卒業研究(中間報告)

06t4027h 斎藤高行

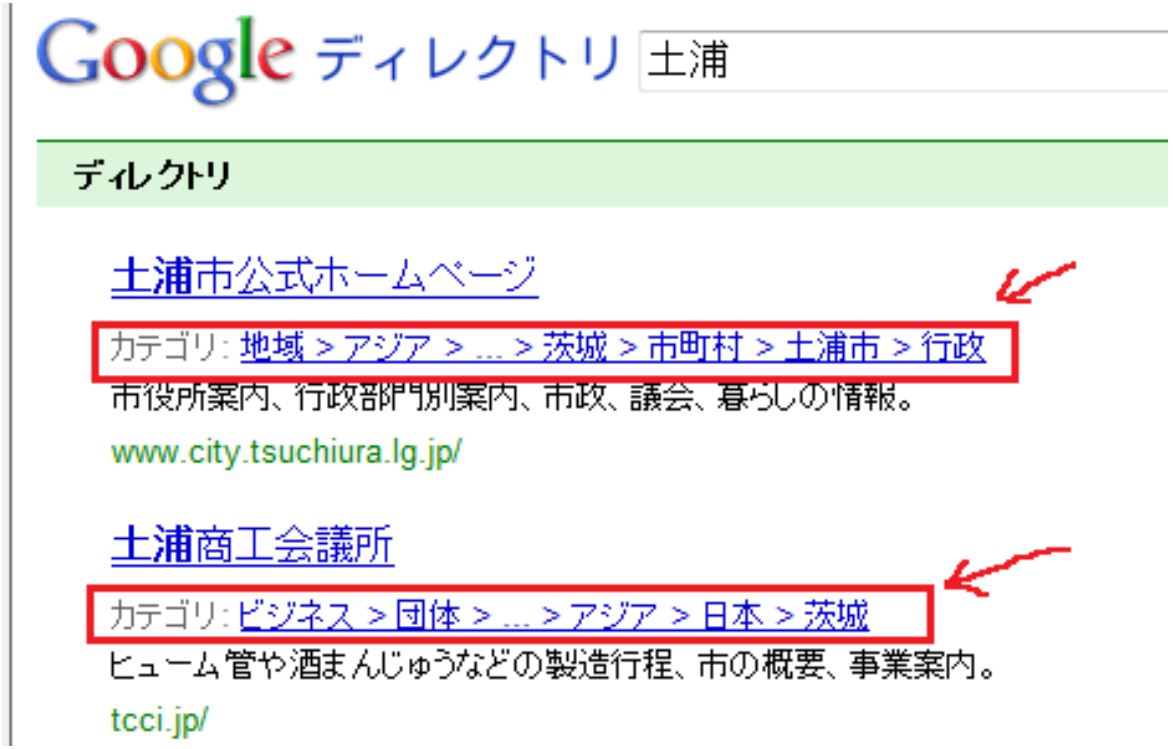
研究内容

- 目的は、単語クラスタリングの精度の向上

- ”Googleディレクトリ”検索を利用し、似た要素をもつ単語どうしをまとめるプログラムの改良

クラスタリングまでのフロー

各単語の”Googleディレクトリ”検索でヒットするトピックの”カテゴリ”に注目して、その単語のカテゴリラベルを種類ごとに数える



クラスタリングまでのフロー

全てのトピックから数え上げたカテゴリを基に、
単語・カテゴリラベル・各カテゴリラベルの合計
の情報として出力し、クラスタリングツールである”Cluto”を用いてクラスタリングを行う

現在の進捗状況

近いカテゴリを多く、遠いカテゴリを少なくするような「重み」を付けることにより、精度の向上を図る

現段階では、クラスタリングの正解率をみながら、収集するカテゴリの数と、重みの適切な値域を探している

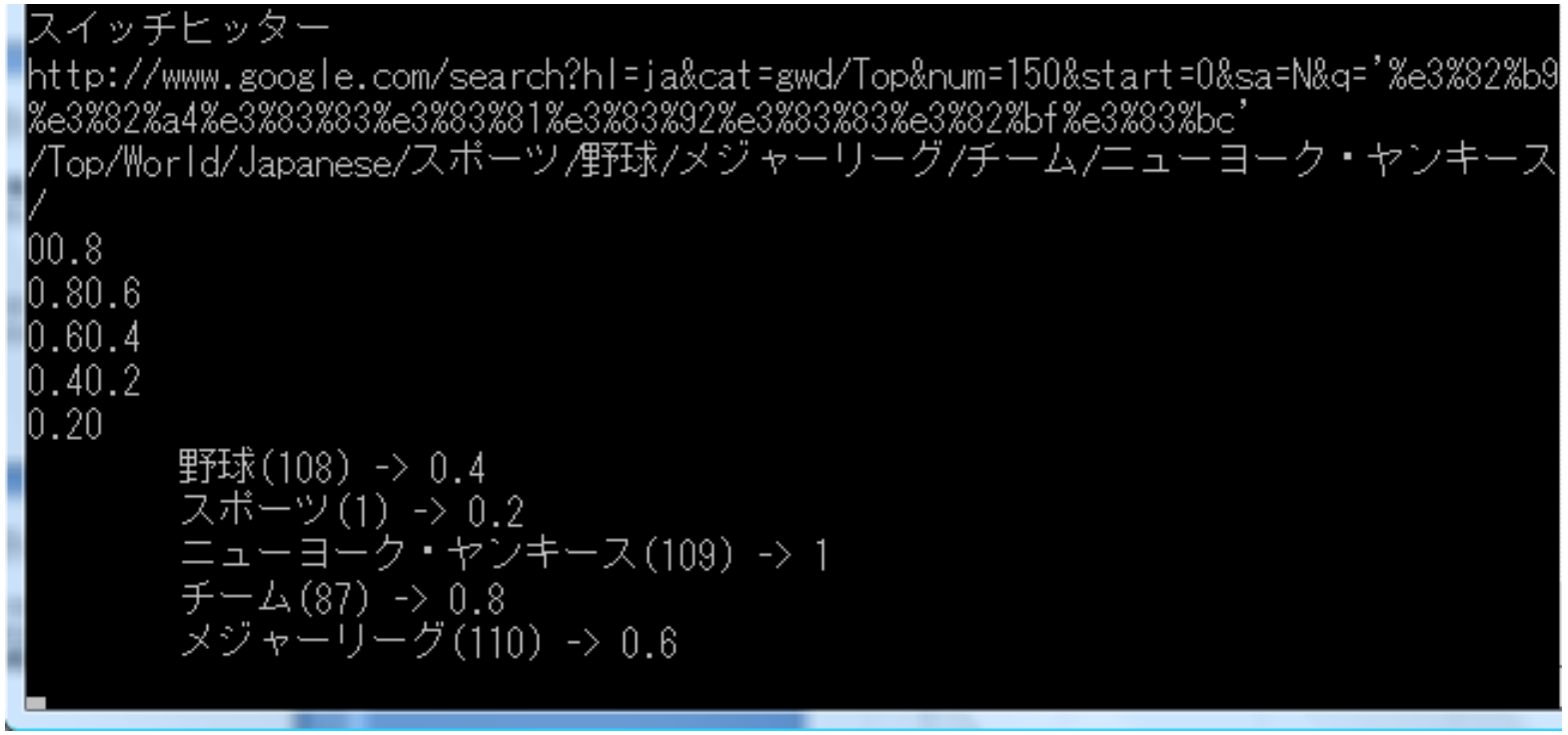
現在の進捗状況

「重み」とは、カテゴリラベルの出現頻度を測定する上で、ラベルがひとつ抽象化される度にその出現頻度から減算する数値を指す

現在の進捗状況

例として、重みw=0.2のとき

単語”スイッチヒッター”に関する1つのトピックのカテゴリラベルは次のように数えられる



問題点

“Googleディレクトリ”を用いてのクラスタリングには、いくつか注意しなければならない問題がある

- 検索ヒット件数が少ない場合
- カテゴリラベルが日本語でない場合

問題点

- 検索ヒット件数が少ない場合

カテゴリラベルの収集量が少なく、精度が低くなる可能性が考えられる

特にヒット件数＝0件の場合にはエラーとなり、その単語はクラスタリングから除外される

問題点

- カテゴリラベルが日本語でない場合

予めデータベース登録した単語(検索する単語のことではない)とマッチしないカテゴリラベルは数えることができない。

その場合、マッチするラベルのみが測定され、クラスタリング精度の低下につながる。

また、先述の問題と絡むとエラーになりやすい。

今後の課題

- 先述の問題を解決するような方法を探す

- 効果的な重み関数を探す
  (曲線の形になると予想……未検証)