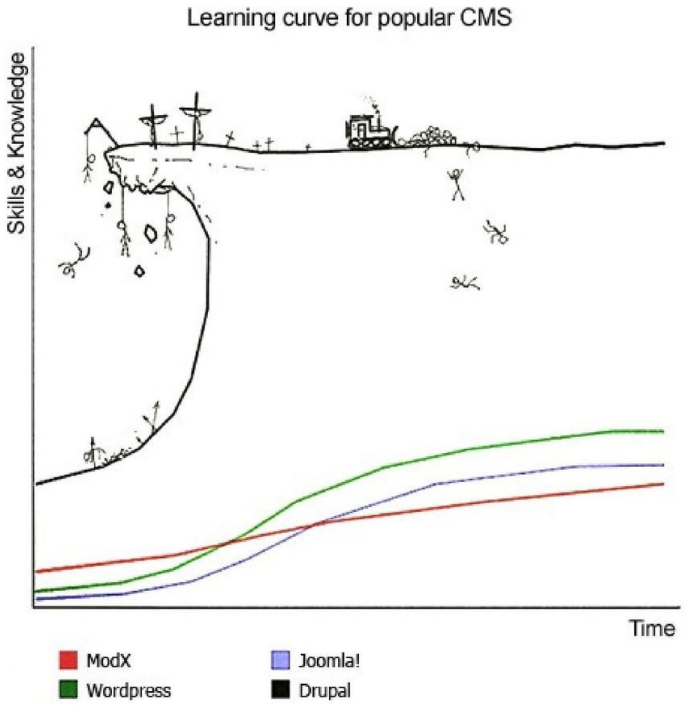
Drupalの日本語検索

2019/12/7 スタジオ・ウミ 新田

自己紹介

- 新田幸子
- スタジオウミのバックエンド
- エンジニア & Drupal歴 1年半
- シリコンバレーに1年留学した時に周囲に影響されプログラマに

Drupal's Learning Cliff

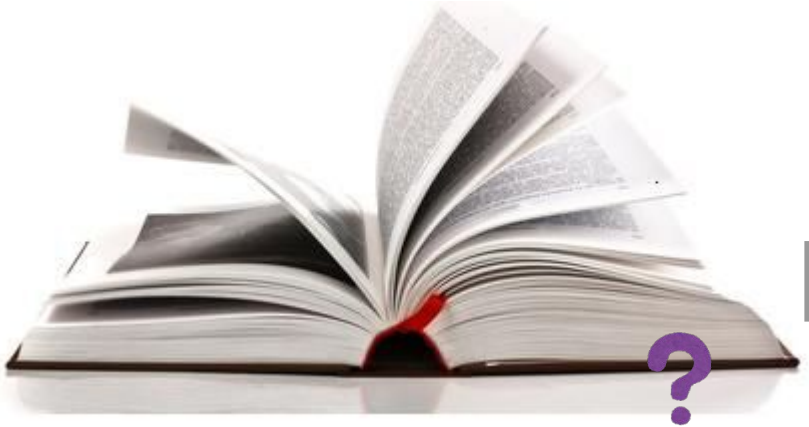


目次

- サイト内検索の仕組み
  - インデックスとは？
  - 日本語のインデックス方法 2種
- Drupalの定番検索モジュール比較
- 開発中のSearch APIプラグインの紹介

サイト内検索の仕組み

インデックス = 索引



○○についてのページはどこだろう？

索　引

数字

1to1　53
9マスの広告　30

アルファベット

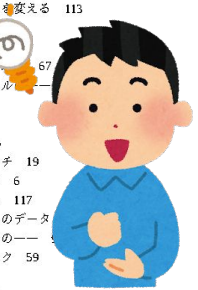
CPC
——12円　29
——15円　31
——4.6円　24
——4円　31
CPC（コスト・パー・クリック）　26
Googleアナリティクス　115
——で見たいデータ　115
——のオーガニック検索キーワード　96, 117, 124
——のコンテンツサマリー　118
——の参照元／メディア　116
——のユーザーサマリー　115
SEO対策　81, 110, 117
WEBサイトが良いのか悪いのか　123
WEBサイトの運用改善　114, 115
WEB制作　161
WEBメディアはクオリティか量か　120
Yahoo!ショッピング　16
——のオンラインレポート　23
——の広告　17
——の広告表示率　20, 27
——の新戦略　12
——の対象ページ表示回数　19

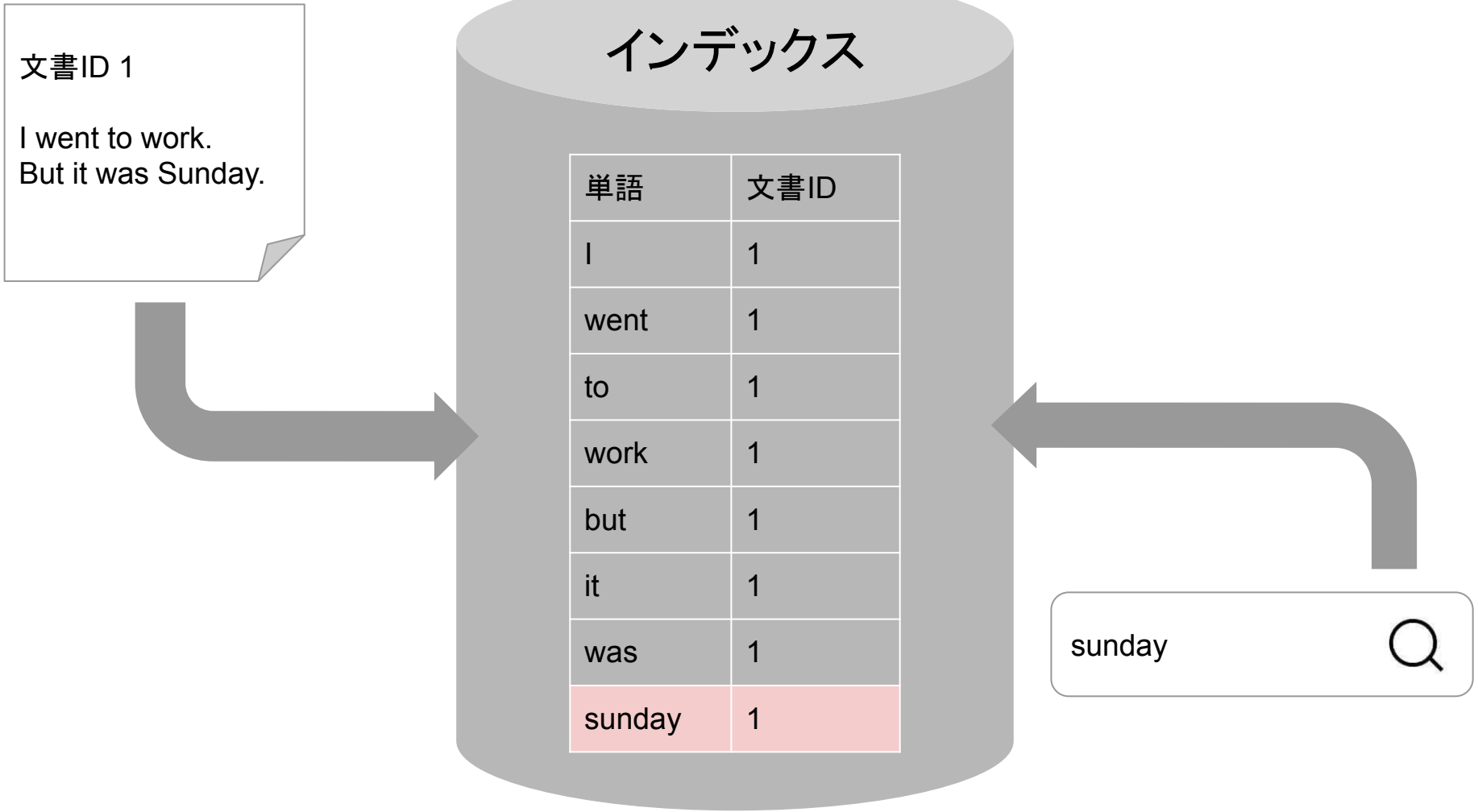
あ

アイデアを絞る　21, 131, 134
頭を使う　129
頭を使って新規顧客を集める　131
新しくお客様に来てもらうには　3
今のお客様をどう回すか　3
インターネットから新規顧客を獲得　63
インターネット広告　1, 42, 45, 81, 125, 174
インターネットといえば検索でなくなる　77
インターネットの活用　108, 124
インターネットの情報メディア　2
売上が上がる　2, 136
オウンドメディア　80, 82, 131
お金を使うと頭を使わなくなる　129

か

画像の加工　158
画像の撮影　158, 160
紙通販　173
——とイーコマース　174
紙とWEB　170, 172
看板広告　43
競合分析　149
——の考え方　152
——の習慣　150
——のミーティング　150
競合のお客様になる　153
競合調査のスケジューリング　156
原因と結果　150
軸の商品の状況　154
自社サイト　151
ファビコンだけを表示　149
口コミ　9, 58, 137
結果と原因　113
原因を用意して結果を変える　113
検索　1
——結果から
——順位が動い
ショッピングモール
検索キーワード
——の編集　72
——を考える　73
——を日々ウォッチ　19
——を見つけ出す　6
オーガニック——　117
オーガニック——のデータ
オーガニック検索の——
検索順位をチェック　59
密着マーク　65
検索順位の変化　59
継続を継続すること　75




文書ID 1
I went to work.
But it was Sunday.
インデックス
単語
文書ID
I
1
went
1
to
1
work
1
but
1
it
1
was
1
sunday
1
sunday

実際に英語でインデックスするときの処理(例)

I went to work. But it was Sunday.　「. , : ; ! ? [ ]」などの記号を除去

I went to work but it was Sunday　大文字を小文字に直す

i went to work but it was sunday　動詞、形容詞、名詞を基本形に直す(ステミング)

i go to work but it is sunday　動詞・名詞・形容詞以外の単語を省く

i go work it is sunday　不要な単語を省く(ストップワード)

go work sunday　スペースを目印に文字列を切り分ける(トークナイズ)

go	work	sunday

日本語のインデックス方法2種

昔むかしあるところにおじいさんとおばあさんがいました。
ある日おじいさんは山へ芝刈りに、おばあさんは川へ洗濯に行きました。

日本語のインデックス方法その１ 形態素解析

昔むかしあるところにおじいさんとおばあさんがいました。

形態素解析器

昔	名詞,副詞可能,*,*,*,*,昔,ムカシ,ムカシ
むかし	名詞,副詞可能,*,*,*,*,むかし,ムカシ,ムカシ
ある	動詞,自立,*,*,五段・ラ行,基本形,ある,アル,アル
ところ	名詞,非自立,副詞可能,*,*,*,ところ,トコロ,トコロ
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
おじいさん	名詞,一般,*,*,*,*,おじいさん,オジイサン,オジーサン
と	助詞,並立助詞,*,*,*,*,と,ト,ト
おばあさん	名詞,一般,*,*,*,*,おばあさん,オバアサン,オバーサン
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
い	動詞,自立,*,*,一段,連用形,いる,イ,イ
まし	助動詞,*,*,*,特殊・マス,連用形,ます,マシ,マシ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。

単語	文書ID
昔	1
むかし	1
ある	1
ところ	1
おじいさん	1
おばあさん	1
いる	1

おじいさん

日本語のインデックス方法その２ N-gram

昔むかしあるところにおじいさんとおばあさんがいました。

文書から決まった文字数の文字列を全て取り出そう
今回は3文字ずつ

おじいさん

おじい
じいさ
いさん

単語	文書ID
昔むか	1
むかし	1
かしあ	1
しある	1
あると	1
るとこ	1
ところ	1
ころに	1
ろにお	1
におじ	1
おじい	1

じいさ	1
いさん	1
さんと	1
んとお	1
とおば	1
おばあ	1
ばあさ	1
あさん	1
さんが	1
んがい	1
がいま	1
いまし	1
ました	1

形態素解析とN-gramのインデックス比較

昔	1
むかし	1
ある	1
ところ	1
おじいさん	1
おばあさん	1
いる	1

昔むか	1
むかし	1
かしあ	1
しある	1
あると	1
るとこ	1
ところ	1
ころに	1
ろにお	1
におじ	1
おじい	1
じいさ	1
いさん	1

さんと	1
んとお	1
とおば	1
おばあ	1
ばあさ	1
あさん	1
さんが	1
んがい	1
がいま	1
いまし	1
ました	1

日本語のインデックス作成 & 検索時に行われる代表的な処理

	例	形態素解析	N-gram
記号を除去	「、。！」など	○	○
全角と半角を統一	ﾋﾟｻﾞ → ピザ、 ９ → 9	○	○
大文字と小文字を統一	Apple → apple	○	○
表記のゆれを統一	ヴァイオリン → バイオリン	○	✕
動詞を原形にする	走った → 走る	○	✕
ストップワードを除去	「私」「こと」「もの」	○	○

Drupalの定番検索モジュール比較

モジュール名	**Search(コア)**	**Search API**	**Google Custom Search**
セットアップ	標準でインストールされている	● 管理画面から使用するプラグインなどの設定 ● (オプション) Apache SolrなどDrupal外の検索サーバーの構築	● 「Googleカスタム検索」に登録 ● キーを管理画面に登録
インデックス先	Drupalのデータベース	● Drupalのデータベース ● Apache SolrなどDrupal外の検索サーバー	Googleの検索サーバー
インデックス対象	ノードかユーザー (ディスプレイモードでフィールド制御)	全てのエンティティ (フィールドの区別あり)	ドメイン内にある Googleがクローリングできるコンテンツ
日本語の検索方法	N-gram	● 形態素解析に対応した検索サーバーを使用しない限りN-gram (Tokenizerプラグインが対応)	形態素解析
部分一致(N-gram、形態素解析両方)	不可	可(非推奨)	Googleのアルゴリズムによる
結果結果画面	カスタマイズ可	● Viewsを利用(パスなど設定する必要あり) ● カスタマイズ可	● Googleのロゴと広告が表示される ● カスタマイズ可
検索窓	ブロックで提供されている	Viewsの外部設置フィルターをブロック化して設置	Searchモジュールの検索窓を利用
パフォーマンス	コンテンツ数数万の大規模サイトには向かない	● Drupalのデータベースを用いる場合 →Searchモジュールと同じ ● Apache Solrなどを使う場合 →大規模サイトでも問題無し	大規模サイトでも問題無し

開発中のモジュールの紹介

開発の動機

Drupalで形態素解析したい...

でもApache Solrはめんどくさい...

形態素解析API使ってみたい...

MeCabも使ってみたい...

開発中モジュールの概要

- Search APIのプロセッサープラグイン
- 文章を受け取って日本語の形態素解析を行い、単語ごとにスペースで区切る処理を行う
- インデックス時と検索時に使用可
- 形態素解析器にはMeCabとYahoo! 形態素解析APIを使用予定
- 以下のオプションを実装予定
  - インデックス対象にする品詞
  - 活用のある単語を基本形にするかどうか