

Wikipediaからの意外性のある情報の抽出

A study for extracting serendipitous Information from Wikipedia

野田陽平 清田陽司 中川裕志 (東京大学)

Yohei Noda, Yoji Kiyota, Hiroshi Nakagawa (University of Tokyo)

Introduction

本研究は、Wikipediaの記事の中から、複数の分野にまたがった意外性のある知識を発見することを目的としている。Wikipediaは各記事が1つ以上のカテゴリに属しており、そのカテゴリネットワークはグラフ構造を成している。それらのグラフ上の構造を特徴量として利用し、機械学習により、各記事に対して意外性を含む記事であるか否かの判定を行う。例えば、「オープンソース」と「コーラ」という一見意味的には離れたカテゴリにも、それらに共通して属する「オープンコーラ」という記事が存在する。本研究では、このような意外性のある関係性をWikipediaの中から自動的に発見する手法を提案する。

Present Search Engines

Query型検索エンジン(Google, Yahoo!等)

- ユーザにより与えられたクエリが含まれている文書を提示する。
- Page Rank等による重みづけ

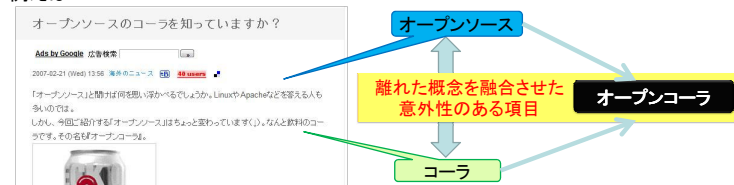
ディレクトリ型検索エンジン(Googleディレクトリ, Yahoo!カテゴリ等)

- 似た内容の文書をまとめ、ディレクトリ形式で格納する。ユーザは自らが求める内容に沿ってディレクトリを辿り、閲覧したい文書を探す。

→「ユーザが予想した記事」は発見できるが...

ユーザが予想できない「意外性」を含む記事の中にも興味深い記事はある

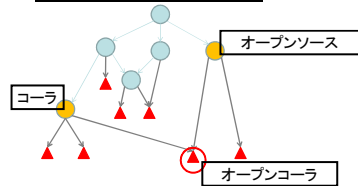
例えば...



Our idea to extract serendipitous information from Wikipedia

意外性...一見関連がなさそうな2つの分野にまたがった知識を発見したときに感じる(関連性の発見)

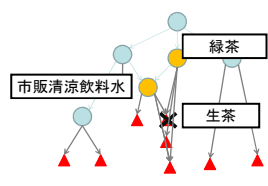
意外性が高い関係性の例



- カテゴリ同士が意味的に遠い
- 「コーラ」と「オープンソース」という、一見離れた概念を結びつけるような関係性→意外性が高い

- 共通項目の希少性が高い
- 「コーラ」と「オープンソース」の両カテゴリに属するような項目は1つしかない

意外性が低い関係性の例



- カテゴリ同士が意味的に近い
- 「市販清涼飲料水」と「緑茶」という意味的に近い概念を結びつけるような関係性→意外性は低い

- 共通項目の希少性が低い
- 「市販清涼飲料水」と「緑茶」の両カテゴリに属するような項目は多数存在する

これらを表現するfeatureを見つけ、機械学習を用いて意外性の判定を行う

Related Works

- 意外性を含む情報を抽出する研究
 - レコメンドエンジンにおける意外性のあるアイテムの提示(Sarwarら)
 - ユーザにとっては探しにくい、重要であるコンテンツを「コンテンツホール」と定義し、それらを抽出することを提案(Nadamotora)
- Wikipediaをグラフ構造として扱う研究
 - Wikipediaのグラフ構造を元にDocumentのクラスタリングを行う研究

Feature

$c_i \dots category, a_i \dots article$

• 親カテゴリA,Bの子供の数

$$Child(c_i) = count(a_k | a_k \in c_i)$$

• 親カテゴリA,Bの階層...rootからの最短距離

$$H(c_i) = \min |root - c_i|$$

• 親カテゴリA,Bの共通子項目(共通インスタンス)の数

$$CoChild = count(a_i | a_i \in C_A \wedge a_i \in C_B)$$

• 親カテゴリA,Bの字面上の距離(現段階では未使用) (Levenshtein距離をカテゴリ名の長さで正規化)

$$Distance(c_A, c_B) = \frac{LevenshteinDistance(c_A, c_B)}{\sqrt{length(c_A)} \sqrt{length(c_B)}}$$

教師データの分布

Positive Labeled Data

Feature	平均	標準偏差
親カテゴリAの子供の数	273.85	643.74
親カテゴリAの階層	3.89	1.20
親カテゴリBの子供の数	118.34	257.37
親カテゴリBの階層	4.30	1.24
共通子項目の数	1.13	0.80
親カテゴリ同士の距離	1.21	0.42

Negative Labeled data

feature	平均	標準偏差
親カテゴリAの子供の数	1077.49	1793.35
親カテゴリAの階層	5.60	1.46
親カテゴリBの子供の数	702.51	1757.90
親カテゴリBの階層	5.66	1.34
共通子項目の数	163.30	390.70
親カテゴリ同士の距離	1.04	0.45

Experiment and Result

SVM(Support Vector Machine)により、学習、分類を行った

結果

使用したfeature	Linear Kernel	Polynomial Kernel	Polynomial Kernel	Gaussian Kernel
すべてのfeature	83.75%	85%	81.25%	85%
親カテゴリAの子供の数以外	82.5%	77.5%	73.75%	78.75%
親カテゴリAの階層以外	72.5%	80%	76.25%	67.5%
親カテゴリBの階層以外	77.5%	77.5%	81.25%	82.5%
共通子項目の数以外	72.5%	71.25%	71.25%	70%

抽出された正例、負例

判別結果	カテゴリA	カテゴリAの子供の数	カテゴリAの階層	カテゴリB	カテゴリBの子供の数	カテゴリBの階層	項目	共通項目数
○	奇跡の水	10	3	機能水	7	3	アルカリイオン水	2
○	オープンソース	374	4	コーラ	26	6	オープンコーラ	1
○	能力開発	63	3	キャラクター	7	5	キャラ立ち	1
×	ヴァッレダオスタ州のムーネ	46	5	イタリアのムーネ	436	7	アルナド	8
×	日本のタレント	7617	6	東京と出身の人物	9825	5	堺正章	2088
×	静岡市の鉄道駅	34	7	葵区	140	7	静岡駅	11

Conclusion and Future Work

- Conclusion
 - Wikipediaの構造を用いて、項目が持つ意外性のある関係性の判定を行う手法の提案
- Future Work
 - 文書中のリンク情報を用いた詳細な情報抽出 (Wikipediaの記事中の意外性情報が記述された部分の特定)
 - 教師データの作り方の工夫
 - 機械学習以外の方法