日本語キーワードの関連度に基づく検索エンジン「Google」「Yahoo!」の性能比較

Comparison between Google and Yahoo! based on similarity method

谷　研究室　　木下 孝史
Takashi kinoshita

概要

Kolmogorov 記述量に基づいたデータ間の類似度に関する距離公式 Normalized Information Distance が発案された。これを応用した、単語間の関連度を求める公式 Normalized Google Distance を提案された。本研究ではこれを用いて、検索エンジン「Google」「Yahoo!」の性能を単語間の関連度から検証する。

1　はじめに

現在、インターネット上には多種多様な情報が散らばっている。インターネット上で必要な情報を得ようとするとき、我々は検索エンジンを用いる。代表的な検索エンジンとして Google,Yahoo!,msn などが挙げられ、これらによって検索エンジンのシェアの大半を占めている。米国での検索エンジンのシェアは 07 年 12 月現在、Google65.98 %,Yahoo!20.88 %,msn7.94 %。日本では 08 年 1 月 20 日現在、yahoo!61.2 %,Google30.3 %,msn2.0 %となっている。

近年、Ming Li 達が Kolmogorov 記述量に基づくデータ間の類似度に関する距離をあらわす公式を発案した。その後、Rubi L. Cilibrasi 達 [1] によって、類似度に関する距離公式を応用し、検索エンジンを用いた 2 単語間の関連の度合を数値化する公式を提案した。

本研究では、この公式を用いて、検索エンジン「Google」[3]「Yahoo!」[4] を対象とし、単語のカテゴリー内でどちらの優れているか比較評価する。

2 節では Kolmogorov 記述量について、3 節では Normalized Information Distance について、4 節では Normalized Compression Distance について、5 節では Normalized Google Distance について、6 節では実験について、7 節では結果考察、8 節では今度の課題について解説する。

2　Kolmogorov 記述量の定義

Kolmogorov 記述量の基本的な定義を述べておく。

2.1　Kolmogorov 記述量

あるデータ x が存在し、データ x の Kolmogorov 記述量とはあるプログラム言語で x を生成する最小のプログラムのサイズである。どんなプログラム言語を選んでも、それが妥当であれば情報量は定数分の差しかないことが証明されている。S をプログラム言語、$|p|$ をプログラムサイズとすると、データ x の Kolmogorov 記述量 $K_s(x)$ は、以下のように定義される。

$$K_s(x) = min|p| : S(p) = x$$

すなわち $K_s(x)$ は、”プログラム言語 S において x を生成する最小のプログラムの長さ ”と考えていいだろう。

次に対象 y が与えられているときの対象 x についての記述量（相対記述量）について考える。文字列 x を、計算可能関数（インタプリタ関数）f、それに文字列 p と y により $f(p, x) = x$ と記述することを考える。インタプリタ関数 f のもとで、補助関数 y に対する x の記述量 K_f を次のように定義する。

$$K_f(x|y) = min\{|p| : P \in \{0, 1\} * f(p, y) = x\}$$

また、x と y を区別できる形で出力させる最短のプログラムの長さ $K(xy)$ とあらわす。$O(logK(xy))$ の誤差範囲では、

$$K(xy) = K(x) + K(y|x)$$

となることが証明されている。

2.2　K のアルゴリズム的性質

$K(x)$ を正の整数から正の整数への関数と考える。

定理

(a) 関数 $K(x)$ は帰納的ではない。しかも、どのような機能的部分関数 $\phi(x)$ を考えても、もしその値が無限個の点で定義されているのならば、その定義域上のどこかの点で $\phi(x) \neq K(x)$ となる。

(b) 引数 t に対しては単調減少で（全域的な）機能的関数 $H(t, x)$ が存在し,$\lim_{t\to\infty} H(t, x) = K(x)$。すなわち、

$K(x)$ の良い近似（ただし一様近似ではない）は計算不可能。

3 Normalized Information Distance

まず距離を定義する。

定義

ある集合 X 上の距離とは、非負実数値関数 $d : X \times X \to R$ で任意の $x, y, z \in X$ に対して次のような性質を満たす。

$$x = y \to d(x, y) = 0$$
$$d(x, y) \geq 0$$
$$d(x, y) = d(y, x)$$
$$d(x, y) \leq d(x, z) + d(z, y) : \text{三角不等式}$$

これを基に、情報の距離を考える。

入力 x から出力 y を生成する最小のプログラムのサイズを information distance という。このとき、x と y の information distance は以下のように表される。

$$E(x, y) = K(x, y) - minK(x), K(y)$$

information distance E(x,y) は任意の x, y, z に対して次のような性質を満たす。

$$E(x, x) = 0$$
$$E(x, y) > 0$$
$$E(x, y) = E(y, x)$$
$$E(x, y) \leq E(x, z) + E(z, y) : \text{三角不等式}$$

3.1 Normalized Information Distance

この information distance には欠点がある。対象の大きさによって、information distance が同じであっても似ているかどうかは変わってくる。そこで、この information distance を正規化する。この正規化した information distance を NID(Normalized Information Distance) といい、以下のように表す。

$$NID(x, y) = \frac{K(x, y) - min(K(x), K(y))}{max((K(x), K(y))}$$

4 Normalized Compression Distance

NID により、x と y の類似度を得ることができるが Kolmogorov 記述量は計算不可能であることから、NID を計算することはできない。しかし実際には、Kolmogorov 記述量を近似する圧縮プログラムを利用することができる。

圧縮アルゴリズムを計算可能な関数と定義する。C を圧縮機、C(x) を、文字列 x を圧縮機 C で圧縮したときのサイズとする。すると、Normalized Compression Distance を以下のように表すことができる。

$$NCD(x, y) = \frac{C(x, y) - min(C(x), C(y))}{max((C(x), C(y))}$$

5 Normalized Google Distance

5.1 Google

コーパスは文字やフレーズの関係頻度を定義する。グーグルには１００万のコーパスがある。グーグルにインデックスされているウェブページは 10^10 であり、ある単語は１００万のページに現れるとされている。これは十分な大きさであり、グーグル検索によって得られる検索結果の数は実際の頻度を近似できると考えられる。これを利用して、関連度を求める。

5.2 Normalized Google Distance

以下に、グーグルを用いた関連度、NGD(Normalized Google Distance) を定義する。

N:グーグルインデックス数
x, y:単語
X, Y:単語 x, y によって得られるそれぞれのウェブページの集合
$|X|, |Y|$:X, Y それぞれの集合の数
$g(x) = \frac{|X|}{N}$
$G(x) = log\frac{1}{g(x)}$

$$\begin{aligned} NGD(x, y) &= \frac{G(x, y) - min(G(x), G(y)}{max(G(x), G(y)} \\ &= \frac{\{logf(x), logf(y)\} - logf(x, y)}{logN - min\{logf(x), logf(y)\}} \end{aligned}$$

5.3 NGD の性質

NGD(x,y) は任意の x, y, z に対して以下の性質を持つ。

$NGD(x, y)$ の値域 : $[0, \infty)$
$NGD(x, x) = 0$
$NGD(x, y) = NGD(y, x)$
$NGD(x, y) \leq NGD(x, z) + NGD(z, y)$ を満たさない

この公式は他の検索エンジンでも同様に扱える。

6 実験

6.1 実験概要

Google と Yahoo!それぞれに単語を送り、その検索によって得られる件数をデータとして扱う。本実験では、

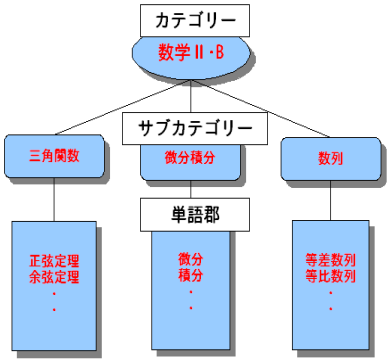
対象を検索エンジンとしているが、検索エンジンの性能を直接評価するのは非常に困難である。そこで、カテゴリーを利用することで Google と Yahoo!を比較する。

6.2 カテゴリー

カテゴリーはサブカテゴリーから構成され、サブカテゴリーはそれに含まれる単語から構成される。
カテゴリーは比較を行う分野を表し、サブカテゴリーはカテゴリーに含まれる分野を表す。

6.3 カテゴリーの例

以下にカテゴリーを「数学　B」にしたときの例を示す。



6.4 得点付け

本実験では得られた NGD を用いて得点付けをする。得点の付け方を以下のものとする。

単語 x,y における Google の NGD を NGD_g,Yahoo!の NGD を NGD_y とする。
$\frac{NGD_g}{NGD_y}$ を Yahoo!の得点とし、その単語間の得点とする
$\frac{NGD_y}{NGD_g}$ を Google の得点とし、その単語間の得点とする
$\frac{NGD_g}{NGD_y} = 1$ のとき、その単語間の得点を 0 とする
各 Google への得点、Yahoo!への得点の総和を求め、それをそのサブカテゴリーの得点とする。

6.5 実験方法

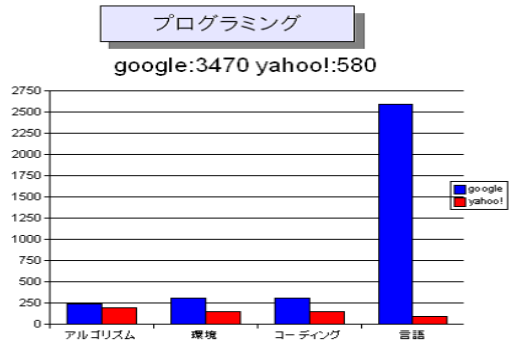
1. サブカテゴリーの単語郡において、全ての単語間における NGD を「Google」と「Yahoo!」を用いて求める。

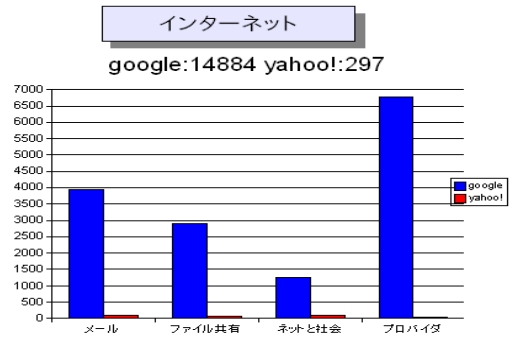
2. 「Google」と「Yahoo!」で求めた NGD を用いて、両者の対応する NGD から得点付けをする。

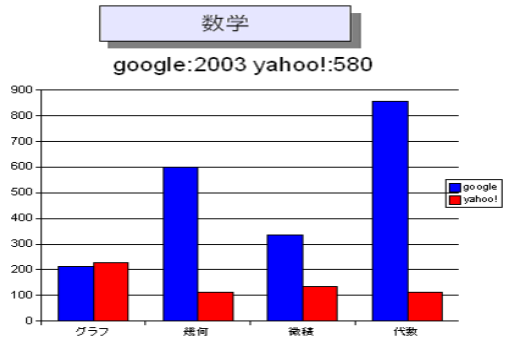
3. Google への得点の総和、Yahoo!への得点の総和をともにサブカテゴリーの得点とする。

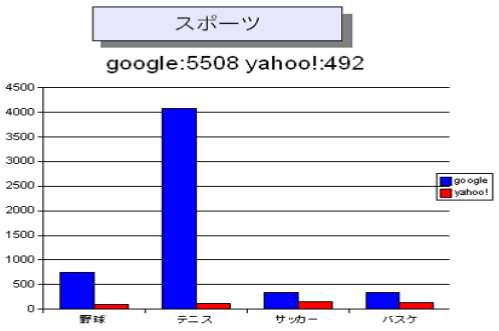
4. 各サブカテゴリーで得られる Google の得点、Yahoo!の得点のそれぞれの総和をカテゴリーの得点とし、評価する。

7 結果考察









今回は４つのカテゴリーと少数ではあるが、明らかな違いが現れている。類似度、という観点から見ると、圧倒

的に「Google」が勝っているということがわかる。各サブカテゴリーでの単語の選び方による違いもあるだろうが、そこを考慮しても検索エンジンとしては「Google」が優勢という傾向はあるだろう。

8　今後の課題

今回の実験ではカテゴリーが４つと少なかったため、より多く、幅広く実験データサンプルを取り、カテゴリーを増やし、再度検証する。また、サブカテゴリーの単語の選び方、評価の仕方によってどのように違いが現れるかを検証する。

参考文献

[1] Rubi L.Cilibrasi and Paul M.B. Vitanyi
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING
The Google Similarity Distance (2007)

[2] 谷 聖一
データの複雑さ・データ表現の複雑さ

[3] Google
http://www.google.co.jp/

[4] Yahoo!Japan
http://www.yahoo.co.jp/