1. More Like This Algorithmus

Der More Like This (MLT) Algorithmus ist, ein Algorithmus, der es ermöglicht Dokumente zu finden, die ähnlich sind zu den Dokumenten einer Ergebnisliste. Das Ganze läuft durch Relevanz-Scoring, wo Dokumente auf Relevanz geprüft werden. Um eine optimale Gestaltung des Relevanz-Scoring zu erzielen, das Algorithmus verringert die Anzahl an Kandidaten Dokumente durch einen Boolean Test, bei dem das gesuchte Dokument vergleicht wird mit der Abfrage. Die durch Boolean Test selektierte Dokumente werden Punkte erhalten. Die Bestimmung des Rangs für die Relevanz wird auf Basis diesen Punkten erfolgen. Ein Dokument kann trotz eine Übereinstimmung irrelevant sein für die Abfrage. Deswegen ist die Einstellung einer Filterung von Übereinstimmungen nach Index, dokumenttyp oder durch kontextueller Logik wichtig zur Lösung dieses Problems.

Die Bewertungsfunktion der Informationsabrufsoftware-Bibliothek Lucene wird bei dem MLT Algorithmus genutzt. Diese entspricht ein auf der Begriffsfrequenz und der Inverse Dokumenthäufigkeit basierende Ähnlichkeitsmodell, das das Vektorraummodell für mehrere Begriffe Abfragen benutzt. Es sind unterschiedliche Ähnlichkeitsalgorithmen verfügbar:

1. BM25 okapi: diese basiert auf die Begriffsfrequenz/inverse Dokumenthäufigkeit. Die Steuerung sowohl von nichtlineare Begriffshäufigkeitsnormierung als auch des Normalisierungsgrades von Begriffsfrequenz-Werte durch die Dokumentlänge.
2. die klassische Ähnlichkeit: basiert auch auf Begriffsfrequenz/inverse Dokumenthäufigkeit. Sie bestimmt, ob Überlagerungstoken bei der Berechnung der Norm ignoriert werden[ELA18].
3. die Abweichung von Zufälligkeitsähnlichkeit
4. Abweichung von der Ähnlichkeit der Unabhängigkeit,
5. Informationsbasierte Ähnlichkeit
6. Skriptähnlich Ähnlichkeit: die Verwendung eines Skriptes zur Angabe wie die Ergebnisse berechnet werden sollen [ELA18].
7. LM Jelinek Mercer Ähnlichkeit: die Erfassung von wichtigen Muster im Text mit Rücksichtlosigkeit von Rauschen.
8. LM Dirichlet Ähnlichkeit

[ELA18] <https://www.elastic.co/guide/en/elasticsearch/reference/6.2/index-modules-similarity.html> (Letzter Zugriff: 16.04.2018)