Estimating the Number of Web Pages of the Environmental Administration in Germany

Franz Schenk, Fred Kruse

Abstract
The German environmental information portal PortalU² is an access point to a large number of information sources, all from public authorities. The user is offered a single point of entry for queries on a large variety of data with environmental relevance. In this respect, one of the most important tasks in the maintenance of the portal is to guarantee a high degree of relevance for all the different kinds of information sources. In addition to relevance, which often cannot be defined exactly, quantity of information is a measure that normally can be obtained more easily. While this is generally true for highly structured information like records in a metadata catalogue or entities in database systems, it can be difficult when dealing with unstructured information sources like web pages. In this paper we show the obstacles that we encountered in our examinations on the size of our web page index and present how the information providers respectively contribute to the web page index of the environmental administration in Germany.

1 Introduction
Typically, web crawlers fetch all web pages starting from a given set of entry points (so-called start URLs or seeds). If only selected domains or subsets of domains should be crawled, not only starting points but also the boundaries of these domains have to be specified. In PortalU (Konstantinidis et al. 2010), all of the web sites of the environmental administration in Germany that we know of can be searched. There are web sites of institutions from both the Federal States and the Federal Government, only local authorities (like communities) and non-governmental organizations are not considered. Domain experts from most of the institutions collaborate in providing the URL definitions that are necessary for the configuration of the search engine.

From this set of URL definitions the web page index of PortalU is being generated by a search engine. Although the number of seeds can be given as an exact number, the number of unique web pages in the resulting index can only be estimated because many pages contained therein are duplicates or have no meaningful content. From the point of view of PortalU as an information provider, it is interesting to obtain more precise knowledge about the number of unique web pages in the index. Information providers themselves normally do not know the amount of pages originating from their domain(s). Comparisons between the different providers and their domains become reliable only when real numbers instead of estimates become available. Moreover, legitimate statements in long term observations on the increase (or, more seldomly, the decrease) of web pages can only be expressed when erroneous pages have been eliminated from the set of pages. A short overview on the technical background is given in the next section, followed by a description of the problems with content management systems in Section 3 and a presentation of the results in Section 4. An outlook towards future work can be found in Section 5.

1 Coordination Center PortalU at the Lower Saxony Ministry of Environment, Energy, and Climate Protection (franz.schenk | fred.kruse)@portalu.de
2 http://www.portalu.de
2 Technical Background

The search engine in PortalU is based on the open source software nutch\(^3\) and lucene\(^4\) and is part of the InGrid software tools that build the technical foundations of PortalU as an information brokering system. The nutch software is modified in several ways, mostly for the convenience of users maintaining the sets URLs that specify the target domains. The technical principals of building a web page index is very simple. Basically, the web crawler starts with the given set of start-URLs by fetching these pages from the web servers and continues to do so recursively with all outgoing links as long as these links point to web pages within the given boundaries. Theoretically, this process consists of a finite number of iterations until eventually no new pages can be found that are not already contained within the web page index. In practise, however, there are a number of complications that often lead to an infinite crawling process where new path-names (URLs) or even new (pseudo-) web pages are constantly being created. As resources in all computer systems are limited this might become a severe obstacle when not being dealt with. Normally, two main groups of problems can be identified: either web pages cannot be fetched at all or web pages are found to be duplicates of already existing pages in the index. In the end, both problems make it difficult to provide for exact numbers about the size of the index.

Some of the more trivial problems are of a technical nature and lie in the communication between the services. Sometimes web pages are no longer available, domain names may have changed, or services are temporarily unavailable. Further, search engines can be blocked completely (where web site administrators have to be asked for permission) or limited in the number of requests per minute. Also, the URLs themselves may cause problems. They are collected from outlinks of pages that are already in the index and become the candidates for the next round of page fetching. However, it does happen regularly that they do not point to a valid destination. The reasons can be manifold: domain names may have changed, path structures on the web server can be altered, content is only available temporarily. Moreover, domain names in the public administration change on a regular basis in the wake of public elections when responsibilities become newly distributed. Although most of the resulting problems are only of a temporary nature, they still contribute to a reduced performance of the search engine and makes it harder to establish the total number of pages in the index.

More serious problems are related to flaws in content management systems. While the aforementioned difficulties can be addressed easily most of the time, it is much harder to deal with the opposite phenomenon where to many pages are added to the index. The reasons are either false configurations or faults in the software. A short overview to the most common problems is given in the next section.

3 Search Engines and Content Management Systems

Web sites with complex and constantly changing content almost always depend on content management Systems (CMSs) for the management of web pages. They make the handling of the pages more convenient and offer many features for structuring and designing the web site that can be used out of the box. For example, feedback forms, interactive tables, or print pages can be embedded easily into a page. Also, social media functions are available that offer recommender services and integration into social web environments. The most popular CMSs that we observed in the environmental sector of the public administration in Germany, were Government Site Builder, Typo3 and CoreMedia.

Although nowadays most CMSs are highly sophisticated software products some still show serious flaws if one takes the point of view of a search engine. In the following we list the most common phenomena.

\(^3\) http://nutch.apache.org , last visited: 01.07.2012
\(^4\) http://lucene.apache.org , last visited: 01.07.2012
3.1 Script and Style Information

Often there are pages containing only Javascript code or style information that are not intended to be displayed as a web page. A CMS needs those pages for page creation, rendering and implementation of page functionality. This information, however, is only intended to be used internally. In a well-configured CMS all code (e.g. CSS, JS, PERL) is kept in separate parts of the web content hierarchy that is accessible only to the system itself and hidden from ordinary users (or search engines).

But javascript files or style sheet schemas are often publicly available and therefore become fetched by search engines. Although search engines are normally configured in a way such that scripts become ignored based on their file suffixes this does only work when proper file extensions are used. In practise, this is not always the case and special path patterns have to be specified in order to exclude script directories on certain web sites. Some examples for such patterns (given as regular expressions) can be found in Figure 1.

3.2 Print Pages

A standard feature of CMSs are print pages that offer slightly different versions of a web page optimized for printing. The idea is to remove all additional properties of the web page (e.g. menus, frames, navigation bars, or search fields) and render only the unique content of the page in a printer-friendly way. The print page is accessible via a simple HTML link or a button, sometimes it is implemented as a scripted feature. Hereby, these print pages become also available to the search engines and are added to the page index. Although the content of a print page is also contained in the original web page search engines not always automatically detect these pages as duplicates. Search engines could easily be prevented from fetching the print pages by a HTML meta parameter (e.g. `<meta name="robots" content="noindex "/>`), but many CMSs do not make use of that parameter. For that reason it is sometimes necessary to exclude print pages explicitly. Some of the manifold varieties of URL parameters that can be seen in print pages are given in Figure 2.
3.3 Recommender Pages

Another common feature in CMSs are recommender pages. This is a feature in web sites that allows users to recommend a page they like to friends via email. Typically, a link is placed on every page (e.g. Tell a friend, Recommend etc.) that opens a new page where the user may enter an email address which sends a link to the page that the user just visited. Although these pages, where the email address can be given, do not provide for any real content, they are often accessible as ordinary web pages. Here, again, HTML meta tags for the exclusion of search engines would be very helpful, but most of the CMSs simply ignore this possibility and do not use them. Figure 3 gives some examples of regular expressions that identify URLs that are dynamically created by CMSs for recommender pages.

3.4 Interactive Table Sorting

There are some more features of CMSs that may cause redundant information in a web page index. The most prominent one is an interactive feature that allows for the sorting of tables. Tables containing for example search results, glossaries, or numerical values can be sorted for every single column in both directions (ascending or descending). If the user interactively sorts a table, the same page is reloaded where the URL of the page carries the sorting information in URL parameters (e.g. ?order=desc&sortby=name). From the point of view of the search engine, these are new pages because the URL is different. Also the content is composed differently, because (parts of) the content of the page is delivered in a different order. Nevertheless, these pages are redundant because they do not contribute new information. Such duplicates are hard to remove generically because each CMS has its own way to communicate the user interaction. Therefore, exclusion patterns have to be defined for each domain independently. Figure 4 gives some examples for different ways of encoding sorting instructions as URL parameters.

3.5 Infinite cycles in Path Extensions

The last class of problems are infinite cycles in the dynamic creation of URL paths. Here, web pages contain links that are created dynamically at the time when the page is generated for delivery to the user (or a search machine). Due to faults in the programming of the software it can happen that the links lead to the re-fetching of the same page, dynamically prolonging all links on that page with the same path fragments as in the last retrieval. As a search engine follows every outlink of a web page that remains in the same domain, this process may last, theoretically, for an infinite number of iterations.
These errors often do not even add duplicates to the index because the new URLs either lead to the same web page or to the same error message. Nevertheless they have to be filtered because they can significantly reduce the performance of the search engine: because the number of web pages that can be fetched is limited (either because of a cost-model in a commercial product or because of limited resources in hardware and bandwidth) the number of failed fetches should be reduced as far as possible. One single failure is certainly no problem. But faulty web sites can easily generate millions of false URLs that efficiently block the search engine. It is not unusual that such web sites comprise of some thousand real pages but quickly generate millions of unfetchable URLs. It is relatively easy to identify such domains: whenever the number of unfetched pages becomes greater than the number of already fetched pages there is a strong indication that the situation is getting out of hand.

3.6 Using Exclusion Patterns

The search engine in PortalU tackles the problems with CMSs using exclusion patterns. These patterns are given in form of regular expressions (see Figure 1-3). All URLs in the URL database (and all newly found outlinks) are filtered against these patterns. Every URL that can be matched against any of the exclusion patterns is removed from the URL database, for example the following URL


would be matched by the pattern www.umweltbundesamt.de/.*boden-und-altlasten.*boden-und-altlasten.* because of the repeated path element boden-und-altlasten in the URL.

Exclusion patterns can be used either explicitly to expunge single web pages by giving full URLs or implicitly in form of regular expressions for the filtering of sets of URLs. The latter are especially useful in this context, because many of the flaws in CMSs cause regular patterns in the creation of URLs. The so-called exclude-URL patterns are widely used in PortalU.

Without exclusion patterns the number of URLs in the URL database would be constantly increasing, only limited by the capacities of the search engine. Problems with a specific domain are often indicated by an abnormal increase of unfetched URLs for that domain. In such situations, all URLs of the respective domain have to be analysed in order to identify the subset of URLs that have been created by faults of the CMS (see again Sections 3.1 to 3.5). The effect of a single exclude URL varies from situation to situation. In some cases a pattern may filter more than a million URLs, sometimes there are just a few hundred URLs that become removed by a filter expression. From this consideration it becomes clear that one has to weigh effort against effect. The purpose of URL filtering is not to remove every single faulty or duplicate URL from the index. Rather, the filtering is used to ensure an efficient web crawling process. Therefore it is reasonable to work on new exclusion patterns when a significant reduction of unwanted URLs can be expected. Pattern detection can be automatized up to a certain extent, but in the end it is always necessary that an expert takes a close look on the index.

The next section gives details about the effect of exclusion patterns on the web page index of PortalU. Moreover the some statistics on the web sites of the information providers is presented.

4 Results

Maintaining a web page index is a continuous task, mainly because there are frequent changes in the set of start URLs, but also because of changing web site structures. In PortalU, the search results are composed from the result sets of six different search engines. The LUBW (Landesamt für Umwelt, Messungen und Naturschutz) in Baden-Württemberg builds a web page index using Google Search Appliance. The InGrid search engine (see again Section 2) is used in five separate installations: the main search engine which is located on the PortalU servers, covering the web sites of the Federal Government, Brandenburg, Berlin, Bavaria, Bremen, Hessen, Hamburg, Mecklenburg-Vorpommern, Northrhine-Westphalia, Schleswig-
Holstein, Saxony-Anhalt, and Thuringia. Moreover, there are installations at the Ministry for Environment, Energy, and Climate Protection in Lower Saxony, at the Ministry for Environment, Agriculture, Nutrition, Viniculture, and Forests in Rhineland-Palatinate, at the LKVK (Landesamt für Vermessung und Kataster) in Saarland, and at the Ministry of Environment and Agriculture in Saxony.

In our presentation we consider only the installation of the InGrid search engine at PortalU. Roughly 1000 start URLs define the entry points to the domains that are targeted by the search engine. Over the period of a last year we continually analyzed the URL database, looking for conspicuous patterns or fragments in the URL database. Before the addition of the first exclude URLs the index comprised roughly 2.3 million web pages. Step by step we added about 650 exclusion patterns, eventually reducing the size of the index to 1.1 million pages. Figure 5 shows how many pages can be related to the individual project partners (the federal states and the federal government).

<table>
<thead>
<tr>
<th>Provider</th>
<th>Seeds</th>
<th>Exclusion Patterns</th>
<th>Web Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brandenburg</td>
<td>1</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Berlin</td>
<td>4</td>
<td>29</td>
<td>37</td>
</tr>
<tr>
<td>Bund</td>
<td>59</td>
<td>182</td>
<td>382</td>
</tr>
<tr>
<td>Bavaria</td>
<td>7</td>
<td>40</td>
<td>27</td>
</tr>
<tr>
<td>Bremen</td>
<td>3</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Hessen</td>
<td>8</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>Hamburg</td>
<td>11</td>
<td>19</td>
<td>40</td>
</tr>
<tr>
<td>Mecklenburg-Vorpommern</td>
<td>31</td>
<td>58</td>
<td>29</td>
</tr>
<tr>
<td>Northrhine-Westphalia</td>
<td>12</td>
<td>586</td>
<td>83</td>
</tr>
<tr>
<td>Schleswig-Holstein</td>
<td>10</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Saxony-Anhalt</td>
<td>23</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td>Thuringia</td>
<td>4</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>173</strong></td>
<td><strong>1004</strong></td>
<td><strong>671</strong></td>
</tr>
</tbody>
</table>

Figure 5

URL Statistics

The table also lists the number of seed URLs and exclusion patterns. Providers are institutions that belong to the domain of a partner, e.g. the umweltbundesamt is a provider of the bund (the Federal Government). Often there is more than just one seed URL for an institution when only certain parts of a domain have environmental relevance.

It is a very striking result that the filtering removed more than a half of all web pages from the index. This emphasizes the need for every search engine to continually monitor the development of pages per domain in order to reliably detect erroneous trends. It also underlines that absolute numbers of web pages have to be considered with care, always keeping in mind that they might not be realistic.

Figure 6 briefly shows the number of web pages from the other search engines. Note that sometimes, as with Baden-Württemberg, the pages from communal authorities are not considered in this statistics. The size of the indexes vary greatly, but here we have no further insights.
The web pages from all the searches together represent public sector information with environmental relevance in Germany. The absolute total of 1,961,540 pages gives only a general direction towards the absolute number of unique pages, certainly there are some more duplicates that could be eliminated with further filter definitions. From our experiences with index maintenance we consider another ten percent of duplicates a fairly conservative estimate, leaving the size of the PortalU index at roughly one million pages. The situation with the other indexes engines (from Figure 6) might be different depending on the efforts that already have been made on their side with respect to URL filtering. A total of 1.5 million web pages might be a realistic assumption.

Should this be considered big or small a number? It is worth looking into the number of pages per web site. Figure 7 shows the ratio between the number of web sites and the number of web pages from that site, grouped by the Federal States or Government. The mean value is 1069, the numbers vary between 276 and 3302, more than one order of magnitude. There is little comparative information to be found on the average number of pages per web site. Estimates on the number of web sites are available, but details about the number of pages are well-kept secrets with all big internet search providers. There are calculations that present 273 as the average number of pages per site in the year 2005. Although this number is several years old and should be regarded with care unless further studies approve of that number, it is nevertheless striking that all page-per-site-values in the PortalU-index as shown in Figure 7 are greater than that number.

Besides that the comparison shows that the number of pages alone does not allow for further conclusions about the potential existence of duplicates, it only gives an indication. However, large numbers of pages per site often point to sites with dynamical structures that can have a seemingly infinite number of pages and are always worth a closer look when analyzing the composition of the index.

It is interesting to see how different the structure and depth of content from various information providers in the public administration are. It confirms that the offers of information range from simple text-based, descriptive content to complex interactive information systems.

For the understanding of these numbers it is also necessary to keep in mind that communal authorities have not been included in the web index. It is very likely that the average number of pages per site would be significantly lower if also web sites of communal authorities would be included in the index. The main rational for that exclusion is that most web sites of the communities do not allow to distinguish, based on the path names (the URLs), between pages with environmental relevance and other pages, making it hard to extract only those pages that are related to environmental topics. One of the main ideas of the web index in PortalU, however, is to deliver search results with high relevance with respect to environmental affairs.

As there certainly is a number of irrelevant pages already in the index it would be very helpful to have an instrument that allows for the distinction between relevant and irrelevant content. An outlook into that problem is given in the next section.

---

5 Outlook

Environmental issues are cross-thematic topics, it is hard to define distinct boundaries that would allow to distinguish them from other topics. For that reason most of the news magazines do not have a separate section Environment. Rather, environmental questions are being discussed in other categories like Politics, Economy, or Science. In the context of PortalU, the situation is different. All web pages in the index have environmental relevance per definition as only the web pages of public institutions dealing with the environment are considered. But to what extent is that assumption true? A web site almost always contains administrative overhead, meaning here: information with little environmental references. Moreover, thematic subsets of a web site can be extracted only if that domain is well-structured in terms of self-indicatory path-names. If such structures cannot be found, which is sometimes the case with content management systems, it is not possible to distinguish between wanted and unwanted thematic subsets of web pages that both come from the same domain. This has serious consequences for thematic search engines.

While general search engines like google or bing just crawl any web page they can reach, a thematic search engine like PortalU offers a search on a selected subset of web pages with high relevance towards a specific theme.

One possible solution to that problem lies in automatic classification. In a previous paper (Schenk, Meyerholt 2011) we presented the results of our experiments with automatic classification of web page content with respect to environmental issues. The results where not overwhelmingly persuasive, though. The main reason is that the training of the algorithms depends on well-selected test sets. Two sets are necessary: one composed of relevant information, the other set should contain solely unrelevant pages. We found that composing the training sets based on the extraction of all sites of a whole domain did not deliver satisfying results. Although we chose domains that seemed to be very well suited (e.g. umweltbundesamt.de for relevant pages, unrelevant pages composed from all pages of some ministries not dealing with environmental questions) we found that the trained algorithms were not able to reliably determine relevance in further test data to a satisfying degree. Improvement seems to be possible only when handpicked training sets could be used.

But automatic classification might become easier with a well-kept web page index. The training could be based on the PortalU web index for relevant pages and on a second set, that still has to be composed by hand. This is the direction that we intend to take in the future, but a second set of web pages representing non-relevant information still has to be prepared. Training the algorithms with these sets could provide for some interesting results on the classification of web sites where the extraction of relevant content is not possible yet.

Bibliography

