

Der PortalU[®] Ranking-Mechanismus

Version 0.2

04.12.2006



Koordinierungsstelle
PortalU

Änderungshistorie

Alle wesentlichen Änderungen und Ergänzungen für jede Version des Dokumentes werden in der untenstehenden Tabelle kurz aufgeführt.

Version	Datum	Änderung
0.1	23.11.2006	Entwurf des Dokumentes, Hr. Bauhardt / 101tec GmbH
0.2	04.12.2006	Überarbeitete Version, Kst. PortalU®

Inhalt:

1	Einführung	3
2	Berechnung des Rankingwerts eines Dokuments unter Verwendung einer Query	5
2.1	<i>Termfrequenz</i>	6
2.2	<i>Inverse Dokumenthäufigkeit</i>	6
2.3	<i>Score</i>	7
2.4	<i>Beispielrechnung</i>	8
3	Anhang	9
3.1	<i>Bildschirmkopie der in Punkt 2 analysierten Seite</i>	9
3.2	<i>Text-Auszug der in Punkt 2 analysierten Seite</i>	9
3.3	<i>Weiterführende Links</i>	11

1 Einführung

Die Berechnung des Rankings für ein Dokument übernimmt in InGrid die Indexierungskomponente *Lucene*.

➔ <http://lucene.apache.org/>

➔ <http://de.wikipedia.org/wiki/Lucene>

Die in PortalU eingesetzte Web-Such-Maschine *Nutch* basiert auf Lucene und fügt weitere, für eine Websuche erforderliche, Bausteine hinzu: Einen Crawler, eine Querverlinkungsdatenbank sowie Parser für HTML, .doc, .pdf und andere Dokument-Formate.

➔ <http://lucene.apache.org/nutch/index.html>

Im Rahmen der automatischen Volltextindexierung durch Lucene werden, mit Ausnahme von definierten Stoppwörtern und HTML-Tags, zunächst alle Wörter eines Dokuments extrahiert und in das Indexfeld *content* gespeichert. Weiterhin wird bei HTML-Seiten der Inhalt des `<title>`-Tags der Seite in das Indexfeld *title* geschrieben. Handelt es sich bei den zu indexierenden Dokumenten nicht um HTML-Seiten, sondern zum Beispiel um PDF-, oder MS Word-Dokumente, so werden die ersten 100 Zeichen des Textes als *title* extrahiert. Alternativ können auch die Titel-Metadaten der proprietären Dokumentformate zur Füllung von *title* genutzt werden. In der Praxis lieferte dieser Ansatz allerdings schlechtere Ergebnisse, da diese Metadaten nicht hinreichend gepflegt sind.

Weitere Indexfelder die aus den zu indizierenden Seiten extrahiert werden und die für das Ranking von Bedeutung sind:

- *host*: Auf Einzelausdrücke reduzierte („tokenized“) Top-Level-Domäne der Dokument-URL. Zum Beispiel: `http www umweltbundesamt de`
- *url*: komplette URL des Dokuments, z.B. `http www umweltbundesamt de abfallwirtschaft index htm` (ebenfalls „tokenized“)
- *anchor*: Dieser Parameter wird für jedes Dokument aus der Querverlinkungsdatenbank extrahiert. Er speichert die Ankertexte aller Links aller anderen Seiten im Index, die auf die aktuelle Seite verlinken. Beispiel für einen Ankertext:

```
<a href="http://www.xxx.de/index.html">Dies ist ein Ankertext</a>
```

➔ http://de.wikipedia.org/wiki/Automatische_Indexierung

Unabhängig von Lucene wird das Dokument vom *Semantic Network Service* des Umweltbundesamtes (SNS) analysiert. Die vom SNS-Webservice (SNS *autoclassify*-Aufruf) zurück gelieferten Terme werden in die Indexfelder *buzzword* geschrieben.

➔ <http://www.semantic-network.de>

Für die Ermittlung des Rankingwertes eines Dokuments im Hinblick auf eine bestimmte Anfrage (Query) wird von PortalU der im *Information Retrieval* häufig genutzte *TF/IDF Algorithmus* benutzt. Der TF/IDF Algorithmus ist eine Gewichtungsmethode für Terme und basiert auf der *Termfrequenz* (TF) und der *inversen Dokumenthäufigkeit* (inverse document frequency, IDF).

➔ http://de.wikipedia.org/wiki/Information_Retrieval

Die TF eines Begriffs in einem Dokument liefert einen Hinweis auf die Bedeutung dieses Terms für das Dokument. Die IDF misst die allgemeine Bedeutung des Terms über alle im Index vorhandenen Dokumente. Zunächst wird die Häufigkeit eines Begriffs in einem Dokument ermittelt. Dieser Wert wird mit der Häufigkeit der Dokumente, in denen der Begriff vorkommt, ins Verhältnis gesetzt. So erhält man einen Hinweis auf den Wert des Begriffs als Deskriptor. Die Gewichtung eines Begriffs ist hoch wenn wenige Dokumente, in denen der Begriff enthalten ist, im Index vorhanden sind, der Begriff im zu indexierenden Dokument aber häufig vorkommt.

➔ <http://de.wikipedia.org/wiki/TF-IDF>

Im Rahmen der Indexierung werden die für das Indexfeld *content* extrahierten Begriffe außerdem mittels *Stemming* auf einen gemeinsamen Wortstamm zurückgeführt, z.B. „wass“ statt „wasser“. Für *title*, *anchor*, *host*, *url* und *buzzword* werden die Originalterme erhalten, eine Wortstammreduktion findet nicht statt.

➔ <http://de.wikipedia.org/wiki/Stemming>

Neben dem TF/IDF-basierten Ranking spielt für die Ermittlung des Rankingwertes eines Dokuments in PortalU eine Rolle, wie häufig dessen URL aus anderen Dokumenten im Index verlinkt wird. Ein derartiges Verfahren kommt z.B. auch bei google mit dem *Page-Rank-Algorithmus* zum Tragen. Die InGrid-Suchmaschine *Nutch* nutzt, analog zu Page-Rank, *OPIC* (Adaptive On-Line Page Importance Computation). Eine detaillierte Beschreibung der Arbeitsweise von OPIC findet sich unter der folgenden Webadresse:

➔ <http://www2003.org/cdrom/papers/refereed/p007/p7-abiteboul.html>

➔ <http://de.wikipedia.org/wiki/PageRank>

Lucene boostet bereits beim Indexieren alle Indexfelder eines Dokuments mit dessen OPIC-Wert.

2 Berechnung des Rankingwerts eines Dokuments unter Verwendung einer Query

Für das Dokument <http://umwelt.schleswig-holstein.de/servlet/is/22624/> soll der Rankingwert für die Query **Wasser** bestimmt werden. Die folgende Auflistung zeigt einen Auszug aus der zugehörigen Indexzeile:

```
segment      = 20060804162310
digest       = 6e0afc66ad3338e5ddee98bb0d9de171
boost        = 1.2027289
buzzword     = gewässer
buzzword     = wasserpolitik
buzzword     = heilwasser
buzzword     = salzwasser
buzzword     = trinkwasser
buzzword     = wasserrecht
buzzword     = grundwasser
buzzword     = wasserschutz
buzzword     = stauwasser
buzzword     = küstengewässer
buzzword     = niederschlagsgebiet
buzzword     = hydrodynamik
buzzword     = schwarzwasser
buzzword     = süßwasser
buzzword     = warmwasser
buzzword     = wasserhaushalt
buzzword     = wasseroberfläche
buzzword     = wasserstand
buzzword     = wasservorkommen
buzzword     = oberflächenwasser
buzzword     = brunnen
buzzword     = speisewasser
buzzword     = tiefwasser
buzzword     = infiltrationswasser (eindringendes wasser)
buzzword     = wasserinformationssystem
buzzword     = brauchwasser
buzzword     = eis
buzzword     = meer
buzzword     = badegewässer
buzzword     = eg-wasserrahmenrichtlinie
partner      = sh
lang         = de
topic        = water
content      = ...
alt_summary  = no_alt_summary
datatype     = www
datatype     = default
datatype     = measure
provider     = sh_munl
primaryType  = text
subType      = html
url          = http://umwelt.schleswig-holstein.de/servlet/is/22624/
title        = Wasser
```

Auf die komplette Auflistung des Inhalts von „content“ wurde aus Platzgründen verzichtet. Die Lucene-Query setzt sich aktuell aus folgenden Bestandteilen zusammen:

```
+(datatype:default)
```

```
+(url:wasser^4.0 anchor:wasser^2.0 content:wass title:wasser^7.0
host:wasser^2.0)
```

Hat der Nutzer in der Personalisierung die Option „*Ähnliche Begriffe in die Suche einbeziehen*“ aktiviert, erweitert sich der untere Teil der Query um den Parameter *buzzword*:

```
+(url:wasser^4.0 anchor:wasser^2.0 content:wass title:wasser^7.0
host:wasser^2.0 buzzword:wasser)
```

Die Werte hinter „^“ sind die aktuellen Boosting-Faktoren. Das „+“ bedeutet, dass der folgende Ausdruck wahr sein muss. Der Indexeintrag eines Dokuments muss damit die folgenden Kriterien erfüllen um in die Ergebnismenge aufgenommen zu werden:

- datatype:default (Pflicht, zusätzliche Datatypes möglich)

UND

- Es muss „wasser“ bzw. „wass“ in (url UND/ODER anchor UND/ODER content UND/ODER title UND/ODER host [UND/ODER buzzword]) gespeichert sein

2.1 Termfrequenz

Die Termfrequenz (TF) gibt die relative Häufigkeit eines Wortes beziehungsweise Termes in einem Dokument an. Sie dient als Indikator der Repräsentativität des Wortes für den Inhalt des Gesamtdokumentes und berechnet sich für PortalU wie folgt:

$$TF(t, d) = \sqrt{\frac{h(t)}{nt(d)}}$$

mit

- $h(t)$: Häufigkeit des Terms t im Dokument d
- $nt(d)$: Gesamtzahl der Terme (ohne Stoppwörter, HTML-Tags etc.) im Dokument d

2.2 Inverse Dokumenthäufigkeit

Die Inverse Dokumenthäufigkeit (IDF) dient zur Bestimmung der Trennfähigkeit eines Wortes bzw. Termes. Ein Wort, das nur in wenigen Dokumenten oft vorkommt hat in diesem Sinne eine höhere Bedeutung als eines, das in sehr vielen Dokumenten oder insgesamt nur sehr selten auftritt. Die IDF wird in PortalU mit der folgenden Formel bestimmt:

$$IDF(t) = 1 + \log\left(\frac{n(d)}{nd(t) + 1}\right)$$

mit

- $n(d)$: Anzahl der Dokumente im Index
- $nd(t)$: Anzahl der Dokumente, die den Term t enthalten

2.3 Score

Die Höhe des Rankingwerts ergibt sich aus folgender Formel:

$$Score = \sum_{tq} qw \cdot fw$$

mit

- tq : terms in query
- qw : query weight
- fw : field weight

wobei

$$qw = queryFieldBoost \cdot idf \cdot queryNorm$$

$$fw = docFieldBoost \cdot tf \cdot idf \cdot fieldNorm$$

mit

- $queryFieldBoost$: Faktor, mit dem das betreffende Feld multipliziert („geboostet“) wird, kommt bei der Suche zum tragen
- $docFieldBoost$: Boostingfaktor eines Indexfeldes, der bei der Indexierung zum tragen kommt, Default: 1
- $query-/fieldNorm$: Normalisierungsfaktoren (s.u.)
- idf : inverse document frequency
- tf : term frequency

Die Normalisierungsfaktoren sorgen für einen sinnvollen Gesamt-Rankingwert:

$$queryNorm = \frac{1}{\sqrt{\sum_{tq} (idf(t) \cdot queryFieldBoost(tq))^2}}$$

$$fieldNorm = \frac{1}{\sqrt{numTerms}}$$

mit

- *numTerms* : Gesamtzahl der Terme im betreffenden Indexfeld, z.B. „content“

2.4 Beispielrechnung

Im Folgenden werden für die Berechnung des Rankingwerts der Beispielseite die Indexfelder *title* und *content* herangezogen. Die Felder *host*, *url* und *anchor* werden nicht berücksichtigt, da in Ihnen das Suchwort „Wasser“ nicht vorkommt.

$$qw = queryFieldBoost \cdot idf \cdot queryNorm$$

$$fw = docFieldBoost \cdot tf \cdot idf \cdot fieldNorm$$

$$qw(content : wass) = 1 \cdot 1.8465904 \cdot 0.022311274 = \underline{\underline{0.041199785}}$$

$$fw(content : wass) = 1 \cdot 2.0 \cdot 1.8465904 \cdot 0.03125 = \underline{\underline{0.1154119}}$$

$$qw(title : wasser) = 7 \cdot 4.9356546 \cdot 0.022311274 = \underline{\underline{0.77084523}}$$

$$fw(title : wasser) = 1 \cdot 1 \cdot 4.9356546 \cdot 1 = \underline{\underline{4.9356546}}$$

$$qw(content : wass) \cdot fw(content : wass) = 0.0047549456$$

$$qw(title : wasser) \cdot fw(title : wasser) = 3.8046257$$

$$Score = 0.0047549456 + 3.8046257 = \underline{\underline{3.8093808}}$$

Der Rankingwert des analysierten Dokuments zur Query „Wasser“ beträgt damit 3.8093808.

3 Anhang

3.1 Bildschirmkopie der in Punkt 2 analysierten Seite

<http://umwelt.schleswig-holstein.de/servlet/is/22624/> [Mozilla Firefox 1.5.0.5]



3.2 Text-Auszug der in Punkt 2 analysierten Seite

[Mozilla Firefox 1.5.0.5]

```

Wasser
Willkommen </servlet/is/11/>
Zum Inhalt springen <#ContentMarker> Zur Fußzeile springen <#Footer>
    
```

Hauptnavigation

[InfoNet-Umwelt SH </servlet/is/127/>](#)
[Startseite </servlet/is/1/>](#)
[Aktuelles </servlet/is/101/>](#)
[Themen </servlet/is/102/>](#)
[Organisationen </servlet/is/103/>](#)
[Dialog </servlet/is/193/>](#)
[Termine </servlet/is/7204/>](#)
[Neu </servlet/is/Global..ShowNews/?daysBack=7>](#)
[Navigator <javascript:popupNavigator\(\)>](#)
[Profisuche <javascript:execSearch\(\)>](#)
[Hilfe </servlet/is/20/>](#)
[Anmelden </servlet/is/12/>](#)
[Impressum </servlet/is/127/>](#)
[Login <javascript:login\(\)>](#)

[Springe zu Boden </servlet/is/23017/>](#)
[Springe zu LuftKlimaschutz </servlet/is/23019/>](#)
[Springe zu Umweltbericht </servlet/is/23286/>](#)
 [[InfoNet-Umwelt </servlet/is/1/>](#) [Organisationen </servlet/is/103/>](#)
[Verwaltung </servlet/is/138/>](#) [MLUR </servlet/is/154/>](#)
[Agrar_u_Umweltbericht </servlet/is/47994/>](#) [Umweltbericht </servlet/is/23286/>](#) [Wasser </servlet/is/22624/>](#)]
 Küstengewässer und Meere: [<../23036/kuestengewaesser.htm>](#)
[Küstengewässerschutz <../23438/kuestengewschutz.htm>](#)
[Biologisches Monitoring <../24017/biomon.htm>](#)
[Algenfrüherkennung <../24568/algen.htm>](#)
[Chemisches Küstengewässermonitoring <../23439/ostnordsee.htm>](#)
[Home <../23286/umweltinfo.htm>](#) [Aktuelles <../22743/aktuelles.htm>](#)
[Behörden <../22744/>](#) [Förderprogramme <../22745/foerderprogramme.htm>](#)
[Kontakt/Impressum <../22746/kontakt.htm>](#) [Landesregierung <http://landesregierung.schleswig-holstein.de/>](#)

Themen: Suche:

[Allgemeine Informationen <../23034/allgeminfos.htm>](#)
[Niederschlag <../23035/niederschlag.htm>](#)
[Flüsse und Bäche <../23436/fliesssgewaesser.htm>](#)
[Seen <../23437/seen.htm>](#)
[Küstengewässer und Meere <../23036/kuestengewaesser.htm>](#)
[Grundwasser <../23037/grundwasser.htm>](#)
[Abwasser <../23038/abwasser.htm>](#)
[Wassergefährdende Stoffe <../23039/gefaehrd.htm>](#)
[Badegewässer <http://www.badewasserqualitaet.schleswig-holstein.de/>](#)
[Wasserrahmenrichtlinie <http://www.wasser.sh/>](#)

Kaum ein anderes Land ist so durch Wasser geprägt wie Schleswig-Holstein. 20.000 Kilometer Bäche und Flüsse durchziehen das Land; 300 Seen mit einer Gesamtfläche 28 000 Hektar sind hier entstanden; rund 1.200 km lang ist die schleswig-holsteinische Küste an Nord- und Ostsee.

Die Wasserversorgung der Bevölkerung wird zu 100 Prozent aus Grundwasser

sichergestellt. Daher müssen wir verantwortungsbewusst mit unseren Gewässern umgehen. Der Schutz und Erhalt der Gewässer als Lebensraum für Tiere und Pflanzen und als Trinkwasserreservoir ist eine wichtige Aufgabe - jetzt und für die Zukunft.

Umweltatlas Schleswig-Holstein <<http://www.umweltdaten.landsh.de/atlas/>>
Datensuche <../22748>

Naturpilot Schleswig-Holstein <<http://www.naturpilot-sh.de/>>

schleswig-holstein.de <http://www.schleswig-holstein.de/>

Umweltportal Deutschland PortalU <<http://www.portalu.de/>>

Umsetzung der Wasserrahmenrichtlinie <http://www.wasser.sh/>

Naturschutzflächen in Schleswig-Holstein <../47459>

Tierschutz <../23024>

Zum Anfang der Seite Top <#top> Version zum Drucken Print

<javascript:doPrint()>

/Letzte Änderung: 15.11.2006/ Zum Anfang der Seite Top <#top>

Version zum Drucken Print <javascript:doPrint()>

3.3 Weiterführende Links

↗ <http://lucene.apache.org/>

↗ <http://de.wikipedia.org/wiki/Lucene>

↗ http://de.wikipedia.org/wiki/Automatische_Indexierung

↗ http://de.wikipedia.org/wiki/Information_Retrieval

↗ <http://de.wikipedia.org/wiki/TF-IDF>

↗ <http://de.wikipedia.org/wiki/Stemming>

↗ <http://lucene.apache.org/nutch/index.html>

↗ <http://www2003.org/cdrom/papers/refereed/p007/p7-abiteboul.html>

↗ <http://de.wikipedia.org/wiki/PageRank>

↗ http://research.microsoft.com/users/nickcr/pubs/craswell_sigr01.pdf

↗ <http://www.emse.fr/OSWIR05/2005-oswir-p31-cutting.pdf>