Optimizing search results: Mixing organic and inorganic results

Eric Schraufnagel
06/17/2016

MGMA
Medical Group Management Association

Lucene

- Full-text index
  - PDFs, HTML, Word docs, etc.
- Process
  - Add the information into fields
  - Add the fields into a “document”
  - Add the document to the index
- Managed by Kentico’s Smart Search

MGMA
Medical Group Management Association

Understand How Lucene Works

- Scoring
- Boost
- Expression visibility
- Luke

MGMA
Medical Group Management Association

Lucene Scoring

- WEIGHT
  - coord * queryNorm * (tf * idf * getBoost * fieldNorm)
- tf
  - sqrt(freq)
  - Term frequency in document.
  - Measure of how often a term appears in the document.
  - The more frequent a term occurs in a document, the greater the score.
  - Documents which contain more of a term are generally more relevant.
- idf
  - ln(numDocs/(docFreq+1)) + 1
  - Inverse document frequency.
  - Measure of how often the term appears across the index.
- coord
  - overlap / maxOverlap
  - Number of terms in the query that were found in the document
- lengthNorm
  - 1 / sqrt(numTerms)
  - Measure of the importance of a term according to the total number of terms in the field.
  - Longer documents are penalized.

MGMA
Medical Group Management Association

Lucene Scoring (cont.)

- queryNorm
  - Normalization factor so that queries can be compared
- boost (index)
  - Boost of the field at index-time
  - “This document title is worth twice as much as the title of most documents”
- boost (query)
  - Boost of the field at query-time
  - “I care about matches on this clause of my query twice as much as I do about matches to other clauses of my query”
- fieldNorm
  - lengthNorm * Document Boost * Field Boost
  - The combination of length of the field with index and query time boosts.
  - Lossy. Computes float to byte and then byte to float.
    - http://grokbase.com/t/lucene/solr-user/127kb4hf1x/frustrating-differences-in-fieldnorm-between-two-different-versions-of-solr-indexing-the-same-document

Determine Your Mix

- Organic?
- Inorganic?

MGMA
Medical Group Management Association

Pros & Cons

Organic

+ Trustworthy
- Less control

Inorganic

+ Full control
- May be non-related

MGMA
Medical Group Management Association

Organic Optimizations

- Boost headline & summary
- Ignore tags
- Incorporate analytics
- Synonyms

MGMA
Medical Group Management Association

Inorganic Results

- Ads
- Featured documents

MGMA
Medical Group Management Association

Enhancements

- Predictive search
- Substring search
- Stemming
- Fuzzy searching

MGMA
Medical Group Management Association

Resources

Lucene
https://lucene.apache.org/

Luke
http://www.getopt.org/luke/

Solr
http://lucene.apache.org/solr/

Elasticsearch
https://www.elastic.co/products/elasticsearch

MGMA
Medical Group Management Association

Questions?

MGMA®
Medical Group Management Association

Advancing Leaders. Advancing Practices.™