Enhancing Navigation in the World Wide Web

Massimo Marchiori
MIT
Laboratory for Computer Science
545 Technology Square, NE43-257
Cambridge, MA 02139, USA
max@lcs.mit.edu

Abstract

The great potential of the World Wide Web is given by its capability of accessing information by navigating via hyper links from one site to another. However, this great potential is currently not properly fulfilled due to a rather poor connectivity, which makes navigation rather difficult and far from linear. This big problem is intrinsic to the web structure, which is maintained and updated in a completely distributed way. In this paper, we propose a way to enhance the web connectivity, focusing on the concept of navigation cooperation among web sites. We study how suitable bonuses for cooperation can nicely lead to improve the World Wide Web connectivity, making navigation much more fruitful, and acting more effectively just where the cooperation problem is more difficult, namely in the case of market competitors.

1 Introduction

The World Wide Web has revolutioned the way people can access information. In a sense, the web is a collection of a multitude of spatially distributed databases. On the one hand, its superior flexibility relies in the opportunity to overcome the spatial bareers, and to freely jump from one site to another via hyper links, with a simple mouse click. On the other hand, the power of the World Wide Web relies on its size and variety, since it collects an enormous amount of different sites, as reported by every recent estimation (cf. [7]). However, this great power also raises what is the major problem of the web: navigating in hyper space is becoming more and more difficult, since the web is nowadays very poorly connected (see for instance [1]). The problem is in appearance unattackable: the power of the web relies in its distributed character, and so there is no possible global control on it. In this paper we propose a solution to this issue, and test its effectiveness via small-scale simulations called *web arenas*. It is true that there is no global control on users maintaining sites in the web, but there is nevertheless a way to *incite* such users to improve the web navigational structure: the idea is to provide suitable "bonuses" to the maintainers of web pages so to foster navigation cooperation among sites. But who is going to provide such "bonuses", and what form can they have? As market studies clearly indicate, in order to survive into the WWW informative jungle, web users have to almost exclusively resort on search engines (automatic catalogs of the web) and repositories (human collections of links usually topics-based). In turn, repositories are now resorting themselves on search engines to keep their databases up-to-date. Thus, the crucial component in the information management chain is given by search engines. Therefore, the idea is that these bonuses should be provided by search engines' score: if the web structure is in some sense improved by a web page, such page will get a higher rank. Indeed, search engines have become so important in the advertisement market that it has become essential for companies to have their pages listed in top positions of search engines, in order to get a significant web-based promotion. Starting with the already pioneering work of Rhodes ([9]), this phenomenon is now boosting at such a rate to have provoked serious problems to search engines (see e.g. [4]), and has revolutioned the web design companies, which are now specifically asked not only to design good web sites, but also to make them rank high in search engines. A vast number of new companies was born just to make customer web pages as visible as possible. More and more companies, like Exploit, Allwilk, Northern Webs, Ryley & Associates, PlanetOcean, SignPost, Did-It, Mentor Marketing, etc., explicitly study ways to rank high a page in search engines. OpenText arrived to sell "preferred listings", i.e. assuring a particular entry to stay in the top ten for some time (for a discussion on the effects of such a policy, see for instance [12]). In this paper we thus study the effects of search engines bonuses on the navigational structure of the web. We will introduce the concept of cooperation bonus, and show how it theoretically provides a good solution. Next, we report on extensive testings with different bonuses, measuring their effect on the navigation in the WWW. These tests shed new light on the bonus approach, and show how the cooperation bonus ranks by far as the best method. The tests also include the important *visibility bonus*, which is currently implemented by many search engines: it is shown how the effects of this bonus on the global navigability of the World Wide Web are deleterious, and so its usage should be avoided by search engines.

2 Notations

In general, we consider in full generality a so-called *web structure*. Intuitively, a web structure is any "web environment", that can be the World Wide Web, a localized part of it, a web-based intranet and so on. Web structures are composed by *web objects* (for instance, a web page). For rigorous technical definitions, we refer the reader to [3]. The World Wide Web structure will be indicated as usual by WWW. In this paper, when talking about a (hyper) link from a web object A to the web object B, we will always consider understood that A and B belong to *different sites*, that is to say (cf. [1]) we will focus on *global navigation*, and not on in-site local navigation (which indeed is not much of a problem). As far as search engines are concerned, a search engine is usually asked to return web objects that are relevant to a certain query, returning a *ranking*, that is a sequence of web objects, ordered with respect to their relevance. For simplicity, we will consider the query as a finite string, called the *key*. In order to produce such rankings, a search engine needs a so called *score function*, which we denote with $\textsc{Score}$: each search engine currently evaluates the relevancy of a web object A with respect to a key K, assigning a certain *score* $\textsc{Score}_K(A)$. This formalizes how much the user looking for information relating to K may be interested in the web object A. Its intuitive meaning is that the more information, the greater the corresponding score. Without loss of generality, we assume that the score returned by $\textsc{Score}$ is a number between 0 and 1.

3 Navigation Cooperation

In order to be really useful, highways in the World Wide Web must not be built by chance. That is to say, there should be a rationale when building a piece of the information highway: if a user is looking at a web object because he is interested in some particular information, he should (also) be offered with links pointing to objects offering related information. Thus, given two web objects A and B (belonging to two different sites), we say that there is *navigation cooperation* from a web object A to another web object B if a user navigating to A seeking for a specific information can proceed with his search by navigating to B. The prime ingredient for navigation cooperation is therefore the *hyper link* from A to B (however, the presence of such a link does not automatically imply navigation cooperation, as we will see later). Note that the notion of (navigation) cooperation is not necessarily symmetric: one object can cooperate with another, but the converse may not hold.

3.1 The Cooperation Bonus

As said, search engines should provide *score bonuses*, giving higher score to web objects that improve in some sense the navigation in the WWW. The solution to this problem can be provided in the following way. If a web object A provides cooperation navigation to another object, say B, the search engine can provide a bonus depending on $\textsc{Score}_K(B)$, that is to say depending on how much the navigational help is useful for a user which is interested in the topics described by K. So, it remains to formalize how to compute such cooperation bonus. Suppose that the web object A has cooperative links towards the web objects $B_1, \ldots, B_n$. A choice could be for example setting a bonus proportional to $\sum_{i \in [1,n]} \textsc{Score}_K(B_i)$. However, this is not correct for a couple of reasons. The first reason is of implementative nature: the bonus must be bounded, otherwise it couldn't be reasonably implemented. Using a bonus like the one seen before, there is no bound since the bonus can grow indefinitely. The second reason is (by far) more important. When we add a bonus to the original score function, the neat effect is that the search engine is using another score function. Now, the score function must measure the relevance of a web object w.r.t. a given key, and so *we cannot arbitrarily modify* the original score function: we have to verify that it is still a good relevance measure. A possible solution to these problems is the following one. We first order the web objects pointed to by A according to their score (w.r.t. the given key K): suppose, without loss of generality, that we have $\textsc{Score}_K(B_1) \geq \ldots \geq \textsc{Score}_K(B_n)$ (that is to say, B_1 is the most relevant web object w.r.t. K, followed by B_2, and so on until B_n, which is the least relevant web object w.r.t. K). In this case, the *cooperation bonus* is given by $F \cdot \textsc{Score}_K(B_1) + F^2 \cdot \textsc{Score}_K(B_2) + \ldots + F^n \cdot \textsc{Score}_K(B_n)$, where F is a constant in $(0,1)$. That is to say, the bonus provided by each pointed web object in a certain sense fades exponentially. Let us see why this solution works well. The first requirement, giving a bounded score function, is solved, since it is easy to verify that the bonus cannot be greater than $F/(1-F)$. The second requirement, providing a reliable relevance measure, is also satisfied. Indeed, such a bonus can be seen as a safe approximation of the so-called *hyper information*, which is a measure of the relevance of the potential information of a web object with respect to the web space. Being the treatment of hyper information (cf. [3]) out of the scope of the paper, we just hint at the connections between it and the cooperation bonus. The intuition is that the information pointed by a link cannot be considered as *actual*, since it is *potential*: for the user there is a *cost* to retain the textual information pointed by a link (click and... wait). Now, the user looking at the web object A cannot retrieve at the same all the web objects $B_1, \ldots, B_n$ that are pointed by A, but has to sequentially select them. In other words, nondeterminism has a cost, which is paid in term of time used to retrieve a web object. Studies in [3] have shown that a reasonable assumption is that *the cost is exponential with respect to time*. This means, coming back to our example, that in the best case the user will retrieve the most informative web object B_1 (bonus of $F \cdot \textsc{Score}_K(B_1)$), and then the second more informative one B_2 (bonus of $F^2 \cdot \textsc{Score}_K(B_2)$), and so on, therefore just giving the overall bonus $F \cdot \textsc{Score}(B_1) + \ldots + F^n \cdot \textsc{Score}(B_n)$. Note that in computing the cooperation bonus we have to assume the user does the best choice (i.e. he first selects the most informative object, and so on), since the key K is variable. This way, links pointing to web objects not related to K (viz., formally, having a low $\textsc{Score}_K$) do not devalue the cooperation bonus. If we would have used something like the bonus given by a random choice of the links, then a link pointing to a web object with no relation to K would devalue the bonus, while instead such a link should be ignored, as the above method, consistently, does. Formally, this bonus can be seen as a "first order" approximation of the overall hyper information (which requires more higher-order terms). Besides the theoretical foundation, extensive tests have shown that adding the hyper information (or a suitable approximation of it, like the bonus that we have described in this paper) still yields a reliable relevance measure (cf. [3]). As far as the computational cost is concerned, it is easy to see that the cooperation bonus can be implemented in a fast way using hashing and parallel architectures (a must choice for nowadays high performance search engines), and even on sequential architectures, at the expense of some data redundancy. According to our tests, speed can be further increased using various other techniques, for instance by setting a fixed low upper bound on the number of cooperative links to be considered, without significantly affecting the results presented in this paper. Another *very important* characteristic of the cooperation bonus as defined in this paper is that it integrates smoothly with existing search engines technology: it can be implemented on top of every search engine, acting as a post-processor. This means that the original score function is treated as a black box, and does not need any internal modification. The neat effect from the perspective of the search engine maintainer is therefore a safe *modular architecture* of the search engine's components: the cooperation bonus can be implemented as a separate module, that can be kept neatly separate from the main evaluation module(s); this way we do not increase the (already high) complexity of the major score function, and the consequent effort in its maintainance; this latter component is indeed by reported experience one of the parts of a search engine that needs more and more updates, due to a number of factors: the (yet) poor quality of the evaluation due to such a large domain; the rapidly varying structure of the data present of the WWW; and, the continuous security struggle of users trying to artificially increase their ranking (cf. Section 6).

4 Navigability

A cooperative link is potentially improving the web connectivity, and the usefulness of navigation for users. However, in order to quantify how fruitful the usage of cooperative links can be, we need a way to measure the *navigability* of a web structure. First, we need what is called a *categorization* (also called *classification*) of the web objects. A categorization classifies each web objects into a certain category. Thus, for example, we could have as categories of interest *Computers* and *Music*, with the intended meaning of indicating those web objects dealing with computers and music, respectively. Then, a categorization would be a set of web objects classified in the category *Computers* (those web objects that we classify as pertaining to computers), and a set of web objects classified in the category *Music* (those web objects that we classify as pertaining to music). Once we have a categorization of the web objects, it is clear how to intuitively measure the navigability of a web structure: a user looking for information in a particular category must be easily able to navigate through all the objects belonging to that category when starting from one of these objects. This means such objects must be tightly connected by hyper links, and that they shouldn't be too much connected to objects belonging to other categories, otherwise one can get lost while navigating. These informal provisos can be formally expressed as follows. Let S be a subset of a web structure W, and $|S|$ and $|W|$ be the number of web objects in S and W respectively. Denote with $\mathcal{I}(S)$ the number of links connecting two objects in S, and with $\mathcal{O}(S)$ the number of links connecting objects in S with objects not in S. Then, the *cohesion* of S is measured as follows:

$$\frac{\mathcal{I}(S) \cdot 2(|W| - |S|) - \mathcal{O}(S) \cdot (|S| - 1)}{2|S|(|S| - 1)(|W| - |S|)}$$

The cohesion, in fact, can be shown to be just the difference between the percentage of "intra" connectivity (to what extent the elements of the subset are connected to each other), and the percentage of "inter" connectivity (to what extent the subset is connected with the rest of the web). So, now we have all the tools to measure how fruitful the navigation in a web structure can be: the *navigability of a web structure W w.r.t. a categorization Cat* is the average cohesion of each category of *Cat*.

5 Web Arenas

In this section we present extensive simulations on the practical behaviour of bonuses. We have tested how various bonuses affects the navigability of a web structure. In our simulations we employed so-called *web arena games*. A population of users, called *the players*, is chosen. A specific site domain is assigned to each player. The "moves" of the game are: each player can build web objects in the web arena having the same domain as the one assigned to him (i.e., can only act on his local site). Once assigned the *goal* of a web arena, each player is supposed to perform moves in order to reach the goal. In our case, we selected twentyeight persons as players: the players included web site designers, advertisement responsibles, workers in private companies, programmers, computer scientists and students. All the players had good knowledge of the World Wide Web and of HTML. In order to measure the navigability of the web arena, we needed a categorization. We chose the Excite Ontology of the WWW. It consists of a layered categorization: the first-layer categorization (the most general), has 20 main categories, ranging from Arts to Travel[1]. Each category is composed by several subcategories, forming the second-layer categorization (which is therefore more precise than the first-layer level) and so on: for instance, the category Art is subdivided into 12 subcategories ranging from Architecture to Theater. In fact, we did not utilize the whole Excite Ontology but only a part of it: we dropped some categories that would have needed a too specialistic competence for our population, or that were out of scope. For instance, among the 20 main categories we dropped the category "Personal Pages", and among the subcategories of Arts we dropped topics like "Architecture", "Arts Magazines", "Ceramics", "Craftworks" and "Fine Arts". We assigned to each member of the population some specific categories drawn from the ontology (at least one for each layer categorization). The important thing was that only the name of the category was given, but *no mention at all was made about the ontology itself*. This means that not only each player was unaware of the existence of the ontology, but also that only the particular category name was provided, and not the whole classification. In order to play a web arena game, there is also the need for a goal. We played the game several times: each time, a specific search engine for the web arena was used. The goal for each player was to rank high in the search engine, for each of the categories he was provided with, just like if he was the responsible of the site in the "real" WWW, and wanted to have its site noticed. Note that, just as in the WWW case, a maintainer of a site devoted to a certain argument does not have a "sure" way to establish that his site will have a high rank for all the users using a search engine and looking for information related to his site (in our terminology, pertinent to that category). Thus, what one can do is only to try to rank high for many keywords that are in all likelihood representative or related to the given category.

We studied the behaviour of four kinds of bonuses.

The first one was the *no-bonus*: that is to say, no bonuses were used. In order to get a situation as realistic as possible, we just employed as search engine a basic module (actually part of a bigger search engine being developed by the author) performing classic weighted scoring based on frequencies and counts (this performs roughly as good as each search engine present nowadays). The same module was then reutilized

[1] As by now, Excite has slightly changed its categorization into "channels"; this anyway does not affect the results of this paper.

in the other four web arena games by adding on top of its score function a specific bonus.

The second one was the *visibility* bonus, which gives bonuses to a web object proportionally to the number of links that point to it. This bonus is of particular importance (cf. [10]) because it is currently employed by many search engines like WebCrawler, Excite, Lycos and Magellan (although not with the purpose to improve the web structure, but just to enhance the evaluation of the score).

The third one was the *cooperation* bonus, as introduced in this paper.

The fourth one was the *connectivity* bonus, that just naïvely assigns a bonus to each link (this can be seen as a degenerate case of cooperation bonus, where the cooperation specification is always satisfied).

In each case, the players were provided with the information about what bonus was employed (in the case of the cooperation bonus, the public cooperation specification was made available). The internal specifics of the basic search engine (no-bonus case) were not given, in accordance with the WWW case. In order to be a realistic simulation, the behaviour of the search engine was made similar to the WWW case: the search engine was not always completely up-to-date, but had a *refresh time* which was fixed to one week. This simulates the actual behaviour of search engines, and is also a form of protection of search engines, since one cannot interactively check whether every modification to a web objects leads to a greater score or not (cf. [4]).

No-Bonus
The evolution of the navigability in the no-bonus case is illustrated by the diagram in Figure 1. For each week (corresponding to the refresh time of the search engine), the diagram reports the navigability of the web arena w.r.t. the first three categorizations of the Excite Ontology. Out of the three bars present for each week, the rightmost bar (pictured with darkest gray) refers to the top-layer categorization, the middle one to the second-layer categorization, and the leftmost one to the third-layer categorization. The outcome of the simulation was a web arena with a very poor connectivity, and (what's worst), as can be seen from the diagram, with a very low navigability too, for all the levels of the ontology. This is not surprising, since no stimulus to improve connectivity is given, and everything is left to the good will of the users. Also, it is clear that, in the case of market competitors, cooperation is in most of the cases unfruitful. In summary, this situation is a good representative of the actual situation of the World Wide Web (cf. [1]).

Visibility Bonus
The situation obtained using the visibility bonus has been the following. Connectivity stayed very poor. There were some examples of mutual cooperation, but they even worsened navigability, since a common trend was to perform mutual cooperation with sites belonging to different market niches. As a result, navigability was worst than the "no-bonus" case, as can be seen from the diagram of Figure 2.

Cooperation Bonus
When the cooperation bonus was applied, connectivity increased a lot, and, as shown in Figure 3, navigability was much improved. This confirms our initial intuitions about the effectiveness of the cooperation bonus. It is interesting to report how the players reacted to the public specification $\mathcal{P}$: they all *ultimately* stick to it. Many tried at the beginning not to follow the $\mathcal{P}$ specification, and after some refreshes realized that this was not worthwhile due to lost bonuses, so at the end they gave up to "break" $\mathcal{S}$, and concentrated on the content of the page and on the cooperative links.

Connectivity Bonus
In this case, the progress made with the cooperation bonus got lost. The connectivity increased by far, but this time there was no rationale of providing "navigationally useful" links. Even, the trend, as in the visibility bonus case, was to link sites not belonging to the same market niche, thus further ruining navigability, as the diagram of Figure 4 reports.

6 Advertisement Impact

Search engine maintainers are well aware that every modification of the score function has a broad impact on the advertisement market, due to the pressure imposed by site maintainers, and web design companies, in order to have their site better ranked than the other market competitors (cf. e.g. [2, 5, 8, 11, 4]), a phenomenon referred to as *sep* (which stands for *s*earch *e*ngines *p*ersuasion). It is well-known that this great problem is doubly dangerous for a search engine, since it affects not only the quality of the search engines, but even its advertisement earnings, which constitute a primary financial entry. The sep phenomenon is nowadays particularly bad even if little or no information on the score function is given by search engines maintainers. Now, the problem with a bonus is that, to work, one must know its existence (and how to get a bonus), which obviously amplifies the risks of sep. Therefore, every concrete implementation of a bonus must take into account also this problem as a primary factor. In the specific case of the cooperation bonus, there is in principle ample possibility for sep, since the presence of a link in a web object does not necessarily mean that there is cooperation. Indeed, as we will see, there are many ways to insert a link into a web object without actually being cooperative at all, making difficult or even impossible for a user to utilize the link for navigation. In some cases, we can actually determine whether or not a link is noncooperative. In other cases, we cannot have this certainty. The adopted strategy is to develop a suitable *cooperation specification* $\mathcal{S}$, that is to say a number of rules such that if a link satisfies them, it can be classified as a cooperative link. This means that in those cases where we are in doubt, we opt for false cooperation, that is to say we do not consider a link as cooperative. Another crucial factor is that *people wanting to get bonuses for cooperation should be able to easily build cooperative links*. This implies that rules on how to build cooperative links should be made publicly available. However, for security reasons it is unsafe to make the whole cooperation specification $\mathcal{S}$ public, since there may be the chance that some fault is found in $\mathcal{S}$, and so a user can construct links that surely pass the cooperation specification, but are not cooperative. The solution is to make public not the complete cooperation specification, but another specification, the *public cooperation specification* (denoted with $\mathcal{P}$). This specification must be such that: 1) if a link passes $\mathcal{P}$, it passes $\mathcal{S}$, and 2) $\mathcal{P}$ must be *easy to understand*. Thus, $\mathcal{P}$ is a kind of easy approximation of $\mathcal{S}$, which helps the user that wants to cooperate to easily build cooperation links, and nevertheless maintains an acceptable security level, keeping secret the (possibly complex) specification $\mathcal{S}$. In the following section we will discuss a possible choice of the specification $\mathcal{P}$ and $\mathcal{S}$: with some differences, they are being implemented in a second-generation search engine under development by the author. As far as the com-

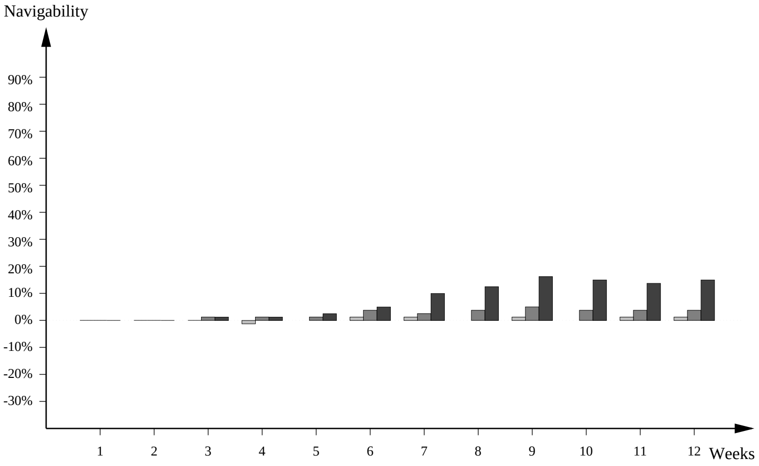

Figure 1: Navigability in the No-Bonus case.

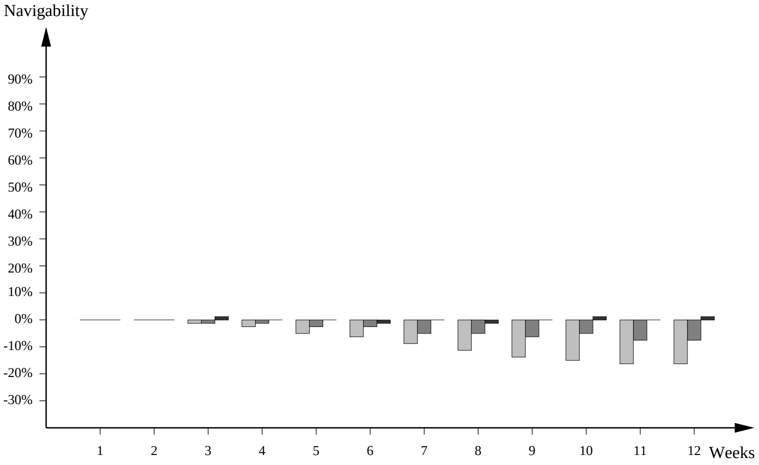

Figure 2: Navigability in the Visibility Bonus case.

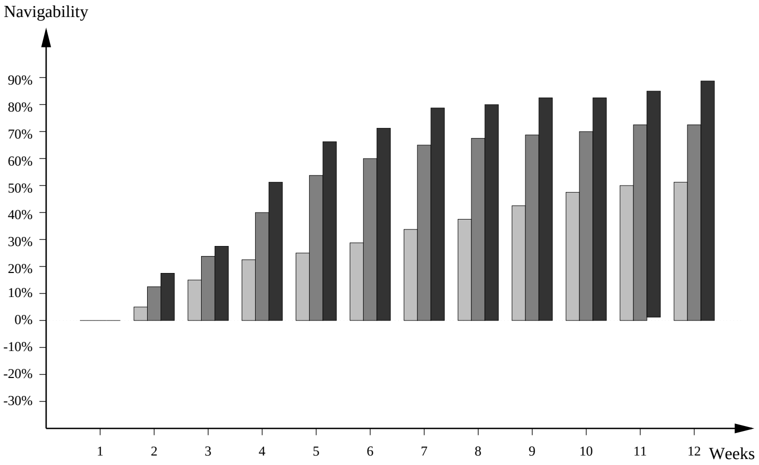

Figure 3: Navigability in the Cooperation Bonus case.

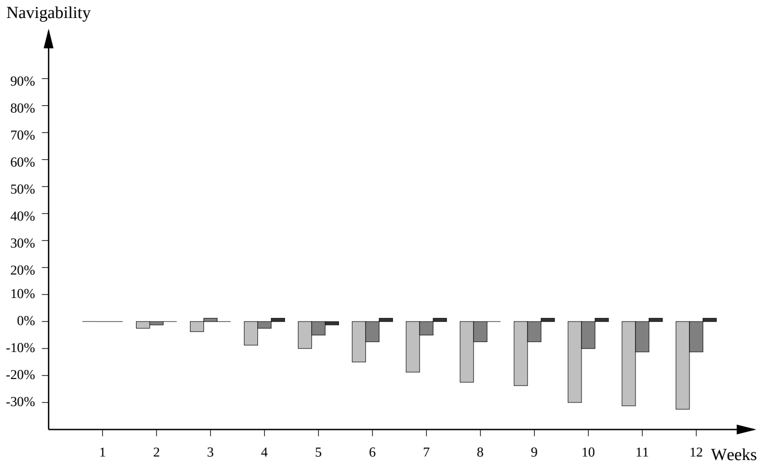

Figure 4: Navigability in the Connectivity Bonus case.

putational complexity is concerned, observe that $\mathcal{P}$ should of course be rather fast to execute (which will be the case for the proposal contained in this paper), and that in any case for optimum performance $\mathcal{P}$ should be applied as preprocessor as soon as the cooperative links list is built, so that it does not affect in any way the final search engine user.

7 Noncooperation

As said, the presence of a hyper link in a web object A does not necessarily mean that a user viewing A can readily utilize it for navigation: there is indeed a number of techniques that can be used to make a link noncooperative. In the sequel, we will list them, and describe for each of them the consequent cooperation specification (and its public version).

7.1 Layout Techniques

The first group of techniques that we will consider can be dubbed as "layout" ones: they allow to hide a link, making it invisible (or hardly visible) by employing HTML layout commands. All these techniques rely on the use of so-called "ghost components" (see [4]), that is to say a part of the code where text can be inserted, that in all likelihood will never be observed by a user when viewing the web object using a web browser: here, however, the problem is not simply to hide text, but to hide a hyper link as well. This implies that the great majority of the ghost components cannot be used, and only four[2] remain:

1. put the link outside the BODY of the web object
2. use an unreadably small font
3. use an unreadable combination text/background
4. put the link in the NOFRAMES part

Let us explain all these techniques, and provide solutions for them. The first case is straightforward: we just do not consider as cooperative a link which is outside the body part of a web object. This information can be fully reported in the public cooperation specification. However, all the layout techniques can be expressed using the following general requirement: *the link must be visible*. We will see this is a correct public cooperation specification for almost all the techniques of this section. Let us now consider Case 2. The change of font size can be done only using the HTML font size command. Therefore, it is straightforward for a parser to exactly determine what font size a certain part of a web object will be actually displayed with. Thus, we can require that for instance links that appear with font size lower than 2 are not considered as cooperative. The details of this part of the cooperation specification can be made public (it is a rather simple condition), or one can put in $\mathcal{P}$ just the global requirement that the link must be visible. In Case 3, we have to distinguish two cases. In the first case, a background image is used. Since analysis of image characteristic would be extremely expensive, a reasonable cooperation specific is that no background image can be used for cooperative links. In the second case, the text/background colors are changed using the HTML color commands. In this case, it can be imposed that there is at least a certain contrast between the ink and background colors, so to make the text readable enough. This test can be relaxed in a variety of ways. The easiest solution would be so to just forbid usage of color commands. Here, we suggest another solution for the cooperation specification, which is more flexible: there cannot be color changing commands, but for the preamble of the web object (the preamble is the part of the web object before the BODY tag). That is to say, the user is allowed to use color commands to set the layout of the whole web object, but not of parts of it. This is in accordance with the widespread usage of color commands, which are in the great majority of the cases used only in this way. The corresponding public specification $\mathcal{P}$ can simply state that *a cooperative link must be visible, and no background images are allowed*. Note that if a more restrictive cooperation specification is chosen, for instance no color changing commands but in the preamble, than this should be reflected in the public specification (recall that $\mathcal{P}$ must imply $\mathcal{S}$). So, for instance, one can state this additional proviso as such, or approximating it by saying that no color commands are allowed. Finally, in Case 4 we just do not consider links in the NOFRAMES part as cooperative. This information can be completely omitted from the $\mathcal{P}$ specification, since if a cooperative link is in the NOFRAMES part, it should be also present in the main frame part.

[2] Future changes in the HTML specification and the ongoing "browser war" continuously adding new facilities will lead to an increase of this number (e.g. the new possibility to change the font via the FACE attribute), but anyway the corresponding solutions should be quite straightforward (and, one could just consider as noncooperative a link where an unrecognized HTML feature is present).

7.2 Context Techniques

Finally, there are other techniques to build noncooperative links that can be dubbed as "context" ones, since they place the link in places that are in a sense not its proper context. These techniques are essentially two:

1. put the link far away from the top of the page

2. put the link in the middle of garbage text

Observe that these techniques can be used in combination. Let us now analyze these two techniques. Case 1 can be coped with in several ways. For instance, one can fade the bonus of a cooperative link proportionally to how far it is placed from the top of the page. A variant is to consider only links within a certain distance from the top of the page, for instance 4000/5000 characters. Now, let us come to Case 2: identifying "garbage" text would require in some sense some semantical knowledge on the text, which is readily a hard task. But if we impose that the cooperative link must be not too far away from the top of the page, then putting garbage text has the side-effect to badly affect the layout of the page, which is well known as a key component in advertisement. Thus, one can indeed build parts of a web object that looks without apparent meaning, and insert a link into them, but since these components must be visible (by the requirements issued in Subsection 7.1), this also ruins the graphical presentation of the web object. In the public cooperation specification, a suitable solution is to require that *a cooperative link must appear not too far away from the top of the page*. Finally, note that in the case of market competitors, inserting cooperative links does not mean to do a nice advertisement to the adversaries, since one can put the link in perfectly meaningful statements like "the following products are by far worse than ours: . . .". Indeed, such links are truly navigationally cooperative, because, even if arbitrarily judging on the content, they allow the user to proceed in the search of information by navigation. Thus, being *navigationally* cooperative does not means to be truly *market* cooperative (if it were not so, the advantages of getting score bonuses would not be so high in the market field).

8 Simulation Scaling

Web arenas do provide a hint on the actual behaviour that the WWW can have. However, like all simulations, it must be clarified to what extent their results are scalable.

An observation which is common to all the simulations, but for the first no-bonus case, is that the incitation to rank higher is here much stronger than an actual one in the WWW would be. In particular, usage of bonuses will presumably affect in a way similar to what shown in this paper areas with *high market pressure*, and areas with advertisement interests. The other areas will presumably follow the same trend, but at a slower rate than that shown by the web arenas. This means that in order to obtain a more precise prediction, scalable to the whole World Wide Web, one may need to "normalize" the positive/negative effects of the bonuses, in the worst case multiplying them by the percentage of commercial-related sites present in the web; note that, eventually, the navigational structure should eventually still improve following the average tendencies shown in the arenas, since the majority of the World Wide Web falls in the commercial-related category, and the trend in the last years has shown a huge increase of commercial-related sites w.r.t. non-commercial ones (see e.g. [6]).

Another observation is how much the population size impacts the simulations. All the cases do not seem to present much problems. In the cooperation bonus case we can expect a slight degradation of the navigability measure for less detailed categorizations (like the first-layer one), due to the size problem. This, however, is not likely to be a problem for the final user, since the need for a detailed navigational aid decreases with the lossiness of the required information.

As far as the number of cooperations is concerned, instead, the situation is a bit different. While the no-bonus, the cooperative bonus, and the connectivity bonus cases reasonably scale up, providing an increase in connectivity, the visibility bonus case does not scale well to a big population size. The problem is that here the "web" is quite little, so everyone knows everything about the others. This way, mutual cooperation given by the visibility bonus can produce some instances of cooperation. However, in the real WWW, where the size is enormous, the chance that the cooperation of a site A to B is noticed by B is very low, and becomes near to zero when A and B belong to different categories. So, scaled up to the WWW, the increase of connectivity observed in the simulation provides only a "best case" which is likely to be by far too optimistic. Moreover, the outcome that we have found on the negative effects of the visibility bonus on the global navigability of the WWW are substantiated by the practical impact that this bonus has had so far: indeed, the real WWW (with many major search engines employing the visibility bonus, like WebCrawler, Excite, Lycos and Magellan), has shown, to the best of our knowledge, *no examples* of *real* inter-sites cooperation; moreover, the effects of the visibility bonus on the quality of the score functions have shown to be rather nefarious as well (cf. e.g. [10]). Finally, the cooperation bonus scales up well to bigger populations, since *the bonus stays the same*.

Acknowledgments

I wish to thank Mike Jory (OpenText), and Larry Budnick (Individual Inc.). Thanks also to the anonymous referees.

References

[1] Tim Bray. Measuring the Web. *Fifth International World Wide Web Conference*, May, Paris, 1996.

[2] Kenneth R. Churilla. Secrets of Searching the Web & Promoting your Website. Mentor Marketing Services, 1996.

[3] Massimo Marchiori. The Quest for Correct Information on the Web: Hyper Search Engines. *Sixth International World Wide Web Conference*, April, Santa Clara, California, 1997.

[4] Massimo Marchiori. Security of World Wide Web Search Engines. *Reliability, Quality and Safety of Software-Intensive Systems*. Chapman & Hall, 1997.

[5] K. Murphy. Cheaters Never Win. *Web Week*, May 1996.

[6] Network Wizards. Internet Domain Survey. July 1997.

[7] Nielsen Media Research. Web Audience Measurement: Issues, Challenges and Solutions. *IPQC Conference on Performance Measurement for Web Sites*, San Francisco, 1996.

[8] Northern Webs. The Search Engine Tutorial for Web Designers, 1997.

[9] Jim Rhodes. How to *Promote* Your Business Web Pages.

[10] Danny Sullivan. Webmaster's Guide to Search Engines. Calafia Consulting, 1997.

[11] Gus Venditto. Search Engine Showdown. *Internet World*, 7(5), pp. 79-86, 1996.

[12] Nick Wingfield. Engine sells results, draws fire. *C|net Inc.*, June, 1996.