency of trying to develop “light” scenarios of Web evolution, where the axes information is extracted with a fast first approximation. But on the other hand, the other side of the coin is much more subtler, and somehow follows a “short blanket” principle: using a light solution on the positive side can bring a negative counteraction on the negative side, for example via information-flow skews. Therefore, every practical use of W1, W2, W3, W4 and higher-level combinations has to take into account the potential danger, that “light” solutions can be necessarily prone to a higher risk in terms of possible deception. This is somehow also, on a different but related level, the same dilemma that information extractors like search engines have to face nowadays: the eternal battle between fast execution versus countermeasures. Nevertheless, in the same way search engine work reasonably well, even though not excelling and in some cases failing, the same should be done in general for the informative Web: extending the limitative view of a W1 world of information and embracing a much more informative world of axes that are tightly interlinked with the Web structure. Doing this also exposes every automated reasoning to the danger that is proper of search engines, trust and deception, but this is part of the game: what matters is to exit the isolated world of logics that state classic truth, and embrace the wider realm of modal logics, taking into account chances, modes, and time.

References

[1] T: Berners-Lee, J. Hendler, and O. Lassila. "The semantic web." Scientific American 284.5 : 28-37, 2001.

[2] G. Antoniou, and F. Van Harmelen. A Semantic Web Primer. MIT press, 2004.

[3] N. Shadbolt, W. Hall, and T. Berners-Lee. "The semantic web revisited." Intelligent Systems, IEEE 21.3 : 96-101, 2006.

[4] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform Resource Identifiers (URI): Generic Syntax", IETF RFC, 1998.

[5] M. Marchiori, "The XML Documentation Markup", The World Wide Web Consortium (W3C), 1999.

[6] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol – HTTP/1.1", IETF RFC, 1999.

[7] M. Marchiori, "The Quest for Correct Information on the Web: Hyper Search Engines", Proceedings of the Sixth International World Wide Web Conference (WWW6), 1997.

[8] P. Blackburn, J.F.A.K. van Benthem, and F. Wolter, eds. Handbook of modal logic. Vol. 3. Elsevier, 2006.

[9] B. Bennett, A.G. Cohn, F. Wolter and M. Zakharyaschev, "Multi-dimensional modal logic as a framework for spatio-temporal reasoning." Applied Intelligence 17.3 : 239-251, 2002.

[10] E.A. Emerson. "Temporal and modal logic." Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics (B) 995: 1072, 1990.

[11] A. Kurucz, F. Wolter, M. Zakharyaschev, and Dov M. Gabbay eds. Many-dimensional modal logics: theory and applications., Elsevier 2003.

[12] M. Marchiori, "Security of World Wide Web Search Engines", Proceedings of the Third International Conference on Reliability, Quality and Safety of Software-Intensive Systems (ENCRESS’97), Chapman & Hall, 1997.

[13] Z. Gyongyi, and H. Garcia-Molina. "Web spam taxonomy." First international workshop on adversarial information retrieval on the Web (AIRWeb). 2005.