

The Effects of Web Logs and the Semantic Web on Autonomous Web Agents

Michael P. Evans¹, Richard Newman¹, Timothy A. Millea¹, Timothy Putnam¹, and Andrew Walker²

¹ Applied Software Engineering Group, School of Systems Engineering,
The University of Reading, Reading, UK

{michael.evans,r.newman,t.a.millea,t.putnam}@reading.ac.uk

² School of Mathematics, Kingston University,
Kingston-upon-Thames, Surrey, UK

Abstract Search Engines exploit the Web's hyperlink structure to help infer information content. The new phenomenon of personal Web logs, or 'blogs', encourage more extensive annotation of Web content. If their resulting link structures bias the Web crawling applications that search engines depend upon, there are implications for another form of annotation rapidly on the rise, the Semantic Web. We conducted a Web crawl of 160 000 pages in which the link structure of the Web is compared with that of several thousand blogs. Results show that the two link structures are significantly different. We analyse the differences and infer the likely effect upon the performance of existing and future Web agents. The Semantic Web offers new opportunities to navigate the Web, but Web agents should be designed to take advantage of the emerging link structures, or their effectiveness will diminish.

Key words: Web crawling, Web graph, Semantic Web, blogs, search engine, Web agent.

1 Introduction

The World Wide Web is an open and distributed system. With content published at the rate of 1.5 million pages per day, expected mean persistence of just 18 days[1], and an architecture that can be freely extended according to need, the Web may also be regarded as an evolving system.

This evolution is reflected in the Web's hyperlink structure, the set of all Web pages and their hyperlinks known as the *Web Graph*, which is fuzzy in nature, and constantly changing over several different time scales[2]. Despite this, several empirical studies have identified many stable properties, which are exploited by search engines in order to classify and rank Web pages, and by navigational agents in order to find information of behalf of users.

Recently, however, two new trends have emerged that may impact upon the performance of these applications:

- The growing use of Web logs, or *blogs* (online journals published by individuals), which foster a different style of Web usage that could change the Web Graph
- The development of the Semantic Web, which aims to enrich the Web with machine-understandable metadata, enabling a new generation of intelligent Web applications.

To determine the effect these technologies will have on Web crawling applications, we performed the following:

- An analysis of the Web Graph for the Web as a whole and for the sub-set comprising blogs, using our own stochastic heuristic-based Web crawler over some 160,000 Web pages

- Based on this analysis, an assessment of the impact the Semantic Web is likely to have on such applications, given the way the technology is currently developing and the challenges facing it.

The paper is organized as follows: Section 2 provides a background discussion on these technologies together with an overview of related work. Section 3 presents the design of our Web crawler and compares its results with similar Web crawls to validate its accuracy. Section 4 presents an analysis of the impact of annotation technologies upon the evolution of the Web. Section 5 looks ahead to the new Semantic Web technologies, and assesses the effect that semantically-rich metadata will have upon existing Web crawling applications. The paper concludes by envisaging the capabilities of the next generation of Web crawlers.

2 Background & Related Work

Since the Web was first created, there has been a need to search it. As the Web has grown, manual browsing has been assisted through increasingly sophisticated automated search engines, which help identify and locate relevant material. Search engines rely upon Web crawlers that systematically navigate and index the Web, and infer meaning from the information content of individual pages and their link structure. As such, any changes to this content or to the link structure, such as those introduced by the use of blogs or the Semantic Web, may therefore have a dramatic impact on the effectiveness of a crawler in producing balanced and pertinent results.

2.1 Current Structure of the Web Graph

The structure of the Web Graph has been well documented with many studies identifying its key properties. For example, Broder *et al.* performed a Web crawl[3] on 200 million nodes and identified what was termed a ‘Bow Tie’ structure, containing four separate, well-defined components:

- A core, forming a Strongly Connected Component (SCC), in which pages can reach others via directed paths
- An IN component, in which pages contain hyperlinks that link to the SCC, but which are not themselves linked to by pages from the SCC
- An OUT component, in which pages contain hyperlinks are linked to by pages from the SCC, but which do not themselves link to any pages in the SCC
- Tendrils, which can be reached by a path from a page in the IN component, or link to a page in the OUT component, but which have no connection to the SCC.

Other studies have discovered that the Web Graph exhibits ‘Small World’ features[4], and a range of properties that follow Power Laws have been identified, such as the distribution of site sizes[5], and the distribution of the number of hyperlinks per Web page[4].

Web crawlers and navigational agents rely on these findings to:

- Search for content across the Web[6,7]
- Identify authoritative pages[8]
- Identify communities of related Web pages[9,10,11]
- Rank pages according to the number of links connecting to them[12].

A significant change in the Web Graph will therefore affect the performance of the Web crawlers that the search engines rely upon. Although existing studies into the Web Graph have provided revealing insights into its structure, none has yet dealt with the changing use of the Web caused by blogs and the Semantic Web. They may therefore have missed a new evolutionary trend in the Web’s underlying link structure.

2.2 The Impact of Blogging on the Web Graph

‘Blogging’ has become something of a new social phenomenon of an increasingly large subset of the Web’s population. Estimates put the number of active blogs at around 1,880,000[13], with a growth rate of 105% per year[14]. However, blogs increase the degree of Web interactivity, with content being more easily authored, linked to and commented upon (using software such as MoveableType (<http://www.moveabletype.org>) and Blogger (<http://www.blogger.com>)). This leads to a different style of web content: more richly linked than traditional web pages. As such, the combination of the dramatic growth in blogs, coupled with the change in content they introduce, may be changing the structure of the Web Graph.

2.3 The Impact of the Semantic Web

In 1989, Tim Berners-Lee proposed a hypertext system for CERN to manage their documents and information in the face of staff changes[15]. This hypertext system, named the ‘Mesh’, would consist of typed nodes (*e.g.* people or software modules) connected by typed links (*e.g.* ‘refers to’, ‘made’) representing relationships between the nodes.

This proposal evolved into the Web as we know it: display-oriented pages containing simple hyperlinks to other pages. The Semantic Web is closer to Berners-Lee’s original vision of the Web: a machine-understandable Web of *meaning*[16], which attempts to address the issues of context, querying, provenance and trust.

The Semantic Web is navigated by software agents, rather than directly by humans. As an example of the likely impact the Semantic Web will have, consider the search for information. Currently, web crawlers must index the blind search of millions of pages to enable a search engine to give some potentially relevant results. Languages of the Semantic Web, such as RDF[17] and OWL[18,19], enable Web resources to be meaningfully annotated in a machine-readable form, with rich meta-data embedded within links. Semantic Web agents are being developed to perform inference upon this data[20,21] and generate, respond to and refine queries[22]. Centralised Web search engines with their huge cached repositories would therefore be replaced by lightweight directory services and local agents able to navigate the Web directly by meaning.

The Semantic Web has enormous implications for users of the Web in the way in which they store, interrogate, share and interact with information. The richer meta-data of its links, and emphasis on autonomous navigation, will also impact the structure of the Web Graph and the operation of crawlers and other agents that exploit it.

2.4 Assessing the Impact on Web Crawlers

Both blogs and the Semantic Web have the potential to transform the structure of the Web Graph, and thus the performance of Web Crawlers. To assess the impact these technologies may have, we designed a Web crawler with random heuristics to sample a random subset of the Web. The data was validated by comparing it with other large-scale crawls, and analysed for any change in the Web Graph caused by blogs. The results are presented in Section 4. To estimate the impact of the Semantic Web, we used existing information on its current state of development, and present our analysis in Section 5.

3 Designing a Web Crawler to Sample the Web

3.1 Crawler Design

For our Web crawler, we represent the Web Graph as the directed graph $G = (V, E)$, where $V = \{v_1, v_2, \dots, v_n\}$, the set of vertices representing Web pages, and E the collection of edges

representing hyperlinks ('links') that connect the Web pages. Thus, G represents the structure of the Web in terms of its pages and links. When sampling Web pages, a random walk across G is required, using a stochastic process that iteratively visits the vertices of the graph G [23].

A random walk across G should generate a finite-state discrete-time Markov chain, in which each variable v_n (where $v_n \in V$) in the chain is independent of all other variables in the chain. Thus, the probability of reaching the next Web page should be independent of the previous Web page, given the current state[2].

The following represents an overview of the crawling algorithm:

1. Crawling:
 - (a) Download the page referenced by the URL submitted to the crawler
 - (b) Search the page for links and other information about the page
 - (c) Resolve local addresses used in links to absolute URLs through a *URL Resolver*, which converts links into the form `http://subdomain.hostname.tld/path/file.ex`
 - (d) Remove links deemed to be of no use (*e.g.* undesired file type, such as executables)
 - (e) Check the database to see if the page has already been crawled.
2. Record in the database all URLs found on the page.
3. Randomly select a resolved URL from the database and repeat the process by submitting the URL to the crawler.

3.2 Validating the Results of the Web Crawl

We ran the crawler between 2003-04-28 and 2003-07-29 with 31 users blindly interacting with it. Some 160,000 URLs were collected. Once the crawl was complete, we analyzed the Web pages referenced by these URLs, and used the statistics obtained to compare the results with those of other Web crawlers from other studies in order to validate the results. Note that the number of Web pages indexed by Google stands at 4,285,199,774, as of April 2004. Of these it is estimated that only 1.88 million are blog pages[13]. Although we suspect that the hyperlink structure of blog pages will be different from that of general Web pages, the number of blogs should not have a noticeable effect on the Web Graph as a whole. Consequently, comparing the structure of the Web Graph from that determined by other crawls is still valid.

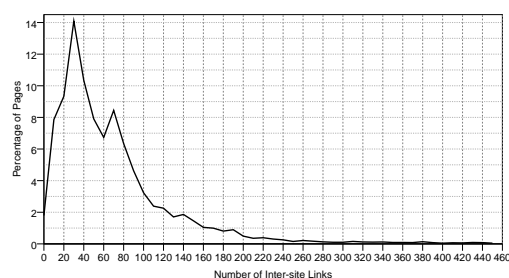


Figure 1. Distribution of inter-site links.

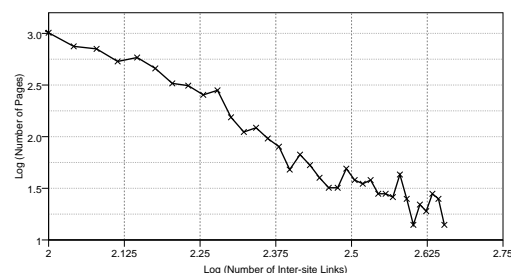


Figure 2. Log-Log plot of inter-site links, revealing Power Law.

Figures 1 and 2 show the distribution of *inter-site links* (*i.e.* those on a Web page that reference another Web page in a different domain) across the different Web pages crawled by our crawler and give a good indication of the underlying structure of the Web Graph. The mean number of links was found to be 48, and the median 77.8. Both Figures 1 and 2 clearly show the power law that exists in the Web Graph and compare well with similar results by Broder *et al.*[3], Henzinger

et al.[24], Barabasi and Bonabeau[25], and Huberman and Adamic[5]. In particular, the line of best fit for Figure 2 reveals an exponent of 3.28, which compares well with Kumar *et al.*'s value of 2.72[26], obtained with a crawl of some 200 million Web pages. Thus, the similarity of the structure revealed by our results to that identified by other, more extensive studies, validates the effectiveness of our Web crawling heuristics in revealing the structure of the Web Graph to a good approximation.

4 Determining the Impact of Blogs

In order to determine the impact of blogs on the performance of Web agents, we separated the Web pages we found into two different categories: general Web pages and those Web pages we determined to be blog pages. For each category we also separated the pages into those we classified as *homepages* (*i.e.* entry pages into the site) and *non-homepages* (*i.e.* all others). We analysed each category to determine the hyperlink structure for each.

4.1 Distribution of Inter-Site Links for Ordinary Homepages and Blog Homepages

Our first analysis focused on the link structure of homepages. Homepages are the front pages of a Web site, and are usually the page through which most visitors to the site will arrive. As such the homepage could be thought of as serving a slightly different purpose to pages on the rest of the site, and so it is possible pages categorised as homepages may exhibit different traits to other pages. For the purposes of our analysis, we classified a homepage as a URL with no query string and one of the following potential file paths:

- /
- /index (with any file extension)
- /default (with any file extension)

Once we had identified the homepages, we then separated those homepages belonging to traditional Web sites from those belonging to blogs, identifying a blog as a Web page with the word 'blog' in either its URL, or page title. Once complete, we charted the Inter-Site Link distribution of each. We found the results to be quite striking.

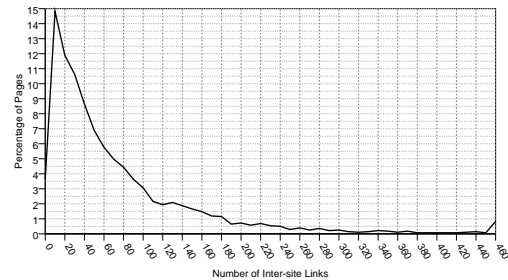


Figure 3. Inter-site link distribution for homepages.

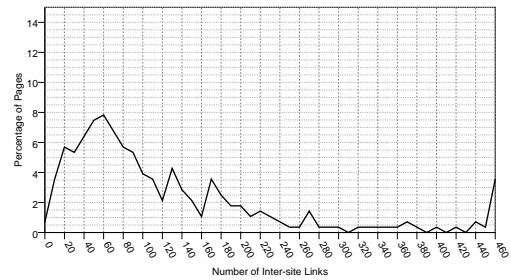


Figure 4. Inter-site link distribution for blog homepages.

Although the asymptotic trend evident in Figure 3 still exists in Figure 4, it is significantly less smooth. Charting a Log-Log plot on each graph reveals that the Power Law that was evident in

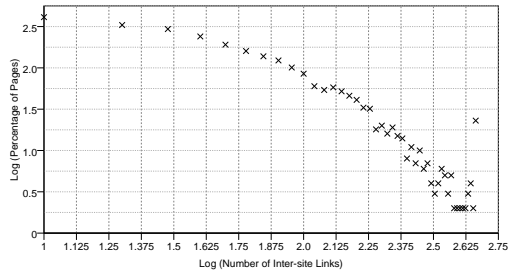


Figure 5. Log-Log plot of inter-site link distribution for all homepages.

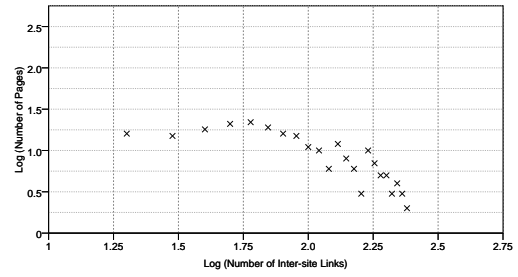


Figure 6. Log-Log plot of inter-site link distribution for blog homepages.

Figure 2 (Log-Log plot of the Inter-Site Link distribution for all Web pages) is less pronounced for all homepages (Figure 5), and has broken down completely for blog homepages (Figure 6).

The reason for this becomes clear when we examine the distribution of the percentage of *Inter-Site Links* and *Intra-Site Links* (*i.e.* those links to other pages within the same site), as shown in Figures 7 and 8.

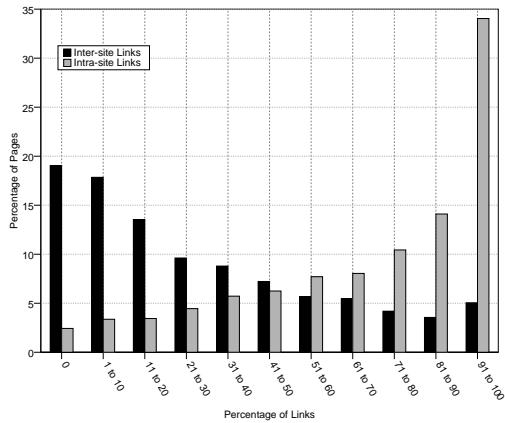


Figure 7. Distribution of the percentage of links on homepages, split into inter-site and intra-site links.

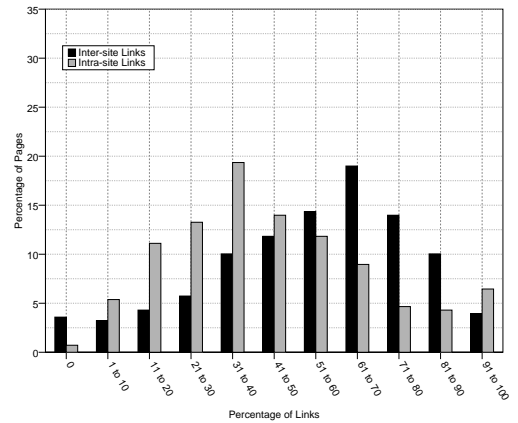


Figure 8. Distribution of the percentage of links on blog homepages, split into inter-site and intra-site links.

As is clearly evident, the pattern of links between a standard Web homepage and a blog homepage are significantly different. In particular, we found that it was most common (34.5%) for traditional Web homepages to have up to 10% of the total Inter-Site Links. In contrast, blog homepages most commonly (18.996% of pages) seem to have between 61% and 70% Inter-Site Links.

The more richly connected nature of the blog homepage becomes even more evident when we chart the distribution of all Inter-Site Links for standard homepages and blog homepages (Figures 9 and 10). We conjecture that the use of blogrolling (a constantly changing list of blogs included on the blog homepage[27]) is a major factor in this disproportionate number of links in a blog homepage.

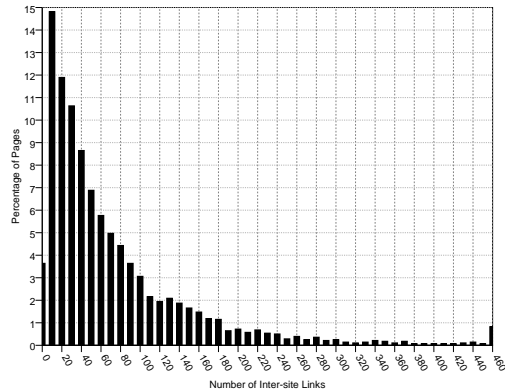


Figure 9. Distribution of total inter-site links for all homepages.

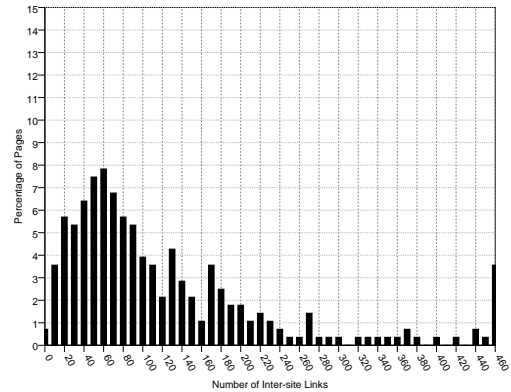


Figure 10. Distribution of total inter-site links for blog homepages.

4.2 Distribution of Inter-Links for Ordinary Non-Homepages

When analysing pages other than homepages, the difference in link structure is less pronounced, as Figures 11 and 13 show. However, although the link structures for non-homepages differ less than for the homepage comparisons, there is still a marked difference between the link structure of a blog non-homepage and that of an general Web page, as the log-log plots show (Figures 12 and 14).

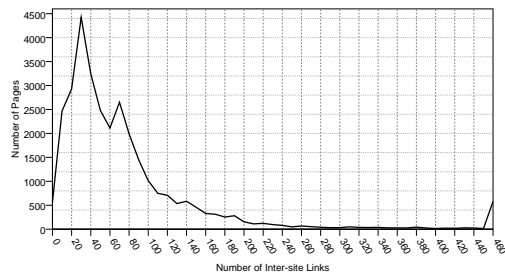


Figure 11. Inter-site link distribution for all non-homepages.

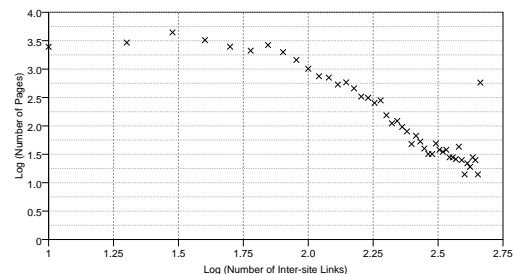


Figure 12. Log-Log plot of inter-site link distribution for all non-homepages.

4.3 Assessing the Impact of Blogs on Current Web Crawler Technology

The difference in link structure between blogs and general Web pages currently has no significant impact on the overall Web Graph due to the relatively low proportion of blog pages. However, if blog pages rise as a proportion of the Web as expected, the change in link structure will become readily apparent. Indeed, anecdotal evidence suggests that blogs are already having an impact on various search engines' ranking algorithms due to the distribution of Inter-Site Links within their pages[27].

Our results show that the link structure for blog pages is significantly different than for general Web pages. We suggest that the richness of the blogs' link structure can be attributed to the fact

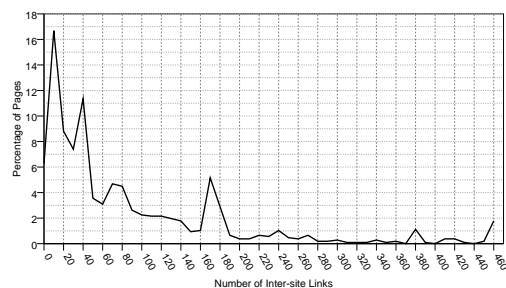


Figure 13. Inter-site link distribution for blog non-homepages.

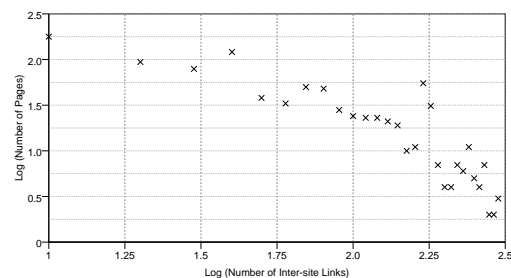


Figure 14. Log-Log plot of inter-site link distribution for blog non-homepages.

that blogs tend to cluster into communities as authors link from their blog to other blogs they find of interest.

Applications that make use of the Web Graph, such as Web crawlers, ranking algorithms and generative models of the Web, must take care to ensure that the sample of Web pages crawled is sufficiently random to be representative of the Web as a whole. Even today, the degree of cross linking between blogs can bias the sample set of pages of a Web crawler considerably.

The rich linking and community-oriented nature of the blogs presents a real problem to Web crawling heuristics that attempt to identify communities by the nature of their links. Such heuristics infer that topics are related whenever a statistically-significant high density of links is found compared with the background density of the Web Graph[2]. However, as we have shown, this change in density may simply represent a set of unrelated blogs, which themselves will exhibit a higher density of links, irrespective of their topical relationship to one another.

More positively, the significant difference between the link structures of blogs and of the general Web could be exploited by Web crawling applications to identify blogs and make adequate compensation for them in returned results. Furthermore, the relative confidence in which they may be identified enables their inclusion, or exclusivity, to be further search parameters.

5 Estimating the Impact of the Semantic Web

Blog posts typically take the form of an annotation: the thoughts of the author on a link, film, person, book, or another post. The richness of linking is as a consequence of the inevitable cross-referencing to the subject of the post, and to related commentaries.

This has clear parallels with the emerging Semantic Web. Consider a blog post discussing a film. That post may link heavily — to a DVD trader, to reviews, to the official site — and will make a number of statements in natural language. The same content could be presented on the Semantic Web through statements expressing formally-specified relationships between the author, the film, and other Web entities³.

These statements act as links on the Semantic Web — for instance, allowing navigation to the film’s official site. Here an agent may find further useful statements about the film.

The Semantic Web is founded on annotation and recombination, with reuse of ontologies and combination of information from different sources being key aspects. In this way, all Semantic Web documents will tend towards the ‘blog model’ — most documents will refer heavily to resources and

³ There are real initiatives to develop ontologies for this kind of purpose: *e.g.* the RVW ontology for reviews (<http://www.pmbrowser.info/hublog/archives/000307.html>).

ontologies on other sites, rather than taking the ‘Web model’ by relying on self-contained content. The Web Graph will change as a result, and the assumptions that crawlers use (*e.g.* that a link to a page has value, and asserts a positive relationship) will become increasingly flawed. However, the assumptions will no longer be necessary to achieve current goals, because the semantics of the ‘links’ will be explicitly provided in the document, as Berners-Lee intended.

Instead, we may look forward to more intelligent crawlers, exploiting the rich link structures to retrieve statements, and performing sophisticated reasoning to satisfy requests.

A possible future of this kind of annotation is shown by the W3C’s Annotea project[28]. Annotea allows comments and other remarks to be attached to Web resources, with these annotations being stored on annotation servers around the Web. On visiting a Website, attached statements can be retrieved from these servers. Useful applications are annotation with bookmarks (‘see also’), comments and reviews, and corrections. Furthermore, these Semantic Web annotations can be used to directly shape the Web Graph; providing direct links for common paths, for example.

Annotations are described in RDF, which means they are extensible and machine-understandable. Annotea is one way in which the Semantic Web could be applied directly to enrich Web agents — the meta-data attached to or comprising pages may provide ‘hints’ or ‘short-cuts’ to crawlers, transmitting additional information beyond the simple link structure of the document.

The Semantic Web will not only alter the Web Graph in a similar way to blogs, by providing a higher density of categorised links, but will also provide new directions for improving the performance of Web agents that can capitalise on the *semantics* of this new Web.

6 Conclusion

We have assessed the impact that blogs and the Semantic Web could have on Web crawling applications, and the subsequent effect on autonomous navigational agents that are used to index, classify, and find information on behalf of human users. We have shown how blogs have a different structure from traditional Web pages, one that could confuse current Web crawling applications, but which ultimately could be harnessed to provide such applications with a greater understanding of the content they index. Further, we have also suggested that the Semantic Web, by emphasising and formalising annotation, description, and information combination, will similarly reshape the Web. Its potential for semantically-rich annotation and description of resources will reduce reliance on simplistic measures of relevance, opening the door to both better indexing and intelligent browsing, with all of the inherent performance implications.

References

1. B. E. Brewington and G. Cybenko, “How dynamic is the Web?,” in *Proceedings of the 9th International World Wide Web Conference on Computer Networks: the International Journal of Computer and Telecommunications Networking*, pp. 257–276, North-Holland Publishing Co., 2000.
2. P. Baldi, P. Frascioni, and P. Smyth, *Modeling the Internet and the Web*. John Wiley and Sons, England, May 2003.
3. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, “Graph structure in the Web,” in *Proceedings of the 9th International World Wide Web Conference on Computer Networks: the International Journal of Computer and Telecommunications Networking*, pp. 309–320, North-Holland Publishing Co., 2000.
4. R. Albert, H. Jeong, and A.-L. Barabási, “The diameter of the World Wide Web,” *Nature*, vol. 401, 1999.
5. B. A. Huberman and L. A. Adamic, “Growth dynamics of the World Wide Web,” *Nature*, vol. 401, p. 131, Sept. 1999.
6. L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, “Search in power-law networks,” *Physical Rev. E*, vol. 64, no. 4, 2001.

7. J. M. Kleinberg and S. Lawrence, "The structure of the Web," *Science*, vol. 294, no. 5548, pp. 1849–1850, 2001.
8. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 668–677, ACM Press, 1998.
9. S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock, "The structure of broad topics on the Web," in *Proceedings of the 11th International World Wide Web Conference*, pp. 251–262, ACM Press, 2002.
10. G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of Web communities," in *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–160, ACM Press, 2000.
11. J. M. Kleinberg, "Hubs, authorities, and communities," *ACM Computing Surveys (CSUR)*, vol. 31, no. 4, p. 5, 1999.
12. S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.
13. NITLÉ, "National Institute for Technology and Liberal Education Weblog Census," 2004. <http://www.blogcensus.net>.
14. J. Henning, "The blogging iceberg," tech. rep., 2004. white paper; <http://www.perseus.com/blogsurvey/thebloggingiceberg.html>.
15. T. Berners-Lee, "Information management: A proposal," 1989. available from <http://www.w3.org/History/1989/proposal.html>.
16. T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.
17. "RDF Concepts and Abstract Syntax, W3C Recommendation," 2004. <http://www.w3.org/TR/rdf-concepts>.
18. G. Antoniou and F. van Harmelen, "Web Ontology Language: OWL," in *Handbook on Ontologies in Information Systems* (S. Staab and R. Studer, eds.), Springer-Verlag, 2003.
19. I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, "From SHIQ and RDF to OWL: The making of a Web Ontology Language," *Web Semantics*, vol. 1, no. 1, pp. 7–26, 2003.
20. T. Berners-Lee, D. Connolly, S. Palmer, and M. Nottingham, "cwm — a general-purpose data processor for the semantic web," 2004. <http://www.w3.org/2000/10/swap/doc/cwm>.
21. V. Haarslev and R. Möller, "Racer: A core inference engine for the Semantic Web," in *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools* (Y. Sure and O. Corcho, eds.), vol. 87 of *CEUR Workshop Proceedings*, (Montreal, Canada), 2003.
22. R. Fikes, P. Hayes, and I. Horrocks, "OWL-QL — a language for deductive query answering on the Semantic Web," Tech. Rep. KSL-03-14, Knowledge Systems Laboratory, Stanford University, Stanford, CA, 94305–9020, USA, 2003.
23. Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz, "Approximating aggregate queries about Web pages via random walks," in *Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 535–544, Morgan Kaufmann Publishers Inc., 2000.
24. M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling," pp. 295–308, 2000.
25. A.-L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, May 2003.
26. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, "The Web as a graph," in *Proceedings of the 19th ACM SIGACT-SIGMOD-AIGART Symposium on Principles of Database Systems, (PODS)*, pp. 1–10, ACM Press, 2000.
27. S. Starr, "Google hogged by blogs," July 2003. <http://www.spiked-online.co.uk/Articles/00000006DE60.htm>.
28. World Wide Web Consortium, "Annotea project." <http://www.w3.org/2001/Annotea>.