アナライズ処理の仕組み
と

2014/04/21

@johtani

自己紹介

- 氏名：大谷　純
- 株式会社シーマーク
- Twitter：@johtani
- lucene-gosenのコミッター
- elasticsearch-extended-analyze作者
- ブログ：http://blog.johtani.info

書籍紹介

- ElasticSearch Server日本語版
  - 対応バージョン：0.90.x
  - 初の日本語書籍
  - 付録にKibana、Kuromojiを追加
- 電子版あり
  - Kindle、達人出版会、角川BOOKWALKER
- ご購入は私のブログから！

Doorkeeper

- Doorkeeper（今後のイベント開催にも利用）
  - http://elasticsearch.doorkeeper.jp
- 日本語用メーリングリスト（Google Groups）
  - https://groups.google.com/forum/#!forum/elasticsearch-jp

アジェンダ

- 転置インデックスとは
- アナライズ処理
  - Analyzer = CharFilter、Tokenizer、TokenFilter
- クエリDSL
  - クエリの種類
  - 注意するクエリ
- elasticsearch-extended-analyzeの紹介

転置インデックスとは？

転置インデックスとは？

1	カツオはサザエの弟
2	サザエはワカメの姉

転置インデックスとは？

1	カツオはサザエの弟
2	サザエはワカメの姉

1	カツオ	は	サザエ	の	弟
2	サザエ	は	ワカメ	の	姉

転置インデックスとは？

1	カツオはサザエの弟
2	サザエはワカメの姉

1	カツオ	は	サザエ	の	弟
2	サザエ	は	ワカメ	の	姉

Term	Id
カツオ	1
サザエ	1、2
ワカメ	2
の	1、2
は	1、2
弟	1
姉	1

アナライズ処理

アナライズ処理とは？

- 転置インデックスの単語（Term）をドキュメントから抽出する処理
- フィールド毎に指定されたアナライザが処理
- アナライザはCharFilter、Tokenizer、TokenFilterから構成
- インデックス時、検索時に実行

アナライズ処理の概要

ドキュメント

アナライズ処理

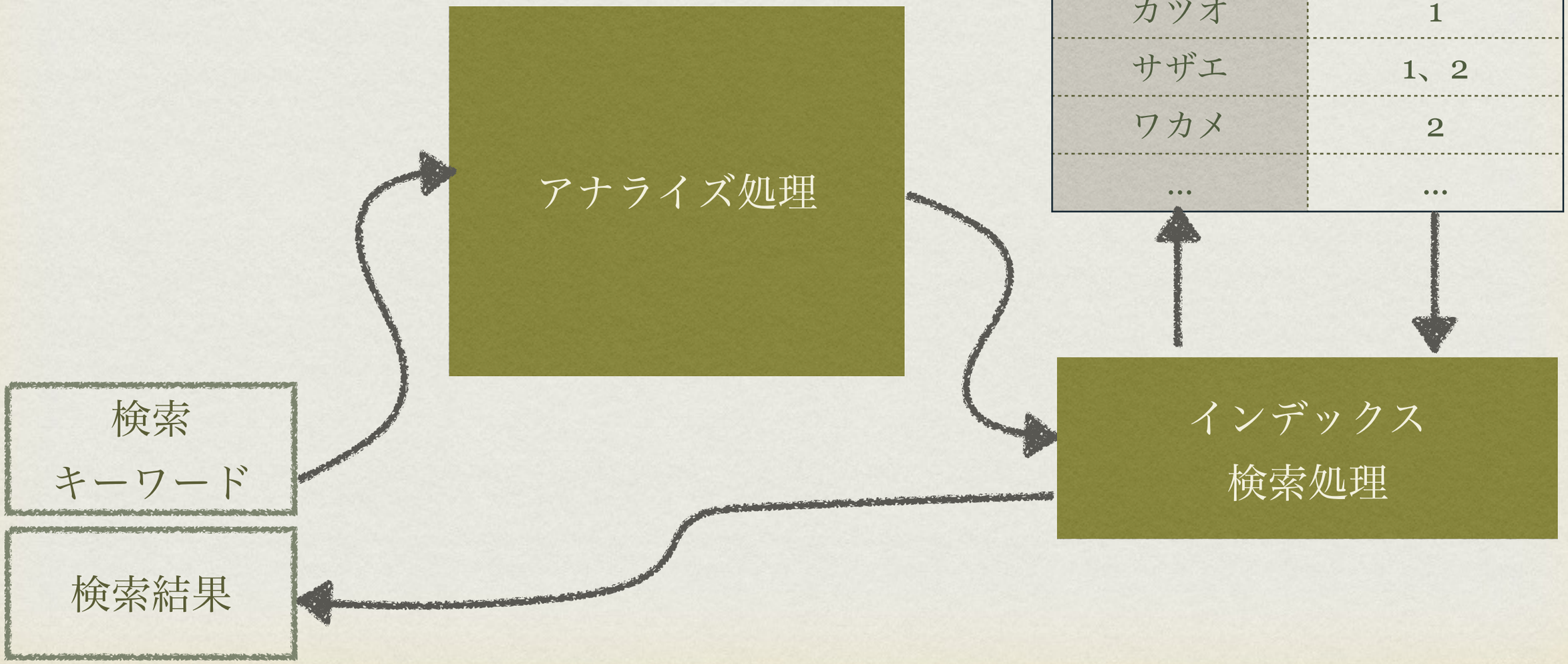
インデックス

Term	Id
カツオ	1
サザエ	1、2
ワカメ	2
…	…

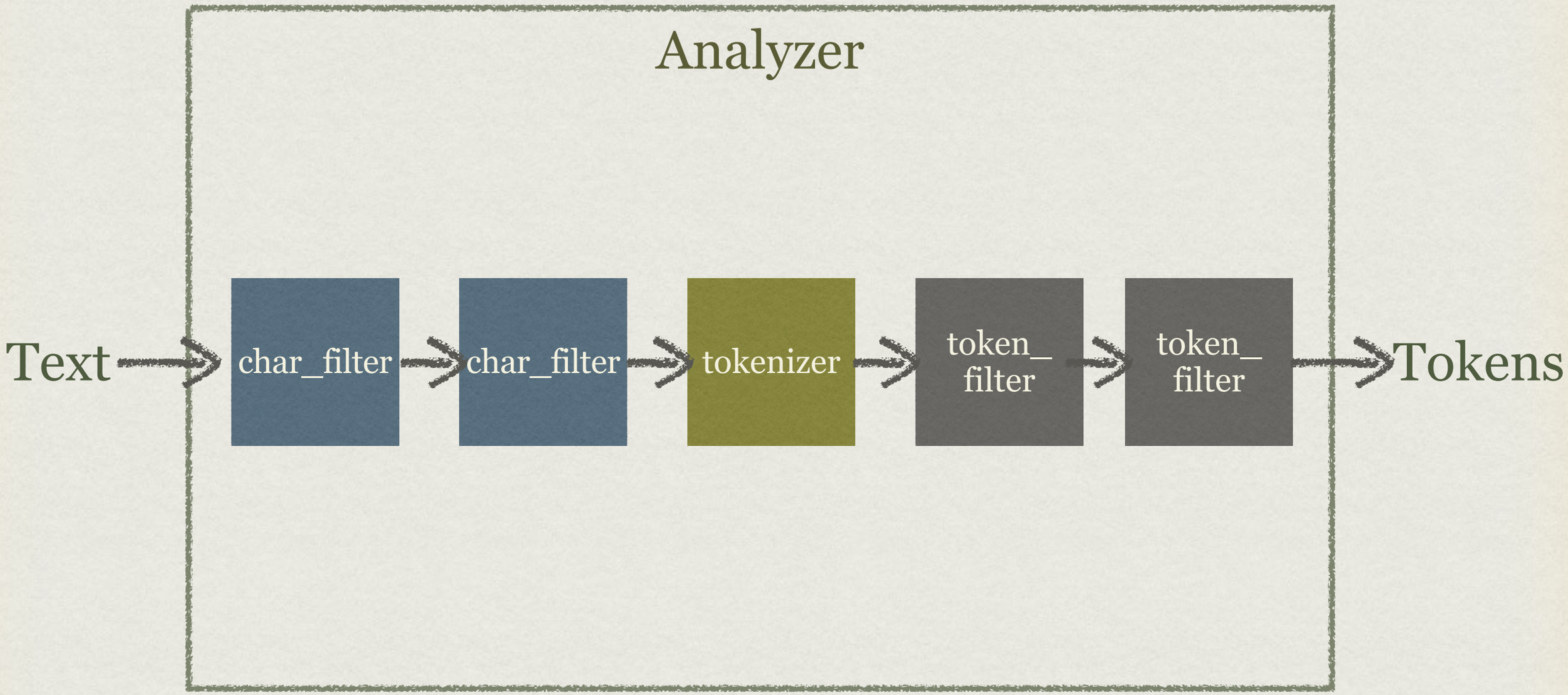
アナライズ処理の概要

インデックス

Term	Id
カツオ	1
サザエ	1、2
ワカメ	2
…	…



アナライズ処理の構成


Analyzer
Text
char_filter
char_filter
tokenizer
token_
filter
token_
filter
Tokens

アナライザの設定

```
{
   "index":{
      “analysis":{
         "analyzer" : {
    "my_analyzer" : {
      "type" : "custom",
       "tokenizer" : “kuromoji_tokenizer”,
      “char_filter" : [“char_filter1”, “char_filter2”…],
      “filter" : [“token_filter1”, “token_filter2”…]
    }
}
```

Char Filter

- 入力文字列を文字単位で処理
- http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/analysis-charfilters.html
- 例：html_strip
  - 入力：<title>Elasticsearch is not a service of AWS</title>
  - 出力：Elasticsearch is not a service of AWS

Tokenizer

- 入力文字列をトークン列に分割
- トークンへの分割するロジックはトークナイザに依存
- http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/analysis-tokenizers.html
- 例：standard
  - 入力：

Elasticsearch is not a service of AWS

  - 出力：

Elasticsearch	is	not	a	service	of	AWS

Tokenizer

- 例：keyword
  - 出力：

```
Elasticsearch is not a service of AWS
```

- 例：kuromoji_tokenizer
  - 入力：

```
寿司が美味しかった
```

  - 出力：

寿司	が	美味しかっ	た

TokenFilter

- Tokenizerにより出力されたToken列に対して処理
- http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/analysis-tokenfilters.html
- 例：lowercase
  - 入力：

Elasticsearch	is	not	a	service	of	AWS

  - 出力：

elasticsearch	is	not	a	service	of	aws

- 例：stop
  - 入力：

Elasticsearch	is	not	a	service	of	AWS

  - 出力：

Elasticsearch				service		AWS

TokenFilter

- https://github.com/elasticsearch/elasticsearch-analysis-kuromoji
- 例：kuromoji_baseform
  - 入力：

寿司	が	美味しかっ	た

  - 出力：

寿司	が	美味しい	た

- 例：kuromoji_readingform
  - 入力：

寿司	が	美味しかっ	た

  - 出力：

sushi	ga	oishika	ta

組み合わせると？

- html_strip+standard+lowercase+stop：
  - 入力：

```
<title>Elasticsearch is not a service of AWS</title>
```

  - html_strip：

```
Elasticsearch is not a service of AWS
```

  - standard：

Elasticsearch	is	not	a	service	of	AWS

  - lowercase：

elasticsearch	is	not	a	service	of	aws

  - stop：

elasticsearch				service		aws

- 最後の単語が転置インデックスのキーとなる

その他の情報も付与

- 単語の他に以下のような情報もTokenに付与
  - position：Tokenが出力される順序
  - start/end offset：テキスト中の文字の位置
- Tokenizer/TokenFilterによる個別情報
  - 例：kuromoji_tokenizer
    - 品詞
    - 読み・発音

クエリ

（今日は主にクエリ）

クエリ

```
{
  "query": {
    "simple_query_string" : {
      "query" : "スパイダーマン",
      "fields" : ["title"]    } },
  "fields": ["title", "category"]
}
```

クエリとフィルタ

```
{
  "query": {
    "simple_query_string" : {
      "query" : "スパイダーマン",
      "fields" : ["title"]   } },
  "fields": ["title", "category"],
  "post_filter": {
    "term": {
      "text": "レオパルドン"  } }
}
```

クエリの数

match query
multi match query
bool query
boosting query
common terms query
constant score query
dis max query
filtered query
fuzzy like this query
fuzzy like this field query
function score query
fuzzy query

geoshape query
has child query
has parent query
ids query
indices query
match all query
more like this query
more like this field query
nested query
prefix query
query string query
simple query string query

range query
regexp query
span first query
span multi term query
span near query
span not query
span or query
span term query
term query
terms query
top children query
wildcard query
template query

種類（検索）

match query

multi match query

bool query

boosting query

common terms query

constant score query

dis max query

filtered query

fuzzy like this query

fuzzy like this field query

function score query

fuzzy query

geoshape query

has child query

has parent query

ids query

indices query

match all query

more like this query

more like this field query

nested query

prefix query

query string query

simple query string query

range query

regexp query

span first query

span multi term query

span near query

span not query

span or query

span term query

term query

terms query

top children query

wildcard query

template query

種類（組み合わせ）

match query

multi match query

bool query

boosting query

common terms query

constant score query

dis max query

filtered query

fuzzy like this query

fuzzy like this field query

function score query

fuzzy query

geoshape query

has child query

has parent query

ids query

indices query

match all query

more like this query

more like this field query

nested query

prefix query

query string query

simple query string query

range query

regexp query

span first query

span multi term query

span near query

span not query

span or query

span term query

term query

terms query

top children query

wildcard query

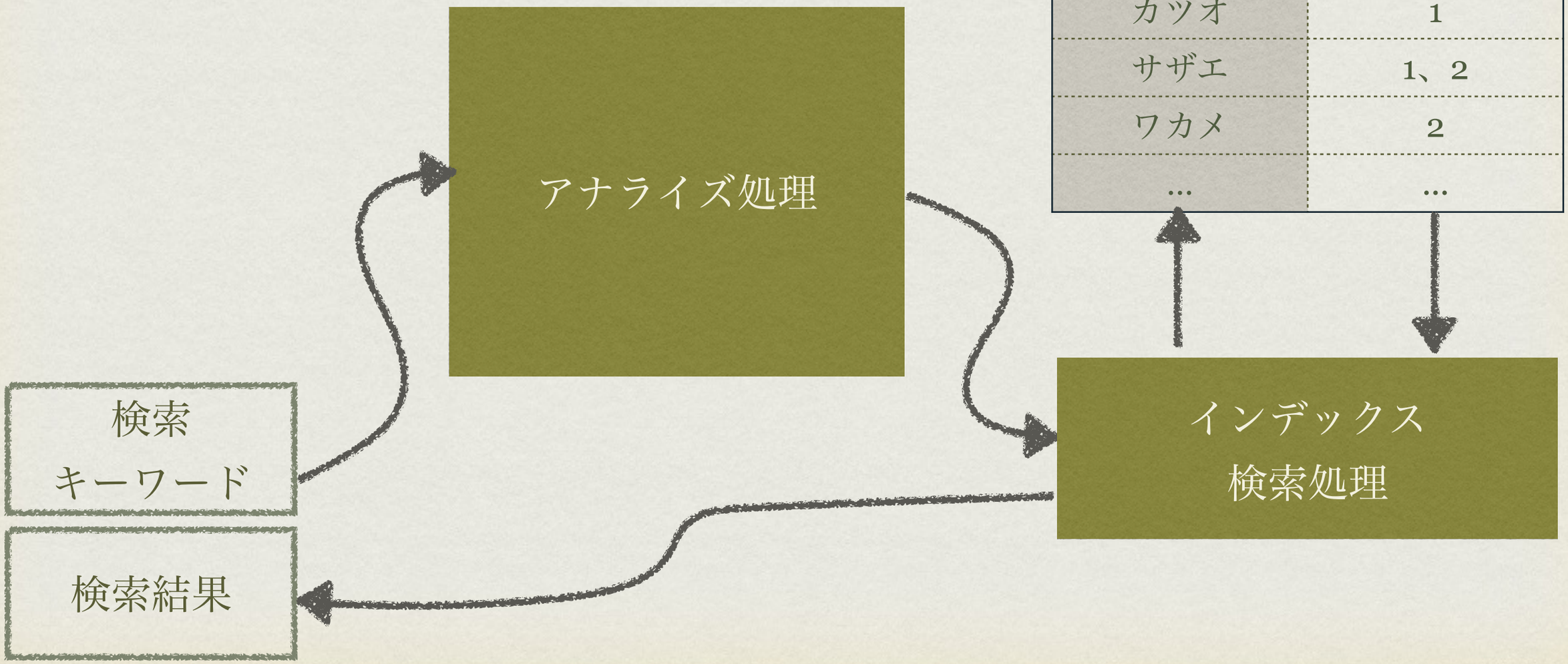
template query

クエリとアナライズ処理の関係

クエリとアナライズ処理の関係

インデックス

Term	Id
カツオ	1
サザエ	1、2
ワカメ	2
…	…



ドキュメントとクエリ

- ドキュメント：
  - 入力：

<title>Elasticsearch is not a service of AWS</title>						

  - アナライズ後：

elasticsearch				service		aws

- クエリ：
  - 入力：

AWS

  - アナライズ後：

aws

ドキュメントとクエリ

- ドキュメント：
  - 入力：

<title>Elasticsearch is not a service of AWS</title>						

  - アナライズ後：

elasticsearch				service		**aws**

- クエリ：
  - 入力：

AWS

  - アナライズ後：

aws

アナライズしないクエリも

インデックス

Term	Id
カツオ	1
サザエ	1、2
ワカメ	2
…	…

検索
キーワード

検索結果

インデックス
検索処理

ドキュメントとクエリ

- ドキュメント：
  - 入力：

<title>Elasticsearch is not a service of AWS</title>						

  - アナライズ後：

elasticsearch				service		aws

- クエリ：
  - 入力：

AWS

- アナライズされないので、「AWS」と「aws」は違う＝ヒットしない

検索クエリの分類（アナライズなし）

match query

multi match query

bool query

boosting query

common terms query

constant score query

dis max query

filtered query

fuzzy like this query

fuzzy like this field query

function score query

fuzzy query

geoshape query

has child query

has parent query

ids query

indices query

match all query

more like this query

more like this field query

nested query

prefix query

query string query

simple query string query

range query

regexp query

span first query

span multi term query

span near query

span not query

span or query

span term query

term query

terms query

top children query

wildcard query

template query

その他クエリの注意点

query_string

- wildcardを使うと、クエリが小文字に変換される。（デフォルト動作）
- wildcardを使うと、アナライズ処理されない（デフォルト動作）

elasticsearch-extended-analyze

extended_analyze

- SolrのAnalysis画面のような情報をJSONで返却
- 画面がない
- アナライズ処理のデバッグのお供に

- インストール方法
  - bin/plugin -i info.johtani/elasticsearch-extended-analyze/1.1.0
- 実行方法
  - curl -XGET "http://localhost:9200/_extended_analyze?tokenizer=kuromoji_tokenizer&filters=kuromoji_baseform&attributes=KeywordAttribute" -d '寿司が美味しかった'

demo

ご静聴、
ありがとうございました。