IFG-Antrag: ID 122733

Bestimmung der Relevanz von Dokumenten im Transparenzportal

A. Hintergrund

IFG-Anfrage: Bitte um Zusendung der Berechnung/ Formel/ Algorithmus/ Datengrundlage, aufgrund welcher die Berechnung der Relevanz eines Objekts/ Datensatz/ Dokuments im Transparenzportal Bremen ermittelt wird.

B. Antwort

Die technische Dokumentation der Suchmaschine des Transparenzportals ist im Transparenzportal veröffentlicht und zu finden unter: https://www.transparenz.bremen.de/sixcms/detail.php?gsid=bremen64.c.45773.de&asl=bremen02.c.732.de.[1]

Die offizielle Projekt-Seite von Solr ist zu finden unter http://lucene.apache.org/solr/.

Solr verwendet ein allgemeines Scoring-Modell, das die Relevanz u.a. wie folgt bewertet:

1. Beispielsweise werden Dokumente höher gerankt, wenn der Suchbegriff öfter darin vorkommt als in anderen Dokumenten (term frequency)
2. Wenn ein Suchbegriff im gesamten Index nur sehr selten vorkommt, wird er den Score, für Einträge, in denen er enthalten ist, maßgeblicher erhöhen, als ein Suchbegriff, der häufiger im Index vorhanden ist (inverse document frequency)
3. Je mehr Suchbegriffe in einem Dokument gefunden werden, desto höher wird es gerankt (coord)

[1] Hinweis: Informationsregister = Transparenzportal

Neben diesem Standard Ranking kann über das Boosting Einfluss auf das Ranking genommen werden, das beim Indexieren oder beim Suchen geschieht.

Im Transparenzportal wird im Wesentlichen beim Indexieren geboostet, die boost-Angabe findet sich dabei in einer spezifischen Solr-Konfiguration. Diese Konfiguration findet sich für alle indexierten Inhaltscontainer in einem Feld-Mapping mit Paaren für das Feld im Contentmanagementsystem (in Bremen SixCMS) zum Feld im SolrIndex.

Ein Beispiel:

> Daten:
> [..]
> thema:
> name: 'thema___s_i_s_m'
> boost: 2

Das heißt, das Feld 'thema' des Containers Daten entspricht dem Feld thema___s_i_s_m in Solr und wird bei der Gewichtung mit dem Faktor 2 berücksichtigt.

Grundsätzlich werden Titel höher als der Rest gerankt, aber auch Felder wie Gliederungsnummer oder amtliche Abk. (bei Gesetzen und Vorschriften) haben ein hohes Boosting, damit Treffer in diesen Feldern den entsprechenden Eintrag mit einer hohen Relevanz versehen.

Die folgende Übersicht zeigt die Felder aller relevanten Container, die durchsucht werden und deren Boosting-Einstellungen.

```
Sachgebiete (bremen59.a.41.de):
  title boost: 10
  thema boost: 2

Schlagworte (bremen59.a.42.de):
  title: boost: 150
  thema: boost: 3

Anwendungen (bremen236.a.138.de):
  title: boost: 10
  untertitel: boost: 2
  beschreibung: boost: 1
  thema: boost: 2

Dokumente (bremen59.a.43.de):
  title: boost: 1000
  untertitel: boost: 20
  beschreibung: boost: 10
```

```
    metadaten_kategorie_r: boost: 2
    metadaten_sachgebietsfelder_r: boost: 2
    thema: boost: 3
    metadaten_schlagworte_r: boost: 2
    bremvor_metadaten_schlagworte: boost: 2
    verantwortliche_stelle:boost: 3
    lizenz: boost: 3
    txt_vt_links_downloads: boost: 1
    bremvor_metadaten_amtl_abk: boost: 10
    bremvor_metadaten_gliederungsnummer: boost: 10
    bremvor_text_normtext: boost: 1

  Daten (bremen236.a.117.de):
    title: boost: 10
    untertitel: boost: 2
    beschreibung: boost: 1
    thema: boost: 2

  IFGAntraege (bremen2014_tp.a.204.de):
    id_antrag: boost: 1
    title: boost: 10
```

Es gibt aber neben dem allgemeinen Scoring-Modell und dem Boosting noch einen weiteren Punkt, der einen entscheidenden Einfluss auf die Relevanz hat: die Filter und Tokenizer, die beim Suchen und Indexieren verwendet werden.

Tokenizer und Filter zerlegen und wandeln je nach Konfiguration Dokumente und Suchbegriffe um. Das ist sinnvoll, um beispielsweise unabhängig von Groß- und Kleinschreibung zu suchen, Treffer mit dem gleichen Wortstamm, den gleichen Wortbestandteilen oder deren Synonyme zu finden. Es ist möglich, dass ein Dokument sehr hoch gerankt wird, ohne das Suchwort überhaupt zu enthalten - z.B. wenn ein Synonym des einen Wortbestandteils des Suchbegriffs, in einem wichtigen Feld (z.B. Titel) gefunden wird.

Dies ist die Hauptursache für die teils schlechte Nachvollziehbarkeit des Rankings und hier lässt sich auch am wirkungsvollsten ansetzen.

Technisch lässt sich die Relevanz-Berechnung wie folgt ausgeben und analysieren. Das Ergebnis kann bei konkreten Beispielen einen Hinweis auf die Ursache für die Relevanz eines Eintrags geben.

Ein Beispiel für die Relevanzberechnung eines Eintrag sieht so aus:
"75881":"\n6.2839894 = sum of:\n 6.2839894 = weight(title:gesetx in 21846) [SchemaSimilarity], result of:\n 6.2839894 = score(doc=21846,freq=3.0 = termFreq=3.0\n), product of:\n 3.2717793 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq + 0.5)) from:\n 1588.0 = docFreq\n 41869.0 = docCount\n 1.9206642 =

tfNorm, computed as (freq * (k1 + 1)) / (freq + k1 * (1 - b + b * fieldLength / avgField-Length)) from:\n 3.0 = termFreq=3.0\n 1.2 = parameter k1\n 0.75 = parameter b\n 16.902458 = avgFieldLength\n 2.56 = fieldLength\n",