Text Linkage in the Wiki Medium – A Comparative Study

Alexander Mehler
Department of Computational Linguistics & Text Technology
Bielefeld University
Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Abstract

We analyze four different types of document networks with respect to their small world characteristics. These characteristics allow distinguishing wiki-based systems from citation and more traditional text-based networks augmented by hyperlinks. The study provides evidence that a more appropriate network model is needed which better reflects the specifics of wiki systems. It puts emphasize on their topological differences as a result of wiki-related linking compared to other text-based networks.

1 Introduction

With the advent of web-based communication, more and more corpora are accessible which manifest complex networks based on intertextual relations. This includes the area of *scientific communication* (e.g. digital libraries as CiteSeer), *press communication* (e.g. the New York Times which links topically related articles), *technical communication* (e.g. the Apache Software Foundation's documentations of open source projects) and *electronic encyclopedia* (e.g. Wikipedia and its releases in a multitude of languages). These are sources of *large* corpora of web documents which are connected by *citation links* (digital libraries), *content-based add-ons* (online press communication) or *hyperlinks* to related lexicon articles (electronic encyclopedias).

Obviously, a corpus of such documents is more than a set of textual units. There is structure formation above the level of single documents which can be described by means of graph theory and network analysis (Newman, 2003). But what is new about this kind of structure formation? Or do we just have to face the kind of structuring which is already known from other linguistic networks?

This paper focuses on the specifics of networking in wiki-based systems. It tackles the following questions: *What structure do wiki-based text networks have? Can we expect a wiki-specific topology compared to more traditional (e.g. citation) networks? Or can we expect comparable results when applying network analysis to these emerging networks?* In the following sections, these questions are approached by example of a language specific release of the Wikipedia as well as by wikis for technical documentation. That is, we contribute to answering the question why wiki can be seen as something new compared to other text types *from the point of view of networking*.

In order to support this argumentation, section (2) introduces those network coefficients which are analyzed within the present comparative study. As a preprocessing step, section (3) outlines a webgenre model which in sections (4.1) and (4.2) is used to represent and extract instances of four types of document networks. This allows applying the coefficients of section (2) to these instances (section 4.3) and narrowing down wiki-based networks (section 5). The final section concludes and prospects future work.

2 Network Analysis

For the time being, the overall structure of complex networks is investigated in terms of *Small Worlds* (SW) (Newman, 2003). Since its invention by Milgram (1967), this notion awaited formalization as a measurable property of large complex networks which allows distinguishing small worlds from random graphs. Such a formalization was introduced by Watts & Strogatz (1998) who characterize small worlds by two properties: First, other than in regular graphs, any randomly chosen pair of nodes in a small world has, on average, a considerably shorter *geodesic distance*.[1] Second, compared to random graphs, small worlds show a considerably higher level of *cluster formation*.

In this framework, cluster formation is measured by means of the average fraction of the number $\nabla(v_i)$ of triangles connected to vertex v_i and the number $\veebar(v_i)$ of triples centered on v_i (Watts and Strogatz, 1998):[2]

$$C_2 = \frac{1}{n} \sum_i \frac{\nabla(v_i)}{\veebar(v_i)} \qquad (1)$$

Alternatively, the cluster coefficient C_1 computes the fraction of the number of triangles in the whole network and the number of its connected vertex triples. Further, the *mean geodesic distance* l of a network is the arithmetic mean of all shortest paths of all pairs of vertices in the network. Watts and Strogatz observe high cluster values and short average geodesic distances in small worlds which apparently combine cluster formation with shortcuts as prerequisites of efficient information flow. In the area of information networks, this property has been demonstrated for the WWW (Adamic, 1999), but also for co-occurrence networks (Ferrer i Cancho and Solé, 2001) and semantic networks (Steyvers and Tenenbaum, 2005).

In addition to the SW model of Watts & Strogatz, link distributions were also examined in order to characterize complex networks: Barabási & Albert (1999) argue that the vertex connectivity of social networks is distributed according to a scale-free power-law. They recur to the observation – confirmed by many social-semiotic networks, but not by instances of the random graph model of Erdős & Rényi (Bollobás, 1985) – that the number of links per vertex can be reliably predicted by a power-law. Thus, the probability $P(k)$ that a randomly chosen vertex interacts with k other vertices of the same network is approximately

$$P(k) \sim k^{-\gamma} \qquad (2)$$

Successfully fitting a power law to the distribution of out degrees of vertices in complex networks indicates "that most nodes will be relatively poorly connected, while a select minority of *hubs* will be very highly connected." (Watts, 2003, p.107). Thus, for a fixed number of links, the smaller the γ value, the shallower the slope of the curve in a log-log plot, the higher the number of edges to which the most connected hub is incident.

A limit of this model is that it views the probability of linking a source node to a target node to depend solely on the connectivity of the latter. In contrast to this, Newman (2003) proposes a model in which this probability also depends on the connectivity of the former. This is done in order to account for social networks in which vertices tend to be linked if they share certain properties (Newman and Park, 2003), a tendency which is called *assortative mixing*. According to Newman & Park (2003) it allows distinguishing social networks from non-social (e.g. artificial and biological) ones even if they are uniformly attributed as small worlds according to the model of Watts & Strogatz (1998). Newman & Park (2003) analyze assortative mixing of vertex degrees, that is, the correlation of the degrees of linked vertices. They confirm that this correlation is *positive* in the case of social, but *negative* in the case of technical networks (e.g. the Internet) which thus prove disassortative mixing (of degrees).

Although these SW models were applied to citation networks, WWW graphs, semantic networks and co-occurrence graphs, *and thus to a variety of linguistic networks*, a comparative study which focuses on wiki-based structure formation in comparison to other networks *of textual units* is missing so far. In this paper, we present such a study. That is, we examine SW coefficients which allow distinguishing wiki-based systems from more "traditional" networks. In order to do that, a generalized web document model is needed to uniformly represent the document networks to be compared. In the following section, a webgenre model is outlined for this purpose.

3 A Webgenre Structure Model

Linguistic structures vary with the functions of the discourses in which they are manifested (Biber, 1995; Karlgren and Cutting, 1994). In analogy to the *weak contextual hypothesis* (Miller and Charles, 1991) one might state that structural differences reflect functional ones as far as they are confirmed by a significantly high number of textual units and thus are identifiable as recurrent pat-

[1] The geodesic distance of two vertices in a graph is the length of the shortest path in-between.

[2] A triangle is a subgraph of three nodes linked to each other. Note that all coefficients presented in the following sections relate by default to undirected graphs.

terns. In this sense, we expect web documents to be distinguishable by the functional structures they manifest. More specifically, we agree with the notion of *webgenre* (Yoshioka and Herman, 2000) according to which the functional structure of web documents is determined by their membership in *genres* (e.g. of *conference websites*, *personal home pages* or *electronic encyclopedias*).

Our hypothesis is that what is common to instances of different webgenres is the existence of an implicit *logical document structure* (LDS) – in analogy to textual units whose LDS is described in terms of section, paragraph and sentence categories (Power et al., 2003). In the case of web documents we hypothesize that their LDS comprises four levels:

- *Document networks* consist of documents which serve possibly heterogenous functions if necessary independently of each other. A web document network is given, for example, by the system of websites of a university.

- *Web documents* manifest – typically in the form of websites – pragmatically closed acts of web-based communication (e.g. *conference organization* or *online presentation*). Each web document is seen to organize a system of dependent subfunctions which in turn are manifested by modules.

- *Document modules* are, ideally, functionally homogeneous subunits of web documents which manifest single, but dependent subfunctions in the sense that their realization is bound to the realization of other subfunctions manifested by the same encompassing document. Examples of such subfunctions are *call for papers*, *program presentation* or *conference venue organization* as subfunctions of the function of *web-based conference organization*.

- Finally, elementary *building blocks* (e.g. *lists*, *tables*, *sections*) only occur as dependent parts of document modules.

This enumeration does not imply a one-to-one mapping between functionally demarcated *manifested* units (e.g. modules) and *manifesting* (layout) units (e.g. web pages). Obviously, the same functional variety (e.g. of a personal academic home page) which is mapped by a website of dozens of interlinked pages may also be manifested by a single page. The many-to-many relation induced by this and related examples is described in more detail in Mehler & Gleim (2005).

The central hypothesis of this paper is that genre specific structure formation also concerns document networks. That is, we expect them to vary with respect to structural characteristics according to the varying functions they meet. Thus, we do *not* expect that different types of document networks (e.g. systems of genre specific websites vs. wiki-based networks vs. online citation networks) manifest homogeneous characteristics, but significant variations thereof. As we concentrate on coefficients which were originally introduced in the context of small world analyses, we expect, more concretely, that different network types vary according to their fitting *to* or deviation *from* the small world model. As we analyze only a couple of networks, this observation is bound to the corpus of networks considered in this study. It nevertheless hints at how to rethink network analysis in the context of newly emerging network types as, for example, Wikipedia.

In order to support this argumentation, the following section presents a model for representing and extracting document networks. After that, the SW characteristics of these networks are computed and discussed.

4 Network Modeling and Analysis

4.1 Graph Modeling

In order to analyse the characteristics of document networks, a format for uniformly representing their structure is needed. In this section, we present *generalized trees* for this task. Generalized trees are graphs with a kernel tree-like structure – henceforth called *kernel hierarchy* – superimposed by graph-forming edges as models of hyperlinks. Figure (1) illustrates this graph model. It distinguishes three levels of structure formation:

1. According to the webgenre model of section (3), *L1-graphs* map document networks and thus corpora of interlinked (web) documents.

In section (4.3), four sources of such networks are explored: *wiki document networks*, *citation networks*, *webgenre corpora* and, for comparison with a more traditional medium, *networks of newspaper articles*.

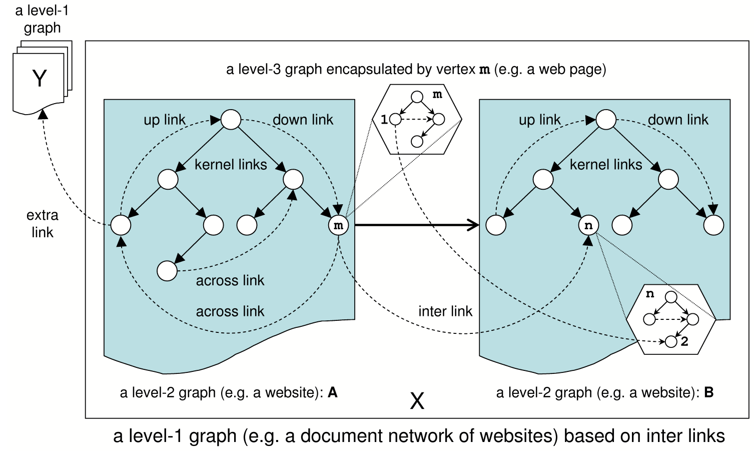
a level-1 graph
Y
a level-3 graph encapsulated by vertex m (e.g. a web page)
m
1
up link
down link
kernel links
extra link
m
across link
across link
inter link
up link
down link
kernel links
n
n
2
a level-2 graph (e.g. a website): A
X
a level-2 graph (e.g. a website): B
a level-1 graph (e.g. a document network of websites) based on inter links

Figure 1: The stratified model of network representation with kernel hierarchies of L2-graphs.

2. *L2-graphs* model the structure of web documents as constituents of a given network. This structure is seen to be based on kernel hierarchies superimposed, amongst others, by *up*, *down* and *across* links (see fig. 1).

In the case of webgenre corpora, L2-graphs model websites. In the case of citation networks, they map documents which consist of a scientific article and add-ons in the form of citation links. Likewise, in the case of online newspapers, L2-graphs model articles together with content-based hyperlinks. Finally, in the case of wikis, L2-graphs represent *wiki documents* each of which consists of a wiki article together with a corresponding discussion and editing page. According to the webgenre model of section (3), L2-graphs model web documents which consist of nodes whose structuring is finally described by L3-graphs:

3. *L3-graphs* model the structure of document modules.

In the case of webgenre corpora, *L3-graphs* map the DOM[3]-based structure of the *web pages* of the websites involved. In the case of all other networks distinguished above they represent the logical structure of single text units (e.g. the section and paragraph structuring of a lexicon, newspaper or scientific article). Note that the tree-like structure of a document module may be superimposed by hyperlinks, too, as illustrated in figure (1) by the vertices m and n.

The kernel hierarchy of an L2-graph is constituted by *kernel links* which are distinguished from *across*, *up*, *down* and *outside* links (Amitay et al., 2003; Eiron and McCurley, 2003; Mehler and Gleim, 2005). These types can be distinguished as follows:

- *Kernel links* associate dominating nodes with their immediately dominated successor nodes in terms of the kernel hierarchy.
- *Down links* associate nodes with one of their (mediately) dominated successor nodes in terms of the kernel hierarchy.
- *Up links* analogously associate nodes of the kernel hierarchy with one of their (mediately dominating) predecessor nodes.
- *Across links* associate nodes of the kernel hierarchy none of which is an (im-)mediate predecessor of the other in terms of the kernel hierarchy.
- *Extra* (or outside) *links* associate nodes of the kernel hierarchy with nodes of other documents.

Kernel hierarchies are exemplified by a *conference website* headed by a title and menu page referring to, for example, the corresponding *call for papers* which in turn leads to pages on the different conference sessions etc. so that finally a hierarchical structure evolves. In this example the kernel hierarchy evidently reflects navigational constraints. That is, the position of a page in the tree reflects

[3] I.e. D*ocument* O*bject* M*odel*.

the probability to be navigated by a reader starting from the root page and following kernel links only.

The kernel hierarchy of a wiki document is spanned by an *article page* in conjunction with the corresponding *discussion* (or *talk*), *history* and *edit this* or *view source* pages which altogether form a flatly structured tree. Likewise in the case of citation networks as the CiteSeer system (Lawrence et al., 1999), a document consists of the various (e.g. PDF or PS) versions of the focal article as well as of one or more web pages manifesting its citations by means of hyperlinks.

From the point of view of document network analysis, L2-graphs and inter links (see fig. 1) are most relevant as they span the corresponding network mediated by documents (e.g. websites) and modules (e.g. web pages). This allows specifying which links of which type in which network are examined in the present study:

- In the case of citation networks, citation links are modeled as interlinks as they relate (scientific) articles encapsulated by documents of this network type. Citation networks are explored by example of the CiteSeer system: We analyze a sample of more than 550,000 articles (see table 1) – the basic population covers up to 800,000 documents.

- In the case of newspaper article networks, content-based links are explored as resources of networking. This is done by example of the 1997 volume of the German newspaper Süddeutsche Zeitung (see table 1). That is, firstly, nodes are given by articles where two nodes are interlinked if the corresponding articles contain *see also* links to each other. In the online and ePaper issue of this newspaper these links are manifested as hyperlinks. Secondly, articles are linked if they appear on the same page of the same issue so that they belong to the same thematic field. By means of these criteria, a bipartite network (Watts, 2003) is built in which the top-mode is spanned by topic and page units, whereas the bottom-mode consists of text units. In such a network, two texts are interlinked whenever they relate to at least one common topic or appear on the same page of the same issue.

- In the case of webgenres we explore a corpus of 1,096 conference websites (see table 1 and 2) henceforth called `indogram` corpus.[4] We analyze the out degrees of all web pages of these websites and thus explore kernel, up, down, across, inter and outside links on the level of L2-graphs. This is done in order to get a *base line* for our comparative study, since WWW-based networks are well known for their small world behavior. More specifically, this relates to estimations of the exponent γ of power laws fitted to their degree distributions (Newman, 2003).

variable	value
number of web sites	1,096
number of web pages	50,943
number of hyperlinks	303,278
maximum depth	23
maximum width	1,035
average size	46
average width	38
average height	2

Table 2: A corpus of conference and workshop websites (counting unit: web pages).

- These three networks are explored in order to comparatively study networking in Wikipedia which is analyzed by example of its German release `de.wikipedia.org` (see table 1). Because of the rich system of its node and link types (see section 4.2) we explore three variants thereof. Further, in order to get a more reliable picture of wiki-based structure formation, we also analyze wikis in the area of technical documentation. This is done by example of three wikis on open source projects of the Apache Software Foundation (cf. `wiki.apache.org`).

In the following section, the extraction of Wikipedia-based networks is explained in more detail.

4.2 Graph Extraction – the Case of Wiki-based Document Networks

In the following section we analyze the network spanned by document modules of the German Wikipedia and their inter links.[5] This cannot simply be done by extracting all its article pages. The reason is that Wikipedia documents consist

[4] See `http://ariadne.coli.uni-bielefeld.de/indogram/resources.html` for the list of URLs of the documents involved.

[5] We downloaded and extracted the XML release of this wiki – cf. `http://download.wikimedia.org/wikipedia/de/pages_current.xml.bz2`.

network	network genre	node	$\|V\|$	$\|E\|$
`de.wikipedia.org`	electronic encyclopedia	wiki unit		
`variant I`		(e.g. article or talk)	303,999	5,895,615
`variant II`			406,074	6,449,906
`variant III`			796,454	9,161,706
`wiki.apache.org/jakarta`	online technical documentation	wiki unit	916	21,835
`wiki.apache.org/struts`	online technical documentation	wiki unit	1,358	40,650
`wiki.apache.org/ws`	online technical documentation	wiki unit	1,042	23,871
`citeseer.ist.psu.edu`	digital library	open archive record	575,326	5,366,832
`indogram`	conference websites genre	web page	50,943	303,278
`Süddeutsche Zeitung 1997`	press communication	newspaper article	87,944	2,179,544

Table 1: The document networks analyzed and the sizes $|V|$ and $|E|$ of their vertex and edge sets.

of modules (manifested by pages) of various types which are likewise connected by links of different types. Consequently, the choice of instances of these types has to be carefully considered.

Table (3) lists the node types (and their frequencies) as found in the wiki or additionally introduced into the study in order to organize the type system into a hierarchy. One heuristic for extracting instances of node types relates to the URL of the corresponding page. Category, portal and media wiki pages, for example, contain the prefix `Kategorie`, `Portal` and `MediaWiki`, respectively, separated by a colon from its page name suffix (as in `http://de.wikipedia.org/wiki/Kategorie:Musik`).

Analogously, table (4) lists the edge types either found within the wiki or additionally introduced into the study. Of special interest are *redirect* nodes and links which manifest transitive and, thus, mediate links of content-based units. An article node v may be linked, for example, with a redirect node r which in turn redirects to an article w. In this case, the document network contains two edges $(v, r), (r, w)$ which have to be resolved to a single edge (v, w) if redirects are to be excluded in accordance with what the MediaWiki system does when processing them.

Based on these considerations, we compute network characteristics of three extractions of the German Wikipedia (see table 1): *Variant I* consists of a graph whose vertex set contains all *Article* nodes and whose edge set is based on *Interlink*s and appropriately resolved *Redirect* links. *Variant II* enlarges variant I by including other content-related wiki units, i.e. *ArticleTalk*, *Portal*, *PortalTalk*, and *Disambiguation* pages (multiply typed nodes were excluded). *Variant III* consists of a graph whose vertex set covers all vertices and edges found in the extraction.

Type	Frequency
Documents total	796,454
Article	303,999
RedirectNode	190,193
Talk	115,314
ArticleTalk	78,224
UserTalk	30,924
ImageTalk	2,379
WikipediaTalk	1,380
CategoryTalk	1,272
TemplateTalk	705
PortalTalk	339
MediaWikiTalk	64
HelpTalk	27
Image	97,402
User	32,150
Disambiguation	22,768
Category	21,999
Template	6,794
Wikipedia	3,435
MediaWiki	1,575
Portal	791
Help	34

Table 3: The system of node types and their frequencies within the German Wikipedia.

4.3 Network Analysis

Based on the input networks described in the previous section we compute the SW coefficients described in section (2). Average geodesic distances are computed by means of the Dijkstra algorithm based on samples of 1,000 vertices of the input networks (or the whole vertex set if it is of minor cardinality). Power law fittings were computed based on the model $P(x) = ax^{-\gamma} + b$. Note that table (1) does not list the cardinalities of multi sets of edges and, thus, does not count multiple edges connecting the same pair of vertices within the corresponding input network – therefore, the numbers in table (1) do not necessarily conform to the counts of link types in table (4). Note further that we compute, as usually done in SW analyses, characteristics of *un*directed graphs. In the case of wiki-based networks, this is justified by the possibility to process *back links* in `Media Wiki` systems. In the case of the CiteSeer system this is justified by the fact that it always displays *citation*

Type	Frequency
Links total	17,814,539
Interlink	12,818,378
CategoryLink	1,415,295
Categorizes	704,092
CategorizedBy	704,092
CategoryAssociatesWith	7,111
TopicOfTalk	103,253
TalkOfTopic	88,095
HyponymOf	26,704
HyperonymOf	26,704
InterPortalAssociation	1,796
Broken	2,361,902
Outside	1,276,818
InterWiki	789,065
External	487,753
Intra	1,175,290
Kernel	1,153,928
Across	6,331
Up	6,121
Reflexive	5,433
Down	3,477
Redirect	182,151

Table 4: The system of link types and their frequencies within the German Wikipedia.

and *cited by* links. Finally, in the case of the newspaper article network, this is due to the fact that it is based on a bipartite graph (see above). Note that the `indogram` corpus consists of predominantly unrelated websites and thus does not allow computing cluster and distance coefficients.

5 Discussion

The numerical results in table (5) are remarkable as they allow identifying three types of networks:

- On the one hand, we observe the extreme case of the `Süddeutsche Zeitung`, that is, of the newspaper article network. It is the only network which, at the same time, has very high cluster values, short geodesic distances *and* a high degree of assortative mixing. Thus, its values support the assertion that it behaves as a small world in the sense of the model of Watts & Strogatz. The only exception is the remarkably low γ value, where, according to the model of Barabási & Albert (1999), a higher value was expected.
- On the other hand, the CiteSeer sample is the reverse case: It has very low values of C_1 *and* C_2, tends to show neither assortative, nor disassortative mixing, and at the same time has a low γ value. The small cluster values can be explained by the low probability with which two authors cited by a focal article are related by a citation relation on their own.[6]
- The third group is given by the wiki-based networks: They tend to have higher C_1 and C_2 values than the citation network does, but also tend to show stochastic mixing and short geodesic distances. The cluster values are confirmed by the wikis of technical documentation (also w.r.t their numerical order). Thus, these wikis tend to be small worlds according to the model of Watts & Strogatz, but also prove disassortative mixing – comparable to technical networks *but in departure from social networks*. Consequently, they are ranked in-between the citation and the newspaper article network.

All these networks show rather short geodesic distances. Thus, l seems to be inappropriate with respect to distinguishing them in terms of SW characteristics. Further, all these examples show remarkably low values of the γ coefficient. In contrast to this, power laws as fitted in the analyses reported by Newman (2003) tend to have much higher exponents – Newman reports on values which range between 1.4 and 3.0. This result is only realized by the `indogram` corpus of conference websites, thus, by a sample of WWW documents whose out degree distribution is fitted by a power law with exponent $\gamma = 2.562$.

These findings support the view that compared to WWW-based networks wiki systems behave more like “traditional” networks of textual units, *but are new in the sense that their topology neither approximates the one of citation networks nor of content-based networks of newspaper articles.* In other words: As intertextual relations are genre sensitive (e.g. citations in scientific communication vs. content-based relations in press communication vs. hyperlinks in online encyclopedias), networks based on such relations seem to inherit this genre sensitivity. That is, for varying genres (e.g. of scientific, technical or press communication) differences in topological characteristics of their instance networks are expected. The study presents results in support of this view of the genre sensitivity of text-based networks.

6 Conclusion

We presented a comparative study of document networks based on small world characteristics.

[6] Although articles can be expected which cite, for example, de Saussure and Chomsky, there certainly exist much less citations of de Saussure in articles of Chomsky.

instance	type	$\langle d \rangle$	l	γ	C_1	C_2	r
Wikipedia `variant I`	undirected	19.39	3.247	0.4222	0.009840	0.223171	-0.10
Wikipedia `variant II`	undirected	15.88	3.554	0.5273	0.009555	0.186392	-0.09
Wikipedia `variant III`	undirected	11.50	4.004	0.7405	0.007169	0.138602	-0.05
`wiki.apache.org/jakarta`	undirected	23.84	4.488	0.2949	0.193325	0.539429	-0.50
`wiki.apache.org/struts`	undirected	29.93	4.530	0.2023	0.162044	0.402418	-0.45
`wiki.apache.org/ws`	undirected	22.91	4.541	0.1989	0.174974	0.485342	-0.48
`citeseer.ist.psu.edu`	undirected	9.33	4.607	0.9801	0.027743	0.067786	-0.04
`indogram`	directed	5.95	× × ×	2.562	× × ×	× × ×	× × ×
`Süddeutsche Zeitung`	undirected	24.78	4.245	0.1146	0.663973	0.683839	0.699

Table 5: Numerical values of SW-related coefficients of structure formation in complex networks: the average number $\langle d \rangle$ of edges per node, the mean geodesic distance l, the exponent γ of successfully fitted power laws, the cluster values C_1, C_2 and the coefficient r of assortative mixing.

According to our findings, three classes of networks were distinguished. This classification separates wiki-based systems from more traditional text networks but also from WWW-based webgenres. Thus, the study provides evidence that there exist genre specific characteristics of text-based networks. This raises the question for models of network growth which better account for these findings. Future work aims at elaborating such a model.

References

Lada A. Adamic. 1999. The small world of web. In Serge Abiteboul and Anne-Marie Vercoustre, editors, *Research and Advanced Technology for Digital Libraries*, pages 443–452. Springer, Berlin.

Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, and Aya Soffer. 2003. The connectivity sonar: detecting site functionality by structural patterns. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia*, pages 38–47.

Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science*, 286:509–512.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Béla Bollobás. 1985. *Random Graphs*. Academic Press, London.

Nadav Eiron and Kevin S. McCurley. 2003. Untangling compound documents on the web. In *Proceedings of the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK*, pages 85–94.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The small-world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265, November.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of COLING '94*, volume II, pages 1071–1075, Kyoto, Japan.

Steve Lawrence, C. Lee Giles, and Kurt Bollacker. 1999. Digital libraries and Autonomous Citation Indexing. *IEEE Computer*, 32(6):67–71.

Alexander Mehler and Rüdiger Gleim. 2005. The net for the graphs — towards webgenre representation for corpus linguistic studies. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as corpus*. Gedit, Bologna, Italy.

Stanley Milgram. 1967. The small-world problem. *Psychology Today*, 2:60–67.

George A. Miller and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Mark E. J. Newman and Juyong Park. 2003. Why social networks are different from other types of networks. *Physical Review E*, 68:036122.

Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review*, 45:167–256.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Mark Steyvers and Josh Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.

Duncan J. Watts. 2003. *Six Degrees. The Science of a Connected Age*. Norton & Company, New York.

Takeshi Yoshioka and George Herman. 2000. Coordinating information using genres. Technical report, Massachusetts Institute of Technology, August.