

中間発表

佐々木研究室

06T4073R 三上健太

はじめに

□ 類似度計算

- 情報検索やデータマイニングでよく用いられる

□ 単語と単語、文書と文書がどれくらい似ているかを判断するポイントとなる

□ 文書中の単語の割合、出現頻度が似ているかどうか比較する

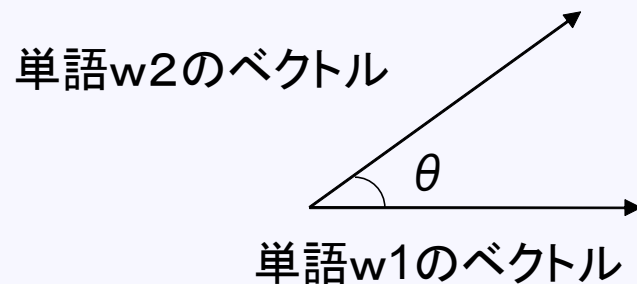
問題点

- 同じ意味だが表記が違うものが類似していないと判断される
 - 表記の揺れに対応していない
 - ex. スパゲティ, スパゲッティ, スパゲッティー

- 概念(単語の意味と意味)を比較して類似度をみる
 - 表記だけではなく、単語が持つ意味に注目
 - 今回はカテゴリを用いる

提案手法

- Googleディレクトリ検索を用いる
- 検索結果の各登録サイトが属するカテゴリ名に注目
- 属するカテゴリの分布が似ていれば、類似度が高いとする
- 後述の計算方法によって類似度を計算
 - コサインを用いる
 - 単語毎にベクトルを作成, そのベクトルに注目





[ウェブ](#) [画像](#) [グループ](#) [ニュース](#) [地図](#) [その他](#)

[表示設定](#)
[ディレクトリヘルプ](#)

カテゴリ別Google!

[アート](#)

[写真](#), [文学](#), [映画](#), ...

[ニュース](#)

[新聞](#), [テレビ](#), ...

[各種資料](#)

[辞書](#)・[事典](#), ...

[ゲーム](#)

[ビデオゲーム](#), [オンライン](#), ...

[ビジネス](#)

[金融](#), ...

[家庭](#)

[ガーデニング](#), [料理](#), [暮らし](#), ...

[コンピュータ](#)

[インターネット](#), [ソフトウェア](#), ...

[レクリエーション](#)

[アウトドア](#), [旅行](#), [車](#)・[バイク](#), ...

[社会](#)

[政治](#), [教育](#), [時事](#), ...

[スポーツ](#)

[サッカー](#), [ゴルフ](#), [野球](#), ...

[健康](#)

[体調](#)・[症例](#), [美容](#), ...

[科学](#)

[天文](#), [社会科学](#), ...

関連カテゴリ:

[Regional > Asia > Japan](#) (7287)

[広告掲載](#) - [ビジネスソリューション](#) - [Googleについて](#)

©2009 - [プライバシー](#)

あなたも「ウェブ最大のディレクトリ作り」に参加しませんか
[URLを登録する](#) - [Open Directory Project](#) - [編集者募集](#)

図1 Googleディレクトリ検索HOME

<http://dir.google.com/>

ウェブ 画像 動画 地図 ニュース グループ Gmail その他 ▼

Google ディレクトリ [表示設定](#)

ディレクトリ サッカー の検索結果 約 4,470 件中 1 - 10 件目 (0.1)

[財団法人日本サッカー協会 公式サイト](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [AFC\(アジア\)](#) > [日本](#)
概要、活動案内、都道府県サッカー協会の連絡先とリンク。
www.jfa.or.jp/

[スポーツナビ | サッカー](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [ニュースとメディア](#)
Jリーグ、日本代表、セリエAなど、様々なサッカー情報を掲載。
sportsnavi.yahoo.co.jp/soccer/

[Yahoo!スポーツ - サッカー](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [ニュースとメディア](#)
新聞・通信社のニュースや関連カテゴリへのリンク。
sports.yahoo.co.jp/soccer/

[Jリーグ公式サイト](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > ... > [日本](#) > [Jリーグ](#)
ニュースや大会概要・公式記録、データ、チケット情報、スタジアムガイドを掲載。
www.j-league.or.jp/

[スポニチ Sponichi Annex サッカー](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [ニュースとメディア](#)
スポーツニッポン紙。国内外のサッカーニュース。
www.sponichi.co.jp/soccer/

[サッカー : nikkansports.com](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [ニュースとメディア](#)
日刊スポーツ紙。Jリーグ・海外サッカー・日本代表関連のニュース。
www.nikkansports.com/soccer/top-soccer.html

[サッカー: スポーツ報知](#)
カテゴリ: [スポーツ](#) > [サッカー](#) > [ニュースとメディア](#)
報知新聞 ニュース記事とコメント、選手毎のバックナンバー

スポンサーリンク
[サッカー nikkansports.com](#)
最新のサッカー情報をあなたのGoogle ページにお届け。
google.co.jp

図2 "サッカー"をキーワードとしてGoogleディレクトリ検索を実行した時の実行結果

計算方法

- (1) 2つの単語w1, w2を入力
- (2) 各単語についてカテゴリ検索
- (3) 検索結果上位N件のカテゴリ名を抽出
- (4) カテゴリ名に含まれる単語集合を抽出
- (5) 検索結果のランク、頻度より重み付け
- (6) ベクトルv1, v2を作成
- (7) ベクトル間のコサインを類似度として出力

$$\cos \theta = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|} = \frac{x_1 y_1 + x_2 y_2 + \dots + x_n y_n}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \cdot \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}}$$

$$(0 \leq \cos \theta \leq 1)$$

進捗状況

- 類似度計算を行うプログラムを作成
 - perlにより実装
- コマンドプロンプトより、単語を2つ指定し実行
 - 各単語のカテゴリ検索実行結果より、登録サイトが属するカテゴリ名を抽出し表示
- コサインにより類似度を計算、表示

実行結果

```
>perl googledir0902.pl サッカー football
レクリエーション -> 1
スポーツ -> 9
AFC(アジア) -> 3
ニュースとメディア -> 5
日本 -> 3
ギャンブル -> 1
サッカー -> 9
Jリーグ -> 2
toto -> 1
ガイドとディレクトリ -> 1
```

```
England -> 1
スポーツ -> 6
選手 -> 1
AFC(アジア) -> 2
News_and_Media -> 3
団体 -> 1
アメリカンフットボール -> 1
日本 -> 2
サッカー -> 5
American -> 2
Jリーグ -> 1
NFL -> 1
学生 -> 1
ガイドとディレクトリ -> 1
UEFA -> 1
```

$\cos \theta = 0.818831145677992$

期待される効果

- 類似度計算がより正確になる
- 文書クラスタリング, 検索などにおいて, より正確な結果が期待できる



検索の精度があがる

現段階の問題点

□ 計算結果において...

- ・ 本当は似ているのに類似度が低いもの
- ・ 本当は似ていないのに類似度が高いもの

これらが検出されることがある



精度の低下

>perl googledir0902.pl 徳川光圀 水戸光圀

地域 -> 1
中国 -> 1
ワイン・果実酒 -> 1
ビジネス -> 2
飲料 -> 1
日本 -> 1
アジア -> 2
食品 -> 2
作家 -> 1
納豆 -> 1
文学 -> 1
西行 -> 1
アート -> 1
世界の料理 -> 1
ビジネス・経済 -> 1
文化施設 -> 1
家庭 -> 1
豆類・豆製品 -> 1
料理 -> 1
日立市 -> 1
茨城 -> 1
農産物・加工品 -> 1
各種資料 -> 1
博物館・美術館 -> 1
市町村 -> 1
検索・リンク -> 1
酒類 -> 1
俳句・和歌・川柳 -> 1

時代劇 -> 1
番組 -> 1
ドラマ -> 1
テレビ -> 1
科学 -> 1
アート -> 1
オルタナティブ科学 -> 1

$$\cos \theta = 0.06213697660012$$

↑
類似度が低い

今後の予定

- 計算結果と実際の結果が異なるものについて改善策を検討
- 類似度計算手法の(5)について実装