Ein PHP-Wrapper für die Internet-Suchmaschine

ht://Dig

oder:

„Google selbstgebaut“

Vortrag für die
PHP-Usergroup Hannover
am 1.7.2004
von Frank Staude und
Udo Schacht-Wiegand

Zusammenfassung

ht://Dig ist eine Suchmaschine für Internet-Seiten. Die Suchergebnisse werden per cgi-Script auf einer HTML-Seite gezeigt. Wir werden sehen, wie man mit einem kleinen PHP-Script die Ergebnisse auch auf dynamisch erzeugten PHP-Seiten zeigen kann. Die genauere Konfiguration und Verwendung von ht://Dig ist sehr komplex und geht über den Umfang dieses Vortrags weit hinaus. Wir verweisen auf die Homepage von htdig.org mit einer ausführlichen Dokumentation und FAQ.

Was ist ht://Dig?

Das Programmpaket ht://Dig (engl.: to dig = „graben“) ist eine komplette Suchmaschine für Intra- und Internet-Dokumente. Die Suche beschränkt sich dabei nicht auf HTML-Seiten, sondern es können auch viele andere Dokument-Arten durchsucht werde, z.B. .doc- und .pdf-Dateien, sofern entsprechende Plugins installiert werden.

ht://Dig liegt derzeit in der stabilen Version 3.1.6 vom 1.2.2002 (!) vor. Es existiert aber auch eine Beta 3.2.0b5 vom 11.11.2003, die wir für Test verwendet haben. Ganz neu: Beta 3.2.0b6 vom 16.6.2004. Merke: Ein gutes Programm muss man auch nicht wöchentlich weiterentwickeln.

Das Paket besteht aus drei Hauptkomponenten:

- htdig: Das Programm zum Durchforsten der Webseiten und zum Sammeln der Daten.
- htmerge: Konvertierung und Indexierung der gefundenen Daten zu einer Datenbank. Typischerweise lässt man htdig und htmerge einmal jede Nacht laufen.
- htsearch: Das eigentliche Suchprogramm, ein cgi-Script, welches durch ein Webformular gestartet wird.

Zusätzlich werden folgende Programme mitgeliefert:

- htfuzzy: Eine „unscharfe“ Suche, die zu gesuchten Wörtern auch jene mit anderen Endungen findet. Je nach Konfiguration können hiermit auch Synonyme gefunden werden. htfuzzy wird standardmäßig verwendet.
- htnotify: Informiert den Webmaster über veraltete Dokumente (Seiten die seit einem festgesetzen Datum nicht mehr verändert wurden).

ht://Dig wurde unter der GNU GENERAL PUBLIC LICENSE veröffentlicht.

Features

- Suche im Intranet bzw. Internet.
- Suche per HTTP oder auf Dateisystemebene möglich.
- Kostenlos (GPL)
- Beachtet robots.txt-Anweisungen
- Boolsche Verknüpfungen (AND, OR, NOT)
- Konfigurierbare Suchergebnisse
- Unscharfe Suche (Fuzzy)
- Durchsucht HTML- und Textdateien (sowie andere mit Plugins)
- In HTML-Dateien können zusätzlich Schlüsselworte eingefügt werden
- Email-Benachrichtgung über veraltete Dokumente
- Ein password-geschützter Server kann indexiert werden (eingeschränkt)
- Suche in Teilbereichen der Datenbanken
- Programm-Quelltext verfügbar
- Tiefe der Suche (Anzahl von Links zum Dokument) kann begrenzt werden
- Unterstützung des ISO-Latin-1 Zeichensatzes

Arbeitsweise

htdig greift (normalerweise) per http-Protokoll auf Dokumente zu und folgt dann den Hyperlinks zu weiteren Dokumenten. Dabei „sieht“ htdig den Quelltext der Seiten, also z.B. auch Kommentare, die aber nur bei besonderer Konfiguration auch für die Suche herangezogen werden können. Per Konfigurationsdatei wird festgelegt, wie vielen Unter-Verzweigungen htdig folgt und ob es die Domain des „Start-URL“ verlassen darf. Durch das http-Protokoll ist die Suche nicht auf den „eigenen“-Server beschränkt. Theoretisch könnte man mit htdig auch „das Netz“ komplett durchsuchen – vorausgesetzt man hätte genügend Zeit und den erforderlichen Plattenplatz. Im Schnitt rechnet man 12 kB pro HTML-Seite.

HINWEIS: Es ist auch möglich, auf Dateisystem-Ebene zu indexieren. Dabei sind htaccess und in PHP geschriebene Anmeldesysteme außer Kraft. Das Script sieht alles - AUCH die Kennwörter für die Datenbank in den PHP-Dateien!

htdig (und htmerge) legen eine Datenbank mit einem Index zu jedem gefundenen Wort an. Dabei wird zu jedem Wort nicht nur die Fundstelle (URL) gespeichert, sondern auch die Position im Text. Daraus kann htsearch bei der Anzeige der Treffer auch die Umgebung des Suchwortes anzeigen.

Installation

Wir beziehen uns hier auf die Debian-Distribution:

```
apt-get install htdig
```

Hier werden folgende Pfade verwendet:

- /etc/htdig/htdig.conf: Konfigurationsdatei
- /usr/lib/cgi-bin/htsearch: cgi-Script
- /var/lib/htdig: Datenbank-Verzeichnis
- /var/www/search.html: Suchformular
- /var/www/htdig: Images und Templates
- /usr/bin: Binaries

Für Windows-User siehe unter Quellen: Idiots Guide to Install htdig on Win32

Inbetriebnahme

Mit dem praktischen Shell-Script „rundig“ (/usr/bin/rundig) kann man htdig und htmerge gemeinsam aufrufen. Beim ersten Aufruf wird die Datenbank (Suchindex) angelegt, weitere Aufrufe ergänzen die Datenbank. Das Verhalten kann natürlich auch über Aufrufparameter gesteuert werden.

Die Debian-Installation legt folgende cron-Jobs an:

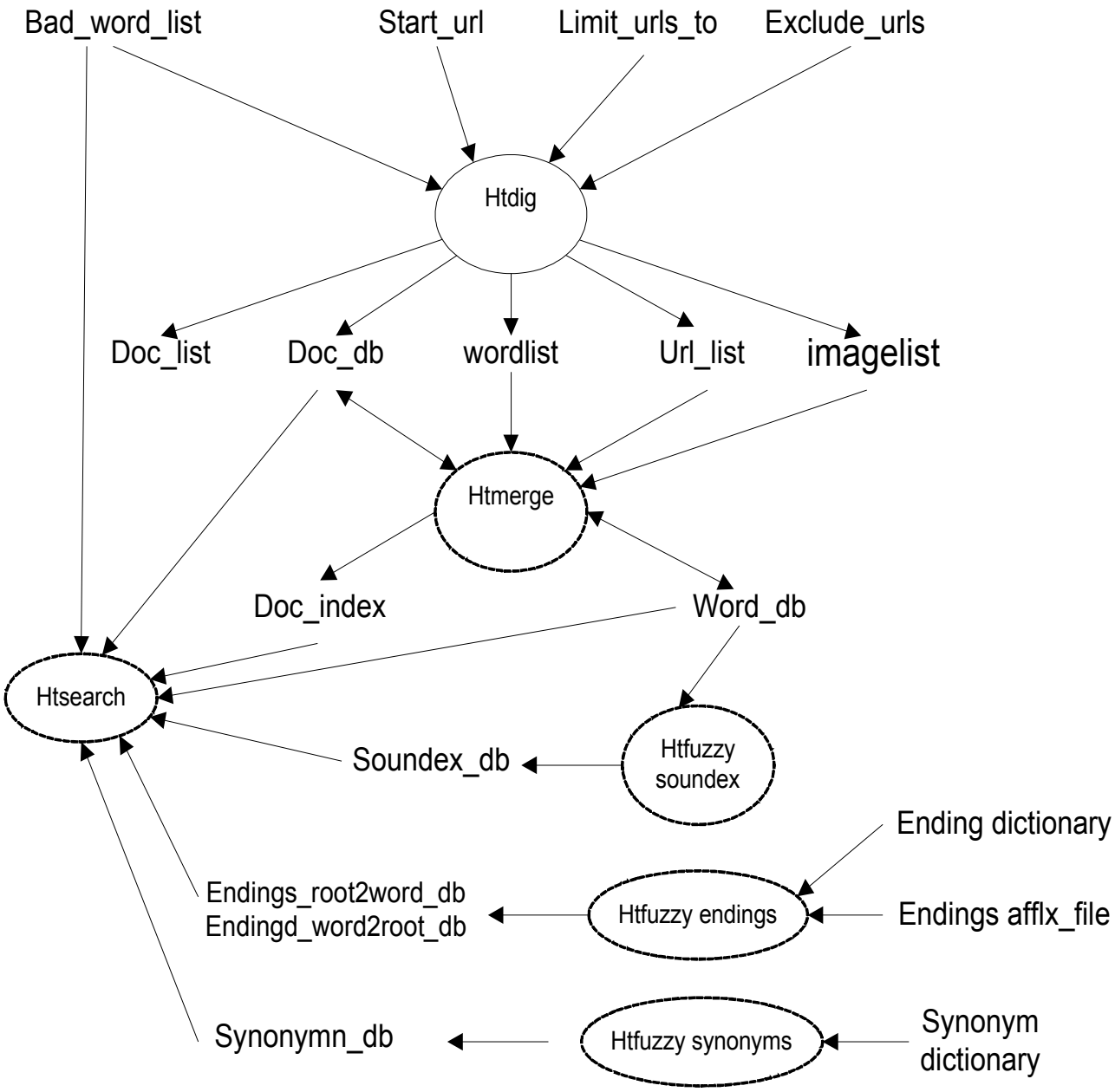
- /etc/cron.daily/htdig - täglich den Suchindex aktualiseren (ergänzen)
- /etc/cron.weekly/htdig – wöchentlich den Suchindex neu aufbauen.

Konfiguration

In /etc/htdig/htdig.conf kann das Verhalten des Programmpakets detailliert festgelegt werden. Eine vollständige Dokumentation findet sich auf der ht://Dig-Webseite. Wir können hier nur die wichtigsten Parameter wiedergeben:

- start_url: Ausgangspunkt(e) der Suche, es können mehrere, durch Whitespace getrennte URLs angegeben werden.
- limit_urls_to: Beschränkung auf die angebenen URLs. Achtung „.“ würde das ganze Internet durchsuchen. Üblich ist z.B.: ${start_url}.
- exclude_urls: Seiten, deren URL diesen String enthalten, werden nicht durchsucht.
- bad_extensions: Dateien, die nicht durchsucht werde, z.B. .gif, .jpg usw.
- max_head_length: Soviele Bytes (ohne HTML-Markup) werden in die Datenbank aufgenommen, um bei der Ergebnisanzeige die Umgebung des Treffers anzuzeigen.
- max_doc_size: Soviele Bytes eines Dokuments werden maximal für die Suche verwendet.
- bad_word_list: Liste unerwünschter Begriffe
- use_star_image: no Zeigt keine Sternchen bei den Ergebnissen an.

Funktionsgrafik ht://Dig

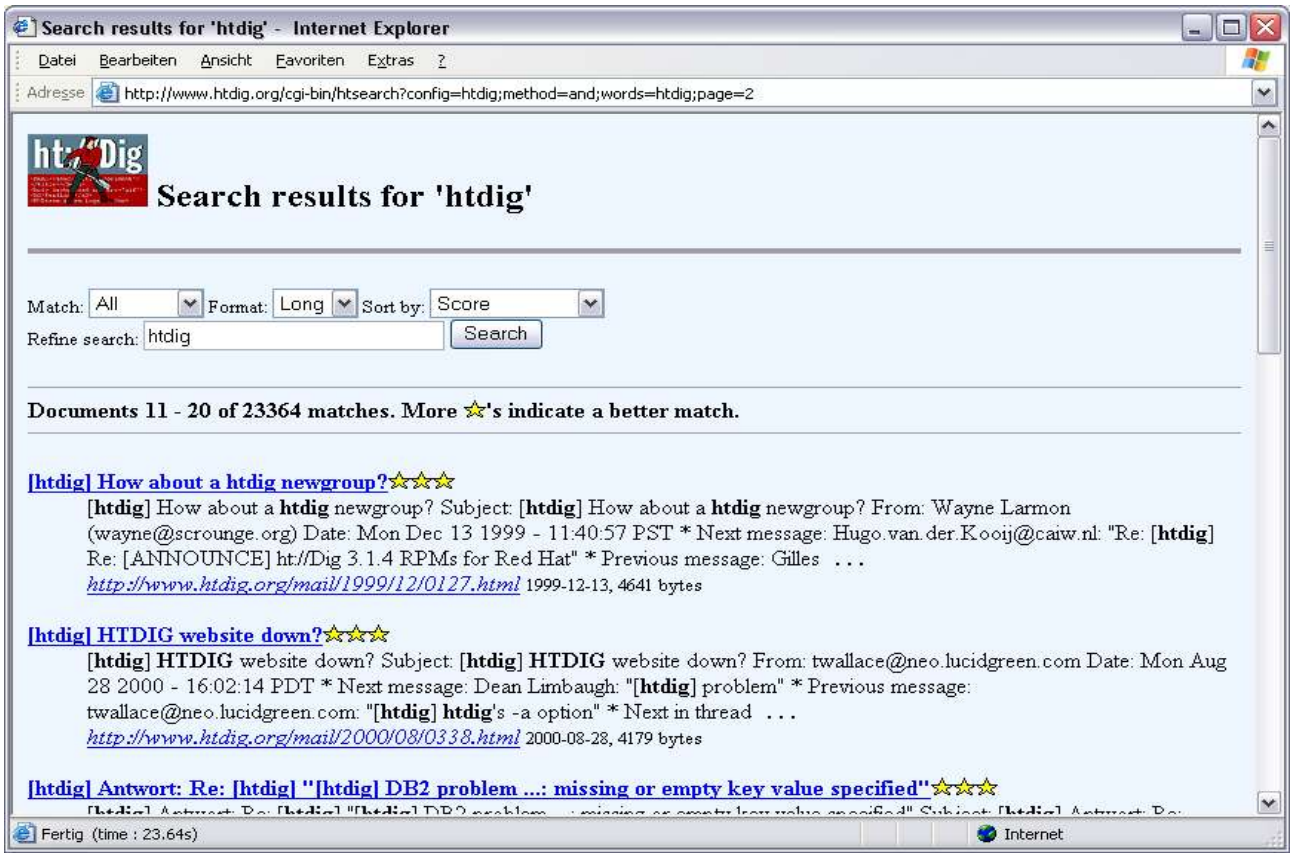


Datenbanken

Doc_list:	Liste der Dokumente
Doc_db/Doc_Index:	Datenbank/Index der Dokumente
wordlist/Word_db:	Liste/Datenbank aller vorkommenden Wörter
Url_list:	Liste der URLs
imagelist:	Liste der Bilder
Soundex_db	Datenbank ähnlich klingender Wörter
Endings	Datenbank mit Endungen
Synonym_db	Datenbank mit Synonymen (Wörter gleicher Bedeutung)

Anzeige der Ergebnisse (htsearch)

Die Anzeige der Suchergebnisse wird über Templates gesteuert. Ohne weitere Konfiguration (default) sieht ein Suchergebnis so aus:



Quelltext der Default-Suchergbnis-Seite:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<html><head><title>Search results for '$&(WORDS)'</title></head>
<body bgcolor="#eef7ff">
<h2><img src="/htdig/htdig.gif" alt="ht://Dig">
Search results for '$&(LOGICAL_WORDS)'</h2>
<hr noshade size="4">
<form method="get" action="$(CGI)">
<font size="-1">
<input type="hidden" name="config" value="$&(CONFIG)">
<input type="hidden" name="restrict" value="$&(RESTRICT)">
<input type="hidden" name="exclude" value="$&(EXCLUDE)">
Match: $(METHOD)
Format: $(FORMAT)
Sort by: $(SORT)
<br>
Refine search:
<input type="text" size="30" name="words" value="$&(WORDS)">
<input type="submit" value="Search">
</font>
</form>
<hr noshade size="1">
<strong>Documents $(FIRSTDISPLAYED) - $(LASTDISPLAYED) of $(MATCHES) matches.
More <img src="/htdig/star.gif" alt="*">'s indicate a better match.
</strong>
<hr noshade size="1">
$(HTSEARCH_RESULTS)
$(PAGEHEADER)
$(PREVPAGE) $(PAGELIST) $(NEXTPAGE)
<hr noshade size="4">
<a href="http://www.htdig.org/">
<img src="/htdig/htdig.gif" border="0" alt="">ht://Dig $(VERSION)</a>
</body></html>
```

Anstelle von `$(HTSEARCH_RESULTS)` wird folgendes Template eingesetzt, welches die einzelnen Suchergebnisse zeigt:

```
<dl><dt><strong><a href="$&(URL)">$&(TITLE)</a></strong>$(STARSLEFT)
</dt><dd>$(EXCERPT)<br>
<em><a href="$&(URL)">$&(URL)</a></em>
<font size="-1">$(MODIFIED), $(SIZE) bytes</font>
</dd></dl>
```

In den Templates können eine Vielzahl von Variablen verwendet werden, um die Ergebnisse darzustellen, üblicherweise als $(VARIABLE). Hier nur ein paar Beispiele, für eine vollständige Liste siehe http://htdig.org/hts_templates.html:

CURRENT
: The number of the current match.

DESCRIPTION
: The first URL description for the matched document.

EXCERPT
: The relevant excerpt for the current match

HOPCOUNT
: The distance of this match away from the starting document(s).

KEYWORDS
: A string of the search keywords with spaces in between, as specified in the *keywords* input parameter.

LOGICAL_WORDS
: A string of the search words with either "and" or "or" between the words, depending on the type of search.

MATCHES
: The total number of matches that were found.

MATCHES_PER_PAGE
: The configured maximum number of matches on this page

MAX_STARS
: The configured maximum number of stars to display in matches.

METADESCRIPTION
: The meta description text (if any) for the matched document.

MODIFIED
: The date and time the document was last modified

NEXTPAGE
: This expands to the value of the next_page_text or no_next_page_text attributes depending on whether there is a next page or not.

PAGE
: The current page number.

PAGELIST
: This expands to a list of hyperlinks using the page_number_text and no_page_number_text attributes.

PERCENT
: The match score as a percentage. Its range is 1 to 100, without a percent sign. The minimum is always 1 so the variable can be used as the value for an HTML WIDTH attribute.

SCORE
: The score of the current match

STARTYEAR, **STARTMONTH**, **STARTDAY**, **ENDYEAR**, **ENDMONTH**, **ENDDAY**
: The currently specified date range for restricting search results.

TITLE
: The title of the document for the current match

URL
: The URL to the document for the current match

VERSION
: The ht://Dig version number

WORDS
: A string of the search words with spaces in between.

Probleme mit ht://Dig

Es gibt einige unvermeidliche Probleme bei der Verwendung von Suchmaschinen in dynamisch generierten Seiten:

- Gleiche Seite unter verschiedenen URLs: So kommt es z.B. bei der Verwendung von Session-IDs im URL dazu, dass htdig eine Seite hundert oder tausendfach auflistet, weil immer wieder eine andere Session-ID generiert wird.
    - Abhilfe: die unerwünschten Teile des URL in bad_extensions angeben: `bad_extensions: ?PHPSESSIONID`

- Navigation und Inhalt in einer Seite: Oft findet man viele Treffer, die das Suchwort in der Navigation, aber nicht im eigentlichen Inhalt (Artikel) enthalten.
    - Abhilfe: Die Navigation wird mit den Tags `<!--htdig_noindex-->` `<!--/htdig_noindex-->` umgeben. Dieser Bereich wird dann nicht von htdig in die Datenbank aufgenommen

Ausgabe in einer PHP-Seite

Zwar lässt sich das Layout der Ergebnisseite in weiten Bereichen anpassen, um jedoch htdig wirklich in dynamische Websites zu integrieren bedarf es einiger Überlegungen. Um die Ausgabe von htdig in PHP-generierten Seiten sinnvoll einbauen zu können, haben wir einen kleinen „Wrapper“ geschrieben.

Funktionsweise

Das Such-Formular steckt in einer PHP-Seite, die sich beim Abschicken der Suche selber aufruft. Sobald die Suche gestartet werden soll, wird das eigentliche htsearch-cgi-Script über die PHP-Funktion fopen() geöffnet.

Fopen() ruft die Webseite (URL) auf. Glücklicherweise kann htdig hier komplett mit GET-Variablen arbeiten, so das einfach der (komplexe) URL übergeben wird. Das Ergbnis wird jedoch nicht sofort sichtbar, sondern über den Filehandle von fopen() zurückgeliefert.

Dadurch können wir den Quelltext der zurückgelieferten Seite in einem String ($content) speichern und mit PHP weiter verarbeiten. Nun haben wir die Möglichkeit, ein paar Anpassungen durch Suchen und Ersetzen vorzunehmen. Schließlich wird der String (die fertige Ergebnisseite) von PHP einfach ausgegeben.

Konfiguration

Zunächst konfigurieren wir ht://Dig so, dass die Anzeige des Ergebnisses unseren Wünschen so nahe wie möglich kommt und möglichst wenig Unerwünschtes angezeigt wird. Dazu müssen wir einige eigene Templates anlegen, am besten in /var/www/htdig.

Eigene Templates anlegen:

/var/www/htdig/ergebnis.php (Komplette Ergebnisseite)

```
$(MATCHES) Dokument(e) gefunden - (Seite $PAGE von $PAGES):
<DL>$(HTSEARCH_RESULTS)
</DL>
<br><br>
<!--SPLIT_HERE-->
$(PREVPAGE) $(PAGELIST) $(NEXTPAGE)
```

/var/www/htdig/treffer.php (Einzelner Treffer)

```
<dt>$(CURRENT))
<a href="$(URL)">$(TITLE)</a>
<dd>$(EXCERPT)
<br>
<span class="klein">
Gr&ouml;&szlig;e: $(SIZEK) kb - Relevanz: $(PERCENT)%</span><br><br>
```

In der Konfigurationsdatei werden die Templates so angegeben:

/etc/htdig/htdig.conf (Ausschnitt)

```
www:                    /var/www/htdig
ergebnis:               ergebnis.php
treffer:                treffer.php

search_results_wrapper: $www/$ergebnis
template_map:           HIT hit $www/$treffer
template_name:          HIT


allow_in_form:          search_results_wrapper \
                        template_map \
                        template_name \
                        script_name
```

Die Variablen www, ergebnis und treffer sind eigene Variaben, die von der Suchform wiederum überschrieben werden können (allow_in_form), z.B. als hidden-Variablen. Dadurch können andere Formulare auf dem Server andere Templates verwenden.

Eine nette Konfiguratiosmöglichkeit ist `script_name`: Damit wird schon von htdig der URL des Scripts in den von htdig generierten Links angepasst. Dieses betrifft u.a. die Seitenverweise am unteren Rand der Ergebnisseite.

Das Suchergebnis sieht jetzt so aus:

```
1735 Dokument(e) gefunden - (Seite 1 von 10):
1) Wiegand.Name - a .name of it's own
      ... Extra Info ... Statistics Member Since 2004/6/16 Rank Webmaster Comments/Posts 0 Last
      Login 2004/6/29 10:26 Signature ... News News Kritik an Skype? (2004/6/29 10:27:26) News
      Skype: Internet-Telefonie vom Feinsten (2004/6/28 23:34:21) News Wiegand.Name goes online
      (2004/6/16 22:16:46)
      Größe: 9 kb - Relevanz: 100%

2) Wiegand.Name - News
      ... Headlines Login Username: Password: Lost Password? Register now! Random Picture Site Info
      Webmasters admin   Recommend Us Voice over IP : Kritik an Skype? Posted by admin on 2004/6/29
      10:29:22 (7 reads) In einigen Open-Source-Foren wurde Kritik an Skype laut. Ob sie berechtigt
      ist, kann ich persönlich ...
      Größe: 10 kb - Relevanz: 78%
```

Dieser Text lässt sich schon viel leichter parsen. Sämtliche Grafiken sind entfallen. Mit dem folgenden PHP-Script erledigen wir jetzt den Rest: Den Aufbau unserer gewünschten dynamischen Seite:

PHP-Code:
Im Suchformular verwenden wir folgenden PHP-Code:

```
<?
// Suche auswerten
if($action=="suchen") {

  $url = "http://UNSER.URL/cgi-bin/htsearch?"
  . "&words=".urlencode($words)
  . "&method=$method"
  . "&sort=$sort"
  . "&treffer=treffer.php"
  . "&ergebnis=ergebnis.php"
  . "&page=$page"
  . "&exclude=$exclude"
  . "&script_name=DIESES-SCRIPT.php?action=suchen";
  // DEBUG
  // echo $url."<p>";

  if($fp = fopen($url,'r')) {
    while(!feof($fp)) {
      $content .= fgets($fp, 1024);
    }
    fclose($fp);
    if(strlen($content)<100) {
      echo "Leider nichts gefunden";
    } else {
      list($oben,$unten) = split("<!--SPLIT_HERE-->",$content);
      echo $oben;

      $unten = str_replace("[prev]", "[vorherige]", $unten);
      $unten = str_replace("[next]", "[nächste]", $unten);

      $trans = array(";" => "&", "suchen?" => "suchen&");
      echo strtr($unten,$trans);
    }
  }
  echo "<br><br>";
}
?>
```

Der $url wird über fopen($url, 'r') aufgerufen. Das Ergebnis wird in $content zurückgeliefert und dann einfach mit split() in zwei Teile zerlegt. Nur im unteren Teil ersetzen wir die englischen Beschriftungen 'prev' und 'next' durch die deutschen Worte.

Eine Besonderheit erledigt noch der Befehl `$trans = array(";" => "&", "suchen?" => "suchen&");` Er wandelt die ';' in '&'-Zeichen um, da htdig die Paramter im URL mit ';' getrennt übergibt.

Wird nichts gefunden oder falls ein anderer Fehler auftritt, so wird einfach „Leider nichts gefunden“ angezeigt, da die Länge von $content weniger als 100 Zeichen beträgt.

Ausblick

Die hier gezeigte Methode, die Ausgabe eines cgi-Scripts mit PHP weiter zu verarbeiten, lässt sich sicherlich auch anderswo anwenden. Für ht://Dig überlegen wir, die Ausgabe des htsearch-Templates noch weiter zu vereinfachen, z.B. ohne Ausgabe irgendwelcher sprachabhängigen Elemente zu gestalten. Die Ausgabe würde damit gewissermaßen „maschinenlesbar“. Die weitere Verarbeitung würde flexibel mit PHP vorgenommen und das Layout mit Style-Sheets gesteuert und damit nahezu beliebig anpassbar und barrierefrei sein. So könnte man z.B. ein htdig-Template bauen, welches das Ergebnis als XML liefert, und somit Suchergebnisse in vielen anderen Programmen weiterverarbeiten.

Quellen

ht://Dig Homepage:
http://htdig.org

Ganz gute deutsche Bschreibung
http://www.suse.de/cgi-bin/print_page_www.pl?NPSPath=/webredesign/htdocs/de/private/support/online_help/howto/htdig/index.html

Durchsuchen von PDFs, deutsche Anleitung:
http://www.linuxkramkiste.de/htdig315.html

Idiot's Guide to installing ht://dig on Win32.
http://www.htdig.org/files/contrib/guides/Installing_on_Win32.html

Zugriff auf HTDig mit Perl-Modulen:
HtDig::Database
Perl interface Ht://Dig docdb and config files
http://search.cpan.org/~ghutchis/HtDig-Database-0.52/Database.pm

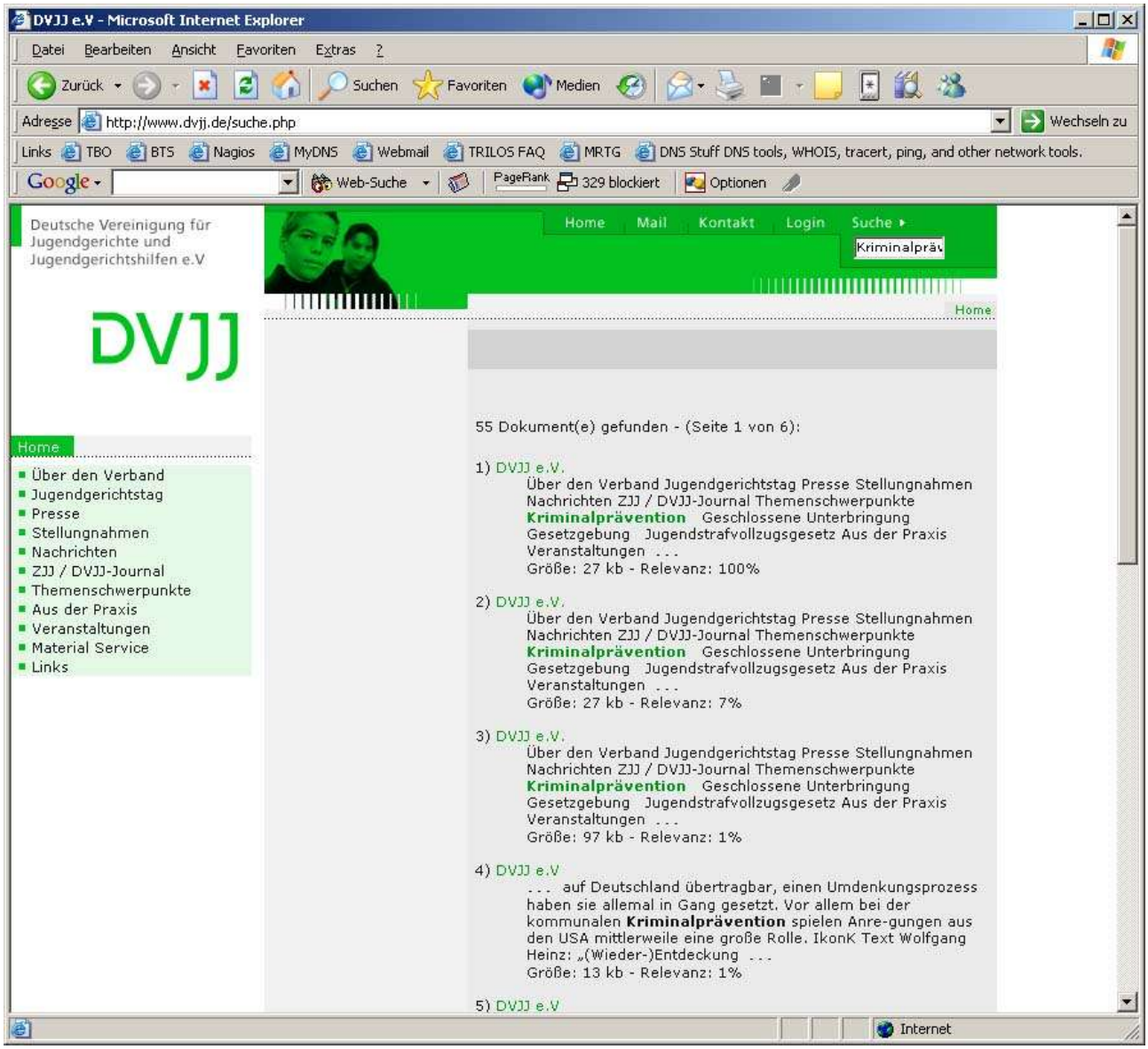
HtDig::Config
Perl extension for managing ht://Dig configuration files
http://search.cpan.org/~jtillman/HtDig-Config-1.01/Config.pm

HtDig::Site
Perl extension for managing a single ht://Dig configuration
http://search.cpan.org/~jtillman/HtDig-Config-1.01/Site.pm

PHP: htdig-php-helper - model an htdig search request as a PHP object:
www.phpclasses.org/browse/package/536.html

Htdiginterface:
This is a PHP class which interfaces with the ht://Dig programs,
allowing you to index and search Web pages from PHP. It is able to setup
a suitable configuration file from a few user-defined parameters, index
Web pages to build the search databases, and search the indexed database
to capture the matches into a PHP data structure ready to be used to
display the results in a PHP-generated page.
www.phpclasses.org/browse/package/26.html

Beispielseite



Ergebnisanzeige auf der Seite von dvjj.de als .php-Seite und im entsprechenden Layout.