1 対多関係を検出する問題における類似度の評価

吉成　未菜里† 　梅村　恭司†† 　岡本圭史† 　山本英子†††

†仙台高等専門学校　情報システム工学科　〒989-3128 仙台市青葉区愛子中央四丁目 16 番 1 号
††豊橋技術科学大学　情報・知能工学系　〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
†††岐阜聖徳学園大学　経済情報学部　〒500-8288 岐阜県岐阜市中鶉 1-38

E-mail: †{s1100640, okamoto}@sendai-nct.{jp, ac.jp}, ††umemura@tut.jp †††eiko@gifu.shotoku.ac.jp

あらまし　出現パターンが類似する 2 つのラベルに関係があると考え, その 2 つの組を求める問題はデータマイニングの基本の問題である. 問題によっては, 2 つのラベルの関係が親と子の関係にあり, 1 つの親ラベルに対し, 多数の子ラベルが対応することがある．この状況において，どのような類似度が検出の性能が高いかについて，実験を行った．実験対象は，補完類似度，条件付き確率，コサイン関数，相互情報量および，この問題で性能が高いという報告のある包含検出関数である．実験の結果，包含検出関数が 1 対 10 以上の正解の場合に，高い性能を示した．ラベルの親と子の対応関係における出現比率は前もってわかることが多いので，比が 1 対 10 以上のときに包含検出関数を使うことに効果があることを報告する．

キーワード　類似度，データマイニング，包含関係

1. はじめに

本稿で取り上げるのは，コーパス中に出現するラベル集合から，ラベル間における 1 対多関係を推定する問題である．1 対多関係とは，対応関係のうち一方が単独，もう一方が複数である関係で，例えば，新聞記事中の地名における都道府県を表すラベルと市郡名を表すラベルの関係などが挙げられる．

語の出現から意味の関係を推定するために相互情報量を用いると効果があるという報告[1]のように，データの出現パターンの類似度を利用して関係を推定する問題は自然言語処理の基本操作である．通常は，出現パターンの距離を構成する類似度を用いることが多い[8][9]. その場合，距離を構成する類似度は引数を交換しても値がかわらない関数（対称性のある関数）となる. これは，検出する関係に，対称律の成り立つことを暗黙に仮定することになる.

一方，データベースのなかから関係を推定するための基本アルゴリズム[2]は，条件付き確率の推定値を用いており，この類似度は対称性がない．自然言語での処理でも，非対称の共起性に注目する報告[4]や，階層的なオントロジを構築する問題[5]がある．この場合には，正解は 1 対多関係であり，正解がこのような関係であっても，暗黙的に 1 対 1 を仮定した関数で関係の推定が行われるのは問題であると考えた.

1 対多の関係を推定するには，非対称性の関数が適していると考えられる．しかしながら，類似度によって性能が異なることはよく知られており[6], 広く使われる条件付き確率よりも検出性能が良い関数が存在する可能性がある. たとえば, 光学的文字認識装置（OCR）で，重畳するパターンに頑健な関数として選択された類似度である補完類似度[3]は，非対称の形の関数であり，関係の抽出の問題とは異なる分野で優れた性質を示した関数である. 山本ら[7]はこの関数（補完類似度）について調べ, 1 対多関係をもつラベル（地名ラベル）の関係の抽出能力を測定し，一般に使われている関数よりも関係の検出能力が高いことを示した.

近年，1 対多関係の推定の問題に関して，性能の高い関数を発見するために，皆川ら[10]は，条件付き確率，相関係数，補完類似度を特別な場合として含む数式集合を定義し，その中の一つに，地名ラベルの関係の検出性能の高い関数があることを示した. 本稿では，この手法により皆川らが提案した関数を包含検出関数と呼ぶ．包含検出関数は，新聞記事に含まれる地名の関係推定においては高い性能を示したが，どのような条件において高い性能を発揮することができるのかは明らかになっていない.

ここで，1 対多関係において，「1」側のラベルを親ラベル,「多」側のラベルを子ラベルと呼ぶこととする. すると，1 つの親ラベルに対応する子ラベルの数は問題によって大きく異なる．対応する子ラベルが高々数個である親ラベルが多くを占める場合，その問題でのラベルの関係推定では，1 対 1 関係のときに高性能である関数の性能が高いと予想される．一方，1 つの親ラベルに多数の子ラベルが対応する問題に適した関数は，1 対 1 関係のときとは違っていると考えられる. 包含検出関数はどちらの問題に適するのか，また，他の類似度と比較してどの程度高い性能を発揮するのかが明らかになれば，問題解決に用いる類似度の選択に役立つ．なぜなら，1 つの親ラベルに対応する子ラベルの数は，問題によっては事前に予測することができ

ることも多いからである．

そこで，親ラベルと子ラベルの数によって各類似度の振る舞いはどう変わるのか調べるため，実験を行った．具体的には，正解集合を，親ラベルと子ラベルの比が 1:9 以下のラベル集合と 1 対 10 以上のラベル集合の 2 つに分ける．次にそれぞれについて，(1)実際の 1 対多関係から人工的に生成したデータ集合(2)現実の新聞記事コーパスから得られたデータ集合を用いて，親ラベルと子ラベルの関係にあると思われるラベルの組を推定し，各類似度の性能を比較・評価する．比較対象は，包含検出関数，補完類似度，条件付き確率，コサイン関数，相互情報量である．また，実験対象のラベルとしては，地名を選んだ．地名は実世界において 1 対多関係をもち，正解集合も実世界で定まっているためである．実験の結果，包含検出関数の性能は親ラベルと子ラベルの比に関係していることが明らかになった．

本稿では，包含検出関数は，親ラベルと子ラベルの数の比が 1 対 9 以下の場合に比べて，1 対 10 以上の場合に高い性能を示すことを報告する．

2. 比較対象となる類似度

本章では，最初にラベルの出現パターンを表すパラメータについて，[7]および[10]に基づいて表現する．次に，パラメータを用いて，包含検出関数と，比較対象となる類似度を表現する．

2.1. パラメータ

2 つのラベル X,Y の出現パターンについて，表 1 に示す．

表 1: ラベルの出現パターンを表すパラメータ

		Y	
		出現する	出現しない
X	出現する	a	b
	出現しない	c	d

表 1 に基づいて定義すると，それぞれ以下のような意味をもつ．

- a: X,Y がどちらも出現するデータ数
- b: X が出現し，かつ Y が出現しないデータ数
- c: X が出現せず，かつ Y が出現するデータ数
- d: X,Y がどちらも出現しないデータ数
- a+b+c+d: データの総数(N とする)

2.2 からはこれらのパラメータを用いて，類似度を表現する．

2.2. 条件付き確率

2 つのラベル間の関係が強い際に大きな値を示す尺度として知られている最も単純なものは条件付き確率である．条件付き確率は，Y が出現するという条件下で X も出現する確率である．パラメータを用いて表すと，次のようになる．

$$\text{条件付き確率} = \frac{\mathrm{a}}{\mathrm{a}+\mathrm{c}}$$

2.3. コサイン関数

コサイン関数は，式が単純であり，もっとも広く知られている類似度の 1 つである．N 個のデータがあったとき，i 番目の X が出現するならば$f_i = 1$，出現しないならば$f_i = 0$とし，X に関する N 次元のベクトル$\vec{F} = (f_1, f_2, \dots, f_N)$を定義する．同様に，Y に関するベクトルも$\vec{T} = (t_1, t_2, \dots, t_N)$と定義する．これらのベクトルのなす角θについての$\cos\theta$がコサイン関数となる．パラメータを用いて表すと，次のようになる．

$$\text{コサイン関数} = \frac{a}{\sqrt{(a+b)(a+c)}}$$

2.4. 相互情報量

相互情報量は，情報理論における基礎的な概念であり[1]，一般的には一方の変数を知ることによってもう一方の変数に関する不確実性がどれだけ減少するかを表す．相互情報量は 2 つの確率変数に対して定義される尺度であるが，本稿では 2 つの確率変数を X の出現確率，Y の出現確率とする．パラメータを用いて表すと，次のようになる．

$$\begin{aligned}\text{相互情報量(X,Y)} &= \frac{\mathrm{a}}{\mathrm{N}}\log\frac{aN}{(a+b)(a+c)} \\ &+ \frac{\mathrm{b}}{\mathrm{N}}\log\frac{bN}{(a+b)(b+d)} \\ &+ \frac{\mathrm{c}}{\mathrm{N}}\log\frac{cN}{(c+d)(a+c)} \\ &+ \frac{\mathrm{d}}{\mathrm{N}}\log\frac{dN}{(c+d)(b+d)}\end{aligned}$$

2.5. 補完類似度

補完類似度は，澤木らの研究[3]によって提案された尺度であるが，1 対多関係の推定のために提案された類似度ではない．本来は，かすれやよごれのある文字，背景やデザインが施された文字などの劣化印刷文字を認識するために提案された尺度である．山本らは，この補完類似度が 1 対多関係の推定に有効であることを示した [7]．パラメータを用いて表すと，次のようになる．

$$\text{補完類似度} = \frac{ad - bc}{\sqrt{(a+c)(b+d)}}$$

2.6. 包含検出関数

ここで，皆川らの研究[10]によって提案された関数を包含検出関数と名付ける．包含検出関数は，パラメータを用いて表すと，次のようになる．

$$\text{包含検出関数} = \frac{ad(b-c)}{\sqrt{(a+1)(a+c+1)(b+c+1)(a+b+c+1)}}$$

包含検出関数の特徴は，包含関係を重視するという点にある．その特徴は特に分子に現れている．ad は，2 つのラベルの出現パターンがどれだけ一致しているかを表しており，b-c は親ラベルの出現パターンが子ラベルの出現パターンをどれだけ包含しているかを表しているといえる．このことから，上記の関数を包含検出関数と名付けた．

また，包含検出関数は，7 年間分の新聞記事データから地名の 1 対多関係を推定する実験において，補完類似度よりも高い性能を示すことが分かっている[2].

3. 各類似度の評価方法

実験方法の前に，各類似度の性能を評価する方法を本章で述べる．

3.1. グラフの作成方法

本研究においては，グラフを作成することで各類似度の性能を比較・考察する．グラフは，データ集合中のラベルの各ペアについて計算された類似度をもとに作成する．グラフの横軸はランク，縦軸は再現度とする．再現度は以下の式で求めることができる．

$$\text{再現度} = \frac{|\boldsymbol{R}_D \cap \boldsymbol{R}|}{|\boldsymbol{R}|}$$

R:実際の正解集合

R_D:データから推定された正解集合

また，グラフの詳しい作成方法は以下に示す．

グラフの作成方法

1: for each 条件付き確率，コサイン関数，相互情報量，補完類似度，包含検出関数
2: for each ラベルのあらゆるペア $\langle l_i, l_j \rangle$ ∈データ集合 D
3: 類似度を計算する
4: end for
5: 計算結果を類似度の降順にソート
6: ソート結果に 1 から 4000 位までのランクを付与する
7: ランク r の初期値を 500 とする
8: while r <= 4000
9: 1 位から r 位までの再現度を求める
10: 横軸 r，縦軸再現度の点をプロットする
11: r ← r + 500
12: end while
13:end for

3.2. 関数の性能とグラフ

3.1 の方法で作成したグラフを用いて関数の性能を考察する．本稿では，プロットした全ての点においてある関数 f_1 よりも別の関数 f_2 が高い再現度を示した場合を「f_2 は f_1 よりも性能が高い」と呼ぶこととする．このとき，f_2 のグラフの線は常に f_1 のグラフの線よりも上にある．

また，グラフの右側ほど再現度計算の対象となるラベルのペアが増えるので，グラフは単調増加となる．再現度の定義上の最大値は 1 であるが，データ集合の中に正解となるラベルのペアがすべて含まれるわけではないため，プロットされた再現度の最大値は 1 になるとは限らない．

4. 人工的に生成したデータ集合を用いた実験

本章では，人工的に生成したデータ集合を用いた実験について述べる．人工的に生成したデータでの実験を行うことで，親ラベルと子ラベルの比の違いと包含検出関数の性能の関係，包含検出関数と他の類似度の特徴の違いについて，自然データに依存しない状態で評価・考察することができる．

実験では，山本ら[7]，皆川ら[10]と同様に，地名(都道府県市郡名)をラベルとして用いた．理由としては，皆川らも[10]で述べているように，地名が実世界において 1 対多関係を持っていることと，地名間の関係が実世界で定まっており推定結果の評価が容易であることが挙げられる．

実験では，各々の類似度について，ノイズあり/なしのデータ集合に出現する全てのラベルの組み合わせについて，類似度の計算を行う．その値を用いて、3.1 の方法でグラフを作成する．

4.1. データ集合の作成

あらかじめラベル集合・正解集合を親ラベル:子ラベル=1:9 以下，親ラベル:子ラベル=1 対 10 以上に分けてからデータ集合を生成した．以後，親ラベル:子ラベル=1:9 以下であることを子が少ない，親ラベル:子ラベル=1 対 10 以上であることを子が多いと表現することと

する．また，子の多い少ないそれぞれについてノイズのないデータ集合とノイズのあるデータ集合の両方を生成した．生成アルゴリズムは，山本らの実験[1]で用いられたものと同一である．

生成アルゴリズムを以下に示す．

ノイズのないデータ集合の生成

1: until 5000 回繰り返すまで
2: until 2 回繰り返すまで
3: 正解集合 R からランダムに，親ラベル l_p と子ラベル l_c のペアを 1 組取り出す
4: 取り出したペア〈l_p,l_c〉をノイズのないデータ集合 D^*に加える
5: end until
6: 改行する
7: end until

ノイズのあるデータ集合の生成

1: while i < 5000
2: D^*の i 行目 d_i を取り出す
3: ラベル集合 L からランダムに 1 つラベル l を取り出す
4: $d_i \cup l$ をノイズのあるデータ集合 D^{**}に加える
5: 改行する
6: i ← i + 1
7: end while

生成アルゴリズムの繰り返し回数にもあるように，それぞれのデータ集合は 5000 行からなる．また，ノイズのないデータ集合の 1 行は，正解集合のペアのうちランダムな 2 組で構成されている．ノイズのあるデータ集合の 1 行は，ノイズのないデータ集合の構成に，ノイズとしてランダムなラベル 1 個を付加したものとなっている．

4.2. 実験結果

実験結果のグラフを図 1 に示す．

子が多く，ノイズが含まれるデータ集合の場合，プロットした点において包含検出関数は実験した類似度の中で最も高い，または補完類似度に次ぐ再現度を見せている．ノイズが含まれない場合も，実験した類似度の中で最も高い，または補完類似度に次ぐ再現度を見せている．

しかし，子が少ないデータ集合での実験では，包含検出関数の性能は相対的には大幅に落ちる．子が少なく，ノイズのあるデータ集合の場合，プロットした全ての点において，包含検出関数の再現度は他のどの類似度よりも低い．ノイズのないデータ集合でも，相互情報量の次に再現度が低いという結果になった．

4.3. 考察

実験結果から，包含検出関数は，子が多いデータで性能が良く，子の数が少ないと性能が落ちることがわかった．この傾向は，乱数を変更して 7 回再実験しても同様であった． よって，この結果は危険率 1 %で有意である．

5. 新聞記事から得られたデータ集合を用いた実験

本章では，実際の新聞記事から得られたデータ集合を用いた実験について述べる．また，その結果を人工的に生成したデータでの実験結果と比較する．

実際に類似度を適用するデータ集合は，人工的に生成したデータ集合とは違った特徴をもつことが多い．まず挙げられるのは,ラベルの出現頻度の偏りである．人工的に生成したデータ集合は乱数を用いるためラベルの出現頻度に偏りが少ないのに対し，実際のデータ集合はラベルの出現頻度に偏りをもつことが多い．次に挙げられるのは,出現数をカウントする基準となる，データの大きさである．人工的に生成したデータ集合は，出現数をカウントする単位となるデータの大きさが全て均一であるが，実際のデータ集合では均一でないことも多い．

このことから，人工的に生成したデータ集合での実験結果だけでは，子の多い 1 対多関係の推定という問題に対して，複雑な要因が存在する状況でも同じ性質を示すかを明らかにすることができない．そこで，現実に沿ったデータ集合による実験が必要であると考え，新聞記事から得られたデータ集合でも実験を行った．

5.1. 使用したデータ集合について

現実に基づいたデータ集合として，91 から 97 年度版の毎日新聞の記事から，都道府県市郡名のみを抽出したものを用いた．1 つの記事が 1 データに対応するため，a, b, c, d, N は記事を単位とした出現数となる．正解集合は，ポスタルガイドから抽出した．

実験方法は 3.1 に倣う．すなわち，各々の類似度について，各年度のデータに出現する全てのラベルの組み合わせについて，類似度の計算を行う．次に，計算された値を用いて，ラベルの組み合わせを類似度の高い順に並べ替える．それから，その上位 4000 組を，横軸に類似度の高さのランク，縦軸に 1 位からその順位までの正解集合の再現度をとったグラフ上にプロットする．

ただし，正解集合は子の少ない 1 対多関係に含まれ

るラベルの集合と子の多い 1 対多関係に含まれるラベルの集合とにあらかじめ分けておき，正解集合の親ラベルと子ラベルの比が実験結果に及ぼす影響を調べる．

各々の類似度について，各年度のデータに出現する全てのラベルの組み合わせについて，類似度の計算を行う．計算された値を用いて，ラベルの組み合わせを類似度の高い順に並べ替える．その上位 4000 組を，横軸に類似度の高さのランク，縦軸に 1 位からその順位までの正解集合の再現度をとったグラフ上にプロットする．

5.2. 実験結果

実験結果のグラフを図 2 から図 8 に示す．図 2 から図 8 より，91 年から 97 年まで，7 つの年について，全ての年について以下の 7 つのことがいえる．

1. 子が少ないときに，包含検出関数と条件付き確率は，同じ傾向，かつ，類似の性能を示す．
2. 子が少ないときは，包含検出関数が 1 番となることはなかった．
3. 子が多いときは，包含検出関数が上記 5 つの関数のなかで一番性能が高い．
4. 子が多いときは，コサイン尺度の性能が一番悪い．
5. 子が多くても少なくても，条件付き確率よりも，包含検出関数の性能が高い．

7 年間分のデータで一貫しているので，上記 1 から 5 は危険率 1%で有意である．

5.3. 考察

本稿で扱った，生成データと新聞記事のデータ，両方の結果により，包含検出関数を使うときの次のことが示唆される．子の数が小さいことが分かっているときには，包含検出関数は推奨できない．一方で，子の数が 10 以上あるような場合には,包含検出関数を使うことが推奨できる．また，子の数が小さいケースと大きいケースが混在しているのが，皆川らによって報告されたケース[10]である．皆川らによると，全体的な性能は包含検出関数が高い．包含検出関数を使用する際は，子の数が小さい正解に対しての振る舞いに注意するべきであることが分かった．

6. おわりに

本稿では，人工的に生成したデータ集合と実際の新聞記事に基づいたデータ集合を用いた実験によって，ラベル間の 1 対多関係を推定する問題のうち，特に 1 つの親ラベルに 10 個以上の子ラベルが対応する問題において包含検出関数が有効であることを示した．包含検出関数とは，次の関数である．

$$\frac{ad(b-c)}{\sqrt{(a+1)(a+c+1)(b+c+1)(a+b+c+1)}}$$

1 つの親ラベルに対応する子ラベルが具体的にいくつであるときに包含検出関数が高い性能を発揮するのか，また子ラベルの数以外にも類似度の性能を左右する要素はあるのか，あるとしたら何かなどは，今後さらなる調査が必要である．

参 考 文 献

[1] Church, K. and Hanks, P., "Word Association Norms, Mutual Information, and Lexicography.", Computational Linguistics, Vol.16, No.1, 1990.

[2] Agrawal R. and Srikant, R.:"Fast Algorithms for Mining Association Rules", In Proceedings of 20th international Conference of Very Large Database, VLDB 1215, pp487-499, 1994.

[3] 澤木美奈子, 萩田紀博, “補完類似度による劣化印刷文字認識”, 電子情報通信学会技術研究報告. PRU, パターン認識・理解 95(43), 1995.

[4] 新納浩幸, 井佐原均, "片方向の共起性による述語定型表現の自動抽出", 自然言語処理 Vol.2, No. 3, pp73-86, 1995.

[5] Caraballo, S.A., "Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text.", In Proceedings of The 37th Annual Meeting of the Association for Computational Linguistics, pp.57-64, 1999.

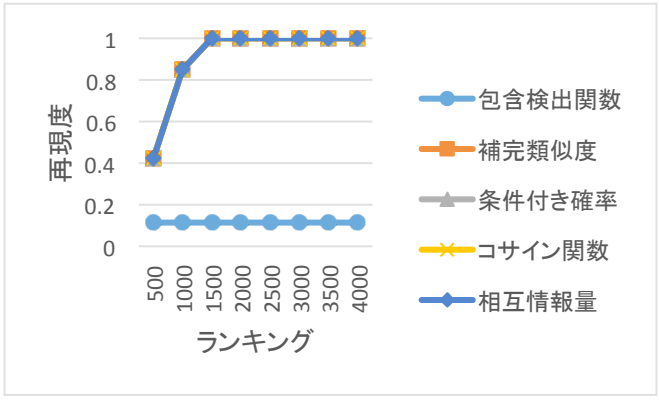
[6] P. N. Tan, V. Kumar and J. Srivastava: "Selecting the Right interestingness Measure for Association Patterns.", In Proceedings of The 8th International Conference on Knowledge Discovery and Data Mining, pp. 32-41, 2002.

[7] 山本英子, 梅村恭司, “コーパス中の 1 対多関係を推定する問題における類似度”, 自然言語処理 Vol. 9 No. 2, 2002.

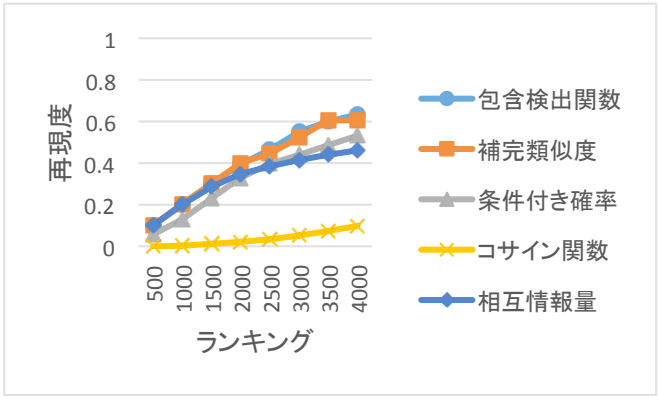
[8] 高村大也, 奥村学, “自然言語処理シリーズ 1 言語処理のための機械学習入門”, コロナ社, 2010.

[9] C.M.ビショップ, 元田浩, 栗田多喜夫, 樋口知之, 松本裕治, 村田昇, “パターン認識と機械学習 上”, 丸善出版, 2012.

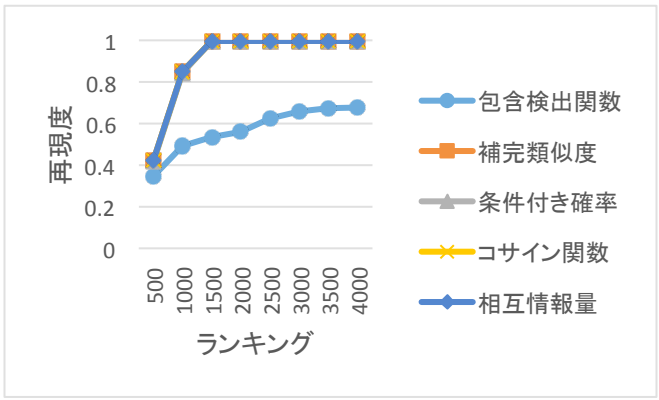
[10] 皆川歩, 岡部正幸, 梅村恭司, “数式の規則的生成による類似度の探索の研究”, 言語処理学会第 19. 回年次大会, pp850－853, 2013.



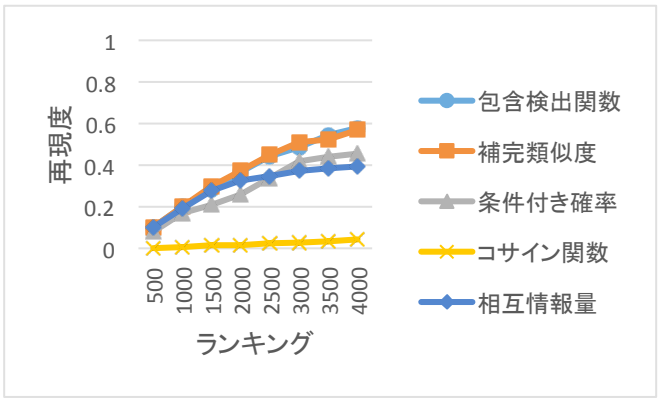
ノイズなし-子が少ない
(補完類似度・条件付き確率・コサイン関数の
グラフは相互情報量のグラフと重なっている)



ノイズなし-子が多い

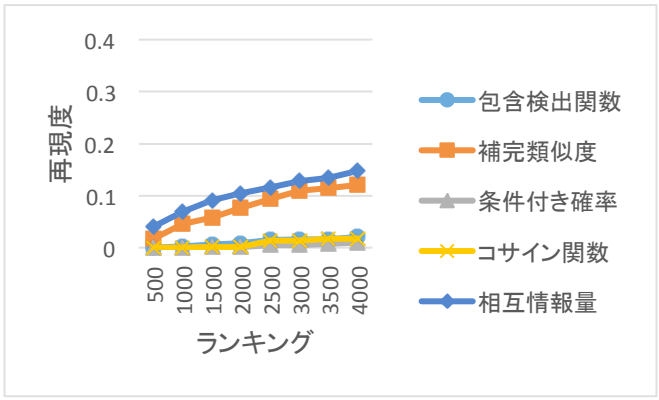


ノイズあり-子が少ない
(補完類似度・条件付き確率・コサイン関数の
グラフは相互情報量のグラフと重なっている)

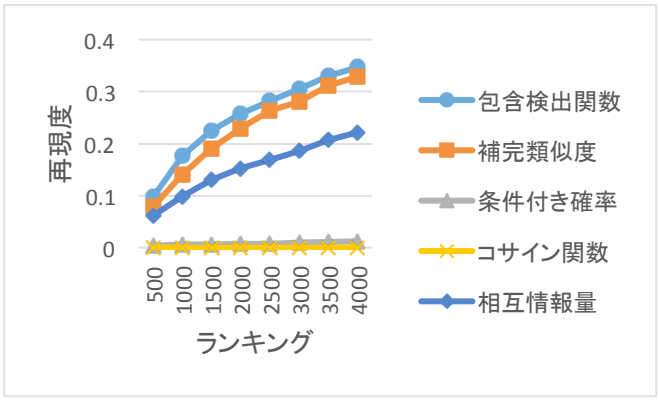


ノイズあり-子が多い
(包含検出関数のグラフは補完類似度の
グラフとほぼ重なっている)

図 1: 人工的に生成したデータ集合による実験の結果

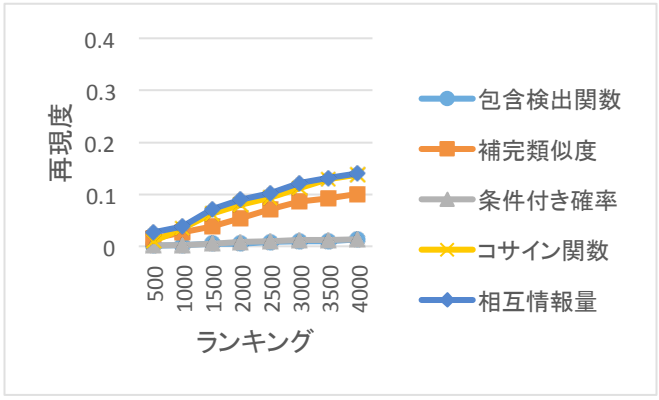


91 年度-子が少ない
(包含検出関数と条件付き確率関数のグラフ
はコサイン関数のグラフとほぼ重なっている)

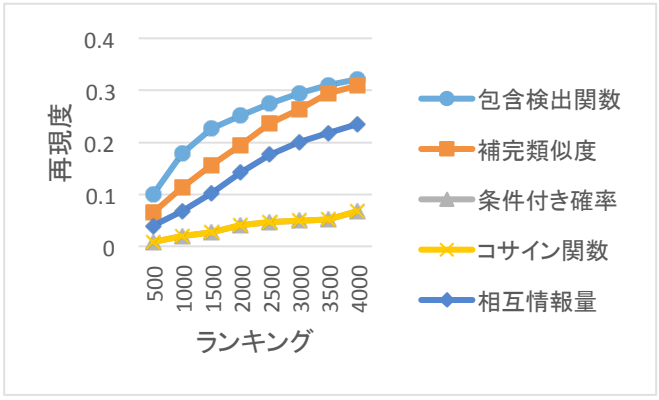


91 年度-子が多い
(条件付き確率のグラフはコサイン
関数のグラフとほぼ重なっている)

図 2: 91 年度の新聞記事から得られたデータ集合による実験結果

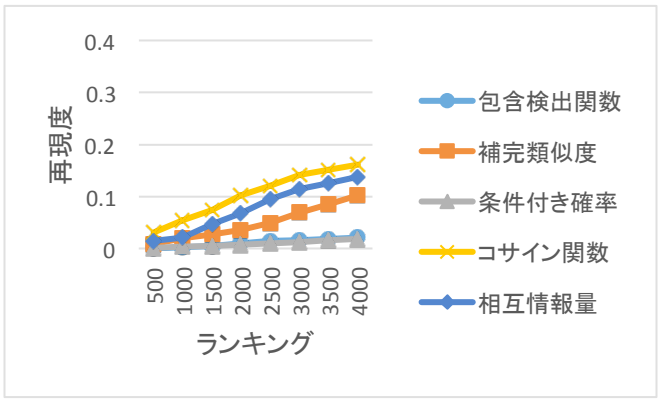


92 年度-子が少ない
(包含検出関数のグラフは条件付き確率のグラフとほぼ重なっている)

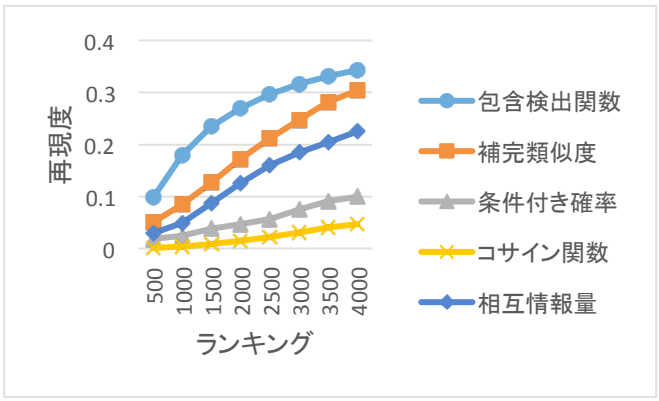


92 年度-子が多い
(条件付き確率のグラフはコサイン関数のグラフとほぼ重なっている)

図 3: 92 年度の新聞記事から得られたデータ集合による実験結果

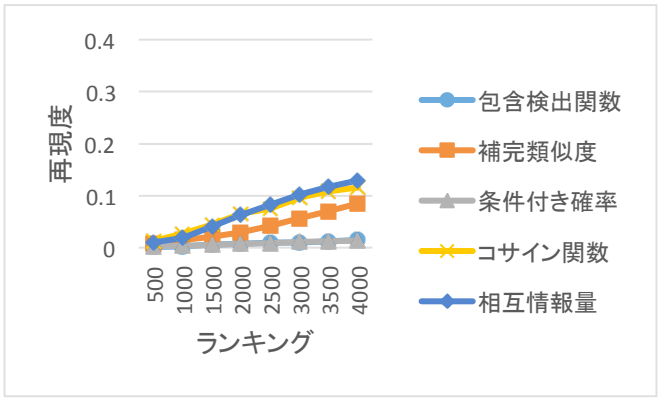


93 年度-子が少ない
(包含検出関数のグラフは条件付き確率のグラフとほぼ重なっている)

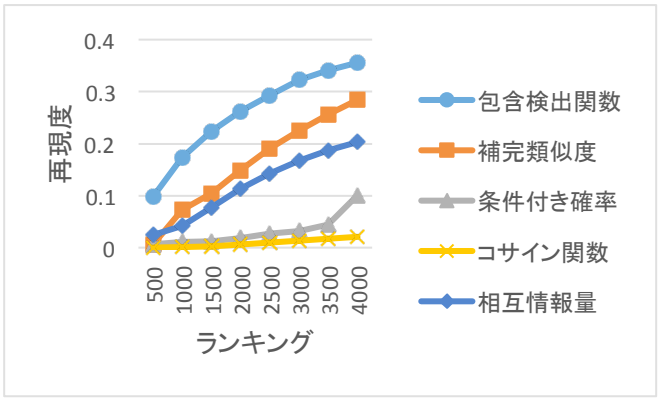


93 年度-子が多い

図 4: 93 年度の新聞記事から得られたデータ集合による実験結果

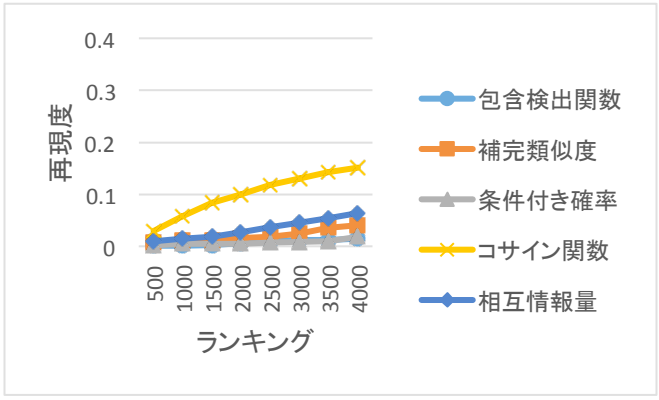


94 年度-子が少ない
(包含検出関数のグラフは条件付き確率のグラフとほぼ重なっている．また，コサイン関数のグラフは相互情報量とほぼ重なっている)

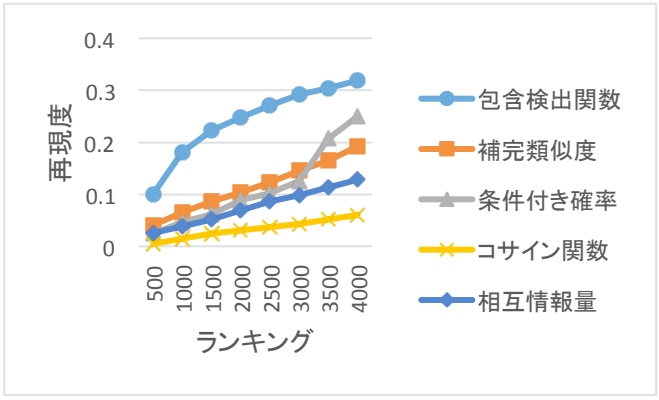


94 年度-子が多い

図 5: 94 年度の新聞記事から得られたデータ集合による実験結果

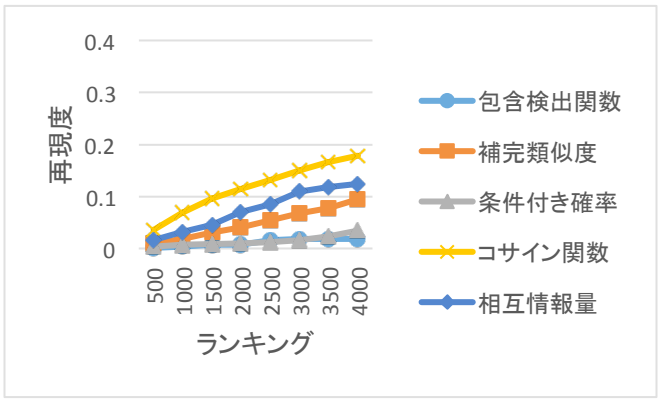


95 年度-子が少ない
(包含検出関数のグラフは条件付き
確率のグラフとほぼ重なっている)

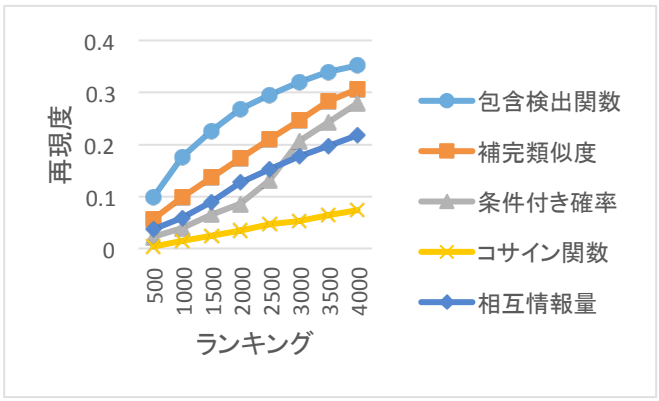


95 年度-子が多い

図 6: 95 年度の新聞記事から得られたデータ集合による実験結果

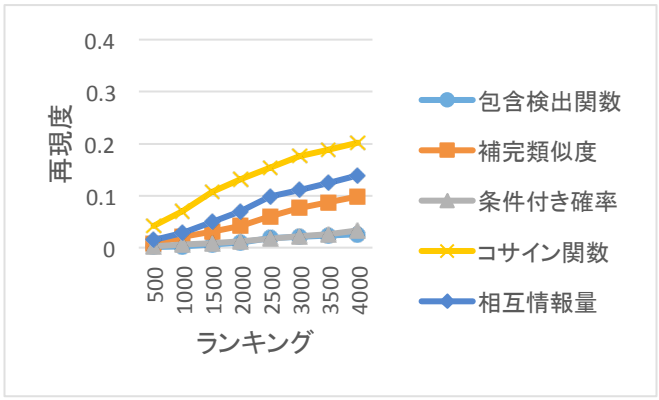


96 年度-子が少ない
(包含検出関数のグラフは条件付き
確率のグラフとほぼ重なっている)

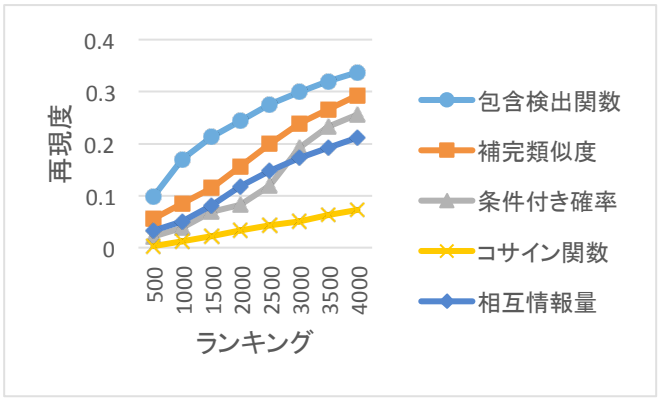


96 年度-子が多い

図 7: 96 年度の新聞記事から得られたデータ集合による実験結果



97 年度-子が少ない
(包含検出関数のグラフは条件付き
確率のグラフとほぼ重なっている)



97 年度-子が多い

図 8: 97 年度の新聞記事から得られたデータ集合による実験の結果