Proceedings of EPNACS 2011 within ECCS'11
Emergent Properties in Natural and Artificial Complex Systems
Vienna, Austria - September 15, 2011

EXPLORATORY ANALYSIS OF WEB DATA: METHODS, TOOLS AND GEOGRAPHICAL DISTRIBUTION

Fabien Pfaender, Nicolas Esposito*, Mathieu Jacomy[†]

Abstract. We propose a methodology for the exploration of Web data, built on the principles of exploratory data analysis and analytical visualisation of information. Our approach aims at combining these two approaches in order to benefit from both of them. This allows us to explore heterogeneous complex dynamic systems such as the Web, and to construct emergent structures and indicators without getting lost. By studying the geographical dimension for a specific Web locality, which is exemplary in many ways, we were able to test our methodology and various visualisation tools, thus validating our theoretical proposals.

Keywords. Exploratory Data Analysis, Web, Geography, Information Visualisation, Methodology

1 Introduction

At a time when we are faced with systems which are becoming ever more complex, dynamic and heterogeneous, we propose a method for engineering the emergence of structures and indicators in complex systems. This engineering comprises a set of original methods and tools which are able to extract the relevant properties. This should enable us to explore any system whatsoever; the Web, which is the object of the first part of this paper, is a prototypical example. As developed in sub section 2.1, constituents of this complex system are not all discovered and known and their interactions are subject to a lot of analysis without leading to an end even if some emergent properties seems to emerge[10]. At the heart of the Web, the geographical dimension is in full boom, and contributes a means for users to anchor themselves in reality while increasing the possible uses. However it is not simple to mobilize the spatial register on the Web, especially when this dimension has not been originally conceived for it. It is then necessary to give birth to meaning, by playing on whatever geographical references lie to hand (geolocalisation of resources, positions of actors, spatial references in the content, etc.); this is exemplary of a heterogeneity which the method we propose makes it possible to deal with, both theoretically and practically, as we shall see in sections 3 and 4.

2 Exploring the Web

The figures that are used to characterize the Web are enough to make one dizzy. It is a question of tens of thousands of millions of resources that are indexed and accessible by research engines, and even this is only the visible part of the iceberg [5]. A short time ago, the total amount of information contained on the Web was estimated to be 17 times the size of the Congress Library [14]; but the figures so defy the imagination that even this analogy is difficult to grasp. Moreover, these figures are based on a conception of the Web as mainly occupied by static documents, which are relatively simple to capture and to analyse. However, what was still the case five years ago is no longer valid. Thus, we have passed from a documentary model to a model of dynamic resources with a wide variety of formats (social networks, sites which regroup contents, micro-publications, etc.) to which one has access according to diverse modalities , and which render the task of systematic exploration even more arduous.

Faced with this mountain of information, which by its size, by its form, by the means which are necessary to measure up to it has the stature of an Everest, any attempt to take the Web as an object of study leaves one with the choice of two attitudes. The first attitude consists of considering the Web system in its totality. In doing this, one equips oneself with a panel of general indicators which require an enormous computational power, and which even so will have difficulty being locally precise [16]. This is the case in particular for those research engines which have to evaluate and to classify the relevance of any resource whatsoever for a given user request. All the potential resources have to be considered as comparable, and the calculations thus use an identical algorithm of classification for each of them; this is the case for example with the Pagerank of Google [9] . By doing this, one transcends local organisations for the benefit of a destructuring universality, somewhat like a satellite view of the Web which reveals the general large-scale principles.

*Fabien Pfaender and Nicolas Esposito are with Department of Cognition Research and Enactive Design, Compigne University of Technology, France. E-mails: fabien.pfaender@utc.fr, nicolas.esposito@utc.fr

[†]Mathieu Jacomy is with Department médialab, SciencesPo, France. E-mail: mathieu.jacomy@gmail.com

Another vision of the Web consists of viewing it as a very large structure composed of smaller localities, noteworthy organisations of resources belonging to different registers. Each locality is organized in a specific way and emergent properties varies from one locality to another. In adopting this vision, one seeks not so much to elucidate general indicators, that can be true for the entire system but too vague to be truly useful, but rather to formulate hypotheses and discover properties concerning each of the sub-spaces encountered. This model gains in complexity (more properties and more interactions to explain that are heterogenous between subspaces) what it loses in universality (as the general model is abandoned); it has the advantage of not distorting its components, making it possible to provide a fine analysis of a locality or a group of localities [11]. We will refer to this below as the "second approach", and it is the one which we will develop in this article.

2.1 Structures with multiple dimensions

Whatever posture be adopted, the reality of the Web translates into a multitude of resources on several intertwined levels. What at the start is only a Web resource, i.e. an object accessible via the HTTP protocol with the help of a URL address, can thus be described on five different levels which, once combined, form the localities that will be mentioned below. These levels are not complex properties in themselves, they are primary categories in which complex properties can be classified.

We first distinguish the level of the resource. Here one finds what the server sends back when one calls a particular URL. The content can take various forms, the most common being text, images or audiovisual documents. An HTML resource (a hypertext page) commonly possesses a structure composed of several linked resources (scripts, style pages, media, etc.) each of which, taken one by one, constitutes a full resource in its own right. The second level is that of the link. This level is naturally very present on Internet, but it is important to distinguish intra-level and inter-level links. The former consists of links between elements of the same level, such as hypertext links between Web resources, links of relationship or friendship on the social level, or yet again temporal links of succession or simultaneity on the temporal plane. By contrast, inter-level links connect together two different levels: for example a link can connect the author of a Web-page to the page in question, just as a link can connect a resource to a geographical site in a number of different ways that we will examine below. These links are very often the source of inaccuracies, since they associate two elements which are heterogeneous by nature. We also will return to this aspect in the case of geo-localisation. The social level is the third level of description that we can invoke. This level describes the human individuals who gain access or who publish the content and the relations they hold between them. This level is very much in fashion nowadays when social networks flourish on the Web; not content with putting actors in touch, this also makes it possible to describe their relations with an ever-increasing richness of detail. Next, the geographical level integrates a territorial component. If the graph of links constructs a multi-dimensional space without any physical referent, the geographical reference is more and more used on the Web. This makes it possible to answer questions such as the physical location of resources, the routes by which information passes, the positions of authors and readers, or yet again the places mentioned in the content itself. Finally, the temporal level describes the evolution of resources, of links, of geography or yet again of actors over time. In a system where the dynamics of resources and links is so strong, there is a lot at stake in integrating this temporal dimension in order to study the system. However, here we come up against technical limits which are still very strong, in particular because of the great mass of data which prohibits taking regular snapshots.

In order to study and explore the Web, we must be able to combine all of these levels, in order to make sense of what is going on. The resources are inscribed in these different levels which all together compose a system which is complex both by its size and by its structure. And even if the number of combinations is so great that it is not conceivable to completely integrate all these levels, we can nevertheless seek to grasp some organisational features in order to highlight the localities we mentioned above as being at the heart of the second approach we adopt here. Of course the crucial question is the method to be employed to reach this objective; but before revealing this method and putting it into practice, we must first specify more completely on one hand the importance of the geographical aspect which will be at the centre of our preoccupations, and on the other hand the choice of the supposed locality that we will analyse on an experimental basis.

2.2 Spatiality in full development

Among the different levels evoked in the vast variety of possibilities available to describe a Web resource, it is difficult to ignore the rise in strength of the level of geography: from participatory cartographies to contents that are geo-tagged , by way of open data that refer to towns and the localisations of resources or actors [19], the geographical reference has become unavoidable. This phenomenon is due both to technical advances which have made geolocalisation easier, and by the consequent fleshing out of the services provided . But it is important to keep in mind that this progress answers a demand which was not hitherto satisfied for a strong spatial anchoring which would bring the virtual reality of the Web back to a more concrete reality. On the Web, spatial information benefits from the natural cultural familiarity that users have with it, and which renders it special. In particular this helps to fight against the disagreeable or painful disorientation that threatens when surfing on the Web [7]; it helps to recover an orientation, to get ones bearings. It is true that the culture of the Web bathes in the idea that the hypertext link always prevails, because it was present right from the conception of the network and largely explains its success. There is thus a tendency to consider the hypertext link as the unique structuring principle of the network; but this is not necessarily the case. Indeed, even if the geographical structure starts out as a mere addition, a supplementary indicator that is not essential, it has the potential to transcend the hypertext link. Geographical information thus becomes a way of structuring this space, and of explaining it [17]; the ideal being to couple together as many dimensions as possible, and in particular hypertext and geography.

Moreover, integrating the geographical dimension is particularly interesting for testing methods of analysing the Web, because it brings into play a great many different aspects. Thus, one can investigate the localisation of the material which makes up the contents, or of the persons in charge of these contents, as well as elucidating the geographical references in the contents themselves. The richness and diversity of the means of analysis renders the integration of this dimension perfect for a heterogeneous methodological study where one wishes to reveal the structure of several different levels present in the same set of data.

2.3 Choice of a set of data

In order to test our second posture and the methodology which will allow us to reveal a certain sort of organisation, we have chosen a particular set of data in order to limit the problems of capture and to focus on the exploration. The question of capture on the Web is a whole problem in itself, once again because of its size and its heterogeneous and dynamic nature. While one may postulate the existence of distinct aggregates of data, it remains problematical to capture them while mastering all the different dimensions. In order to eliminate a part of the problem, the system we have chosen is the on-line shop for Apples iOS software programmes `http://itunes.apple.com/us/genre/ios/id36?mt=8`. One can find there software for iPhone, iPad and iPod touch. This is a system which represents a bounded portion of the Web with known limits, since it contains a list of software units which evolve with a dynamic that can be measured. Thus, for a very large number of pages (close to 500,000 unique addresses) one can extract a title, a description, user evaluations, a price, a classification by category; links to other similar software, links with Internet sites, etc. We thus have a relevant terrain, very broad, with information of all sorts, hypertext links towards the exterior of the Apple site, and internal links towards other software which make it possible to efficiently test an exploratory chain.

3 A hybrid methodology

The exploration of complex systems and the emergence of knowledge concerning them is a fundamental problem which is born from the confluence of two factors: on one hand, the abundance of data in a digital form which means that they can be computed and exploited using algorithms; on the other hand the wish to consider ever more factors in the study of phenomena, in a quest for exhaustive measures taken as a synonym for truth. It is a question of understanding systems as wholes, facing up to their complexity. Two major routes have been considered for studying complex systems in this way, and they echo two conceptions of weak emergence[4]: low-level first or micro-level first emergence and high-level or macro-level first emergence.

The first route is thus to consider a low-level first emergence, where the property of emergence is carried by the elements which make up the system and deducible from them. It is a question of putting the autonomous elements that one masters into interaction with each other, so as to observe their behaviour compared to that of the natural system that one wishes to analyse. However, this approach does not make it possible to discover new hypotheses on the basis of the real system. It is rather a question of simulating the real situation by trial and error, making hypotheses and testing them, which produces more hypotheses.

The second route, by contrast, considers a high-level first emergence where the emergence is hardly reducible to the components of the system. The major drawback is that in this case, it is necessary to discover and to reveal the structure by considering the system as a whole (or what one imagines may be its whole), in favour of a model which is certainly approximate, but functional. This allow to discover low level properties than can be simulate in a second time. The majority of studies which concern networks (be they the Web, co-citations in scientific articles, or social networks) [2] take this route, using visualisation by means of graphs as revealing notable patterns and as generating indicators which are specific to the network under observation. For this route there are very few tried, systematic methods. The EDA (Exploratory Data Analysis) is one of them, whereas a method with similar objectives appeared at the same time as the visualisation of information [20] under the name of Visual Analytics [21].

3.1 Augmentation of the EDA and combination with visual analyses

One of the rare existing methodologies, the EDA [24], thus allows for the instrumentally equipped discovery of new structures in a complex system, by means of a systematic recourse to the visualisation of information in all the phases of the exploration of an unknown complex system; this has the effect of maximizing intuitions during the formulation of hypotheses [1]. In other words, it is a question of providing the user with the means of grasping a mass of data and becoming familiar with it, in order to discover noteworthy patterns which can be the source of questioning and the elaboration of hypotheses concerning the system. The exploration is divided into precise subtasks in order to end up with a relevant, coherent model of the original system. We can summarize the succession of tasks as follows: visualize the whole; zoom and concentrate on a part; pay attention to particular details. This makes it possible to overcome the classical dichotomy between the analysis of macro-structures and the analysis of micro-interactions. Thus, the juggling between the aggregation and the disaggregation of the data, and the back-and-forth between the global and the local, makes it possible to detect and to follow the emergent patterns.

This process of generating hypotheses concerning the system by abduction finds a similar echo in Visual Analytics where the summary of tasks becomes: first view the general level; then zoom and filter; finally enter into detail as needed. This method, scientifically less rigorous because it contents itself with an overview of the data before filtering it in an analytical process, is nevertheless very useful to equip one-self with innovative visualisations which are lacking with EDA. To the extent that visualisations are at the heart of the process, it is important to have a maximum of choice in the conceptions or in the possibilities of interactions, that one will then adapt as needed. Moreover, visualisations of information are particularly well adapted to the heterogeneous, dynamic data that they make it possible to present together, whereas the EDA is difficult to use on non-homogeneous digital data. On the other hand the visualisations of information are generally ad hoc constructions, conceived with a unique aim which makes it possible to explore a particular aspect in a single interface; whereas ever since its origin the EDA pleads in favour of a pluralistic use of statistical visualisations [15] in order to cover a maximum of cases.

The EDA thus brings a degree of rigour to the process and the faculty of linking simple statistical visualisation; whereas the visualisation of information and its analytical branch contributes complex visualisations more adapted to exploration. There is thus an interest in combining the advantages of both these two domains without losing either rigour, or creativity.

3.2 A semiological component as an integral part of the analysis

Since the visualisations are at the basis of the approach, in order to be effective they must be framed by a theoretical scheme of graphical semiology. The choice of the visualisations, which are the principal link between the human and his object, is crucial. The EDA is based on existing semiologies to help it make this choice [22, 6]; but the latter are limited to informational graphics and are only really effective for systems that are already constituted. Thus they are of little help when faced with novel systems of which one knows little or nothing.

In order to solve these problems and to dispose of a basis as solid as possible in order to construct a process for the exploratory analysis of data, we have developed and used our own semiology based on experimental studies [18]. We have thus succeeded in showing that according to the structures used, the visualisation in question does not lead to the same reasoning and knowledge. According to their degrees of constraint, they can either offer a wide range of possibilities in perceptual terms and thus in terms of information and knowledge, or on the contrary restrict the range of possibilities in order to focus on a precise element of information. It is thus possible to combine different types of visualisation, as a function of their characteristics, according to the needs of different stages in the elaboration of indicators. We can distinguish two extreme cases: on one hand one wishes to have an open visualisation in order to formulate hypotheses concerning the global organisation and to maximize the insights; on the other hand, one may wish to perceive and to use the result concerning the organisation, of a hypothetical particular descriptor while minimizing fresh insights. The open visualisations which maximize the insights will be those endowed with structures that are as free as possible. Structures such as graphs or maps, for example, fill this role very well. A balance between global reading and local reading will provide insights of different sorts. If regularities are perceived in these visualisation and one finds points of perceptual anchoring whatever their nature, be they the orientation of links, groupings, spread-out displays, routes, incongruous patterns, etc. they become potential indicators that will have to be analysed in order to formulate a clear hypothesis. The choice of the visualisation and its mode of presentation correspond already to a generic and very imprecise way of formulating a hypothesis which amounts to a vague intuition. The

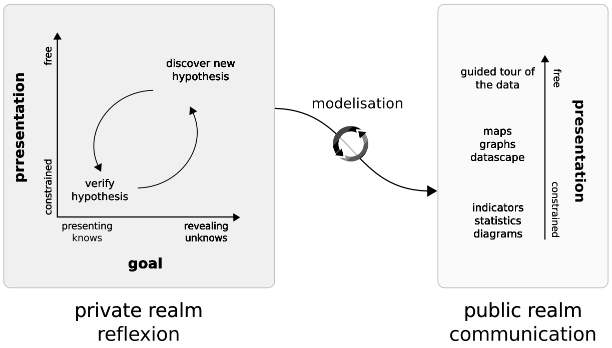

Figure 1: The EDA process adapted to very large sets of heterogeneous data. The exploration is presented on the left and is conducted by a single explorer. The presentations, free and constrained, result from a taxonomy of visualisation as a function of the degrees of freedom that they allow in their interpretation

formulation of a clear, precise hypothesis should have the aim of verifying this vague hypothesis by refining it and presenting it by means of the second sort of visualisation, the constrained visualisations. The latter are constructed from the basic data to which one applies filters or data processing algorithms, corresponding to a specific hypothesis formulated in advance which is spatialized in order gain knowledge of it. It is a question of transmitting a specific element of knowledge to the system, and the presentation must be perceptually constrained so that it is not possible to draw erroneous conclusions, but only an indication concerning the specific hypothesis that one wishes to test. The structure of a diagram in two dimensions is an example of a visualisation of this type, on condition that it respects the semiological canons of the genre.

We have been able to test these phases, and the proposed methodology as a whole, on the set of data Apple iOS Application Store US. This provided the opportunity to confront widely diverse tools and visualisations with a real system.

4 Heterogeneous tools for exploration

Now that we have laid down the theoretical basis, it is possible to begin the exploration we have chosen to carry out, by testing the numerous tools which are available free; these run from visualisations to data bases, including in between data-processing algorithms and other programs which are necessary for carrying out this study. There exist today a large number of means, which are made available to research scientists, journalists and others interested in the exploration of data [12, 25]; the drawback is that each uses its own format of exchange or posting. It is therefore necessary to jungle permanently with heterogeneous tools; the toolkit becomes a heterogeneous assembly of diverse scripts and it is necessary to continually pass from one to another. In this context, it is important to keep a dynamic connection between the intuitions one may have and their application, in order to conserve efficiency. Thus, latency resulting from concentrating on the technique leads us away from understanding the system and from an intimacy with the data, which is a key to success in exploration.

Before plunging into the exploration, it is necessary to carry out a capture of the set of data which is as correct as possible. This means, concretely, launching an organizing robot which will recuperate all the links to pages, and which will then take care of launching a robot to capture the content of each page. The problem here is to recuperate all the pages while adapting to their structures which often vary, and to do this in a minimum time in order to have a snapshot which is not too extended in time, and without being banished from the site under visit (if one generates too much traffic on a site, the latter can choose to no longer serve the pages). The clich of iOS which stores US Apple, slowing down the capture robots as little as possible, nevertheless takes seven complete days for near to 500 000 referenced pages and 277 997 unique applications. Between the beginning and the end of the operations, 15 000 applications have disappeared. Each page follows an identical pattern, with minimal variations which contain structural elements that are stored as one goes along in a free relational data-base (PostgreSQL) which will subsequently allow for easier extractions. Each capture and each insertion is visualized in the form of a very long list, where in the case of an error the cause is immediately apparent (change in colour and length of line) in order to permit visual surveillance. The crawler robots (robots that download a web resource to analyze its content and extract relevant information) are programmes developed in Java especially for the occasion in order to render them more precise, because specifically adapted to a particular site, but which are not re-usable on other portions of the Web. Once this work is completed (provisionally at least), it is time to pass to the exploration itself. To do this, we have recourse to the methodology which consists of getting to grips with the system with the help of free visualisations of the graph type, and to validate preliminary intuitions by constrained visualisations. In order to see the large picture, we have tried out two approaches. The first was to observe the distribution of applications in the various categories, to see which categories were the best represented and if there were substantial disparities. A histogram, resumed here as a sparkline[23], makes it possible to isolate two categories which are particularly well represented, games and books, with close to 40 000 applications each. We then find categories that are less well furnished, such as news, finance or weather, with an average of only 5 000 applications each. Concerned to see how the applications within each of these categories were organized, we opted for a second approach in parallel which consisted of viewing the whole set of applications in the form of a graph, considering the suggestions for purchase situated at the bottom of each application as a link towards another application. The nature of the link is a matter of discussion and cannot be completely trusted as this is a black box process provided by Apple. Our goal here is as much to reveal or at least have some hints about the suggestion algorithm as to describe the system structure in itself. Further work should include more links like similar authorship or hypertextual link between applications but outside of the Apple IOS store. An interactive visualisation of data of such size in the form of a graph was only possible using the free software Gephi [3], and only after waiting 40 minutes for displaying the latter.

4.1 Division in levels

This graph revealed two notable patterns. The first was a heart which was very dense, but relatively disconnected

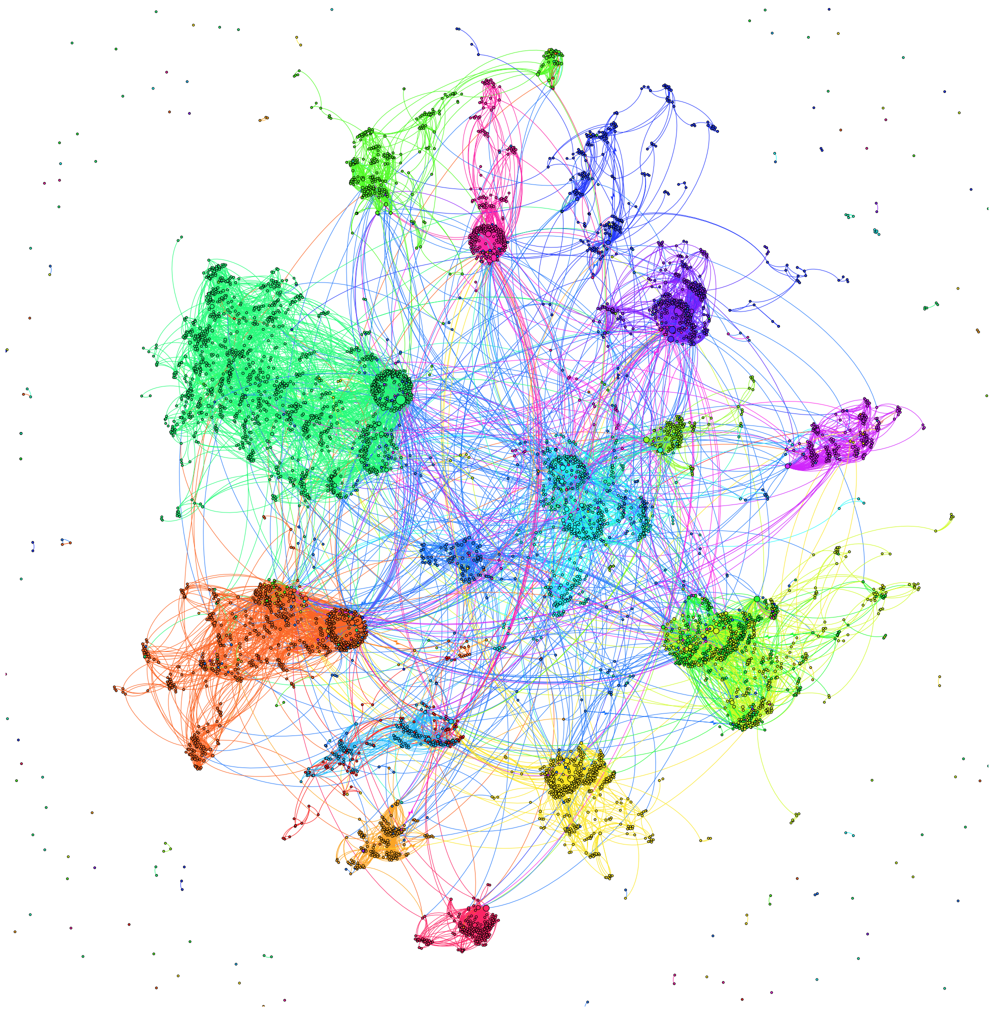

Figure 2: Graph of explorations of the heart. The size of the nodes depends on their degree (number of entering and leaving links), whereas the colour is associated with a category. It will be seen that the category indicators are relevant, since one finds localities on the graph that correspond to them.

from the rest of the graph. In fact this corresponded to applications destined for iPad, an Apple tablet for which applications can be dedicated. Their small number (10 000) leads to a large number of redundant links, which thus creates a sub-graph within the complete graph. In order to try and see more clearly what is happening, it was thus necessary to remove from the graph these specifically dedicated applications which obscure the overall exploration. This made it possible to discover an impressive quantity of applications which were not linked (too few purchases or not well displayed), others which were never visited but which had exit links, and finally a relatively small number which had both entry and exit links. This made it possible to visually exhibit three levels, which are found pretty systematically, in terms of thematic localities [13]; these are an exterior, a periphery, and a heart. We then confirmed this visual hypothesis by another visualisation, constrained this time in order to analyse the composition of the system with respect to these three levels. This figure was constructed with the visualisation library Web Protoviz [8], which allows a rapid and easy display of dynamically generated diagrams. This is useful for rapidly conceiving visualisations which respect elementary semiological principles and which concentrate on seeing rather than doing.

It turns out that the heart here is constituted of applications which are promoted by Apple, since we find many links towards them. This heart represents only 2.5% of the whole. A back-and-forth at the micro level for the on-line Apple shop reveals that a good number of these applications are promoted (or had been promoted at some previous time) on the home-page of the dedicated software iTunes. The visualisation revealed that the heart was composed of thematic localities corresponding to the categories, the most central one being the category utilities . Next, only about a quarter of the applications are on the exterior. It is interesting to note the criteria which led to their exclusion. The Figure 2 reveals an interesting pattern according to which one particular category, books, has a very large number of these excluded applications (more than 25,000). At issue here are books edited in the form of applications, which compete with general applications which are free and very successful. In a general way, we find on the exterior applications which are too specialised, which have therefore not attracted the crowds and which are thus lost in the mass. By the way, we may remark that actually they are not particularly less well noted than the others.

This characterisation of the system in three levels of different natures makes it possible to concentrate the analysis on a particular zone, as recommended by EDA when it comes to passing from the micro-level to the meso-level. Among the multitude of questions which arise at this stage of the analysis, we have chosen to concentrate on a connection between the social plane and the geographical plane.

4.2 Fragmented geographical information

Indeed, one of the questions at the origin of this study is to understand how the developers of applications are distributed geographically, in order to couple this information with a study on the social plane. The system is closed and the applications are not free, so that we cannot content ourselves only with the pages that have been captured. Matters would be easy if the system already included geographical information or information concerning the authors of the applications, but this is unfortunately not the case. One finds neither geo-localisations, nor the names of authors or societies, only a pseudonym which renders a research of the Yellow Pages type inoperative. We thus come up against the problem of geography on the Web, and methodologies making it possible to obtain information concerning localisation. In the absence of any data of this sort in our original system, we are forced to have recourse to alternative indicators to obtain information concerning geographical distribution. The first of these indicators is to systematically localise the Website of the application. To do this there are mainly two techniques which function differently: the Geoip and the Whois. The Geoip consists of finding the place which could correspond to the physical location of the server which sends back the Webpage of the application. For a Web address of the type www.utc.fr, we create a correspondence with an IP address (role of the Domain Name Server) which follows a route made of various network equipments around the world and from which one can induce a localisation. For this we used the free correspondence base of Geoip which is reliable at the level of the country (99.5%), less so at the level of the town (79% in the USA) (`www.maxmind.com/app/geolitecity`). Thanks to the latter we were able to obtain in 24 hours (the time the program takes to run) a localisation for each of the applications, on condition that the site was correctly informed on the Apple page of the application.

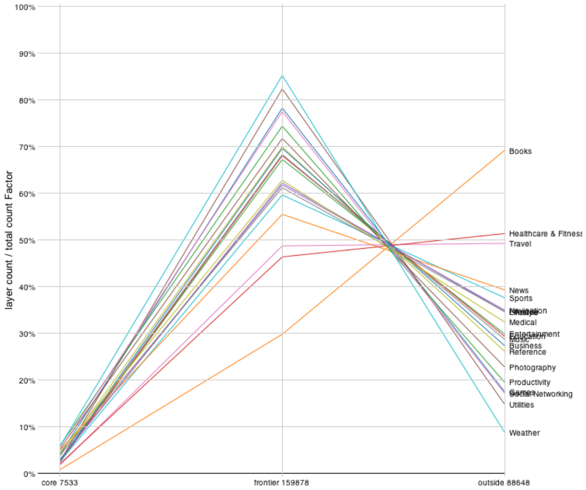

Figure 3: Distribution of applications in the heart, the border and the exterior, according to their category visualized in parallel co-ordinates.

We find in Figure 4 that the countries of the G8 and China are the most represented, and that the United States is far in front of the others. (They represent as much as the rest of the world combined although we should recall that we are on a US shop). The distribution also shows the domination of the East coast. There is very little material to compare with other web geoip datasets distribution. Nevertheless preliminary studies indicate that applications web sites are spread all over the world with a rather strong concentration in western Europe wereas applications are proposed in a US only store. It suggests that developers are concentrated in highly developed countries with a good engineering level and that we have to face an internationalization of developing applications in the analysis.

However, this visualisation must be taken with caution and must be confirmed, because there are many sources of inaccuracy: the first is that anyone at all can buy a name of a domain anywhere. We can count on the wish of the authors to choose a provider near to them or who speak the same language, but this is not at all certain. Next, some hosts of content optimize the requests by orienting towards a server close to the applicant with replicas all over the world. The simple fact of requesting the IP address of a site from a given site, as is the case here where one computer requested 277,000 IP addresses from the same place, is already enough to tamper with the results. Finally, the components of the sites (style page, pictures, etc.) can be hosted on several servers, and the Geoip only localises the master-page, and thus reveals only a part of the truth.

In order to face up to these limitations, it is possible to obtain the address of the person responsible for the Web

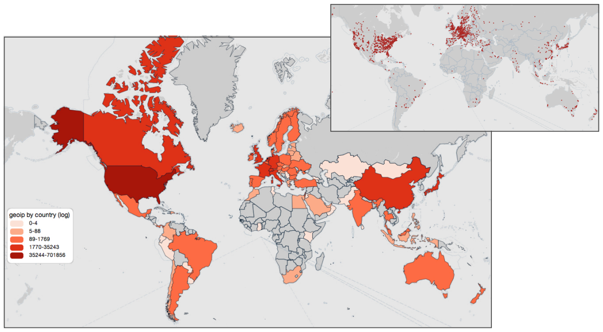

Figure 4: Chloropleth map of the geographical distribution of servers which host the Web sites of the applications (logarithmic shading scale). These visualisations are realised in SVG on an underlying map openstreetmap (www.openstreetmap.org) and generated dynamically by the javascript library polymaps (polymaps.org).

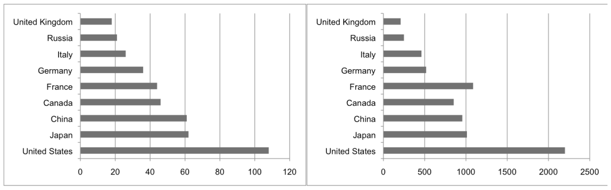

Figure 5: Number of citations of the names of G8 countries plus China in titles (on the left) and in the descriptions (on the right).

site by means of a service called Whois. This service provides information of a better quality, since it links a Web site to a real entity (a society, a person). For example, the applications of the game editor Gameloft reply via Geoip with a localisation in Canada, whereas Whois indicates that the seat of the enterprise is in France. On the other hand, nothing makes it possible to know directly whether the IP corresponds to a Canadian subsidiary of the enterprise. There remains the case where the provider of access puts up a screen with his own name in order to protect the information concerning his clients, which can happen. Moreover, whereas the Geoip base is available for free, a large number of Whois requests cannot be obtained for free. Thus for example it would cost about 800 euros to localize the 277,000 applications of our system. All the more so since after that it is necessary to convert the addresses into geographical locations with other services such as Google Maps, which also have limitations on the use that can be made of them. All this does not totally forbid the undertaking, but reduces the scope of our actions. A compromise thus consists of using the division in layers that we mentioned previously to focus on the 7000 largest applications, but we lose a lot of information. This study is still being carried out. Another approach to studying a geographical distribution was to note the various languages in which the applications are available, since this information can be related to the country in question. Of course since we are dealing here with the iOS Store US, English is used in 95% of cases. But among the other main languages used we find again the countries of the G8 together with China (in 4-6% of applications). This information was confirmed by analysing the text available in the applications themselves. Since the G8 countries turn up often, we looked to see how these countries are quoted in the applications, in the titles and descriptions. This makes it possible to measure the interest of the country in the system in general. Figure 5 shows the results of this distribution, which is very close to the distribution calculated in a sample of 5 million books (see `http://bit.ly/eulTDK`).

5 Conclusion

It is clear that we have not yet exploited all the information that could be extracted from this data-set, particularly concerning geographical information. The method has nevertheless allowed us to test the exploratory chain, and to explore a complex system quite effectively. Some emergent properties have been hypothetized and validated with visualizations like the heart, periphery and exterior. However, the use of the various softwares for visualisation and data-processing remains delicate when spatiality comes into play, because of the very great heterogeneity in format or in precision of the data that are available. Geographical properties seems to exists and the analysis will be pushed forward to let them emerge. This highlights how useful it would be to have a software capable of aggregating the various tools, in order to allow an exploratory analysis of data that would be less costly and less dependent on computer skills. As a result of the present study, our team has recently undertaken the development of software of this sort.

Acknowledgements

The research for this work was supported by the Scientific Council of the Compigne University of Technology.

References

[1] Andrienko, N., Andrienko, G., *Exploratory analysis of spatial and temporal data: a systematic approach*, Springer, 2006

[2] Barabasi, A. L., *Linked : The New Science of Networks*, Perseus Publishing, 2002

[3] Bastian M., Heymann S., Jacomy M., *Gephi: an open source software for exploring and manipulating networks*, International AAAI Conference on Weblogs and Social Media, 2009

[4] Bedau, M. A., *Weak Emergence.* In James Tomberlin, ed., Philosophical Perspectives: Mind, Causation, and World, vol. 11 (Blackwell Publishers), 1997, pp. 375-399.

[5] Bergman, M. K., *The Deep Web: Surfacing Hidden Value*, The Journal of Electronic Publishing, volume 7, issue 1, 2001

[6] Bertin, J.,*La smiologie graphique*, Mouton-Gauthier Villard, 1963, Paris

[7] Boullier, D., Ghitalla, F., Gkouskou-Giannakou, P., Le Douarin, L., Neau, A., *L'Outre-lecture,Manipuler, s'approprier, interprter le Web*, Centre Pompidou Library Editions, 2003, Paris

[8] Bostock, M., Heer, J., *Protovis: A Graphical Toolkit for Visualization*, IEEE Transactions on Visualization and Computer Graphics, 2009, p. 1121-1128

[9] Brin, S. and Page, L., *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Seventh International World-Wide Web Conference (WWW 1998), 1998, Brisbane, Australia.

[10] Cilliers, P., *Complexity and Postmodernism: Understanding complex systems.* Taylor & Francis, first edition, 1998

[11] Ghitalla, F. *La gographie des agrgats de document sur le Web*, Research White Paper, Published online, 2004, http://www.scribd.com/doc/44666231/Geographie-des-agregats-Web

[12] Janert, P. K., *Data Analysis with Open Source Tools*, O'Reilly Media, Inc, 2010

[13] Kleinberg, J., M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. S., *The Web as a Graph: Measurements, Models and Methods*, Lecture Notes in Computer Science, Vol. 1627, 1999

[14] Lyman, P. and Varian, H. R., *How Much Information*, 2003. Published online http://www.sims.berkeley.edu/how-much-info-2003

[15] NIST/SEMATECH *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 2010

[16] Pfaender, F., Jacomy, M., Fouetillou, G., *Two Visions of the Web: from Globality to Localities*, Proceedings of Information and Communication Technologies, 2nd edition, 2006, Damas, p. 566-571.

[17] Pfaender, F., Jacomy, M., Ramm, M., *Percevoir le territoire Web Picard*, in Proceedings of SAGEO06 , 2006, Strasbourg, France

[18] Pfaender, F., *Spatialisation de l'information : Lire, inscrire et explorer les systmes informationnels*, PHD thesis, Universit de Technologie de Compigne, 2009

[19] Scharl, A., Tochtermann, K. (diteurs), *The Geospatial Web, How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*, Advanced Information and Knowledge Processing Series, 2007, London: Springer

[20] Shneiderman, B., Card, S.-K., MacKinlay, J.-D., *Readings in Information Visualization, Using Vision to Think*, Morgan-Kaufmann Publishers, 1999, New York

[21] Thomas, J. J., Cook, K. A., *Illuminating the Path: The R&D Agenda for Visual Analytics*, National Visualization and Analytics Center, Published online, 2005, http://nvac.pnl.gov/docs/RD_Agenda_VisualAnalytics.pdf

[22] Tufte, E., *Visual Display of Quantitive Information*, second edition, Graphic press, cheschire, Connecticut, 1993

[23] Tufte, E., *Beautiful Evidence*, Graphic press, cheschire, Connecticut, 2006

[24] Tukey, J. W., *Exploratory data analysis*, Addison-Wesley, 1977

[25] Warden, P., *Data Source Handbook*, O'Reilly Media, Inc, 2011