F-008

オフライン正規化検索エンジン距離による文字列間の類似度推定

佐藤 哲 †

楽天株式会社 †

1. はじめに

近年の情報爆発社会の中で必要な情報を探し出すために，探したい文字列そのものだけではなく，その文字列と似た文字列を含む情報を検索する研究が盛んである．さらに，従来は表記ゆれや文字コードの正規化を処理するのに留まっていたが，文字列の意味に基づいた距離を考慮した類似度計算手法が研究されている．そのような研究の一つとして，本発表では Web 上の辞書サイト Wikipedia のデータを利用して文字列の意味を考慮し，文字列間の類似度を計算するシステムの試作結果について述べる．

2. 文字列間の類似度計算

探したい文字列と表記的な類似性を計算する手法には，莫大な研究結果が既にライブラリとして実装されている [1]．しかし，それらは文字列の意味の類似性を考慮したものではない．任意のオブジェクトの意味を考慮して類似度を計算する手法としては，芸術作品等の解説文をオブジェクトに対応する意味オブジェクトとして比較分類する手法が存在する [2]．また，それらとは独立して，ユニバーサル符号の理論を応用した汎用的な記号列間類似度測定の研究がある．しかし，汎用的になるほど計算コストが高く，実用性に乏しくなる．

本研究では，汎用性を考慮して Vitányi らによるユニバーサル符号化理論に基づく手法 [3] を利用する．文献 [3] では, キーワード x 及び y の間の距離は次のように定義される：

$$NGD(x,y) = \frac{\max\{\log f(x), f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}}. \quad (1)$$

ここで，$f(x)$ は単語 x を検索した場合のヒット数を，$f(x,y)$ は 2 単語を同時に検索した場合のヒット数を表す．式 (1) は，検索エンジンに登録されいるページ数 N が十分に大きいことが前提となっており，登録ページ数が少ない場合や，あるキーワードに対する登録ページ数またはそのページのコンテンツが十分でない場合は精度が落ちることが知られている．また，多量のデータを保持している検索エンジンサービスの利用も前提とされており，利用目的に特化するようカスタマイズすることは考慮されていない．そこで本研究では，検索エンジンとして Apache Lucene[4] を用いてあらかじめ検索対象コンテンツのインデックスを作成しておき，コンテンツは自由に編集できる組み込み型の検索エンジンを用いた文字列間の類似度計算の試験結果について報告する．以下，NGD とは Lucene の検索結果から式（1）を用いて距離を計算する手法を指し，検索エンジンである Google は用いない．

Similarity estimation between the words using off-line normalized search engine distance

†Tetsu R. Satoh, Rakuten Inc.

3. 正規化 Lucene 距離

オープンソースソフトウェアである Lucene を用いていることで，ここでは提案手法を正規化 Lucene 距離（Normalized Lucene Distance:NLD）と呼ぶことにする．検索対象とするコンテンツとしては，Wikipedia 日本語版の XML 形式データ [5] を用いる．コンテンツ [5] は多くの冗長な情報を含むため，プログラミング言語 Ruby の XML パーサライブラリである Nokogiri[6] を用いて検索に必要なデータを抽出して Lucene でインデックスを作成する．Nokogiri では，次のような処理を行う：

処理:`<!-- XXX --!>`等の不可視なコメント，`<br\/>`,`<small>`, `style="..."`等の意味に無関係な表示整形タグ，`[[en:XXX]]` 等の多言語のページに対するリンクタグ，`<ref>{{PDFlink...` 等の PDF ファイルへのリンクタグ，`<ref>[http...` 等の外部サイトへのリンクタグ，`#REDIRECT...` 等の他のページのリダイレクトのみのタグを除去する

また，NGD は値については正規化されていず，かつ検索エンジンサービスではなく Wikipedia データを用いることで情報量が減ることを考慮し，値が 0 から 1 になるよう正規化し，かつ検索結果のヒット数を用いて情報量の偏りを補正するよう修正した以下の正規化 Lucene 距離（Normalized Lucene Distance:NLD）を提案する：

$$NLD(x,y) = 1.0 - e^{-\alpha NGD(x,y)} \quad (2)$$

ただし

$$\alpha = \log\left(\frac{\min(f(x), f(y))}{f(x,y)}\right) \quad (3)$$

である．式（２）は値域に ∞ を含む NGD を値域 {0,1} に写像しているため，一般に類似度が高くなる（距離が近くなる）傾向があることが予想されるので，次節にて NGD と NLD の比較実験を行う．

4. 文字列間類似度測定実験

本節では，異なるカテゴリに属すると人間が判断すると思われるキーワード 5 個× 2 種類の 10 キー

表 1: 実験対象芸能人名称

女性有名人	スポーツ
akb48	ゴルフ
少女時代	JRA
大島優子	野球
前田敦子	サッカー
KARA	阪神タイガース

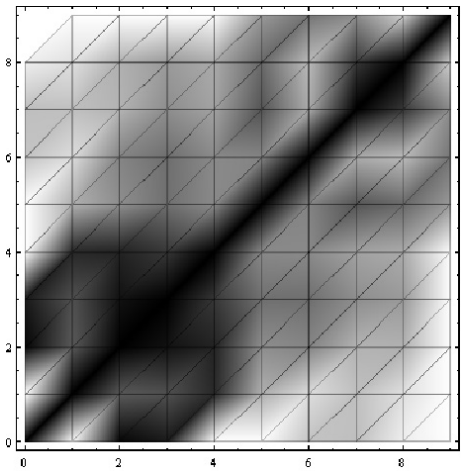

図 1: NGD による類似度測定

ワードの間の類似度を測定し，カテゴリ分類結果が人間の主観と近いかどうかを検証する実験結果を紹介する．本実験では，Lucene が検索を実施する際のアナライザと呼ばれる文脈解析手法として，辞書である NAIST-JDIC[7] と Igo-Analyzer[8] を導入した．また，NAIST-JDIC には「akb48」等の新しい語を加えるカスタマイズを施した．

類似度測定を行う対象としては，カテゴリ女性有名人及びカテゴリスポーツに属する単語を用いた．利用した単語を表 1 に示す．なお，これらの単語は単純に原稿執筆時に楽天 Infoseek キーワードランキング [9] から取得したもので，並び順は 2011 年 6 月度の順位である．この全ての文字列データに対し，NGD を適用して類似度を測定した結果を図 1 に，NLD を適用して類似度を測定した結果を図 2 に示す．色が濃いほど意味的な関係が深いことを表す．番号 0 から 4 が女性有名人を表し，5 から 9 がスポーツ名称を表す．従って濃度グラフは 2 つずつ重複する 4 つの領域が現れ,「女性有名人」と「スポーツ」の同ジャンルに属する単語同士は濃い色が現れることが期待される．

結果を解析すると，NGD/NLD のいずれにおいてもともに（akb48，少女時代）を表す座標 (0,1) 及び (1,0) の色が薄くなっており，これは同ジャンルとは言え日本と韓国のグループという違いがあるので，妥当な結果である．他にも (大島優子，野球)（大島優子，サッカー）などが図 2 では異なるカテゴリで

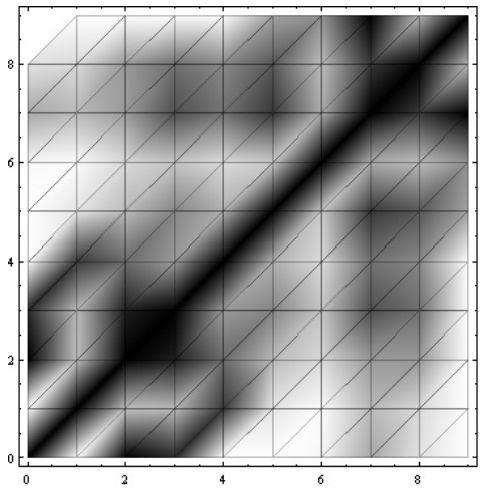

図 2: NLD による類似度測定

はあるが色が濃くなっている．これは単語としてカテゴリが異なるようにも思われるが，Wikipedia 上の「大島優子」の項に「またサッカーファンでもあり、自身のブログに度々サッカー日本代表についての記事を書いているほか、」などの記述があるから関連性があると判断されたものである．ヒット数が少ないために図 1 では色が薄くなっているが，本提案手法では補正されていることが分かる．

5. おわりに

本発表では，ユニバーサル符号化理論を元にした汎用的な類似度計算法を応用した文字列間の類似度測定法を提案し，カスタマイズ可能な汎用性及び実用性について論じた．今後は，NGD 及び NLD が文字列の検索ヒット数のみにより計算され，文字列長や文字コードには依存しないことを考慮してより高速に計算可能な実装を検討する予定である．

参考文献

[1] S. Chapman. Samś string metrics. *http://staffwww.dcs.shef.ac.uk/people/S.Chapman/stringmetrics.html*.

[2] 津久井めぐみ, 高田明典. メディアアートの現代的意義と分類に関する試論. 第 73 回情処全大, Vol. 4, No. 6ZD-10, pp. 611–612, 2011.

[3] R. L. Cilibrasi and P. M. B. Vitányi. The google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, Vol. 19, No. 3, pp. 370–383, 2007.

[4] Apache Foundation. Apache lucene. *http://lucene.apache.org/*.

[5] Wikipedia Foundation. Wikipedia 日本語版. *http://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2*.

[6] A. Patterson et. al. Nokogiri. *http://nokogiri.org/*.

[7] henrich and masayu-a. Naist japanese dictionary. *http://sourceforge.jp/projects/naist-jdic/*.

[8] http://d.hatena.ne.jp/sile. Igo analyzer. *http://es.sourceforge.jp/projects/igo/*.

[9] *http://ranking.infoseek.co.jp/keyword/mon/ranking_monthly_all.html*.