Characteristics of Scientific Web Publications: Preliminary Data Gathering and Analysis

Erik Thorlund Jepsen, Piet Seiden, Peter Ingwersen, and Lennart Björneborn
Department of Information Studies, Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S, Denmark. E-mail: {etj, ps, pi, lb}@db.dk

Pia Borlund
Department of Information Studies, Royal School of Library and Information Science, Sohngaardsholmsvej 2, DK-9000 Aalborg, Denmark. E-mail: pb@db.dk

Because of the increasing presence of scientific publications on the Web, combined with the existing difficulties in easily verifying and retrieving these publications, research on techniques and methods for retrieval of scientific Web publications is called for. In this article, we report on the initial steps taken toward the construction of a test collection of scientific Web publications within the subject domain of plant biology. The steps reported are those of data gathering and data analysis aiming at identifying characteristics of scientific Web publications. The data used in this article were generated based on specifically selected domain topics that are searched for in three publicly accessible search engines (Google, AllTheWeb, and AltaVista). A sample of the retrieved hits was analyzed with regard to how various publication attributes correlated with the scientific quality of the content and whether this information could be employed to harvest, filter, and rank Web publications. The attributes analyzed were inlinks, outlinks, bibliographic references, file format, language, search engine overlap, structural position (according to site structure), and the occurrence of various types of metadata. As could be expected, the ranked output differs between the three search engines. Apparently, this is caused by differences in ranking algorithms rather than the databases themselves. In fact, because scientific Web content in this subject domain receives few inlinks, both AltaVista and AllTheWeb retrieved a higher degree of accessible scientific content than Google. Because of the search engine cutoffs of accessible URLs, the feasibility of using search engine output for Web content analysis is also discussed.

Introduction

The Web has a significant impact on the practice in scientific publication. According to Cronin and McKim (1996, p. 170), the Web is reshaping the ways in which scholars communicate with one another, i.e., new kinds of scholarly and proto-scholarly publishing are emerging, which means that work-in-progress, broadsides, early drafts, and refereed articles are now almost immediately sharable. Nevertheless, only a minor part of these scientific publications is accessible through the Web, because they are difficult to control (i.e., to find, identify, access, and assess). In February 1999 Lawrence and Giles (1999, p. 107) estimated that only 6% of randomly selected Web sites contained scientific or educational content, defined as university, college, and research lab servers. It is an open question whether content of a scientific nature should solely be found in those domains. In response to this question, Björneborn and Ingwersen (2001, p. 69) explain how the nature of the Web has generated a reality of freedom of information. On one hand, the Web allows for everybody to express themselves, practically without control from authorities, to become visible worldwide, and by linking to the pages that one wants to link to, to assume credibility by being *there,* and to obtain access to data, information, values, and knowledge in many shapes and degrees of truth. On the other hand, the Web increasingly becomes a Web of uncertainty to its users. The ease of publication calls for quality watch and assessments, as indicated by the results from an investigation of the reliability of biology related Web sites. In a study by Allen, Burke, Welch, and Rieseberg (1999), 500 Web sites were retrieved through simple searches in the search engine Northernlight.com of the three topics "evolution," "genetically modified organism," and "endangered species." Two expert referees then examined the search results sequentially, until each referee had reviewed approximately 60 sites containing information pertinent to each of the topics. The structured reviews showed that between 12% and 46% of the sites were considered informative. Among the informative sites 10%–34% were judged "inaccurate," that is, containing factually incorrect information; 20%–35% sites were judged "misleading,"

Accepted January 23, 2004

 ● Published online 13 August 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20079

referring to misinterpreting science or blatantly omitting facts supporting an opposing position; and more than 48% sites for each topic were "unreferenced" (did not build on refereed papers). Although these results to some degree could be influenced by the simplistic search strategy employed and the generality and popularity of the selected topics, they underline the need for the development of filters or similar tools to aid both the scientist as well as the general user in obtaining qualified scientific information through the Web.

WebTAPIR is a research project concerned with webometric and Web IR issues, aiming at developing real-time filtering and ranking mechanisms, which will prioritize qualified scientific material above inferior publications. According to this goal, it is of obvious importance to identify Web publication attributes, which can be employed in harvesting, filtering, and ranking processes. In our opinion this can best be achieved in a subject field exhibiting a large diversity of Web publications, a condition that is met by the chosen subject domain of plant biology. We use *Web publication* synonymously with *Web page,* and we employ the definition of Web publication in accordance to that of Whatis?com (Whatis?com, n.d.): "Each page is an individual HTML file with its own Web address (URL)." Further, in this respect, "[a] Web site is a collection of Web files on a particular subject that includes a beginning file called home page" (Whatis?com, http://whatis.techtarget.com).

In this article, we report on the initial steps taken toward the construction of a test collection of scientific Web publications within the subject domain of plant biology. The test collection is intended to serve as a platform for webometric studies of scientific communication and behavior. The steps reported are those of data gathering and data analysis aiming at identifying characteristics and attributes of scientific Web publications. Thus, the ambition of the article is to share partly our experimental experiences gained in the preliminary process of data gathering and partly the empirical findings of characteristics and attributes of scientific Web publications.

In the section on Some Characteristics of the Web, we sketch out characteristics of the Web that hinder the control of Web publications. The section on WebTAPIR briefly describes the WebTAPIR project, and how the project endeavors to overcome some of the obstacles described in the section, Some Characteristics of the Web. In the section on Research Design, we describe the design of the present study. In the section on Distributions: Language, Search Engines, and Formats, we present the initial search results and the set of retrievable URLs according to topics, search engine results, and languages. Furthermore, this section includes a description of the distribution of formats in the accessible URLs of the Google search results. In the section on Content Classification and Indicators of Scientific Quality, we describe the classification scheme developed to classify the Web publications according to their degree of scientific qualities. We also present the sample that was used to reach the classification. In the Summary and Conclusion, we sum up our empirical findings and discuss briefly how these findings are to be employed in our future work.

Some Characteristics of the Web

The exponential growth, as well as geographical as lingual diversification, of Web publications impedes the control (retrieval, identification, access, and evaluation of publications) of the Web. Another obstacle for obtaining control is caused by "the hidden Web," that is, "pages that are not normally considered for indexing by Web search engines, such as pages with authorization requirements, pages excluded from indexing using the robots exclusion standard, and pages hidden behind search forms" (Lawrence & Giles, 1999, p. 107).

The freedom of publication on the Web, along with the absence of filtering and reviewing procedures, causes huge differences between Web publications with regard to quality, stability, site structure complexity, formats, and the amount and quality of metadata. Although the location and the content of Web publications are easily changed, a thorough and stable registration of informative publications is less easily achieved. One approach might be to retrieve upper-level Web pages within the site structure to minimize the dangers of relocation; whereas lower-level pages probably contain a larger amount of informative scientific information in the form of work in progress, papers, articles, drafts, and the like. Further, scientific material is characterized by publications published in formats like PostScript, PDF, and MSWord (Lawrence, Bollacker, & Giles, 1999). This type of binary encoded formats poses problems, because they seldom contain structural information or metadata, which might serve filtering and retrieval purposes. In fact, although different metadata initiatives (e.g., the Dublin Core Metadata Initiative, n.d.) seems promising, the actual use of metadata is still very rare. In an analysis of corporate Web sites reported on by Drott (2002, p. 211), it is shown that no more than 43% of corporate Web sites contain even a minimum of Metadata in the form of keyword or description tags on their home page, and only 0.3% of the sites contained metadata using the Dublin Core standard (Lawrence & Giles, 1999, p. 109).

Search Engine Coverage

According to Lawrence and Giles (1999, p. 107), search engines do not index sites equally, they may not even index new pages for months, and no engine indexes more than about 16% of the Web as of February 1999. This lack of comprehensive indexing by the search engines is not necessarily an obstacle for obtaining qualified information, but studies (e.g., Lawrence & Giles, 1999)—including the study described in this paper—reveal low overlaps in search engine coverage. The retrieval uniqueness of each engine calls for researchers to employ several search engines to accomplish exhaustive results.

Commercial search engines, such as AltaVista and Google, may in fact perform well, but the procedures and

algorithms used for harvesting, indexing, and ranking Web publications are usually proprietary information not available to the IR community. Because of these circumstances, we cannot rely on the search engines to provide us with anything but a skewed selection (sample) of the Web. This sample could to some degree be considered a representation of the typical information available to the general user, because search engines are popular access roads to the information on the Web. However, in relation to webometrics and IR research, it is necessary to have access to the *raw* Web, to make reliable interpretations and develop solid IR solutions.

WebTAPIR

WebTAPIR is a research project at the Royal School of Library and Information Science, Denmark, that aims at improving access to scientific Web-based information in Scandinavian and English languages. In this article, we describe the preliminary process of gathering information about the Web information content within the subject domain of plant biology and discuss the problems and consequences identified.

The next step will be the generation of a database of Web publications, which will function as a test collection. As pointed out by Sparck Jones and van Rijsbergen (1976), test collections must reflect the variety of real retrieval environments. However, some level of homogeneity is required. This duality between variety and homogeneity holds for "content" (of publications and requests), "source types," "sources," "origins," "time," and "language" (Sparck Jones & van Rijsbergen, 1976, p. 64). Web publications are published in different formats with great variations in structures as well as in the amount, type, and quality of embedded metadata. Building a test collection depends heavily on the possibilities of identifying and extracting (or creating) representations of identical attributes for all publications. An attribute could be "author," and the value of that author attribute could be, for example, Sparck Jones and van Rijsbergen. In other words, values must reflect homogeneities as well as varieties, but attributes must be identical—or at least comparable. Robertson (1981, p. 25) explains how quite sufficient numbers of documents normally are easy to get hold of and input into the system(s), and that is most easily done if the documents are available in a suitable form; most difficult, if some fundamentally new form of indexing has to be applied to them. To obtain comparability of Web publications in a test collection, publications must be parsed into similar formats, and information should be extracted to establish uniform representations. We cannot hope to obtain the same amount of identifiable, isolable, indexable, and controllable data as in traditional records or uniformly structured full-text publications. But scientific Web publications potentially offer several indications, which could be used to improve retrieval performance by *cognitive overlaps* (also known as "polyrepresentation," Ingwersen, 1992, 1994, 1996) of attributes' values present in different parts of Web publications, for example, in outlinks, citations, and bibliographic references in texts. Web publications themselves contain attributes of different kinds of nature, for example, title, metadata, bibliographic references, and outlinks, as well as the text (e.g., paragraphs, images, graphs, and bulleted lists). Additional representations can be obtained from other sources with different cognitive origin and contexts, for example, databases, search engines, and server structures (local links and structural position), as well as other sources producing inlinks or citations to the publication at hand. Hence the future research agenda of WebTAPIR is used for different cognitive origins in the research and development of filtering, ranking, and linking algorithms. (For definitions of inlinks and outlinks, see the section, Research Design).

Still, the first step toward the building of a test collection of scientific Web publications, as well as achieving the future research agenda of Web TAPIR, is to carry out the preliminary data gathering and identification of characteristics of scientific Web publications as reported on in the following section, Research Design.

Research Design

In this section, we describe the research design and experimental procedure applied. The research objectives for our study were

1. To identify characteristics and attributes of scientific Web publications that can support harvesting, filtering, and ranking mechanisms; and
2. To obtain experimental experience about the data gathering of scientific Web publications.

The gathering of Web publications within the subject domain of plant biology was made by searching three domain specific topics: "photosynthesis," "herbicide resistance," and "plant hormones" (for the exact search terms and search operators used, see Tables 1, 2, and 3. A domain expert (plant biologist) was consulted, and so were various subject lists and vocabularies (in English, Danish, Swedish, and Norwegian), to broaden the search statements with relevant synonyms, quasi synonyms, and spelling variations. The three

TABLE 1. Search results for "photosynthesis." The first five terms target Scandinavian publications, the remaining publications in English.

Search terms "photosynthesis"	Google	AllTheWeb	AltaVista
fotosyntes	1,540	885	1,161
fotosyntese	2,100	1,177	1,002
fotosyntesen	2,740	1,634	1,743
fotosyntetisk	71	65	46
fotosyntetiske	72	76	43
photosynthesis	238,000	119,286	79,388
photosynthetic NOT photosynthesis	42,500	26,984	25,183
Scandinavian as % of English	2.3%	2.6%	3.8%

TABLE 2. Search results for "herbicide resistance." The first eight terms target Scandinavian publications, and the remaining publications in English.

Search terms "herbicide resistance"	Google	AllTheWeb	AltaVista
ogräsmedel AND resistens	40	17	21
ogräsmedlen resistens	10	4	4
ugrasmiddel AND resistens	13	10	9
ugrasmidler resistens	13	13	12
herbicidresistens	111	149	91
ukrudtsmiddel resistens	45	54	35
ukrudtsmidler AND resistens	65	94	55
herbicidresistent	63	—	65
herbicide resistance	53,2000	28,630	23,547
herbicide resistant NOT resistance	19,900	11,731	9,993
herbicide resistence	218	233	128
Scandinavian as % of English	0.5%	0.8%	0.9%

topics were chosen to retrieve pages ranging from a popular to a more specialized scientific level. The three search topics were searched for in the following three search engines: Google, AllTheWeb, and AltaVista. We used AltaVista and Google because of their popularity. Furthermore, Google was the only well-known search engine that indexed formats like MSword and PDF at the time of our study. AllTheWeb (and to some degree Google) was employed, because it gave actual access to the highest number of URLs. In fact, we started out by employing two other search engines (Hotbot and Northernlight), but limited access to retrieved URLs and peculiarities in search results (huge differences in hits with the same queries repeated with less than a 1-minute interval) forced us to omit these search engines.

The searches were carried out during the period of November 14 to December 19, 2001. A simple search strategy was employed in the search of the three topics, in that no advanced search facilities like spelling control mechanisms, exact match, or truncation offered by the three search engines were used. The reason is that we wanted partly to retrieve as many URLs as possible, and partly to be able to compare the search results from the different engines. Support for the simple search strategy is given by Bar-Ilan (2001, p. 13), who points out that search engine facilities like stemming and Boolean commands are constantly added, removed, and refined, which hindered comparison between different search engines as well as the evaluation of single search engines over time. Also in support of the simple search strategy employed is the limited outcome of the Scandinavian searches (see Tables 1–3). The search terms (including synonyms and spelling variations) were searched separately, as depicted in Tables 1–3, because of the cutoffs of accessible URLs of the search engines. Possible problems arising from default control of spelling variations or stemming were neglected, because later removal of duplicates would eliminate this problem.

The actual accessible URLs were extracted, duplicates were removed, and language and search engine distribution was analyzed. Furthermore, the distribution of formats in the Google results was analyzed. These distributions are presented in the section on Distributions: Language, Search Engines, and Formats.

Ideally, the search results provided a qualified starting point for the creation of a sample. Unfortunately, but nevertheless an interesting observation, the search engines—although listing the number of indexed Web publications—did not facilitate actual access to all the so-called retrieved

TABLE 3. Search results for "plant hormones." The first five terms target Scandinavian publications, and the remaining publications in English.

Search terms "plant hormones"	Google	AllTheWeb	AltaVista
plantehormon	40	47	18
plantehormoner NOT plantehormon	72	116	61
plantevækstfaktorer	50	145	46
växthormon	34	12	25
växthormoner NOT växthormon	48	38	9
plant growth regulator	42,500	3,355	21,969
plant hormone	126,000	3,429	57,392
plant growth regulators NOT plant growth regulator	67,600	5,358	23,816
plant hormones NOT plant hormone	63,800	5,498	31,119
Scandinavian as % of English	0.1%	2.0%	0.1%

URLs. In fact, only AllTheWeb provided access to several thousands of the indexed and retrieved Web publications (4,100), whereas Google's cutoff was close to 1,000 URLs, and AltaVista only allowed access to 200 URLs. Bar-Ilan (2001, p. 22) emphasizes how this might also be a problem to the informetrician who is interested in the whole set of results for a given query (or at least in the size of this set), whereas it might be less problematic to the average user who only needs a few "most relevant" URLs.

The present pilot study also aims at uncovering quality levels and indicators that will be encountered in the future development of filtering and ranking mechanisms. Because only the highest-ranked (prioritized) Web publications are accessible, we were unable to access the lower-ranked publications. The problem then is whether such publications actually are inferior in a qualitative sense, for example, if they were nonscientific. The results, presented in the section on Content and Metadata Correlation and the section on Content of Web Publications Correlated to Inlinks, Outlinks, and Bibliographic References, show that this indeed may not be the case. For the purpose of informetric analysis, the test collection must contain all types of Web publications to develop appropriate filtering mechanisms. Search engine overlaps could prove to be well suited as a quality indicator and as a potential filtering feature for the test collection. Unfortunately, as a result of the skewness in accessible URLs, the data material does not legitimate an analysis of the correlation between search engine overlaps and scientific quality.

In the section on Content Classification and Indicators of Scientific Quality, we describe how the sample of 600 pages were analyzed and heuristically classified by the domain expert. The classification consists of six categories regarding the scientific potentiality of the content, namely, "scientific," "scientifically related," "teaching," "low-grade," "noise," and "unavailable." The classification scheme was designed specifically for the present study, inspired by categories employed by Almind and Ingwersen (1997) and by Kleinberg (1999). To analyze whether some formats are a better indicator of scientific content than others, we picked a random sample of 50 URLs in PDF formats and correlated these pages with the content classification. Furthermore, a subsample ($n = 88$) was generated containing all the scientific Web publications and an equal number of randomly picked Web publications from the other categories. This was done to analyze to which degree information arising from metadata, links, bibliographic references, and indexing levels correlated with the content categories. The 88 Web publications were analyzed with reference to indexing levels, meta information, bibliographic references, and outlinks (only links to external publications were counted). We defined *outlinks* as hyperlinks present in a given Web publication, pointing (out) to other Web publications. We were, however, not able to identify the exact number of inlinks to the publications. *Inlinks* are defined as hyperlinks that point (in) to a given Web publication. Instead, we used the counting facilities in all three search engines to identify the amount of inlinks for each publication. These facilities depended on the size of the search engine database, and in most cases Google had the highest amount of inlinks. The search engine that counted the highest number of inlinks to a publication was chosen as the best indicator of the *real* number of inlinks.

Distributions: Language, Search Engines, and Formats

Tables 1–3 show how the Scandinavian (lingual) dimension seems to be negligible, although some Scandinavian publications are written in English. "Photosynthesis," which is the broadest topic, has a larger percentage of Scandinavian written publications (2.3%–3.8%) than the two more specialized topics.

Generally speaking, AltaVista seemed to cover Scandinavian publications better than both Google and AllTheWeb. More detailed analysis of the Web publications will reveal if the Scandinavian dimension is as negligible as the intermediate results indicate. Regarding language and location, the initial searches indicated that there are Scandinavian publications that should be included in the test collection. However, lingual information would probably not improve filtering and ranking algorithms, unless the more thorough analysis of the final sample shows a generally higher scientific quality of Scandinavian publications written in English. The relatively low number of hits by AllTheWeb for the English searches on plant hormones can only be explained by the AllTheWeb harvesting.

Language Distributions and Search Engine Coverage

The actual accessible URLs were extracted, and duplicates were removed. As explained in the section on Research Design, the search engines did not facilitate actual access to all the retrieved URLs, hence the expression "actual accessible URLs."

As shown in Tables 4, 5, and 6 the cutoff of accessible URLs resulted in a skewness of the lingual distribution, which was taken into account when generating the subsample (see the section on Content and Metadata Correlation and the section on Content of Web Publications, etc.), which was used to correlate content classification with metadata, inlinks, outlinks, and bibliographic references. This was done by normalizing the Scandinavian dimension in accordance to the initial search results depicted in Tables 1–3.

The findings showed that very few Web publications are multilingual. This is evident in the row "language overlap" in Tables 4–6.

Although there are large search engine overlaps (i.e., the number of Web publications that are retrieved and accessible in two or three of the search engines), the number of unique accessible URLs is also high for each search engine. This means that results from all three search engines must be included in the final sampling. In fact, 42% unique hits in AltaVista for the topics "photosynthesis" and "herbicide resistance" is surprisingly high, considering that AltaVista

TABLE 4. Accessible URLs for the topic "photosynthesis."

"Photosynthesis" URLs	Google	AllTheWeb	AltaVista	Total	% of Total
Scandinavian	2,238	3,273	630	6,141	37%
English	1,712	8,200	400	10,312	63%
Total	3,950	11,473	1,030	16,453	100%
Language overlap	1	12	0		
Search Engine overlap					
with Google		1,326	351		
with AllTheWeb	1,326		518		
with AltaVista	351	518			
with (AllTheWeb + AltaVista)	1,404				
with (AllTheWeb + Google)			596		
with (Google + AltaVista)		1,466			
overlapping publications				3466	21%
% unique URLs	64%	87%	42%		

TABLE 5. Accessible URLs for the topic "herbicide resistance."

"Herbicide resistance" URLs	Google	AllTheWeb	AltaVista	Total	% of Total
Scandinavian	286	281	232	799	7%
English	1,838	7,707	522	10,067	93%
Total	2,124	7,988	754	10,866	
Language overlap	0	2	0		
Search Engine overlap					
with Google		855	239		
with AllTheWeb	855		358		
with AltaVista	239	358			
with (AllTheWeb + AltaVista)	935				
with (AllTheWeb + Google)			438		
with (Google + AltaVista)		1054			
overlapping publications				2,427	22%
% unique URLs	56%	87%	42%		

TABLE 6. Accessible URLs for the topic "plant hormones."

"Plant hormone" URLs	Google	AllTheWeb	AltaVista	Total	% of Total
Scandinavian	243	363	157	763	4%
English	3,256	12,863	795	16,914	96%
Total	3,499	13,226	952	17,677	100%
Language overlap	4	30	1		
Search Engine overlap					
with Google		703	138		
with AllTheWeb	703		240		
with AltaVista	138	240			
with (AllTheWeb + AltaVista)	1,326				
with (AllTheWeb + Google)			281		
with (Google + AltaVista)		878			
overlapping publications				2,485	14%
% unique URLs	62%	93%	70%		

only facilitates access to 200 URLs for each search statement. Thus, the sample must be normalized according to the distribution among the search engines in the initial search. We discovered, when analyzing the subsamples that the large amount of unique hits was caused primarily by the diversity of the search engines' ranking algorithms. The higher percentage of unique hits for AllTheWeb compared to Google and AltaVista, is explained by the larger number of accessible Web publications.

Formats

Google was the only search engine that indexed publications in PDF and MSWord formats at the end of 2001. Figure 1 shows the distribution of formats of the Google search results.

Evidently, PDF as well as MSWord are formats that must be included in the test collection, because 14% of all the publications found in Google were PDFs, and 2.5% were MSWord publications. The subsamples must be normalized in accordance with the Google results concerning formats. However, the PDF and MSWord publications cause problems, because these binary encoded formats rarely embed accessible metadata, as can be the case for HTML publications. Thus, software that is able to recognize and parse elements like titles, authors, citations, and links must be employed. A solution to the problem of poorly structured PDF publications has been demonstrated in the development of an autonomous citation indexing system (CiteSeer, n.d.). The CiteSeer Research Index downloads Postscript or PDF files and convert with great success the Postscript and PDF formats into text and subsequently extract information like URL, title, author, abstract, introduction, citations, citation context, and full text (Giles, Bollacker, & Lawrence, 1998; Lawrence, Giles, & Bollacker, 1999). In our future attempts to employ the idea of *polyrepresentation* we will try to add link information to the binary encoded publications by experimenting with employing information about in- and outlinks from their parent (HTML) pages.

Content Classification and Indicators of Scientific Quality

The domain expert evaluated a selection of the retrieved URLs to obtain a preliminary indication of the proportion of the scientific quality and the distribution of levels of the retrieved URLs. For the evaluation, we heuristically developed a six-category classification scheme. For each topic, 100 URLs from each language group were selected at random and assessed—in total, 600 URLs. The results are presented in Table 7.

The first category, "scientific," was assigned to content that was deemed to be of scientific quality, for example, preprints, conference reports, abstracts, and scientific articles.

The second category, "scientifically related," was assigned to materials of potential relevance for a scientific query, such as directories, CVs, institutional reports. This kind of information appears to be abundant on the Web and thus will interfere with a search for specific information on a given scientific subject. Further, this kind of information could be potentially relevant as leads to more specific information on the subject searched for, even though the information need is not directly met.

The third category, "teaching," contains content that is developed in relation to teaching, for example, textbooks, fact pages, tutorials, student papers, and course descriptions.

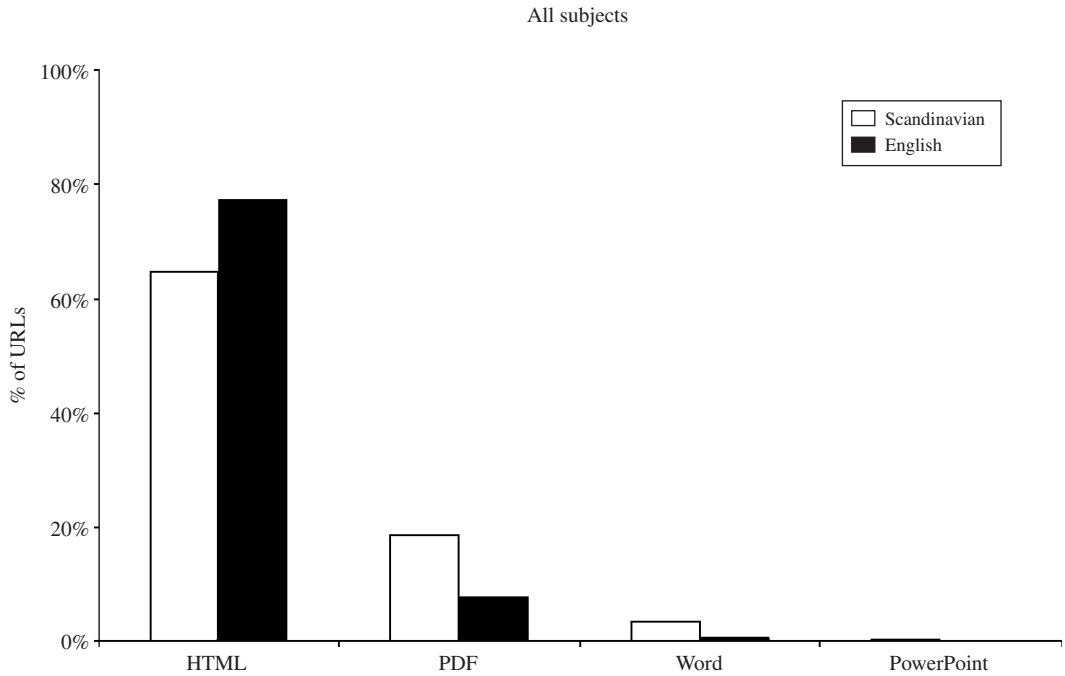

FIG. 1. Distribution of formats of the Google search results (n = 9,573).

TABLE 7. Classification of randomly picked subsamples from each language group within each topic investigated. All numbers are in percent of subsample size ($n = 200$ for each topic).

	"Plant hormones"		"Photosynthesis"		"Herbicide resistance"	
Category	English	Scandinavian	English	Scandinavian	English	Scandinavian
Scientific	5%	1%	9%	0%	6%	1%
Scientifically related	17%	14%	25%	11%	24%	13%
Teaching	12%	37%	20%	19%	11%	16%
Low-grade	27%	12%	15%	16%	45%	48%
Noise	17%	15%	17%	41%	3%	1%
Unavailable	22%	21%	14%	13%	11%	21%

This is a particularly interesting category, because many of the traits found in scientific papers can be seen mimicked in student papers, sometimes blurring the distinction with formal and presumably accurate scientific papers.

The fourth category, "low-grade," is reserved to content that fail to meet the criteria of the three previous groups, but it is still concerned with the topic. This category contains content that is either of only commercial interest or deemed inaccurate or misleading.

Content failing to fit within the inclusion criteria for the mentioned classes and thereby not pertinent to the topic was assigned to the fifth category labeled "noise."

Some URLs were not accessible. Hence, a sixth category, consisting of "unavailable" URLs, was also created.

Although this classification in some ways resembles the classification of document types used by Almind and Ingwersen (1997, p. 412), the classes are not directly comparable. Almind and Ingwersen's category, "subject defined homepages," can in fact be present in all of our categories, and publications belonging to their "pointer" category, that is, a Web page whose function is primarily to make a number of hyperlinks available (Almind & Ingwersen, 1997, p. 412) are identified in two of our categories, "scientifically related" (subject directories) and "noise" (commercial directories).

The actual amount of content deemed to be of scientific relevance was found to be very small for all topics and almost nonexisting for the Scandinavian languages. The amount of other materials considered to be of potential relevance, the very heterogeneous "scientifically related" category, was significantly larger with also a fair representation in the Scandinavian languages. This was also the case for the "teaching" related materials, and the "low-grade" materials. The differences within the "noise" category, with only a very low percentage in relation to "herbicide resistance," could be because of the availability of nonambiguous search terms for this topic. Not surprisingly, "herbicide resistance" also showed the largest proportion of materials in the "low-grade" category, because this subject is constantly debated in political terms, very often through Web channels. The expectation that the term "photosynthesis" should lead to an increased amount of teaching materials held true for the English language results, though it was notable that the parallel searches of this topic within the Scandinavian languages resulted in a very high "noise level."

The Scandinavian language group is almost nonexistent with respect to scientific content, which indicates that Scandinavian plant biology research is not published in any of those languages. Further, it signifies that the English language can be used as a (robust) discrimination criterion for scientific content on the Web—at least for Scandinavian research in this area. Some differences because of subject domain specific publication patterns can be expected, but in general, the mining of scientific quality content will require a rigorous discrimination procedure.

Content and File Format Correlation

A large proportion of scientific material is published electronically in PostScript (.ps), TeX (.tex), or PDF (.pdf) formats. Because PostScript and TeX in particular are used in connection with disciplines relying heavily on mathematical notations, those formats were not included in this study of the plant biology domain. The PDF format, however, was found initially to be significantly occurring (Figure 2) and a random subsample ($n = 50$) was therefore categorized along the lines as described above. The results, shown in Table 8, indicate that the PDF Web publication file format together with the English language does indeed correlate with scientific content. But for the Scandinavian language group scientific PDF files do not appear. This result indicates that file format can beneficially be employed combined with other attributes in ranking algorithms, which to some degree must prioritize publications published in PDF and similar formats.

Content and Metadata Correlation

In our analysis of metadata of the retrieved Web publications, we have divided metadata into topical meta tags including keywords as well as descriptions, and metadata containing authoritative information about authors, contributors, publishers, and other corporations. Authoritative metadata was furthermore divided into regular meta tags and authoritative information present in an identifiable form in the publication body text.

The kind of authoritative information in the text differs among the categories. Typically, scientific publications simply name the author, whereas the "scientifically related"

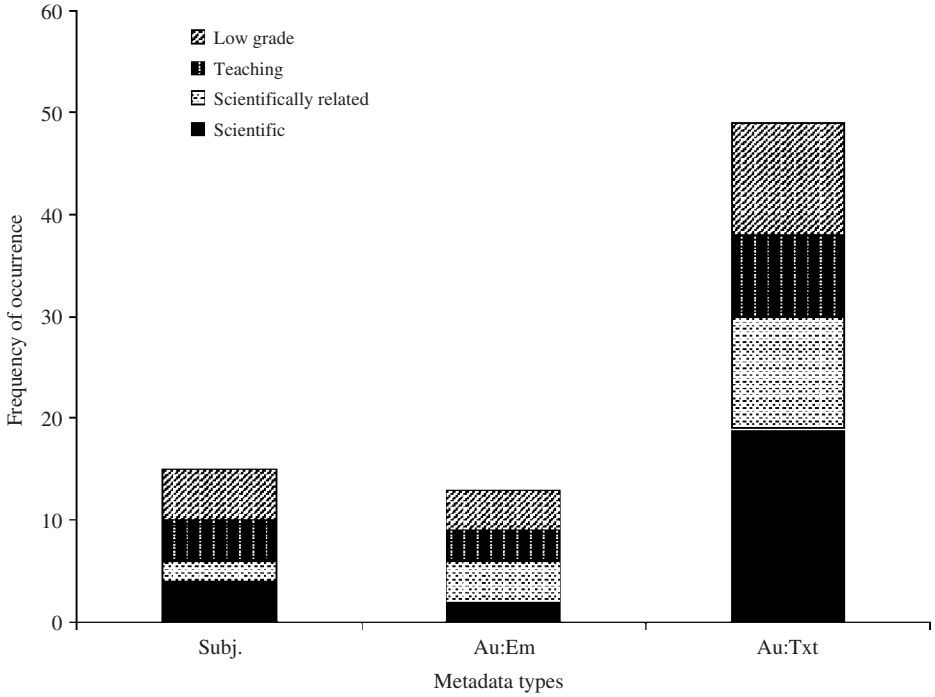

FIG. 2. Occurrence of metadata in the publications sampled. Metadata was registered as either topical meta tags (Subj.), authority describing meta tags (Au:EM), or authority describing information located in the visible publication text (Au:TXT) ($n = 88$).

publications include various information about authors, publishers, and corporate sources. In "low-grade" publications, authoritative information typically informs about a commercial company, although they did not all belong to the .com domain.

Content of Web Publications Correlated to Inlinks, Outlinks, and Bibliographic References

Several studies and search engines try to use the nature of the Web. Kleinberg (1999) describes the Web as consisting of "hubs" and "authorities." Kleinberg (1999, p. 5) defines authority Web pages: "the most prominent sources of primary content, are the authorities on the topic; other pages, equally intrinsic to the structure, assemble high-quality guides and resource lists that act as focused hubs, directing users to recommended authorities. Hubs link heavily to authorities, but hubs may themselves have very few incoming links, and authorities may not link to other authorities." Kleinberg's "hubs" are very similar to Almind and Ingwersen's (1997) "pointer" category. Kleinberg employs outlinks in the manifestation of his idea by assigning pages hub weights (proportional to the sum of the authority weights of the pages that it links to) and authority weights (proportional to the sum of the hub weights of pages that it links to) ideally only topic-specific inlinks are taken into account. In contrast to Kleinberg's use of outlinks, Google ranks publications according to the volume and weight of inlinks. Google (Google Technology, n.d.) explains how PageRank relies on the uniquely democratic nature of the Web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But Google looks at more than the sheer volume of votes or links a page receives; it also analyses the page that casts the vote. Votes cast by pages that are themselves *important* weigh more heavily and help to make other pages *important*.

TABLE 8. Classification of randomly picked subsample of PDF files from each language group independent of subjects. All numbers are in percentage of sample size ($n = 50$).

Category	English	Scandinavian
Scientific	24% (12)	0% (0)
Scientifically related	18% (9)	20% (10)
Teaching	22% (11)	20% (10)
Low-grade	24% (12)	16% (8)
Noise	2% (1)	28% (14)
Unavailable	10% (5)	16% (8)

In our sample the publications judged to be scientific were all very similar to traditional journal articles or parts of articles. Preprints, articles, abstracts, and conference reports are typically structured in a scholarly way, including bibliographic references. In our case, that was 59% of the scientific publications, as shown in Table 9. Only 19% of the scientific publications contained outlinks, and only 14% were linked to by other publications. These findings suggest that ranking based solely on inlinks should be avoided, if one wants to prioritize scientific information.

TABLE 9. Relative distribution within the different publication classifications of publications containing inlinks, outlinks, and bibliographic references ($n = 88$).

	Inlinks			Outlinks			
Link count	0	1–3	>3	0	1–3	>3	Biblio. ref.
Scientific	64%	23%	14%	82%	14%	5%	59%
Scientifically related	45%	30%	25%	55%	25%	20%	55%
Teaching	57%	29%	14%	86%	10%	5%	24%
Low-grade	62%	14%	24%	52%	33%	14%	10%

The "Scientifically related" category contains two types of publications: Web publications similar to the scientifically categorized Web publications that nearly reach this category and pages that may be "hubs." In fact, if we divided the sample into a research/educational domain and a commercial domain, then both domains would seem to have their own hubs. The research/educational hubs were located in the "scientifically related" publications with more than three outlinks (20%), and the commercial hubs were found among the 14% of "low-grade" publications with more than three outlinks.

Apparently, hubs characterized by many outlinks can be identified and employed in the harvesting of scientific publications. Yet, it seems a more difficult task to identify single attributes that indicate scientific content, because the scientific publications themselves receive very few inlinks and donates very few outlinks.

Because of the skewness in our sample, with publications (including the scientific Web publications) retrieved from AllTheWeb, we cannot evaluate the significance of search engine overlaps.

Almost none of the scientific publications originated from the Google searches. However, when we searched directly on the URLs of the 88 publications in our sample, Google had actually indexed all of the publications. This means that these publications were outside the 800 top-ranked and physical accessible Google publications. Consequently, algorithms based on inlink information (like PageRank) are hardly suitable for identification and ranking of scientific content on the Web. The finding of Google's insufficient ability to retrieve or rank the most important pages (from a scientific perspective) is similar to the results of a recent study by Thelwall (2003). Thelwall (2003, p. 215) concludes that "PageRank is not an effective method for identifying the 'best' Web pages in a university system because of its domination by internal links, an argument that would still apply even if all mirror sites had been removed from the data." Scientifically orientated search engines and indexes like the CiteSeer Research Index and Scirus employ varying link and citation information in sophisticated ranking and linking algorithms. Unfortunately, we have not been able to identify more detailed information about how publication harvesting and filtering is performed in these search engines. Regarding indexing levels (structural position), we found no remarkable differences in our searches with reference to how thorough the three search engines harvested and ranked the different levels in the structure of sites.

Summary and Conclusion

In this article, we report on the preliminary data gathering of scientific Web publications intended for a future test collection of subject domain specific Web publications. Thus, the idea and ambition of the article is to share our experimentally gained experiences as well as the empirical findings of scientific Web publication characteristics. As such, the reported study has provided a starting point for further investigation of which Web publication attributes that should be taken into account when monitoring scientific communication on the Web, and developing multi-evidence-based filtering, ranking, and linking algorithms for scientific Web publications.

Three topics were searched against three publicly accessible search engines in both English and Scandinavian. Because of search engine cutoffs of accessible URLs, we were only able to analyze the top-ranked URLs, which means that only a small percentage of the so-called retrieved hits (approximately 45,000 to 1,100,000) could be employed in the sampling and further analysis. The accessible URLs were extracted, duplicates were removed, and a sample ($n = 600$) consisting of URLs from each topic and each language group was heuristically classified by a domain expert according to the scientific potential of the analyzed publications.

We found that language is an important discriminating attribute, because most researchers tend to publish their research in English, whereas, for example, teaching materials often are published in the local language. The existence of scholarly references signifies scientific Web publications. Bibliographic references constitute a robust and objective indicator, which, combined with link information, seems capable of making a valid discrimination between Web publication types.

Another interesting finding is the correlation between PDF files and content classification; that is, PDF files contain a higher proportion of scientific materials than compared with other analyzed Web publication formats. This emphasizes the importance of further development of methods capable of capturing and analyzing PDF files. The PDF file format combined with other structural features of content, like in- and outlink information, as well as the English language, may function as evidence of scientific Web material.

The findings regarding embedded metadata and linkage revealed no direct pattern that could be used as a central component in Web publication ranking, but such data may still serve as secondary discriminators. In fact, in combination with URL-domain information, embedded metadata

could be useful as multievidence in real-time filtering, by sorting out commercial pages. It is evident within the observed scientific subject domain that the amount of inlinks should not be employed as the only feature in filtering or page-ranking algorithms. This is because inlinks mainly favor nonacademic pages.

Hubs that link to large amounts of scientific content can be identified by employing structural and outlink information. Along with URL domains as ".edu" (education) or ".ac" (academia), these hubs can act as well-suited sources for the harvesting of potential scientific content.

The scientific publications analyzed whether PDF or HTML files were all typically structured in a scholarly way, containing titles, statements of responsibility, and bibliographic references in identifiable and isolable parts of the text. Thus filtering, retrieval, and ranking must be based on combinations of such formal textual elements and topical indicators derived from structures in the publication text.

Our future work will include analyses of the semantic structures of scientific Web publications aimed at revealing additional characteristics of scientific Web publications. Characteristics we consider essential will be reflected in the test collection of scientific Web publications, because the characteristics may prove useful for future webometric studies as well as for further development and refinement of IR techniques for retrieval of scientific Web publications.

Acknowledgments

This work is sponsored by research grants from the Research Council of the Danish Cultural Ministry (ref. no. A2002 06–021) as well as from NordInfo, and it forms part of the overall TAPIR research project (Text Access Potentials for interactive Information Retrieval).

References

Allen, E.S., Burke, J.M., Welch, M.E., & Rieseberg, L.H. (1999). How reliable is science information on the Web? Nature, 402, 722.

Almind, T.C., & Ingwersen, P. (1997). Informetric analysis on the World Wide Web: Methodological approaches to "webometrics." Journal of Documentation, 53(4), 404–426.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes—A review and analysis. Scientometrics, 50(1), 7–32.

Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. Scientometrics, 50(1), 65–82.

CiteSeer. (n.d.). The NEC Research Institute Scientific Literature Digital Library. Retrieved January 12, 2004, from http://citeseer.com/

Cronin, B., & McKim, G. (1996). Science and scholarship on the World Wide Web: A North American perspective. Journal of Documentation, 52, 163–172.

Drott, M. (2002). Indexing aids at corporate Websites: The use of robots.txt and META tags. Information Processing & Management, 38, 209–219.

Dublin Core Metadata Initiative. (n.d.). Retrieved January 12, 2004, from http://dublincore.org/

Giles, C.L., Bollacker, K.D., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In Digital Libraries 98–Third ACM Conference on Digital Libraries (pp. 89–98). New York: ACM Press.

Google Technology. (n.d.). Google searches more sites more quickly, delivering the most relevant results. Retrieved January 12, 2004, from http://www.google.com/technology/

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: Elements of a cognitive theory for information retrieval interaction. In W.B. Croft & C.J. van Rijsbergen (Eds.), Proceedings of the 17th ACM Sigir Conference on Research and Development in Information Retrieval (pp. 101–110). Dublin, 1994. London: Springer Verlag.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, 52(1), 3–50.

Kleinberg, J. (1999). Hubs, authorities and communities. ACM Computing Surveys, 31(4).

Lawrence, S., Bollacker, K., & Giles, C.L. (1999). Indexing and retrieval of scientific literature. In Eight International Conference on Information and Knowledge Management, CIKM '99, Kansas City, MO, November 2–6, 1999 (pp. 139–146).

Lawrence, S., & Giles, C.L. (1999). Accessibility and distribution of information on the Web. Nature, 400, 107–110.

Lawrence, S., Giles, C.L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. IEEE Computer, 32(6), 67–71.

Robertson, S.E. (1981). The methodology of information retrieval experiment. In K. Sparck Jones (Ed.), Information Retrieval Experiment (pp. 9–31). London: Butterworth.

Scirus. (n.d.). Retrieved from About Us. Retrieved January 12, 2004, from http://www.scirus.com/about/

Sparck Jones, K., & van Rijsbergen, C.J. (1976). Information retrieval test collections. Journal of Documentation, 32(1), 59–75.

Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages? Journal of Documentation, 59(2), 205–217.

Whatis?com. (n.d.). Retrieved January 12, 2004, from http://whatis.techtarget.com/