Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar & Co.

Joeran Beel
UC Berkeley
School of Information
jbeel@berkeley.edu

Bela Gipp
UC Berkeley
School of Information
gipp@berkeley.edu

Erik Wilde
UC Berkeley
School of Information
dret@berkeley.edu

ABSTRACT
This article introduces and discusses the concept of academic search engine optimization (ASEO). Based on three recently conducted studies, guidelines are provided on how to optimize scholarly literature for academic search engines in general and for Google Scholar in particular. In addition, we briefly discuss the risk of researchers' illegitimately 'over-optimizing' their articles.

Keywords
academic search engines, academic search engine optimization, ASEO, Google Scholar, ranking algorithm, search engine optimization, SEO

1. INTRODUCTION
Researchers should have an interest in ensuring that their articles are indexed by academic search engines[1] such as Google Scholar, IEEE Xplore, PubMed, and SciPlore.org, which greatly improves their ability to make their articles available to the academic community. Not only should authors take an interest in seeing *that* their articles are indexed, they also should be interesting in *where* the articles are displayed in the results list. Like any other type of ranked search results, articles displayed in top positions are more likely to be read.

This article presents the concept of *academic search engine optimization* (ASEO) to optimize scholarly literature for academic search engines. The first part of the article covers related work that has been done mostly in the field of general search engine optimization for Web pages. The second part defines ASEO and compares it to search engine optimization for Web pages. The third part provides an overview of ranking algorithms of academic search engines in general, followed by an overview of Google Scholar's ranking algorithm. Finally, guidelines are provided on how authors can optimize their articles for academic search engines. This article does not cover how publishers or providers of academic repositories can optimize their Web sites and repositories for academic search engines. The guidelines are based on three studies we have recently conducted [1-3] and on our experience in developing the academic search engine SciPlore.org.

[1] In this article we do not distinguish between 'academic databases' and 'academic search engines'; the latter term is used as synonym for both.

2. RELATED WORK
On the Web, search engine optimization (SEO) for Web sites is a common procedure. SEO involves creating or modifying a Web site in a way that makes it 'easier for search engines to both crawl and index [its] content' [4]. There exists a huge community that discusses the latest trends in SEO and provides advice for Webmasters in forums, blogs, and newsgroups.[2] Even research articles and books exist on the subject of SEO [5-10]. When SEO began, many expressed their concerns that it would promote spam and tweaking, and, indeed, search-engine spam is a serious issue [11-26]. Today, however, SEO is a common and widely accepted procedure and overall, search engines manage to identify spam quite well. Probably the strongest argument for SEO is the fact that search engines themselves publish guidelines on how to optimize Web sites for search engines [4, 27]. But similar information on optimizing scholarly literature for academic search engines does not exist, to our knowledge.[3]

2.1 Introduction to Academic Search Engine Optimization (ASEO)
Based on the definition of *search engine optimization* for Web pages (SEO), we define *academic search engine optimization* (ASEO) as follows:

> *Academic search engine optimization (ASEO) is the creation, publication, and modification of scholarly literature in a way that makes it easier for academic search engines to both crawl it and index it.*

ASEO differs from SEO in four significant respects. First, for Web search, Google is the market leader in most (Western) countries [28]. This means that for Webmasters (focusing on Western Internet users), it is generally sufficient to optimize their Web sites for Google. In contrast, no such market leader exists for searching academic articles, and researchers would need to

[2] E.g. http://www.abakus-internet-marketing.de/foren
http://www.highrankings.com/forum
http://www.seo-guy.com/forum
http://www.seomoz.org/blog
http://www.seo.com/blog
http://www.abakus-internet-marketing.de/seoblog

[3] Google Scholar offers some information for publishers on how to get their articles indexed by Google Scholar and ranked well [35]. However, this information is superficial in comparison to other SEO articles, and the information is not aimed at authors.

optimize their articles for several academic search engines. If these search engines are based on different crawling and ranking methods, optimization can become complicated.

Second, Webmasters usually do not need to worry about whether their site is indexed by a search engine: as long as any Web page is linked to an already indexed page, it will be crawled and indexed by Web search engines at some point. The situation is different in academia, where only a fraction of all published material is available on the Web and accessible to Web-based academic search engines such as CiteSeer. Most academic articles are stored in publishers' databases; they are part of the 'academic invisible web,' [29] and (academic) search engines usually cannot access and index these articles. A few academic search engines, such as Scirus and Google Scholar, cooperate with publishers, but still they do not cover all existing articles [30-32]. Researchers therefore need to think seriously about how to get their articles indexed by academic search engines.

Third, Webmasters can alter their pages by adding or replacing words and links, deleting pages, offering multiple versions with slight variations, and so on; in this way they can test new methods and adapt to changes in ranking algorithms. Scholarly authors can hardly do so: once an article is published, it is difficult and sometimes impossible to alter it. Therefore, ASEO needs to be performed particularly carefully.

Finally, Web search engines usually index all text on a Web site, or at least the majority of it. In contrast, some academic search engines do not index a document's full text but instead index only the title and abstract. This means that for some academic search engines authors need to focus on the article's title and abstract, but in other cases they still have to consider the full text for other search engines.

2.2 An Overview of Academic Search Engines' Ranking Algorithms

The basic concept of keyword-based searching is the same for all major (academic) search engines. Users search for a search term in a certain document field (e.g., title, abstract, body text), or in all fields, and all documents containing the search term are listed on the results page. Academic search engines use different ranking algorithms to determine in which position the results are displayed. Some let the user choose one factor on which to rank the results (common ranking factors are publication date, citation count, author or journal name and reputation, and relevance of the document); others combine the ranking factors into one algorithm, and, more often than not, the user has no influence on the factor's weighting.

The *relevance* of a document is basically a function of how often the search term occurs in that document and in which part of the document it occurs. Generally speaking, the more often a search term occurs in the document, and the more important the document field is in which the term occurs, the more relevant the document is considered[4]. This means that an occurrence in the title is weighted more heavily than an occurrence in the abstract, which carries more weight than an occurrence in a (sub)heading, than in the body text, and so on. Possible document fields that may be weighted differently by academic search engines are:[5]

- Title
- Author names
- Abstract
- (Sub)headings
- Author keywords
- Body text
- Tables and figures
- Publication name (name of journal, conference, proceedings, book, etc.)
- User keywords (Social tags)
- Social annotations
- Description
- Filename
- URI

The metadata of electronic files are especially important for academic search engines crawling the Web. When a search engine finds a PDF on the Web, it does not know whether this PDF represents an academic article, or which one it belongs to; therefore, the PDF must be identified, and one way to do this is by extracting the author and title. This can be done by analyzing the full text of the document or the metadata of the PDF.

It is also important to note that text in figures and tables usually is indexed only if it is embedded as real text or within a vector graphic. If text is embedded as a raster graphic (e.g., *.bmp, *.png, *.gif, *.tif, *.jpg), most, if not all, search engines will not index the text (see Figures 1 and 2 for an illustration of differences between vector and raster/bitmap graphics).[6] To our knowledge, none of the major academic search engines currently considers synonyms. This means that a document containing only the term 'academic search engine' would not be found via a search for 'scientific paper search engine' or 'academic database.' What most academic search engines do is stemming: words are reduced to their stems (e.g., 'analysed' and 'analysing' would be reduced to 'analyse').

2.3 Google Scholar's Ranking Algorithm

Google Scholar is one of those search engines that combine several factors into one ranking algorithm. The most important factors are relevance, citation count, author name(s), and name of publication.[7]

[4] Some algorithms, such as the BM25(f), saturate when a word occurs often in the text [36].

[5] Some of the data could be retrieved from the document full text, other from the metadata (of electronic files)

[6] Theoretically search engines could index the text in raster/bitmap graphics, but they would have to apply optical character recognition (OCR). To our knowledge, no search engine currently does this, although some are using OCR to index complete scans of scholarly literature.

[7] Google Scholar offers different search functions. For instance, it is possible to search for 'related articles' and 'recent articles.' In this article we focus on the normal ranking algorithm, which is applied for the standard keyword search.

2.3.1 Relevance

Google Scholar focuses strongly on document titles. Documents containing the search term in the title are likely to be positioned near the top of the results list. Google Scholar also seems to consider the length of a title: In a search for the term 'SEO,' a document titled 'SEO: An Overview' would be ranked higher than one titled 'Search Engine Optimization (SEO): A Literature Survey of the Current State of the Art.'

Although Google Scholar indexes entire documents, the total search term count in the document has little or no impact. In a search for 'recommender systems,' a document containing fifty instances of this term would not necessarily be ranked higher than a document containing only ten instances.

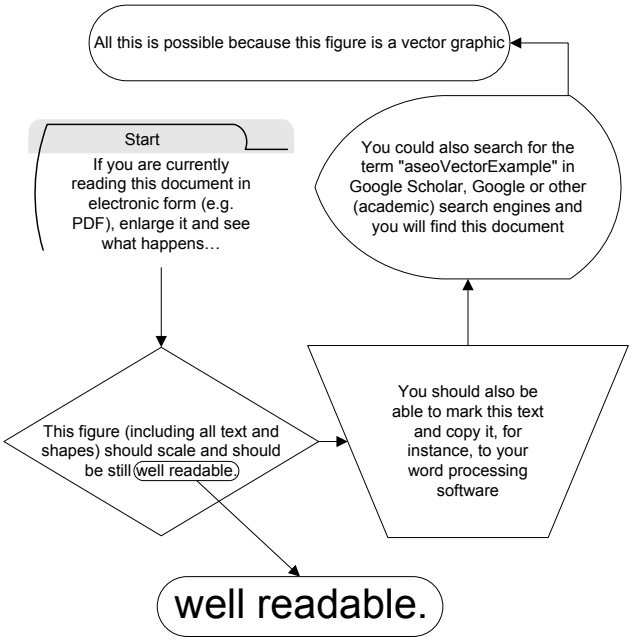

Figure 1: Example of a Vector Graphic

Like other search engines, Google Scholar does not index text in figures and tables inserted as raster/bitmap graphics, but it does index text in vector graphics. It is also known that neither synonyms nor PDF metadata are considered.

2.3.2 Citation Counts

Citation counts play a major role in Google Scholar's ranking algorithm, as illustrated in Figure 3, which shows the mean citation count for each position in Google Scholar.[8] It is clear that, on average, articles in the top positions have significantly more citations than articles in the lowest positions. This means that to achieve a good ranking in Google Scholar, many citations are essential. Google Scholar seems not to differentiate between self-citations and citations by third parties.

[8] On average, articles at position 1 had 834 citations, articles at position 2 had 552, articles at position 3 had 426, and articles at position 1000 had fifty-three. The study was based on 1,032,766 results produced by 1050 search queries in November 2008. For more detail see [1].

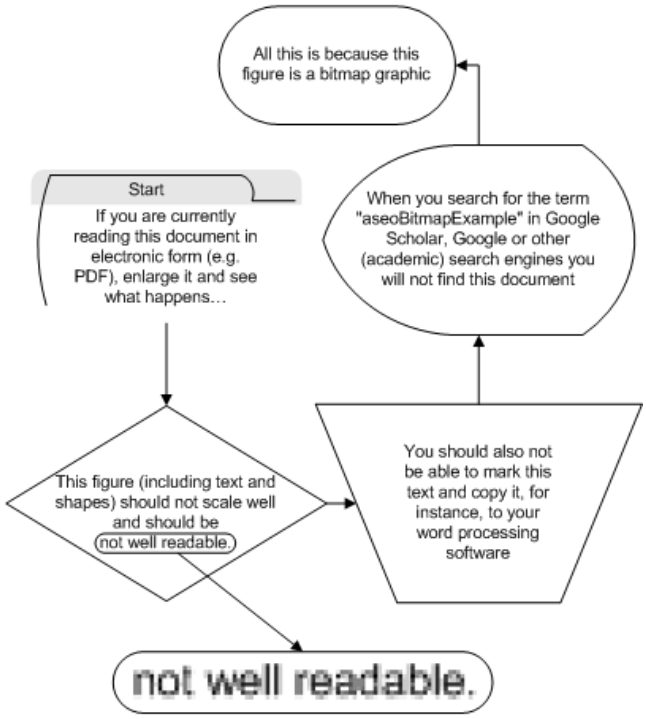

Figure 2: Example of a Bitmap Graphic

2.3.3 Author and Publication Name

If the search query includes an author or publication name, a document in which either appears is likely to be ranked high. For instance, seventy-four of the top 100 results of a search for 'arteriosclerosis and thrombosis cure' were articles about various (medical) topics from the journal Arteriosclerosis, Thrombosis, and Vascular Biology, many of which did not include the search term either in the title or in the full text [2].

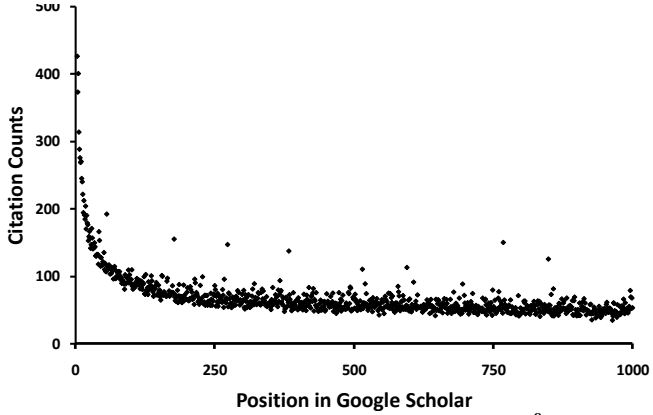

Figure 3: Mean Citation Count per Position[8]

2.3.4 Other factors

Google Scholar's standard search does not consider publication dates. However, Google Scholar offers a special search function for 'recent articles,' which limits results to articles published within the past five years. Furthermore, Google Scholar claims to consider both publication and author reputation [33]. However, we could not research the influence of these factors because of a lack of data, and therefore we do not consider them here.

2.3.5 *Sources Indexed by Google Scholar*

Bert van Heerde, a professional in the field of SEO, uses the term 'invitation based search engine' to describe Google Scholar: Only articles from trusted sources and articles that are 'invited' (cited) by articles already indexed are included in the database [34]. 'Trusted sources,' in this case, are publishers that cooperate directly with Google Scholar, as well as publishers and Webmasters who have requested that Google Scholar crawl their databases and Web sites.[9]

Once an article is included in Google Scholar's database, Google Scholar searches the Web for corresponding PDF files, even if a trusted publisher has already provided the full text.[10] It makes no difference on which site the PDF is published; for instance, Google Scholar has indexed PDF files of our articles from the publisher's site, our university's site, our private home pages, and SciPlore.org. PDFs found on the Web are linked directly on Google Scholar's results pages, in addition to the link to the publisher's full text (see Figure 4 for an illustrative example).

Figure 4: Linking database entries with external PDFs

If different PDF files of an article exist, Google Scholar groups them to improve the article's ranking [35]. For instance, if a preprint version of an article is available on the author's Web page and the final version is available on the publisher's site, Google indexes both as one version. If the two versions contain different words, Google Scholar associates all contained words with the article. This is an interesting feature that we will discuss in more detail in the next section.

3. OPTIMIZING SCHOLARLY LITERATURE FOR GOOGLE SCHOLAR AND OTHER ACADEMIC SEARCH ENGINES

3.1 Preparation

In the beginning it is necessary to think about the most important words that are relevant to the article. It is not possible to optimize one document for dozens of keywords, so it is better to choose a few. There are tools that help in selecting the right keywords, such as Google Trends, Google Insights, Google Adwords keyword tool, Google Search–based keyword tool, and Spacky.[11]

It might be wise not to select those keywords that are most popular. It is usually a good idea to query the common academic search engines using each proposed keyword; if the search already returns hundreds of documents, it may be better to choose another keyword with less competition.[12]

3.2 Writing Your Article

Once the keywords are chosen, they need to be mentioned in the right places: in the title, and as often as possible in the abstract and the body of the text (but, of course, not so often as to annoy readers). Although in general titles should be fairly short, we suggest choosing a longer title if there are many relevant keywords.

Synonyms of important keywords should also be mentioned a few times in the body of the text, so that the article may be found by someone who does not know the most common terminology used in the research field. If possible, synonyms should also be mentioned in the abstract, particularly because some academic search engines do not index the document's full text.

Be consistent in spelling people's names, taking special care with names that contain special characters. If names are used inconsistently, search engines may not be able to identify articles or citations correctly; as a consequence, citations may be assigned incorrectly, and articles will not be as highly ranked as they could be. For instance, *Jöran*, *Joeran*, and *Joran* are all correct spellings of the same name (given different transcription rules), but Google Scholar sees them as three different names.

The article should use a common scientific layout and structure, including standard sections: *introduction*, *related work*, *results*, and so on. A common scientific layout and structure will help Web-based academic search engines to identify an article as scientific.

Academic search engines, and especially Google Scholar, assign significant weight to citation counts. Citations influence whether articles are indexed at all, and they also influence the ranking of articles. We do not want to encourage readers to build 'citation circles,' or to take any other unethical action. But any published articles you have read that relate to your current research paper should be cited. When referencing your own published work, it is important to include a link where that work can be downloaded. This helps readers to find your article and helps academic search engines to index the referenced article's full text. Of course, this can also be done for other articles that have well-known (i.e., stable and possibly canonical) download locations.

3.3 Preparing for Publication

Text in figures and tables should be machine readable (i.e., vector graphics containing font-based text should be used instead

---

[9] Visit http://www.google.com/support/scholar/bin/request.py to ask Google Scholar to crawl your Web site containing scholarly articles.

[10] Google Scholar also indexes other file types, such as PostScript (*.ps), Microsoft Word (*.doc), and MS PowerPoint (*.ppt). Here we focus on PDF, which is the most common format for scientific articles.

[11] Google Trends http://www.google.com/trends
Google Insights http://www.google.com/insights/search/
Google Adwords
https://adwords.google.com/select/KeywordToolExternal;
Google keyword tool, http://google.com/sktool/
Spacky, http://www.spacky.com

[12] For example, keywords such as 'Web' and 'HTML' may be of limited use because there are too many papers published in that space, in which case it makes more sense to narrow the scope and choose better-differentiated keywords.

of rasterized images) so that it can easily be indexed by academic search engines. Vector graphics also look more professional, and are more user friendly, than raster/bitmap graphics. Graphics stored as JPEG, BMP, GIF, TIFF, or PNG files are not vector graphics.

When documents are converted to PDF, all metadata should be correct (especially author and title). Some search engines use PDF metadata to identify the file or to display information about the article on the search results page. It may also be beneficial to give a meaningful file name to each article.

3.4 Publishing

As part of the optimization process, authors should consider the journal's or publisher's policies. Open-access articles usually receive more citations than articles accessible only by purchase or subscription; and, obviously, only articles that are available on the Web can be indexed by Web-based academic search engines. Accordingly, when selecting a journal or publisher for submission, authors should favor those that cooperate with Google Scholar and other academic search engines, since the article will potentially obtain more readers and receive more citations.[13] If a journal does not publish online, authors should favor publishers who at least allow authors to put their articles on their or their institutions' home pages.

3.5 Follow-Up

There are three ways to optimize articles for academic search engines after publication.

The first is to publish the article on the author's home page, so that Web-based academic search engines can find and index it even if the journal or publisher does not publish the article online. An author who does not have a Web page might post articles on an institutional Web page or upload it to a site such as Sciplore.org, which offers researchers a personal publications home page that is regularly crawled by Google Scholar (and, of course, by SciPlore Search). However, it is important to determine that posting or uploading the article does not constitute a violation of the author's agreement with the publisher.

Second, an article that includes outdated words might be replaced by either updating the existing article or publishing a new version on the author's home page. Google Scholar, at least, considers all versions of an article available on the Web. We consider this a good way of making older articles easier to find. However, this practice may also violate your publisher's copyright policy, and it may also be considered misbehavior by other researchers. It could also be a risky strategy: at some point in the future, search engines may come to classify this practice as spamming. In any case, updated articles should be clearly labeled as such, so that readers are aware that they are reading a modified version.

Third, it is important to create meaningful parent Web pages for PDF files. This means that Web pages that link to the PDF file should mention the most important keywords and the PDFs metadata (title, author, and abstract). We do not know whether any academic search engines are considering these data yet, but normal search engines do consider them, and it seems only a matter of time before academic search engines do, too.

[13] The main criteria for selecting a publisher or journal, of course, should still be its reputation and its general suitability for the paper. The policy is to be seen as an additional factor.

4. DISCUSSION

As was true in the beginning for classic SEO, there are some reservations about ASEO in the academic community. When we submitted our study about Google Scholar's ranking algorithm [2] to a conference, it was rejected. One reviewer provided the following feedback:

> *I'm not a big fan of this area of research […]. I know it's in the call for papers, but I think that's a mistake.*

A second reviewer wrote,

> *[This] paper seems to encourage scientific paper authors to learn Google scholar's ranking method and write papers accordingly to boost ranking [which is not] acceptable to scientific communities which are supposed to advocate true technical quality/impact instead of ranking.*

ASEO should not be seen as a guide on how to cheat academic search engines. Rather, it is about helping academic search engines to understand the content of research papers and, thus, about how to make this content more widely and easily available. Certainly, we can anticipate that some researchers will try to boost their rankings in illegitimate ways. However, the same problem exists in regular Web searching; and eventually Web search engines manage to avoid spam with considerable success, and so will academic search engines. In the long term, ASEO will be beneficial for all – authors, search engines, and users of search engines. Therefore, we believe that academic search engine optimization (ASEO) should be a common procedure for researchers, similar to, for instance, selecting an appropriate journal for publication.

ACKNOWLEDGEMENTS

We thank the SEO Bert van Heerde from Insyde (http://www.insyde.nl/) for his valuable feedback, and Barbara Shahin for proofreading this article.

ABOUT THE AUTHORS

The research career of Jöran Beel and Bela Gipp began about ten years ago when they won second prize in Jugend Forscht, Germany's largest and most reputable youth science competition and received awards from, among others, German Chancellor Gerhard Schröder for their outstanding research work. In 2007, they graduated with distinction at OVGU Magdeburg, Germany, in the field of computer science. They now work for the VLBA-Lab and are PhD students, currently at UC Berkeley as visiting student researchers. During the past years they have published several papers about academic search engines and research paper recommender systems.

Erik Wilde is Adjunct Professor at the UC Berkeley School of Information. He began his work in Web technologies and Web architectures a little over ten years ago by publishing the first book providing a complete overview of Web technologies. After focusing for some years on XML technologies, XML and modelling, mapping issues between XML and non-tree metamodels, and XML-centric design of applications and data

models, he has recently shifted his main focus to information and application architecture, mobile applications, geo-location issues on the Web, and how to design data sharing that is open and accessible for many different service consumers.

REFERENCES

[1] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In André Flory and Martine Collard, editors, *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science (RCIS'09)*, pages 439–446, Fez (Morocco), April 2009. IEEE. doi: 10.1109/RCIS.2009.5089308. ISBN 978-1-4244-2865-6. Available on http://www.sciplore.org.

[2] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: An Introductory Overview. In Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 1, pages 230–241, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935. Available on http://www.sciplore.org.

[3] Jöran Beel and Bela Gipp. Google Scholar's Ranking Algorithm: The Impact of Articles' Age (An Empirical Study). In Shahram Latifi, editor, *Proceedings of the 6th International Conference on Information Technology: New Generations (ITNG'09)*, pages 160–164, Las Vegas (USA), April 2009. IEEE. doi: 10.1109/ITNG.2009.317. ISBN 978-1424437702. Available on http://www.sciplore.org.

[4] Google. Google's Search Engine Optimization Starter Guide. PDF, November 2008. URL http://www.google.com/-webmasters/docs/search-engine-optimization-starter-guide.pdf.

[5] Albert Bifet and Carlos Castillo. An Analysis of Factors Used in Search Engine Ranking. In *Proceedings of the 14th International World Wide Web Conference (WWW2005), First International Workshop on Adversarial Information Retrieval on the Web (AIRWeB'05)*, 2005. http://airweb.cse.lehigh.edu/2005/bifet.pdf.

[6] Michael P. Evans. Analysing Google rankings through search engine optimization data. *Internet Research*, 17 (1): 21–37, 2007. doi: 10.1108/10662240710730470.

[7] Jin Zhang and Alexandra Dimitroff. The impact of metadata implementation on webpage visibility in search engine results (Part II). *Cross-Language Information Retrieval*, 41 (3): 691–715, May 2005.

[8] Harold Davis. *Search Engine Optimization*. O'Reilly, 2006.

[9] Jennifer Grappone and Gradiva Couzin. *Search Engine Optimization: An Hour a Day*. John Wiley and Sons, 2nd edition, 2008.

[10] Peter Kent. *Search engine optimization for dummies*. Willey Publishing Inc, 2006.

[11] AA Benczur, K Csalogány, T Sarlós, and M Uher. SpamRank – Fully Automatic Link Spam Detection. In *Adversarial Information Retrieval on the Web (AiRWEB'05)*, 2005.

[12] A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. *Adversarial Information Retrieval on the Web (AIRWEB), Seattle, Washington, USA*, 2006.

[13] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. *Lecture Notes in Computer Science*, 3720: 96, 2005.

[14] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. pages 1–6, 2004.

[15] Q. Gan and T. Suel. Improving web spam classifiers using link structure. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 20. ACM, 2007.

[16] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proceedings of the 31st international conference on Very large data bases*, page 528. VLDB Endowment, 2005.

[17] H. Saito, M. Toyoda, M. Kitsuregawa, and K. Aihara. A large-scale study of link spam detection by graph algorithms. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 48. ACM, 2007.

[18] B. Wu and K. Chellapilla. Extracting link spam using biased random walks from spam seed sets. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, page 44. ACM, 2007.

[19] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, page 430. ACM, 2007.

[20] G.G. Geng, C.H. Wang, and Q.D. Li. Improving Spamdexing Detection Via a Two-Stage Classification Strategy. page 356, 2008.

[21] I.S. Nathenson. Internet infoglut and invisible ink: Spamdexing search engines with meta tags. *Harv. J. Law & Tec*, 12: 43–683, 1998.

[22] T. Urvoy, E. Chauveau, P. Filoche, and T. Lavergne. Tracking web spam with HTML style similarities. *ACM Transactions on the Web (TWEB)*, 2, 2008.

[23] T. Urvoy, T. Lavergne, and P. Filoche. Tracking web spam with hidden style similarity. In *AIRWeb 2006*, page 25, 2006.

[24] Masahiro Kimura, Kazumi Saito, Kazuhiro Kazama, and Shin ya Sato. Detecting Search Engine Spam from a Trackback Network in Blogspace. *Lecture Notes in Computer Science: Knowledge-Based Intelligent Information and Engineering Systems*, 3684: 723–729, 2005. doi: 10.1007/11554028_101.

[25] Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. Detecting spam web pages through content analysis. In *15th International Conference on World Wide Web*, pages 83–92. ACM, 2006.

[26] Baoning Wu and Brian D. Davison. Identifying link farm spam pages. In *14th International Conference on World Wide Web*, pages 820–829, 2005.

[27] Yahoo! How do I improve the ranking of my web site in the search results?, July 2007. URL http://help.yahoo.com/l/us/-yahoo/search/indexing/ranking-02.html.

[28] Alex Chitu. Google's Market Share in Your Country. Website, March 2009. URL http://googlesystem.blogspot.com/-2009/03/googles-market-share-in-your-country.html https://-spreadsheets.google.com/-ccc?key=pLaE9tsVLp_0y1FKWBCKGBA.

[29] D. Lewandowski and P. Mayr. Exploring the academic invisible web. *Library Hi Tech*, 24 (4): 529–539, 2006.

[30] Nisa Bakkalbasi, Kathleen Bauer, Janis Glover, and Lei Wang. Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3, 2006. doi: 10.1186/1742-5581-3-7.

[31] John J. Meier and Thomas W. Conkling. Google Scholar's Coverage of the Engineering Literature: An Empirical Study. *The Journal of Academic Librarianship*, 34 (34): 196–201, 2008.

[32] William H. Walters. Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, 43 (4): 1121–1132, July 2007. doi: doi:10.1016/j.ipm.2006.08.006.

[33] Google. About Google Scholar. Website, 2008. URL http://-scholar.google.com/intl/en/scholar/about.html.

[34] Bert van Heerde. RE: Pre-print: Academic Search Engine Optimization. Email, 3 September 2009.

[35] Google Scholar. Support for Scholarly Publishers. Website, 2009. URL http://scholar.google.com/intl/en/scholar/-publishers.html.

[36] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM New York, NY, USA, 2004.