

Solrの重複uniqueKeyの扱い

- 単一のshard
- 複数のshard
- uniqueKeyを重複させたい場合
- 拡張

単一のshard

uniqueKeyの重複はない

複数のshard

DistributedSearch - Solr Wiki

- Distributed Searching Limitations
 - The unique key field must be unique across all shards.

複数のshard: 実際の処理

src/java/org/apache/solr/handler/component/
QueryComponent.java で処理されている

- uniqueKeyの重複を検出したら...
 - 一番最初に処理されたものが結果に入る
それ以外は結果に含まれない

uniqueKeyを重複させたい場合

- 一部の文書だけ頻繁に更新したい
 - 最新情報の反映
 - ...
- 更新は重たい
 - 冗長性確保のために分散させるとさらに面倒
- 更新用のデータ数の小さいshardを作って
 - 特定のshardを優先させたり,
更新日時が新しいものを使う

拡張

- shardの名前の大小で優先順位を付ける
- timestampのフィールドで優先順位をつける
- 作ってみた
 - <https://github.com/haruyama/solr/>
 - duplicate_unique_key
 - 実装はちょー適当
 - Apache Licence 2.0