Metodološki zvezki, Vol. 1, No. 2, 2004, 469-489

Hypertextuality of the Slovenian World Wide Web

Gregor Petrič[1]

Abstract

The substantial concern of this article is a question to what extent does the contemporary World Wide Web as an information retrieval system reflect key attributes of ideal hypertextual systems. The topic is relevant, since in the literature notions of hypertext and hypertextual systems are accompanied with strong implications not only for the ease and efficacy of access to information, but also for fostering democratisation, augmenting creativity and cooperativeness of human beings. After the brief presentation of the problem the paper focuses on the methodology of analysing this problem – definition of relevant dimensions of hypertext in the World Wide Web, their operationalisation and empirical verification. The latter is presented most thoroughly since it includes a procedure of generating a network of web sites in the Slovenian World Wide Web on the basis of approximately 1.8 million of web pages, identified by search system Najdi.si. After the definition of units and relations, relevant methods and their results are presented in order to assess the hypertextuality of the Slovenian World Wide Web. It is shown that a relatively great proportion of web sites do not follow the expectation of the designers of the World Wide Web technology for it to be a globally interconnected "Docuverse", however, a large minority of web sites are in aggregate reflecting the attributes of ideal hypertext systems. The results can be informative for the global World Wide Web since one of the essential characteristics of the Slovenian World Wide Web have similar distribution to the one assessed in other researches on significantly larger - although not adequate for complete network analysis - proportions of the World Wide Web.

[1] University of Ljubljana, Faculty of Social Sciences; gregor.petric@uni-lj.si

1 Introduction

One of the most important building blocks of the World Wide Web (WWW) is a hypertext-markup language (HTML), where the prefix "hypertext" implies that information is organised into a set of documents and links between them. This, however, does not mean that the actual structure of web sites and links between them is hypertextual nor that the concept of hypertext appeared with the development of the WWW. The idea of hypertext appeared in 1945 already, when Vannevar Bush in the article “As we may think” offered a scheme of an electro-optical machine, Memex, which would enable storage of information in such a way that pieces of text could be arbitrarily interconnected. This machine could help to overcome the problem of information retrieval in the age of information explosion due to specific structure of information. Information in hypertext is organised on the basis of associations between pieces of text and such a structure is more natural to the functioning of a human mind than rigid linear and hierarchical systems. The access to information would thus be faster and more efficient than before. The idea of hypertext intensively developed in the 60’s and the 70’s, primarily with the work of Theodore Nelson and Douglas Engelbart (Bardini, 2000), but the circumstances for its fullest development and realisation became ideal with Berners-Lee's system of World Wide Web (1989), which would enable global information space of interconnected documents.

Hypertext is not only a mode of information organisation, which enables fast and efficient access to information, but also carries many implications for democratisation of society (Landow, 1997), augmentation of individual mind and stimulation of cooperation (Bardini, 2000). Taken for granted the positive effects of hypertextual organisation of information, the main concern of the paper is the question to what extent does the contemporary WWW incorporate the essential features of hypertextual systems due to existence of many social and other factors that are changing the nature of this technology. Although the following study is mostly applicative and primarily tries to answer the above question, a significant space is dedicated to the methodological issues in analysing the problem. After short insight into the nature of hypertext and relevant factors, influencing the hypertextual characteristics of the WWW, essential dimensions of hypertext are defined, which are also measured on the Slovenian part of the WWW. The latter was analysed as a complete network of web sites, where relations correspond to links between them. To realise this analysis, a complex procedure, consisting of aggregating web pages on the level of web sites, parsing web links and generating a network, had to be accomplished. This unique procedure can also be used for many other issues from web metrics to the advancement of search systems. By using tools of network analysis, the hypertextuality of the World Wide Web is assessed.

2 Hypertext and the World Wide Web

2.1 Social and political consequences of hypertext

The common thread of various and diverse discussions on hypertext from the scientific fields of information retrieval, communication research and critical theory is a notion of hypertext as an electronic form of writing in a non-linear system of individual pieces of text and links between them. The key attributes of hypertextual systems are said to be decentrality, non-linearity, active role of reader and multivocal dynamics[2] (Landow, 1997).

Hypertextual system with its essential features is not only an advanced form of information organisation for enabling fast and efficient access to the data, but has many other important implications. It can be said that hypertext is in a certain sense physical reflection or realisation of poststructuralist (Barthes, Derrida) interpretations of mental processes in reading texts. Text as a mental construction is always incorporated in a network of relations with other texts and it is not a product of individual mind but derived from the plurality of voices, words and expressions of others (Allen, 2000). Surprising similarity, which was already discovered by Landow (1997), exists between Barthes' idea of ideal text and hypertext. For Barthes (1974) the ideal text is unlimited by any physical structure, it is interlinked, reversible, with many entry points, it has no beginning and no end and none of the texts can claim their authority over others. While poststructuralist ideal text is more a mental construct, hypertext on the other hand refers to manifest, actual embeddedness of physical textual units into networks of related texts, yet the social and political implications are very similar. Common both to poststructuralists and inventors of hypertext is the idea of the collapse of classical modernist conceptual scheme, which is based on the notions like centre, margins, hierarchy and linearity, and its supplementation by concepts such as network, non-linearity and relations.

The organization of information in hypertextual systems and communication of ideas stimulated by them overcomes the consumerist idea of a user as a reader, who consumes the meaning of a single, isolated document, constructed by author in the absence of any other related writings. Furthermore, by manifestation of invisible links in mental processes, hypertext transcends ordinary ways of

[2] The first two are discussed below, while the latter two do not refer to the structure of information organisation but to the role of user in it. The active role of reader means that the reading path is not predefined by the authors of text, but is decided actively by the reader. Multivocal dynamics refers to the feature of hypertextual systems that each user can add his own text and link to the existing information space.

understanding (Landow, 1997) and even poses implications for epistemological processes (Deibert, 1997). A change in communication, implied by hypertextual organization of information, is supposed to have an effect on social organization and ways of thinking. Although the state of discussion on social effects of hypertextual organization of information is still predominantly speculative, this paper supposes that it has certain positive effects on its user.

One of the main characteristics of hypertextual systems is the freedom of its users to move around the information space by choosing different pieces of texts, following diverse links between them. Due to such non-linearity, a user can choose his unique path through the information space by deciding on each text piece, which link to follow. This reflects an overall decentralised organisation of information, since the user chooses his own centre of exploring information and thus no text can claim its centrality over others. The ultimate reading path of the user is independent from the author, liberated of power relations and authority. In this sense Nelson claimed that associative structure of information not only stimulates effective access to information, but also liberates them from fixed, rigid linear structure of print documents and allows users to form meaning independent of the author (in Bardini, 2000). This is the consequence of the fact that in electronically connected texts the documents of individual authors are dispersed in the vast space of other documents and thus lose their physical and intellectual distinction from others.

2.2 Hypertextual ideas in the WWW

World Wide Web was developed in close connection with hypertextual information systems, since its inventor Tim Berners-Lee was well educated in existing hypertextual systems as well as the pioneer ideas of Bush and Nelson (Berners-Lee, 1995). The idea of WWW stems from similar considerations as Bush had in 1945 on the need for new forms of information organisation in the light of rapid knowledge growth. The WWW was supposed to be a system of unlimited dimensions and multiplicity of usage, structured as a large network of interconnected, associated documents by the logic of reference. Unlimited possibility of linking any kind of information would enable users to find information, even when they don't know what exactly they are searching (Berners-Lee, 1989).

As sociology of technology (Bijker et al., 1987) reminds us, technology in a society does not possess its characteristics only or necessarily with regard to the features, envisioned by its inventor(s), but acquires them in a complex process of interdependency of societal and technological processes. WWW incorporated ideas of hypertextual systems in its basic technical structure, but the realisation of this potential was greatly influenced by the network of social actors, who exert influence on the nature of technology (Callon, 1987). It is not the goal of this paper to thoroughly identify and evaluate relevant social actors, who have some effect on the hypertextuality of the WWW, but only to point out that there are many factors, which inform us that linkages between web sites, decentrality and non-linearity are not taken for the granted attributes of contemporary WWW. There are not only external factors, which count for this, but the inherent feature of the WWW - a functional relationship between individual parts and the whole - takes the leading role in the hypertextual structure of the WWW. World Wide Web reflects hypertextual systems only when the majority of its parts, that is web sites, actualise its ideas. Every author of a web site should provide multiple links to web sites of other authors, if we would want that the WWW as a whole would retain its essential hypertextual features. Currently no such control mechanism exists in the WWW, which would stimulate the realisation of this kind of “hypertextual ethics” (Petrič, 2003a). Further on, taking into account many various interests of different authors of web sites, coming from the whole spectrum of society, it is unrealistic to expect from the WWW to function as a hypertextual system in its totality. Various social actors design their web sites to satisfy their needs and the needs of their users, but forget to maintain the latent dimension of WWW, its hypertextuality.

Today there are speculations, that hypertext will eventually disappear from the WWW, become its missing link (Bieber et al, 1997), mostly due to commercialisation. Specifically, this refers to the rapid increase of commercial enterprises, using their web sites for commercial activities. Also the information space of the WWW is becoming commercialised, since portals and search systems, which are owned by big corporate actors, are becoming the dominant points of information retrieval. On the level of user experience they are usually the only gateway to the information space in the WWW. This attributes to the reduction of the experience of browsing through the hypertextual structure of web sites and links between them to the repeated exploration of individual, unconnected web sites. Search systems that support hierarchical access to information are overthrowing hypertextual experience and becoming the most important tool for information retrieval (Retallack, 1999). Commercialisation also radically changes the nature of links between web sites by introducing the promotional “banners”, the function of which does not conform to the idea of cooperation between the texts as envisioned by Nelson or Engelbart (In Bardini, 2000).

While the birth of the WWW was accompanied by optimistic voices of democratisation and fostering of human creativity and cooperation, there is more and more anxiety expressed with the recent track of development, which greatly resembles the commercialisation of traditional mass media (Mosco, 2000). On the other hand, there are more optimistic visions for the hypertext in WWW, especially with the Semantic Web project (Berners-Lee, 1997), which incorporates ideas of a more ordered and transparent typification of web sites and links between them, which would support the separation of hypertextual parts of the WWW from others, for instance, the commercial part. In the light of the fact that there are many more factors, which are dissolving the hypertextuality of the WWW, than those, which are stimulating its development, it is expected that WWW is currently low in hypertextuality.

The empirical research does not allow us to investigate the effects of different social processes on the hypertextuality of the WWW, but only to analyse and assess the realisation and incorporation of characteristics of hypertextual systems in the WWW.

3 Operationalisation and Data Collection

3.1 Essential dimensions of hypertext

Hypertextuality of the WWW is measured on the following key dimensions of hypertextual systems: (a) interconnectedness of WWW as a basic condition of hypertextual organisation of information: existence of links between web sites is the fundamental characteristic of hypertextuality, a distinct feature of hypertext as an information system. (b) non-linearity of WWW as a variety of links from each web site to other web sites, as an opposite to the linear sequence of print documents and (c) decentrality of WWW as an absence of central axis of organisation of web sites and a mode of organisation of information in a network structure, where no web site can claim its central position. All three dimensions were measured on the Slovenian part of the World Wide Web, using social network analysis methods (Wasserman & Faust, 1994). Several necessary steps had to be accomplished before an analysis could be performed: defining units and relations between them, generating a network and investigating its "representativeness". There have been a few studies in the past that generated networks of web pages and links between them (Broder et al., 2000, Kumar et al., 1999), but none of them defined the network of web sites as a whole network, which is language limited and whose web sites correspond to distinct social actors.

3.2 Network of web sites

If a network is a set $N=\{E,R_1,R_2,....,R_r\}$, defined by a finite set of units $E=\{X_1,X_2,.....X_n\}$and links between them, which are described by single or multiple relations $R_t c E x E$, for $t=1,2....r$, (Wasserman in Faust, 1994) then the network under investigation is composed of a finite set of static web sites and single relation R, which is a hyperlink from one web site to another. Notation $X_i R X_j$ stands for X_i in relation with X_j, namely, from the i-th web site exists at least one hyperlink on the j-th web site. Although the terms web site and hyperlink have become part of common language, these two essential elements of the network of WWW need exact specification since the structure of the network depends on this process.

A unit in the network is a web site, that is, a set of web pages that theoretically correspond to distinct designer, social actor3, who decides on the content and form of its web site. A network of web sites theoretically represents a network of documents of various authors, but this correspondence is not perfect due to the fact that a single social actor can decide on the content of several web sites[4]. On the empirical level a web site is determined as a set of web pages corresponding to the same domain name – all directories and files that exist in unique domain name. This rule is not universally applicable, since web sites of more than one social actor can exist in the same domain name. This is most common in cases when a server is hosting disk space for users who can be autonomous authors of web sites. In such cases web sites are not distinguished on the level of domain name, but on a certain level of directory structure, usually first or second.

Relation in a network is a hyperlink from one web site to another. Specifically this means that in the HTML file(s) of a certain web site there is at least one mark <A HREF="URL address">, which manifests itself in a browser as part of the web site (picture, paragraph, sentence, word) and by activating it, the content of the target web page, which is authored by another social actor, is represented. Relation is asymmetric, so the relation xRy does not imply yRx. In the network of web sites relation has a nonnegative value, which represents the number of hyperlinks from one web site to another. In the network that was analysed, the values of relations were reduced to 0 and 1, where 0 means existence of zero hyperlinks and 1 at least one hyperlink from web site of author A to web site of author B. The reduction is necessary because of the research problem: what only matters is the existence of hyperlinks between documents of different authors and not their quantity.

3.3 Procedure of generating a network

The network of web sites was generated on the basis of URLs of Slovenian web pages, which were provided by company Noviforum5. The Slovenian part of the WWW, which was investigated, consists not only of pages that correspond to the top domain .si (52%), but also of pages that are written in Slovenian language or were published in various Slovenian search systems. On the basis of a list of more than 1.8 million of URLs, which was formed by Noviforum's web robot from 1st to

[3] Individual, club, association, company, organisation or any other kind of formal or nonformal social organisation.

[4] This information is usually not transparent in a way that would enable its automatic consideration.

[5] This company owns a very popular search system Najdi.si and they also claim that they have the biggest database of Slovenian web pages.

5th of April, 2002, a network of web sites was generated using a semiautomatic aggregation procedure.

Units of the network were formed by aggregating corresponding individual web pages on the basis of above defined criteria[6]. The procedure resulted in a list of 30083 addresses of web sites, which correspond to different social actors[7].

In the next step the relations, that is, hyperlinks, between units in the set of 30083 web sites were parsed. First, a program HTTPGet[8] was created to access each web site and parse its content in ASCIII code in a single database. This procedure lasted from 7th to 13th of May, 2002. Next, a procedure was created to parse hyperlinks from web sites, identify the target web sites and produce a NET file, which is an input file for Pajek, a program for analysis of large networks (Batagelj & Mrvar, 2002).

The original network of 30083 units was reduced to 25247 for two reasons. Firstly, the units were aggregated if more domain names correspond to single social actor and secondly, unreachable or nonexistent web sites[9] were removed from the network. There were 3128 (10.5%) web sites of the latter type and it can be shown (Petrič, 2003b) that the majority of these web sites are actually nonexistent, thus these missing web sites do not represent true missing values and do not pose any serious problem for the analysis. The final network used for analysis consists of 25247 web sites and 139459 directed relations "at least one hyperlink".

3.4 Completeness and "representativeness" of the network

The theoretical unit of the research problem is the whole WWW, which means that for a valid interpretation of results a complete network of all web sites and hyperlinks between them would be needed. That, of course, is almost unfeasible regarding the scale of the WWW and its fluent dynamics. Missing units in a complete network analysis are a severe problem, since with each unit all corresponding hyperlinks are lost. In this research, however, it is supposed that Slovenian part of the WWW is a complete network, which, it can be shown, is a valid supposition. At first sight it seems unreasonable to limit the research of a space by language criteria, where providing a hyperlink to another country is as easy as to the closest neighbour. The question of network barriers is not specific to

[6] Web sites were aggregated on the level of domain names or on the level of first directories, if its name contained a tilde "~".

[7] This correspondence is not perfect due to already mentioned reasons

[8] The program, created in Perl by Matej Kovačič, is intended for parsing the content of web pages. It accesses a certain URL, recognises its type and if it is a HTML document, it saves it in a tab-delimited ASCIII file.

[9] Unreachable are those web sites, whose all web pages were not accesible by web robot, because the server was down at that time. Nonexistent web sites are those, whose files in HTML form do not exist any more on the designated URL address.

the research of WWW, but it is one of the main concerns in the network analysis (Laumann et al., 1983), since it is difficult to set clear criteria, which would differentiate the suitable and unsuitable units. The barriers of the network under investigation are set by language and this way a sense of a natural complete network can be retained. On the one hand, large proportion of web sites are in Slovene language and secondly, users of the Slovenian part of WWW often remain in its limits, since these are the limits of their language.

Further on, two empirical arguments support the decision to analyse Slovenian part of WWW as a complete network. First, only 25% of web sites provide hyperlinks to non-Slovenian web sites, while there are almost 50% of web sites, which offer hyperlinks to web sites within the network. Secondly, the Slovenian part of WWW shows similar distributions of essential characteristics as those found by researchers on the significantly larger parts of WWW.

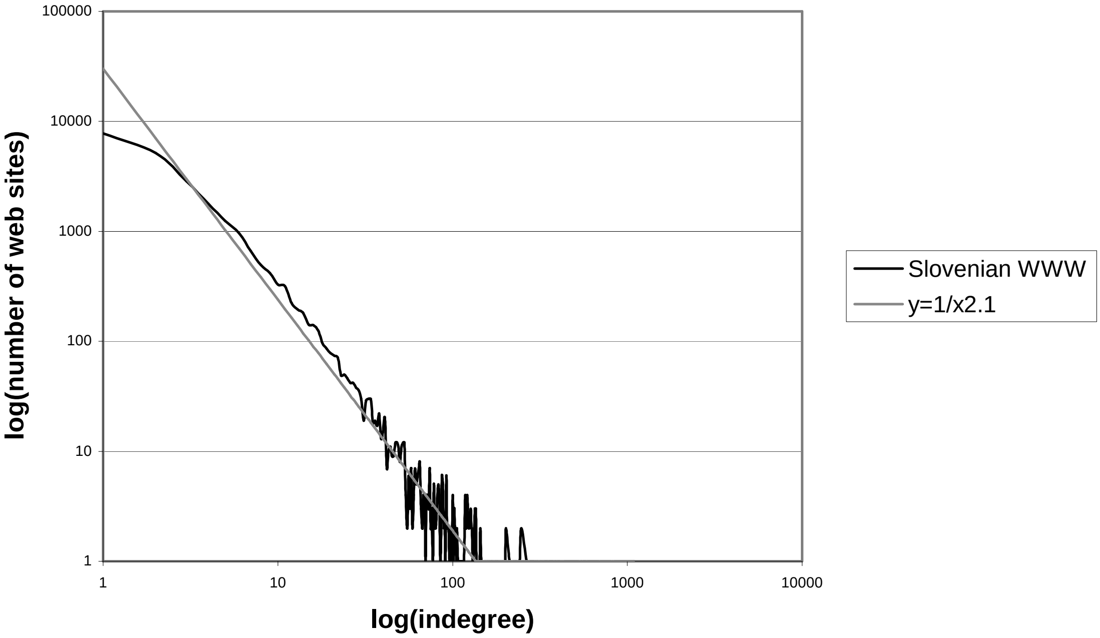

Figure 1: Distribution of number of incoming hyperlinks in Slovenian part of WWW and function $y=1/i^{2.1}*f_i$.

Figure 1 clearly shows that the distribution of indegree on a network of 25247 Slovenian web sites resembles the power law function that was proved in other researches (Barabasi & Albert, 1999; Broder et al., 1999; Kumar et al., 1999) to be representative of the global World Wide Web. Although representativeness is defined as the degree of resemblance of population characteristics in the units of a sample (O'Muircheartaigh, 1997), it can be said that the complete network on

Slovenian web sites is representative of much larger complete network of global web sites. Notwithstanding its smallness, the characteristics of the Slovenian part of WWW resemble greatly the features of the global WWW.

3.5 Basic characteristics of the Slovenian WWW

The population of web sites is aggregated on more than 1.8 million of web pages. A Web site consists in average of 67 individual web pages, but the dispersion is very high (standard deviation = 765). Specifically, 39.8% or 10048 Slovene web sites consist of only one web page and obviously they don't follow hypertextual ideas since the site itself is not hypertextually organised; furthermore, almost 80% of these sites do not offer any hyperlink to other web sites. On the other hand, a minority of web sites consists of a large number of web pages, which are, not surprisingly, portals, search systems, government agencies and other important actors in society.

Slovenian web sites are found not only under the top domain .si, but under many others:

Table 1: Number and percentage of web sites by type of top domain.

Top domain name	Number of web sites	Percentage
si	13403	53.1%
com	7343	29.1%
net	3148	12.5%
org	848	3.4%
to	79	0.3%
info	34	0.1%
other (tv,uk,it,itd.)	392	1.5%
total	25247	100%

Only a little more than a half of web sites in Slovenian part are registered under a domain name .si, which clearly implies that language criteria of selection is much more relevant than domain name selection. At the beginnings of the WWW, a domain name system was more strictly followed and could enable certain inferences, while today a great disorder exists. This topic exceeds the subject matter of this paper, but implies that hypertextuality is only one part of the successful information retrieval in the contemporary WWW.

4 Assessment of Hypertextuality

The network of Slovenian web sites can be treated as a large, sparse complete network, since the number of relations does not exceed much the number of units. Suitable tool for analysis of such networks is program Pajek (Batagelj & Mrvar, 2002), which enables the analysis of more than a million of units. To assess the dimensions of hypertextuality of the WWW, several methods were used, which provide different characteristics of the structure of connectedness of web sites and also graphical presentation of the parts of the network to support interpretation.

4.1 Methods

The method of strong components identifies all strongly connected components in the network, which represent sets of units that are mutually accessible by following the hyperlinks. In the language of graph theory this means that between all the pairs of vertices there is at least one path. The method of weak components is very similar to the previous one with the difference that weak components are selected on the basis of assumption that relations in the network are undirected (Wasserman & Faust, 1994). Further on, method of k-cores identifies groups of vertices, which are connected at least with k-vertices from the same group (Batagelj & Mrvar, 2002).

To measure decentralisation and non-linearity, various measures of unit centrality and network centralisation were used. Degree of a unit $C_D(x)$ is equal to the number of hyperlinks, targeting or emerging from it, while outdegree is a number of hyperlinks, targeting this unit. Relative measure of degree is $C_D(x)/ (n-1)$, where n is number of units in a network (Wasserman & Faust, 1994). Closeness is a measure of centrality, which doesn't take into account only the closest neighbours, but also indirect ones: $C_C(x) = (n-1)/d(x,y)$, for each y from the set of units, where d(x,y) is a distance between two units. The next measure of unit centrality is betweenness $C_V(x)$, which is computed as a relation between the number of shortest paths between y and z, going through x, and the number of shortest paths between y and z. A special measure of unit centrality, hubs&authorities was developed within the research of important web pages in WWW (Kleinberg, 1998) and adapted to network analysis by Batagelj & Mrvar (2002). This method is based on the idea that hubs are units, providing links to other units, while authorities are units, which are targeted by many good hubs.

Network centralisation index shows the overall inequality between the units, based on their unit centrality, and is measured on an interval [0,1]. Higher values correspond to networks, where only one unit is dominant, while others are on the periphery. Various measures of network centralisation exist, depending on the measure of unit centrality (Wasserman & Faust, 1994).

4.2 Interconnectedness

The results of weak component analysis offer a basic insight into connectedness of the whole network of Slovenian web sites.

Table 2: Size and number of weak components in the Slovenian part of WWW (n=25247).

Number of units in weak component	Number of weak components
1	1710
2	21
23495	1

Regarding the fact that the membership in a weak component demands at least one hyperlink pointing to a web site or at least one deriving from it, it can be claimed that in the Slovenian part of WWW there are 1710 web sites, which offer no hyperlinks to other web sites, nor are they "cited" by any other web site. In other words, there are 6.8% of isolates.

Table 3: Size and number of weak components in a network without search systems (n=25212).

Number of units in weak comp.	Number of weak components
1	6505
2	156
3	17
4	4
6	1
8	1
18314	1

In the original network several search system are included, which offer many hyperlinks on other web pages, but are in the context of the research problem more a barrier to the hypertextuality of the WWW than its stimulator. Links from search systems resemble more classical hierarchical systems and don't correspond to the idea of associative link between logically connected documents. For this reason, several major search systems were excluded from the network and not surprisingly the method of weak components now results in 26.8% of isolates. Approximately a quarter of all web sites do not provide any hypertextual experience for the user at all and these web sites are not reachable by any other means than by search systems.

For a detailed analysis of interconnectedness of Slovenian WWW, the method of strong components seems more suitable since it identifies groups of mutually accessible web sites and considers the direction of hyperlinks, which makes it more realistic a method than weak components.

Table 4: Size and number of strong components in the network of Slovenian web sites (n=25212).

Number of units in strong comp.	Number of strong components
1	18431
2	95
3	12
4	3
5	2
6	1
8	2
10	1
6501	1

A strong component of size 6501 represents a group of web sites, which are at least indirectly mutually accessible A user can start browsing the information space of the WWW on any of these web sites and by following “outside” hyperlinks he can reach any other web site from this component. This strong component reflects at best the idea of interconnectedness of documents in hypertextual systems and it can be claimed that 26% of web sites are realising this essential dimension of the hypertext. On the other hand, however, there are 73% of web sites in a component of size 1, which means that they’re either “oneways”, “no ways”, or “isolates”10. Strong components of lesser sizes in general do not offer a true browsing experience of hypertextually organised information, since they are only small islands of interconnected web sites.

A relevant indicator of interconnectedness of web sites in the largest strong component is also diameter. Value of diameter in this case is 13 and shows that at most thirteen mouse clicks are needed to go from one web site of the component to the other, which implies that the component is quite densely connected. Albert et al. (1999) similarly concluded for the majority of web pages in the WWW that the distance between them is in average less than 20 hyperlinks. Although these results might confirm the phenomenon of “small world”, they should be taken with great care, since the results of this research clearly show that a large proportion of web sites are completely isolated, that is, they do not offer any hyperlinks nor are they pointed to by any hyperlink.

[10] "No way" represents a web site, which does not provide any outgoing links, while "oneway" represents a web site, to which no links from other web sites are pointing.

The global structure of the Slovenian part of the WWW could best be represented by the graph below, which somewhat resembles the well-known webgraph by Broder et al. (1999). The graph was computed by procedure of shrinking the whole network according to the partitioning, which was defined by linkage characteristics web sites. This way we arrive to a few basic subgroups of web sites.

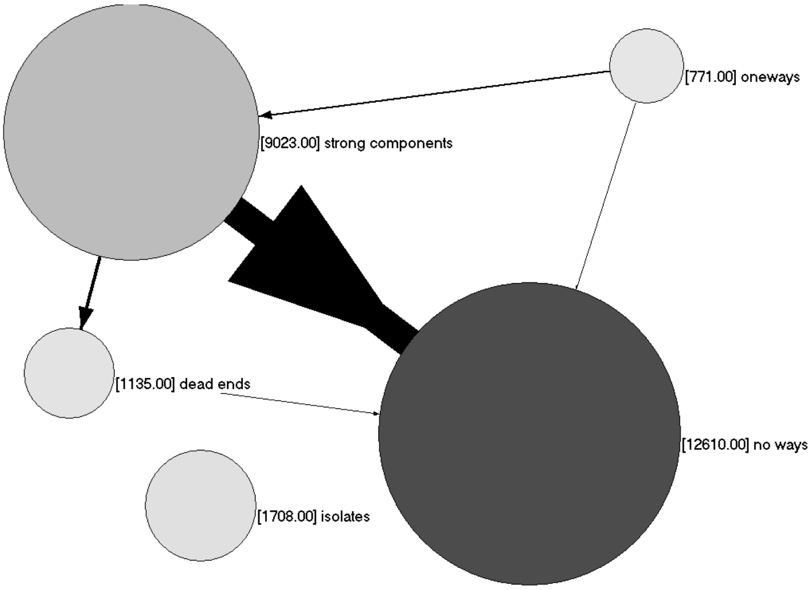

Figure 2: The structure of the Slovenian WWW (n=25247)

As already stated, quite large proportion of web sites (35.7%) are a part of the most connected part of the Slovenian WWW and these are pointed by a small proportion (3%) of "oneways". The largest group of web sites (49.9%) are "no ways", which are only pointed by other web sites and do not offer any outlinks. "Dead ends" is an acyclic group of web sites, which are pointed by strong components, but eventually lead to "no ways". The structure of the Sloveninan WWW is largely determined by web sites that disable hypertextual access to information. A user, who arrives to such a web site, cannot continue searching information by browsing through hyperlinks.

4.3 Decentralisation

The above results show that at least a small proportion of the Slovenian part of the WWW resembles structural properties of hypertextual systems and thus in principle enables users' liberation from rigid linear and hierarchical structures, yet more detailed insight into this question is needed by an analysis of the decentralisation of the WWW. A relevant indicator of decentralisation is the distribution of various measures of unit centrality, which should in ideal hypertextual systems be approximately equal. The distribution of the most basic measure of centrality, indegree centrality or the number of incoming hyperlinks, is following the power law and indicates a large asymmetry in the centralities of individual web sites (Figure 1).

The indegree centrality of web sites is in 90% of cases less than 10, meaning that the majority of web sites are pointed by at most 10 of other web sites. On the other hand, there are only 1% of web sites that have indegree more than 55 and 0.1% of web sites that have it larger than 500. In the whole network there are 130 web sites, which are cited by more than 100 other web sites and only 14, which are pointed by 400 or more web sites: that clearly indicates high centralisation of the network. This is, however, not supported by the centralisation index, which is 0.042, and implies that there is not only one single web site with a large indegree centrality. A user of WWW is not limited to a central axis of information organisation, but is nevertheless limited in information retrieval on a minority of more or less equal web sites, since the majority of web sites are on the periphery and only a small probability exists for a user to browse them.

Additional insight into (de)centralisation of the WWW and higher validity of results can be obtained by analysis of additional centrality measures. Betweenness indicates to what extent is a web site located on the path between other web sites and in this way exposes web sites, which a user will very probably visit in browsing for information. Web sites with high betweenness have high “control” over user, since they are located between pairs of web sites, which wouldn't be closely connected without them. Centralisation index (C_V=0.22), based on betweenness, indicates that the Slovenian part of WWW is quite centralised, which is the consequence of the very outlaying value of a single web site. This web site is not surprisingly search system Mat'kurja, due to which Slovenian web sites seem to be more connected than they really are, as was already indicated in the weak components analysis. Users of Internet, who search for information by browsing, sooner or later bump into a hyperlink to search system Mat'kurja. This, however, breaks their browsing path, since search system do not offer any associative hyperlinks but resemble more traditional modes of information organisation. The search systems rank high on all centrality measures, while web sites of several companies, which have high indegree, rank lower on measures, which take account of the broader vicinity of other web sites. This is a consequence of the fact that many authors provide hyperlinks on company web sites, but the latter don't provide any further hyperlinks.

Measures of centrality, which are based on the whole structure of the network, show that the importance of web sites is not as asymmetrically distributed as the data on indegree centrality would suggest. Distribution of nearness and authority weights shows some asymmetry, but doesn't point to high centralisation of the whole network.

Table 5: Distribution of authority weights of web sites in the Slovenian part of WWW (n=25247).

Authority weight	Number of web sites
less than 0,006	16306
0,006–0,008	5367
0,008–0,010	1984
0,01–0,02	1511
0,02–0,03	65
0,03–0,05	14

This measure is probably the most important measure of centralisation, since it is gives high authority weights to web sites, which are informative and thus reflect the idea of a informatively rich document in a hypertext. Data shows that some "documents" are more important than others, but generally it cannot be claimed that the Slovenian part of WWW is centralised, that there are only a few web sites, which get all the users. It is true that majority of web sites are on the periphery, but there is a large number of relatively equally important web sites. This is further investigated by the use of the k-cores method.

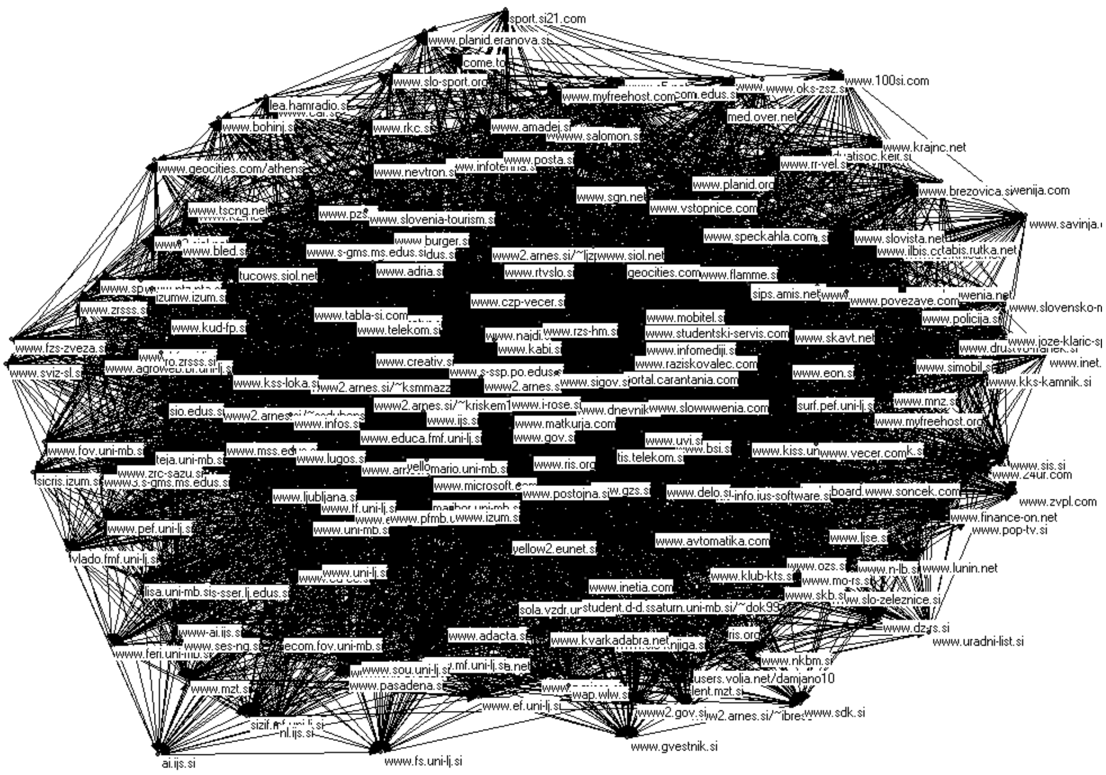

Figure 3: 16-core on a Slovenian part of the WWW (n=256).

The method of k-cores confirms the above findings: there is a large minority of web sites, which conform to the idea of hypertextuality, since they are densely connected and quite decentralised. The results show that there exists a 16-core of size 256, which means that these web sites are pointed by at least 16 other web sites in the group.

The group of 256 web sites shown above represents the most central and the most densely connected part of the Slovenian WWW. In terms of user experience, a user will very probably arrive at one of the above web sites in browsing the web and his further path of exploration of information space will be greatly determined by the linkage structure of these web sites. In the group of most densely connected web sites there are search systems, portals, university and government web sites, some companies, but also web sites of many other "smaller" social actors, which are on the periphery in terms of power relations. In short, these results do not offer any strong basis for making conclusive statements on the nature of decentralisation of the WWW, yet they implies that, notwithstanding the general asymmetric distributions of key characteristics, the WWW in certain parts offers a user an associative, creative and democratic experience of accessing information.

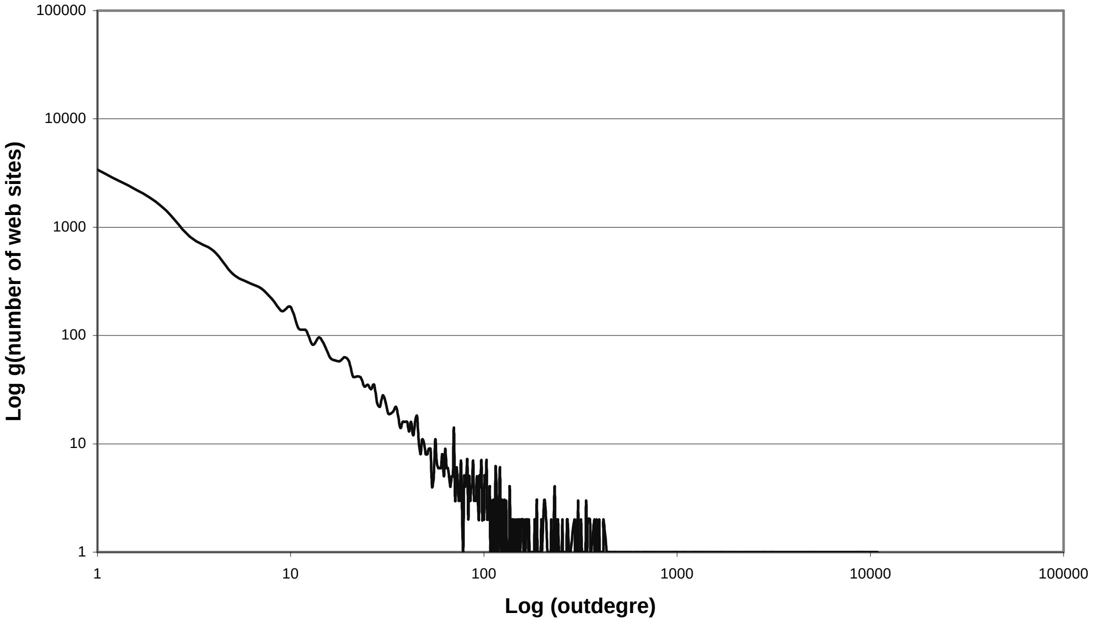

Figure 4: Distribution of outdegree in the Slovenian part of the WWW (n=25247).

4.4 Non-linearity

The third dimension of hypertextuality is mentioned mostly in contrast to the linearity of print documents, where user's path through the document is largely determined by the communication channel and author's intentions. The concept is simply measured as a proportion of web sites, offering a non-linear experience of browsing. The most suitable is the measure of outdegree, that is, the number of hyperlinks emerging from a web site.

Among 11909 web sites that offer any hyperlink to other web sites, there are 4491 offering only one hyperlink, which implies that user experience is to a large extent reduced to linear browsing, although the logic of WWW organisation is inherently non-linear. There are 7301 web sites, which offer from 2 to 100 hyperlinks to other web sites, and they reflect the ideal of non-linearity, so as the remaining 150 web sites, which have more than 100 outgoing hyperlinks. In the whole network this sums to 29.5% of web sites, which potentially stimulate freedom in access to information and independency from the author. Obviously, authors of web sites largely ignore "hypertextual ethics" (Petrič, 2003a) and expect from the user only to consume the content of their web site and discontinue search for information. However, structural information on non-linearity does not allow any strong conclusions. An important step further would be a thorough analysis of modes of hyperlinking by including the content of web sites (Jackson, 1997; Pajares-Tosca, 2001).

5 Conclusions

Although the methods used for assessment of hypertextuality of WWW on three dimensions do not allow any unambiguous conclusions, the data implies that only a minor part of the World Wide Web is interconnected, decentralised and non-linear. This is not so surprising due to several reasons. Firstly, it is a fact that the WWW is a large aggregate of a mass of various web sites, which are realisations of ideas of an unlimited set of social actors with particular intentions and interests. Links to other web sites are formed (if they are) with specific intentions and these intentions only rarely reflect the idea of a hyperlink connecting two logically associated texts. Secondly, many web sites are very small, consisting of only one web page and no out-links. These web sites were probably not meant to be published for large audiences but mostly for own reasons of experimentation with HTML and are as such not an important part of the WWW; yet in aggregate there are many of them (39.8%). Thirdly, the structure of the WWW is today largely determined by specific web sites, such as search systems and portals, and because of their large-scale connectivity, the results show that the structure of the WWW is more centralised than it probably really is.

It is probably true what several authors suggest (Bieber et al, 1999) that search systems and portals enable access to information that is the same as in classical rigid hierarchical systems and that they are discouraging the browsing experience. There seems to be a self-fulfilling prophecy in the process of hypertext disappearance from the WWW, since many of the authors of web sites do not provide links to substantively related web sites, but to the web portals, search systems and large corporations. Nevertheless, as the results already implied, the WWW should not be analysed in its totality but in its smaller, more informatively rich parts, which suggests that we cannot accept the hypothesis that the hypertext is totally absent from the WWW. If search systems and portals are excluded from the network, a third of all web sites are still strongly connected, mutually accessible. This group consists of various social, also marginal actors, whose documents on web sites are more or less equally important and likely accessible by users of the WWW who can in this way experience some freedom and creativity in accessing information.

The research of the structure of the Slovenian WWW offers only a limited insight into the question of hypertextuality and deals with it in a specific manner of analysing its essential three dimensions. The problem of hypertextuality and its implication for user experience is more complex than this analysis suggests, since it is dependent not only on the structure of information, but also on other factors, like browsers. One of the main aims of the research, however, was to give a basic insight into organisation of such a large information space as the WWW and primarily present several methodological issues that accompany such analysis, like unit definition, question of representativeness and network generation: issues, which are essential for any kind of structural analysis of the WWW.

References

[1] Albert, R., Jeong H., Barabasi, A. (1999): Diameter of the World Wide Web, *Nature*, **401**, 130–131.

[2] Allen, G. (2000): *Intertextuality*. London: Routhledge.

[3] Barabasi, A., Albert, R. (1999): Emergence of scaling in random networks. *Science*, **286**(509).

[4] Bardini, T. (2000): *Bootstrapping: Douglas Engelbart, Coevolution, and the Origins of Personal Computing.* Stanford: Stanford University Press.

[5] Barthes, R. (1974): *S/Z: An essay*. Paris: Editions de Seuil.

[6] Batagelj, V. and Mrvar, A. (2002): Pajek.
http://vlado.fmf.uni-lj.si/pub/networks/pajek

[7] Berners-Lee, T. (1989/1990): Information Management: A Proposal. http://www.w3.org/History/1989/proposal.html.

[8] Berners-Lee, T. (1995): Hypertext and Our Collective Destiny. http://www.w3.org/pub/WWW/talks.

[9] Berners-Lee, T. (1997): Axioms of Web architecture. http://www.w3.org/Architecture.

[10] Bieber, M., Vittali F., Ashman, H., Oinas-Kukkonen, H. (1997): Fourth generation hypermedia: Some missing links for the World Wide Web, International *Journal of Human Computer Studies*, **47**, 31–65. http://www.hbuk.co.uk/ap/ijhcs/webusability/.

[11] Bijker W.E., Hughes T.P., and Pinch T. J. (1987): *The Social Construction of Technological Systems: New directions in the Sociology and History of Technology*. London: MIT Press.

[12] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener J. (1999): Graph structure in the web. *Proceedings of the 9th WWW Conference*. http://www.almaden.ibm.com/cs/k53/www9/final/.

[13] Bush, V.. (1945): As We May Think. *Atlantic Monthly*, **176,** 101–108. http://www.isg.sfu.ca/~duchier/misc/vbush.

[14] Callon, M. (1987): Society in the making: The study of technology as a tool for sociological analysis. In E. Wieb, T.P. Bijker, T.J. Hughes, and J. Pinch (Eds.): *The Social Construction of Technological Systems: New directions in the Sociology and History of Technology*. London: MIT Press, 83–107.

[15] Deibert, R.J. (1997): *Parchment, Printing and Hypermedia: Communication in World Order Transformations*. New York: Columbia university press.

[16] Jackson, M. (1997): Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication*, **3**. http://www.ascusc.org/jcmc/vol3/issue1.

[17] Kleinberg, J. (1998): Authoratitve sources in hyperlinked enivorment. *Proceedigns of 9th ACM-SIAM SODA.*

[18] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999): Trawling the Web for cyber communities. *Proceedings of the 8th WWW conference. h*ttp://www8.org/w8.papers/trawling/.

[19] Landow, G.P. (1997): Hypertext 2.0: *The Convergence of Contemporary Critical Theroy and Technology*. London, Baltimore: The Johns Hopkins University Press.

[20] Laumann, E.O., Marsden, P.V., Prensky, D. (1983): The boundary specification problem in network analysis. In S. Ronald, M.J. Burt, and J. Minor (Eds.): *Applied Network Analysis: A Methodological Introduction*. Sage.

[21] Mosco, V. (2000): Political Economy. In T. Swiss (Ed.): *Unspun: Key Concepts for Understanding the World Wide Web*. New York: New York University Press, 51–66.

[22] O'Muircheartaigh, C. (1997): Measurment error in surveys: A historical perspective. In Lyberg et al. (Eds.): *Survey Measurment and Process Quality*. Wiley & Sons.

[23] Pajares-Tosca, S. (2001): A pragmatics of links. *Journal of Digital Information*, **1**. http://jodi.esc.soton.ac.uk/articles/v01

[24] Petrič, G. (2003a): Erozija hipertekstovne etike med avtorji spletnih mest. *Družboslovne razprave*, **19**, 119-142.

[25] Petrič, G. (2003b): Družbeno delovanje v omrežju svetovnega spleta: Individualni in strukturni vidik. Doctoral dissertation. Ljubljana: FDV.

[26] Retallack, D. (1999): In M. Bieber (ed.): Will hypertext become the Web's missing link. http://www7.scu.edu.au/programme/panels/1943/.

[27] Wasserman, S. and Faust, K. (1994): *Social Network Analysis. Methods and Applications.* New York: Cambridge Universiy Press.