Two Visions of the Web: from Globality to Localities

Fabien PFAENDER Mathieu JACOMY Guilhem FOUETILLOU

COSTECH Laboratory
University of Technology of Compiegne
France

fabien.pfaender@utc.fr mathieu.jacomy@gmail.com guilhem.fouetillou@utc.fr

Abstract

This paper presents two different perspectives of the web : a global one that corresponds to the classical approach of search engines and a the local one that we propose as an alternative approach. The search engines perform their indexation operation on the whole web in an automatic way and display their results according to it by proposing a perfectible visualization. We will review the usability of these visualizations while examining the way search engines build their hierarchies. That leads us to reconsider the notion of context and the way models of the web influence our vision of it to finally propose a new model strongly related to its perception through alternative visualizations.

Keywords— *Webming, web structure, information visualization, usability*

1. Introduction

We propose to discuss the approach of the web by search engines considering the opportunity to build maps of the web into observing its shapes. Both approaches are based on a common understanding of the web as an open, heterogeneous and large-scaled network that can be worthily projected as a graph.

We will first quickly present the way algorithms of search engines utilize known shapes of the web, and the way they represent it with its limits. Then, we will stand up for the idea that the notion of locality is essential even to establish a hierarchy of resources. Finally, we will get onto the notion of context on the web considering the major role of localities and cartographic visualizations with an experimental map.

2. Apprehending the shapes of the web : the case of search engines

This part confronts the search engines, which ranking produce lists, and the possible shapes of the web that maps can visualize. Two complementary aspects of the web as a graph are discussed : algorithms and visualizations.

2.1. Principles of web graphs calculation by search engines

Initially the observable shapes of the web allowed to conceive the now classical algorithms on which is based a large part of the search engine technology. However, the evolution of this technology was provided by the observation and the analysis of the behavior of internet users. Knowing that the projection as a graph is a foundation of classical approach of the web, we point divergences between exploiting the shapes of the web and producing empirical hierarchies.

2.1.1. The first ranking algorithms : measure the support of web navigation

First ranking algorithms, aiming to return the most pertinent resources, are developed into exploiting the properties of the web as a graph. The graphs can be easily handled as matrix, and search engines make the most of this way of coding data[1]. Kleinberg's HITS algorithm exploits the bipartite structures of web graphs to build a measure of authority. The measure of authority is a measure of pertinence only if bipartite structures are global properties of the web. For Kleinberg "the creator of page p, by including a link to page q, has in some measure conferred authority on q". The success of Kleinberg's algorithm and his derived give him reason. However, the social aspect of the hypertext link is not completely analyzed today, and that's why it is difficult to evaluate an algorithm thanks to sociological arguments. On the other hand, it is possible to observe the web into bringing out invariant properties, like bipartite properties, and building appropriate algorithms while analyzing these invariant properties in the field of social sciences.

In 1998 too, Sergey Brin and Larry Page publish the paper that describes their search engine Google[2]. Google uses the PageRank algorithm to build a hierarchy of resources. Once again, the intuitive justification proposed by authors is based on the forms of the web. We insist on the fact that besides argumentation, the algorithm is not conceptualized by the internet users, but by the shapes of the web supporting the navigation. Today, the social aspect of the web is exploited without theory, and like in 1998, social tools appear on the web thanks to the only intuition. The algorithms of search engines are based on shapes of the web that are possibly social, and not on the analysis of internet practice.

2.1.2. The search engines released themselves from the shapes of the web

The massive use of search engines by people at large changed the situation. For a search engine, the efficiency of results returning is the main issue : algorithms don't need to refer to shapes of the web. Of course, the model of random surfer is widely used. But search engines, to be the most efficient possible, use different optimizations that make algorithms difficult

to interpret. The proposals to improve the PageRank are mainly mathematical, and the properties of the web intuited are secondary [3]. The Topic-sensitive PageRank[4] classifies the web thanks to the Open-Directory portal : the structure of the web certainly doesn't conform to such directories, and so this algorithm doesn't refer to invariant properties of the web.

The algorithms that rank resources are legitimated by their efficiency. The empirical worth of these technologies, that have the success we know, obliges us to ask the following question : Is there a gap between the hierarchies of search engines and the observable shapes of the web ? To answer this, we will now address the issue of the visualization of the shapes of the web.

2.2. representing the web as a graph : manifesting shapes

The web models have been build as extensions of the graphs theory. The use of such theoretical tools has consequences on the vision or the grasping of the web we can get today through the various possible interfaces that are the web browser, the search engines and their list or the synoptic devices like maps.

2.2.1.graph is the core of the web

The projection of the web structure into a graph is considered as natural. The graph presentation and its use to modelize and spatialize the web correspond with the common perceptive and cognitive experience of the web browser. The browser displays web pages and allows to follow hypertextual links that lead to other pages. Thereby, the web pages appear to be linked one to another in a causal manner. The forward and back buttons are a basic way to move into what becomes navigation space[5][6]. The natural shape of this space made of web pages and causal links between them is a graph. Pages in it are nodes and links are edges. This allows the graph space to be mathematically representable. One can then apply all algorithms and properties dedicated to graph computations. However it is also possible to spatialize it, i.e. to present it graphically[7] although this mode of presentation is still largely uncommon to people at large. This ambiguity between logical structure and spatial structure leads to two different uses of graphs.

2.2.2.two uses of graphs

According to the objective, one can use the graph either in an analytical way or in a synoptic way. The analytical use of the graph take advantage of mathematical properties to automatically reduce its complexity following known properties. This method is automatizable as it consists only in formal calculations. That is the reason why this solution has been adopted by the search engines to provide fast answers to any kind of questions while covering the largest web area possible. But the time argument is not the only reason search engines favored this method. Il is also because that corresponds to a global model of the web on which a calculation can be applied. The synoptic use of graph is used at two ends. The first is a use similar to the analytical method that is to equip a reduction of complexity. Unlike an automatic calculated analysis, the synoptic use of graphs helps a human user to detect visual regularities and salience facts which would not be perceive otherwise[8]. The reduction of complexity is thus visual. In the same way this method is employed thereafter to display in the graph visualization the result of this reduction of complexity. Visualization is use both as an exploration tool and a final representation for user. This procedure or method is essentially manual and requires a human user. Even if it uses some automatic calculations, parameters control and sequence of algorithms are always human managed. The model of the web in that method is not predetermined. On the contrary it lets new interpretations arise according to the location of the web that is studied.

2.2.3.From graph to list and maps

The double use of the graphs leads in direct consequence to two different spatializations of the resources. First we will consider the case of the graph mathematically represented and its the use by the search engines, i.e. to propose the most relevant resources for a given request. Since the analysis is analytical (resources, properties and algorithms are used one after another) and that the model which was used for the analysis is total, the spatialization which results from it retains these two characteristics even if it is not systematic. The construction of a visualization is only the consequence of the concepts which governed its development. That is, a relevance analysis applied on the whole web, result of a graph calculation. The most obvious visualization in this case is the list of results. It is a logic structure that is drawing on the plan but does not call upon metric properties to make sense. There is no principle of distance between the elements. The only important thing is the spatial unity in a column or vertical line which makes it possible to see the list in the middle of other visualizations. To reinforce the visual impression of unity and logic design, the list elements have are of the same uniform size (figure 1). Thus no element of the list has

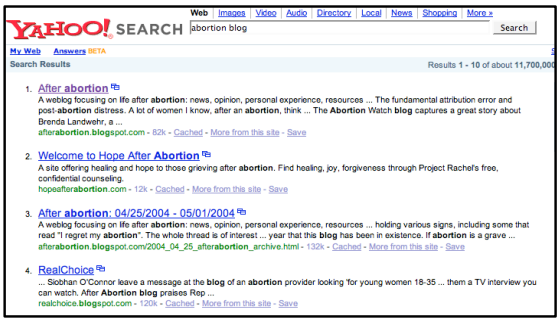

Figure 1 Yahoo ! results page for request *abortion blog* displaying 5 first results of 11,500,000

more importance than another, they are equivalent,

visually and also logically speaking. This principle is significant since without it, the list would lose its logical character and would become cartographic like a tag cloud visualization[14]. Moreover, the apparent order of the results classified from more to least relevant results only comes from the occidental practice of reading top downwards. Nothing in the list itself makes it possible to suppose such a classification. Yahoo thus added a number in front of each resource to make this classification (figure 1) visible whereas it is only an hypothesis in the majority of the other search engines. The fact that all the results share the same appearance and an equal distance between them without any metric is in agreement with the idea of globality and analytic conveyed during the analysis. The map on the contrary is based on the 2 dimensional placement in x and y. The logical order is not applicable. It is the disposition of the logical elements on the map in a particular place in a measurable space that makes sense. The principle here is to show individualize each and every elements the ones compared to the others by their position. The salient perceived elements become the sources of a questioning[9] which is not constrained by a reading direction. Cultural habit and knowledges can influence the reading but it is a minor effect. The resources separate themselves by their differences of position. This spatialization much freer for the reading and interpretation is in conformity with a local vision of the web while letting the singularities of this space appear. It also corresponds to an evolution of the perceptive use of the graphs by equipping it with a metric adapted to the properties which one wishes to observe. The maps are also much less interactive than the graphs to respect the synoptic criterion which must be stable for the reading. A graph is calculated when displayed while a map is drawn. Maps are the final step of a visual process of exploration and gains in globality and stability what it loses in flexibility and freedom of interpretation[13]. It is designed for helping an end-users public to receive a navigation help, limiting their confusion at the same time. 65000

3. Tensions between notions of locality and hierarchy during the observation of the shapes of the web

We will here defend the idea that the hierarchies of the search engines depend on the web localities, although the list of results does not show this property. The localities influence the ranking algorithms themselves and leads The hierarchies of the web to be re-conceptualized.

3.1. A hierarchy is at the same time a measure and a fundamental way to return information

From a technical point of view, a fundamental way to treat a large amount of data is to get a measure, that is to enrich data with scores. Sorting data function of a measure is sufficient to build a hierarchy, and that's why organizing information in a hierarchy is, and will be, a necessity to return data. Nevertheless, there is an important difference between the use of hierarchies by an expert of the web and by an internet user. The expert can compare several measures of the web into analyzing it, and in this case he interests in the whole set of resources : the statement of a measure is the important aspect of the hierarchy. On the contrary, people at large use only one measure (like pertinence) and only interest in top-rated resources. It is the case of search engines and of all the "top ten" that we find in rich-content sites (for example : best sellers, most viewed, latest news...). In that case, the visualization as a list is the most important aspect of the hierarchy.

3.2. Hierarchies of the connectivity depend on subgraphs

With only a part of a graph, it is not possible to know if the nodes that are strongly connected locally are also strongly connected globally. It's easy to understand. We take a non-oriented graph G, and look at nodes' degree : the degree of the node $n \subset G$ written down as $d(n)^G$ is the count of edges of the node n in the graph G. We consider now the graph G and its subgraph $G1$, and the following logical sequence :

$$\forall n, n' \subset G1\ (d(n)^{G1} \geq d(n')^{G1}) \Rightarrow (d(n)^G \geq d(n')^G))$$

It exposes that if a node is strongly connected in the subgraph, then it is also strongly connected in the whole graph. In other ways, organizing the subgraph $G1$ in a hierarchy organizes the respective part of the graph G in a hierarchy. When a hierarchy is calculated on a subgraph of the web, it is expected that this hierarchy applies also to the nodes in the whole web. Unfortunately, the logical sequence above is not true in general, because it depends on G and $G1$. With a simple graph, for example a small crawl, it is possible to determine localities with representative

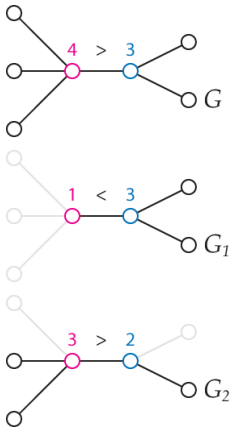

Figure 2 Subgraph G1 has not the same hierarchy as G, while subgraph G2 has the same.

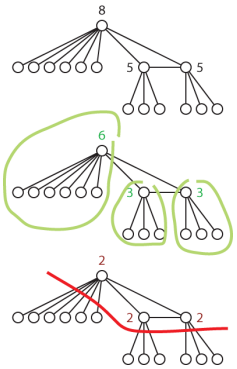

Figure 3The first hierarchy represents the whole graph, while the second doesn't. It is sometimes possible to determine subgraphs with representative hierarchies.

hierarchies, though. The recent developments of the theory of scale-free networks demonstrate that with a scale-free tree, the branches (or subtrees) verify this property[10].

3.3. The ranking strategy of search engines is empirical

According to the theory of scale-free networks, the only networks that have a global hierarchy of connectivity are those that are close to a scale-free tree. In other ways, in some subgraph of the web, the local connectivity isn't representative to the global connectivity[10]. So, the algorithms that calculate the web as a scale-free network can be wrong. In particular, the belief according to which the local and the global are similar on the web thanks to "motifs"[11] is wrong. L. Li et al. use the $s(g)$ function that "measures the extent to which the graph g has a 'hub-like' core and is maximized when high-degree nodes are connected to other high-degree nodes" into evaluating the self-similarity of a graph :

the Internet has extremely low $s(g)$ and thus minimally self-similar at the motif level.

The algorithms of search engines calculate hierarchies on a part of the web, even if this part is very large. Nothing tells us that these scores correspond to a global law (because the whole web isn't indexed) or a local law (because the index is larger than localities), due to the likely self-dissimilarity of the web. Thus it is not possible to claim that the algorithms of search engines like PageRank justify themselves by topological properties of the web. Their indexation and ranking strategy is mainly empirical and aims to satisfy web users.

3.4. On the Web, the hierarchies of connectivity depend of localities

As there is no frame that allows evaluating the global value of a measure in a subgraph of the web, we search local measures allowing a local validation. But the important differences observed between different subgraphs of the web, extracted by the same way but from different resources, reinforce the idea there is no generic measure to organize web pages in a hierarchy. Nevertheless, some local hierarchies can emerge from the web. In the French domain of the "culture of sciences, techniques and industry" (CSTI), the four national institutions have a central role on the web as in reality. Not only their web sites have rich contents, but also they are more generic than the other sites, and they have the best scores of authority in the corpus of resources manually selected. Our experimentations show that sometimes several measures come together so as to delimit a domain. So we retrieve the typical case of the theory of aggregates[12] where a topic and the connectivity come together, although the differences are so important between all observed cases that we prefer to speak of "locality" better than "aggregate". These local hierarchies are not returned by search engines, most of the time because there's no simple query to define a domain. In our example, the major institutions don't have a "CSTI" label, so that they do not appear when we ask search engines using "CSTI" or "Culture Scientifique Technique Industrielle". Generally different terms describe a domain in its generality and its specialties, although resources are strongly connected on the web and cover the same topic.

Topical localities compel their own hierarchies, which are incompatible with a global hierarchy. The localities sets of themes impose their own hierarchies, which are incompatible with a global hierarchy of the Web. In terms of resources relevance that manifest itself through the importance of the context in the search for information. We now will return to this generic question to show its importance in the search for information and to outline the design of the Web which requires interface to be readable by net surfers.

4. Context

Any search for information of quality requires to be able to check the relevance of the resources mobilized. This work usually goes without saying with the knowledge of the context of information. If it is relatively easy to find the context of the resources in the traditional media, it is much more difficult on the Web. The reasons are deep and starts with with the definition of what is the context of a resource on the Web.

4.1. documentary vision and thematic vision

There are two possible visions of the question of the context on the Web which corresponds to two levels of granularities : a local one and a general one. The local level is the level of the document itself. In the case of search engines it is usually a web page. This page exists in a space in relation to other pages and shares with them hypertextuals links. The problem is to find where the document stops because one can follow the links indefinitely and thus extends the document itself. The question of the document bounds is crucial to make a judgment and build a thought[15]. Let's consider the example of the web site `www.abortionfacts.com` (figure 4) : At first sight this site is considered as a neutral site regarding the abortion but if one considers its links (especially outgoing) it becomes a prolife site. These two interpretations are not compatible and the way the reader will read the pages will be totally different. One way of addressing this issue is to perform a topological and semantic analysis of web pages. If a page is too far topologically and thematically then one can consider that we reaches a bound of the document. By repeating this procedure in every direction it is possible to enclose the document and make its analysis. However this local problem of the document is found on the global level with the sets of documents. The context becomes an overview of the topologic and semantic organization of all documents relative to a particular theme. It is significant to know in what is the neighbors of a document to include/understand the

initial intentions of the author in an hermeneutic analysis. But this work is nearly impossible on the web is impossible taking into account its plasticity and dynamicity. Consequently the context becomes a very fuzzy concept of both semantic and topological area which delimits a locality organized by actors/authors. These two definitions echo back to two contributions of knowledge of the context on the Web and argue in favor of local model of the web to respect its local structure of meaning organize in sets of documents and group of actors[16]

4.2. Contribution of the knowledge of the context on the Web

Not only contributing to navigation[6], the context is especially significant to help the reader determining the relevance of a resource or a set of resources. Is the document a hub or an authority, a bridge between two communities, what are its neighbors and friends, etc. (see section4)4This is quite the same on the global level. In order to make sense with a set of resources, one has to be able to deduce the remarkable facts from their organization. What interests us then is not so much to have precise results but to know what is emerging or remarkable in the theme or the whole of sites/pages suggested. To find only one prochoice site against 10 prolife sites brings the reader to reconsider the importance of the prochoice web site. To make these saliences perceptible will allow the reader or sailer to take into account the context and thus increased the relevance and the effectiveness of his navigation.

4.3. The list and its alternatives

Usually the search engines provide their results in the shape of list. The list is a graphical structure which presents on the same space a whole of resources as seen in section 2.2.3However a graphical display in one same space is equivalent to set a context *de facto*. The Co-present resources in this space makes a set of the resources the search engine considered to be most relevant. Indeed, if search engines were sure to be able to answer the request in a relevant way, they would only propose a couple of results and not a full list. They would not specify either that the engine found thousands of results corresponding to the request. Instead of that, it is the reader that have to find what interests him in the set of results among a very large number of pages always displayed in the result page. His task is to make sailiences emerged from the list and then to contextualize the results. It is far from being easy insofar as lists are not conceived for this purpose. their analytical reading makes it impossible to see how the pages are articulated. The display of a category for each link in the list of results is a step in this direction but the difficulty to mentally build an overall picture reduce considerably the effectiveness of the context making. Another problem is these lists is the page by page display. The results are presented 10 by 10 and thus form different page sets instead of one ordered list whereas it is what search engines actually return.

An other way of displaying a set of links is to design an appropriate synoptic view. Graph are one way to do it and maps are another. They give an overall picture of the chosen theme if they are well designed. That is a display where the resources are easily recognized and located ones compared to the others. A links heap is less useful than a list. On the opposite well design maps or graphs that clearly distinguish resources and provide information on the web structure is much better than a list to give immediate context. The relevance criterion can still be added to the map in many ways. In figure4, the most relevant site according to google are displayed in larger square than the others. Moreover this visualization has been designed according to strong design principles [9] in order to be as much effective as possible in the contextualization.

4.4. Experimental map and contexualisation

Figure 4 Collection of 40 sites containing the 50 pages returned by Google to the query "abortion", 28 february 2006

To illustrate the fundamental differences between list and map developed in paper and the importance of context, we have decided to reconstruct the graph of the websites returned by Google using a particular query. The topic we choose for our query is "abortion". With such a topic, search engines return websites with different points of view. We limited the corpus to the 50 first web pages returned by Google, which we categorize into three groups : top3, top10, top50. This categorization is due to the internet users attitude with the results returned by search engines : internet users seldom click on the results after the three best rated websites and much more rarely go past the first page of the ten first results. The 5 emerging categories can be seen on the map. To create it we used a crawler able to reconstitute the hypertext structure of a set of web pages with a focus on our theme to avoid obviously non relevant website. Each Node is placed close to its neighbors that is nodes sharing links with it. A remarkable fact is that the synoptic characteristic of this visualization allows to immediately perceive web organization of the topic and its actors that would not arise otherwise.

Lists are not designed to help users contextualize but they have to do it anyway precisely because of the spatiality of the list. This visualization comes from a global model of the web but it is not assumed to the end. Users need to get an idea of the organization of the web hidden behind their results to make some sense. This means both a local model of the web that reveals this particular structure and a visualization designed to display the structure according to the model.

5. conclusion

Search engines ignore in part the shapes of the web. Topical localities, also called "aggregates" are such cast-off shapes. Not only search engines refuse to consider localities, but localities partially belie the existence of global hierarchies on the web (like PageRank). The localities issue shows search engines' difficulties to return the context, at the algorithm level as at the visualization level. We bear that the principle of global hierarchy justifies itself by common internet practice. We think that the search engine technology cannot address an issue to context problem because its proper principle is to clear the localities to consider each resource or local organization equally. People at large need a global access to the web, and search engines provide it but the localities play a role in common internet practice especially navigation and context making that is choose the resources that fit their needs. Because of the self-dissimilarity of the web, the global and the local level do not match and contextual issues have to be left aside by search engines. We think that the web needs tools to provide its localities as a context. Having regard for this discussion, we claim that these tools cannot just derive from search engines. They have on the contrary to be based on a different model of the web, in design as in practical application. Such tools should permit web users to be less disoriented during navigation by having a guide that allow him to build its own context through localities map and we think that it can be the role of digital libraries in a close future.

6. References

[1] J. M. Kleinberg, *"Authoritative sources in a hyperlinked environment,"* Journal of the ACM, Volume 46 Number 5, p.604-632, 1999.

[2] S. Brin, L. Page, *"The anatomy of a large-scale hypertextual Web search engine,"* In Ashman and Thistlewaite [2], p. 107-117. Brisbane, Australia, 1998.

[3] T. H. Haveliwala, S. D. Kamvar, G. Jeh, *"An Analytical Comparison of Approaches to Personalizing PageRank,"* Preprint, June, 2003.

[4] T. H. Haveliwala, *"Topic-sensitive PageRank,"* in Proceedings of the Eleventh International World Wide Web Conference, 2002.

[5] F. Ghitalla, C. Lenay, *"Les Territoires de l'Information : Navigation et Construction des Espaces de Compréhension sur le Web,"* in *La Navigation, Les Cahiers du Numérique*, Paris : Hermès Editions, 2003.

[6] Dillon, Vaughan, *"It's the journey and the destination : Shape and the emergent property of genre in digital documents,"* in *New Review of Multimedia and Hypermedia*, 3, p. 91-106, 1997.

[7] M. Dodge, R. Kitchin, *"Atlas of Cyberspace,"* England : Addison Wesley, Harlow, 2001.

[8] S.K. Card, J.D. Mackinlay, B. Shneiderman, *"Information visualization,"* in *Readings in Information Visualization : Using Vision to Think,* San Francisco, CA : Morgan Kaufmann, pp. 1-34, 1999.

[9] J. Bertin, *"Semiology of Graphics : Diagrams, Networks, Maps,"* Madison/Wisconsin : University of Wisconsin Press, 1983.

[10] L. Li, D. Alderson, R. Tanaka, J. C. Doyle, W. Willinger, *"Towards a theory of Scale-Free Graphs : Definition, Properties, and Implications,"* (Extended Version), arXiv :cond-mat/0501169, 2005.

[11] R. Milo and al., *"Network motifs : simple buildingblocks of complex networks,"* Science 298, p.824-827, 2002.

[12] R. Botafogo, B. Shneiderman, *"Identifying aggregates in hypertext structures,"* in Proceedings of *Hypertext'91*, p. 63-74, New York :ACM, December 1991.

[13] F. Pfaënder, M. Jacomy, *"Explorer et appréhender l'internet,"* in *13èmes journées de Rochebrune : Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels,* ENST 2006 S001, 2006.

[14] T. Hammond, T. Hannay, B. Lund, J. Scott, *"Social Bookmarking Tools (I) A General Review,"* D-Lib Magazine, Volume 11 Number 4, April 2005.

[15] B. Bachimont, *"Arts et Sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle,"* Mémoire d'Habilitation à Diriger les Recherches, Université de Technologie de Compiègne, 2004.

[16] F. Ghitalla, A. Le Berre, M. Renault,*"Des documents, des liens et des acteurs,"* in proceedings of H2PTM conference, 2005.