情報処理学会論文誌　データベース　Vol. 2　No. 2　110–125 (June 2009)

関係の類似性に基づくWebからのオブジェクト名検索

加藤　誠[†1]　大島裕明[†1]
小山　聡[†1]　田中克己[†1]

本稿では，関係を入力として与え，その関係との類似度に基づいてオブジェクト名を検索する手法についての提案を行う．一般的な検索エンジンを用いた場合，“京都と八ツ橋の関係と類似するような，ニュージーランドに対するもの” を検索することは，以下の 2 つの点で困難である．1 つは，ユーザはニュージーランドに関してある程度の知識を必要とする点であり，もう 1 つは，京都と八ツ橋の関係を言語や数値などで表現する必要がある点である．これらの条件を必要とすることなく，入力として A，B，C が与えられた場合，A と B で成り立つ関係の集合 Relation と，C と D で成り立つ関係の集合 Relation′ が類似するような D の名称を検索する手法を本稿では扱う．我々は，この検索の実現方法の 1 つとして，共起する語の出現分布の差分に基づく手法を提案する．提案手法は 2 つのプロセスからなる．まず，Web 検索エンジンの結果として得られるテキストから，2 つの語 A，B を強く結び付けるような語を発見する．次に，得られた語集合と語 C を用いて検索を行うことにより，A と B の関係と類似する関係を持つ，C に対する語 D を発見する．実験では，33 の関係，854 のテストセットを用いて，提案手法とベースライン手法，言語パターンを用いた手法との比較を行った．提案手法はベースライン手法，言語パターンよりも優れた精度を示し，上位 20 件に正解を含めることができたテストセットの割合は 49.8%であった．

Object Name Search from the Web Based on Relational Similarity

Makoto P. Kato,[†1] Hiroaki Ohshima,[†1]
Satoshi Oyama[†1] and Katsumi Tanaka[†1]

In this paper, we propose a method that searches for object names based on similarity of relations input by users. For example, it is difficult to search for what is to New Zealand as Yatsuhashi is to Kyoto by using a traditional information retrieval method because of the following two reasons. First, users need to have much information on New Zealand. Second, users have to represent the relation between Kyoto and Yatsuhashi by words or values. Object name search based on relational similarity is to search for D which is to C as B is to A without the two restrictions. We propose a method to realize it using the difference between distributions of co-occuring terms. From Web search results, our method finds terms which strongly connect two terms A and B, searches for Web pages with a term C and them, and discovers a term D which is to C as B is to A from them. We experimented with 33 relationships and 854 test sets to compare the proposed method, a baseline one and a lexicon pattern-based one. The proposed method got more precise results than the baseline and the lexicon pattern-based ones, and the percentage of test sets which obtained answers in top 20 was 49.8%.

1. はじめに

近年，インターネットと Web 検索エンジンの普及により，文書や画像，動画などの様々なメディアを容易に検索することが可能となった．

しかし，従来のキーワードや範囲指定などのクエリを用いた検索には様々な制約が存在する．このような検索クエリを用いる場合，検索対象にはあるキーワードが含まれている，ある属性が存在してその属性値は整数である，などといったように，我々は検索するデータに対して十分な知識を持っていなくてはならない．しかし，ユーザがある特定の事物に対して情報を得ようとして検索を行う際には，当然，それに対して十分な知識を持っていないことが想定される．また，ユーザは自分の検索意図を明確な言語表現や数値表現に変換しなければならず，人間同士の意図伝達で効果的に用いられるような，例示やメタファといった表現手段を使用することはできない．クエリの表現能力の問題から，ユーザはしばしば自分の複雑な検索意図をクエリで表現できない場合がある．

また，本稿では扱わないが，文書検索よりも画像や音楽，動画といった人間の感性に依存するような検索で，これらの問題は顕著となる．キーワードによる文書検索では，それらの構成要素である言語でクエリを表現すればよいが，これらマルチメディアコンテンツは言語よりも多くの情報量を持つうえ，まったく性質の異なる単語で表現しなければならない．そのため，我々がそれらをキーワードで表現することは難しい．また属性を指定する検索においても，画像のヒストグラムを指定するものや，音楽の波形を指定するような検索方法は直感的でなく，ユーザが指定するのは現実的でない．また日時情報や制作年などの情報もあ

†1 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

くまでも検索をするうえでは補助的な役割を担うものでしかない．そのため，ユーザがある特定のマルチメディアコンテンツを要求しようとしても，それらをキーワードや属性指定で表現することは文書を検索するよりも困難であると考えられる．

我々が現実世界において未知なる事象について説明を与えるとき，また，未知なる事象を理解しようとするときに，しばしばアナロジという認知過程を経る．アナロジは冒頭にあげた従来の検索手法の問題点を解決する可能性を秘めている．第 1 に，アナロジは既知の知識を未知の知識に対して適用し，推測することを可能とする．たとえば，音の性質が解明された際には音と水のアナロジが用いられた．水の波紋の性質を音も有しているだろうという類推がなされたのである．このように，アナロジによって未知の知識を既知の知識で置き換えることができる．アナロジを検索に利用することにより，ユーザは検索するデータに対して知識を必要とせず，既知の知識を利用してクエリを生成することができる．第 2 に，アナロジは既知の知識を元にした豊かな表現能力を有しており，それらは日常的によく用いられている．音と水の例を使えば，音は水のようである，というアナロジを用いることによって，我々は音について多くの属性の情報を簡略に伝えることができる．検索に適用すれば，ユーザは「ベートーヴェン」の「第九（喜びの歌）」のような風景画像や，「レオナルド・ダ・ヴィンチ」の「最後の晩餐」のようなクラッシック音楽といった，多くの情報を含む比喩を用いたクエリを用いることができる．

そこで本稿では，アナロジの中でも比較的単純な部類に属する比例アナロジを利用した情報検索手法，すなわち，関係に基づく情報検索手法を提案し，その 1 つの実現手法として共起を用いた手法について述べる．比例アナロジは，4 つの項，A，B，C，D において，A と B の関係と C と D の関係が等しいときに成り立ち，このとき $A : B :: C : D$ と表現されるものである．例として，月:地球::エウロパ:木星や大工:木::石工:石などがあげられる．

関係に基づく情報検索手法を説明するために，京都の八つ橋にあたるものはニュージーランドでは何であるか，という検索要求を考える．ユーザがニュージーランドについての知識をまったく持っていない場合，キーワードとして “ワイン” や “蜂蜜” などといった具体的な条件を指定して検索することはできない．また，京都と八つ橋という関係は「土産」といったキーワードに凝縮されるほど単純な関係ではなく，それは食品でなくてはならない，であるとか，嗜好品の類でなくてはならない，といった複雑な意図が込められていることに注意しなくてはならない．これらの理由から，既存の検索手法でニュージーランドの八つ橋にあたるものを検索することは非常に困難であることが分かる．

上述の例を実現するような検索は以下のようになる．

“データ A，B と C を入力として与え，A と B で成り立つ関係の集合を Relation，C とデータ D で成り立つ関係の集合を Relation′ としたとき，Relation と Relation′ の類似度が高い順にランク付けをした D を出力として返す検索”

京都と八つ橋，ニュージーランドの例にあてはめれば，A = 京都，B = 八つ橋，C = ニュージーランドを入力として，D = ワイン，蜂蜜などを出力として返す検索であると考えることができる．関係に基づくとは，すなわち，関係集合 Relation と Relation′ の類似度を考慮するということにほかならない．これに対して，一般的なキーワード検索は入力された語自体と検索対象となった文書に含まれる語との類似度，完全一致も含めた意味の類似を考慮した検索である．本稿では，関係に基づく情報検索手法の 1 つの実装として，入力と出力にオブジェクト名を想定したシステムを提案する．我々は，このような検索を Web 文書中の単語の共起を利用して実現し，それに対して評価実験を行った．

2 章では，関係に基づく情報検索に関係する研究のうち，アナロジと比例アナロジについての研究，関係発見，そして，本提案と類似した相対的情報検索手法について述べる．3 章では，関係に基づく情報検索についての定義を与え，4 章では Web 文書中の語の共起によって実現する手法を提案する．5 章では，我々が提案した実現手法についての定量的評価と考察を，6 章では本稿のまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 関係の類似性

紀元前 350 年，アリストテレス[20] によって比例的アナロジが言及されて以来，アナロジについて数多くの研究がなされてきた．

Structure-Mapping Engine[6] はアナロジをある領域から他の領域への知識の対応付けであると見なし，事例ベースの推論を行うシステムである．このシステムの元となる Structure-Mapping Theory では，要素どうしの類似性よりも，要素間の関係どうしの類似性に重点を置いている．我々はこの点を検索の分野において利用しようとしている．

Turney ら[19] は語 A と語 B の組が与えられたときに，あらかじめ用意された複数の語 C と語 D の組のうち，A と B の関係と最も近い C と D の組を選択するという比例アナロジの問題に対して様々なアプローチによる研究を行ってきた．その中でも最も効果的であることが示されたのは，ベクトル空間モデルを用いた手法である．この手法は，あらかじめ用意された複数の言語パターン，たとえば，“A of B” や “B to A” などに対応した次元を持つベクトルを用意し，その要素をそれらの言語パターンを満たす文書数としたもので A と B

の組を表現する．同様にして，用意された C と D の各組をベクトルによって表現し，これと A，B の組を表すベクトルとのコサイン類似度をとる．最後に，ベクトル間のコサイン類似度を関係の類似度と見なし，最も類似度が高い C，D の組を正解とするのが，ベクトル空間モデルを利用して提案された手法である．また，Turney [16),18)] はベクトル空間モデルを用いた手法を，Latent Relational Analysis（LRA）によって拡張した手法を提案しており，その正答率は 56.4%にものぼる．それ以外にも，与えられた語 A と B の関係を表すような言語パターンをコーパスから取得する手法についても考案されている[17)]．Bollegala ら[3)] は，LRA では 8 日以上かかっていた同問題 374 問を，Web からの言語パターン抽出と SVM による学習で，6 時間以下に短縮し高速化を行った．これらの研究は，すでに与えられた解候補の中から最も適したものを選択する問題に焦点を当てており，入力として A と B，C を与えて，D を検索する我々の手法とは，関係の類似に着目する面では共通するものの，根本的に異なるものである．

一方，入力として語 A，B，C を与え，A と B の関係が C と D の関係と類似するような語 D を発見する研究が大島ら[21)] によって始められている．これは語 A，B 間の言語パターンを抽出し，その言語パターンに対して語 C を適用することによって，高速に語 D に該当する語を発見することを目的としている．言語パターンを用いた手法では入力 A，B，C が与えられたときに，まず語 A と語 B を含む文書を検索する．次に，検索結果のタイトル，および，スニペットから B の前後にある言語パターンを取得する．得られた言語パターンの出現回数をもとに評価を行い，最適な言語パターンを決定する．最後に，最適な言語パターンと C を含むような文書を検索し，B の前に現れる言語パターンの後ろの文字列，かつ，B の後に現れる言語パターンの前の文字列を抽出する．両方で得られた文字列はその出現回数に応じて評価値が計算され，評価値順に出力される．

言語パターンにより文の意味に基づいて発見する方法に対して，我々は統計的に類似関係にある語を発見している．これは，精度の面で劣ると思われるが，明示的に表現されにくいような関係を特定することにおいては優れていると考えられる．この点を確認するために，提案手法と大島らによる手法の比較実験とその考察を 5 章において述べる．

関係自体を発見することを目的とした研究が，Luo ら[12)] によって行われている．これは与えられた 2 つのエンティティを結び付けるような語に高い重みを与え，2 つの関係がよく記述されているページが上位に来るようにソートすることにより，ユーザに 2 つのエンティティの関係を提示するシステムである．

我々が提案する関係に基づく情報検索に近い検索手法として，中島ら[22)] の相対的情報検索手法がある．相対的情報検索とは，サンプルデータ集合 S 中のデータ x を指定することにより，S における x の相対的関係を満たすようなターゲット集合 T 中のデータを検索するものである．相対検索はサンプルデータ集合 S の要素とターゲット集合 T の要素のマッピングで実現される．まず，データ集合 S のすべての要素と x の特徴ベクトルの差分を連結したものを $relative(x, S)$ と表現する．そして，T のある要素 t に対しても同様に $relative(t, T)$ を求め，これが最も $relative(x, S)$ と類似するような t を答えとして出力する．

本稿で提案する関係に基づく情報検索と，相対的情報検索との違いは 2 つある．1 つは，相対的情報検索がデータ集合全体とそれに属するデータの関係に基づくのに対して，我々は 2 つの独立したデータの関係に基づいて検索を行う点である．もう 1 つの相違点として，相対的関係が限定的であることがあげられる．中島らは相対的関係を特徴ベクトルの差分としてとらえている．相対的関係は，A は B よりも X という点で上回る，といったデータ間の差に基づくものであり，両者が敵対関係にある，原料と生成物の関係にある，などといった差分では表現できない関係を扱うことはできないと考えられる．本稿で扱う関係は，相対的関係のみでなく，あらゆる関係を含むものである．

2.2　特定の関係にある語の発見

入力された語と関連した語を発見する手法について，いくつかの提案がなされている．Church ら[4)] は相互情報量を用いて 2 語の関連度を測った．また，Turney [15)] や Baroni ら[1)] は与えられた 2 語が同義語であるかどうかを，Web 検索エンジンが検索結果として返す文書数で判定する方法を提案している．これらの手法は，語の共起や相互情報量を用いたものである．Bollegala ら[2)] は Web 検索結果のスニペットから，2 語の意味的類似度を判定する言語パターンを自動的に抽出し，これに加え，様々な類似度計算手法をサポートベクタマシンによって組み合わせ，頑強な意味的類似度の計算を行っている．

WordNet [13)] は人手で構築されたシソーラスである．しかし，これは固有名詞や新出語，また，語間の多様な関係に対応していないため，自動的にシソーラスを構築する方法が必要とされる．

テキストコーパスやデータセットから，特別な関係にある語を自動的に抽出する研究は数多く行われている．Hearst [8)] は上位語や下位語を発見するために，“such as” のような言語パターンに着目した．Bayesian Sets [7)] はベイズ推定を用いて，同位語を共起テーブルのような大規模データから発見するものである．このアルゴリズムは単純かつ高速であるが，EachMovie や Grolier encyclopedia のような大量のデータセットを必要とする．Lin [11)] は係り受け解析をした大量のテキストデータから類義語を発見する手法について提案している．

特別な関係にある語を発見する研究もいくつか行われている．Oyama ら[14] はある語の話題について詳細に述べている語を Web から取得した．これらの語の関係は一種の “part-of” 関係にあるといえる．Hokama ら[10] は人物の呼称を Web から抽出している．これは，人物の名称の前に現れる言語パターンを用いて候補となる言語表現を発見した後に，それが呼称であるかの評価を行っている．

このように，特定の関係にある語の抽出について，多くの研究が行われている．我々の手法は，入力 A と B により様々な関係を例示することによって，類似関係にある語を取得することができるため，これらの研究の一般化であるともいえる．

3. 関係に基づく情報検索

関係に基づく情報検索とは “データ A，B と C を入力として与え，A と B で成り立つ関係の集合を Relation，C とデータ D で成り立つ関係の集合を Relation′ としたとき，Relation と Relation′ の類似度が高い順にランク付けをした D を出力として返す検索” である．

これを情報検索の要素にあてはめると以下のような式で表すことができる．

$$\mathbf{Q} \ni q_i = (a_i, b_i, c_i) \tag{1}$$

$$\mathrm{Relation}(\mathrm{a_i}, \mathrm{b_i}) = \{R | R \in \mathcal{R} \wedge R(a_i, b_i)\} \tag{2}$$

$$\mathrm{Rank}(q_i, d_j) = \mathrm{Sim}(\mathrm{Relation}(a_i, b_i), \mathrm{Relation}(c_i, d_j)) \tag{3}$$

ただし，$\mathbf{Q}$ はクエリ集合，$\mathbf{D}$ は検索対象のデータ集合を表しており，$d_j \in \mathbf{D}$ である．Rank は $\mathbf{Q} \times \mathbf{D} \to \mathbb{R}$ の関数であり，クエリ q_i が与えられたときに，データ集合 $\mathbf{D}$ に対して順序をつける役割を果たしている．$\mathcal{R}$ は関係集合であり，Sim は関係集合の類似度を実数 $\mathbb{R}$ で返す関数である．この定義には，関係集合の類似度関数 Sim が含まれていないため，関係に基づく情報検索を実現するためにはこれらを決定する必要がある．

関係に基づく情報検索が有効に働くシナリオとして以下のものがあげられる．

- **未知の分野に対する検索を既知の分野の情報で検索したい場合**
  暖かで力強いクラッシック曲を探したいが音楽は詳しくない．しかし，絵画にはある程度の心得がある．そこで力強いタッチで描かれた，ゴッホの絵画に相当するような音楽を検索したい．この場合，A = 絵画，B = ゴッホ，C = 音楽である．
- **オブジェクト名が思い出せず，比喩でしか表現できない場合**
  あるハリウッド俳優の名前の名前を忘れてしまった．その俳優は，T 研究室でいえば O 先生にあたるような人である．そこで，T 研究室でいえば O 先生にあたるようなハリウッド俳優を検索したい．この場合，A =T 研究室，B =O 先生，C = ハリウッド俳優である．
- **すでに明確なイメージを持っているが，キーワードで表現することが困難な場合**
  生き生きとした自然の写真を検索したい．それは,「生き生き」というキーワードで表せるようなものでなく，まるで「ベートーヴェンの第九（喜びの歌）」のようなイメージである．そこで，第九のような風景画像を検索したい．この場合，A = ベートーヴェン，B = 第九，C = 風景画像である．

以上のように，関係に基づく情報検索は既存の検索の問題点を補い，特に，画像と音楽，音楽と画像など異種コンテンツを横断的に利用して検索することが可能であると考えられる．

4. 提 案 手 法

本稿では，関係に基づく情報検索のうち，入力 A，B，C および出力 D がオブジェクト名である場合について行う．我々は Web 文書中に出現する名詞をオブジェクト名としてとらえ，共起する語の出現分布の差分を利用した手法を用いる．2 章であげたいくつかの関連研究において用いられている言語パターンを，我々が利用しない理由は，関係に基づく情報検索は言語のみに依存するものでなく，マルチメディアへの応用を考えているためである．また，検索に利用することを考えた場合，学習が必要になるような手法は処理に時間がかかるため利用しにくい．さらに，あらかじめ用意された解候補の中から選択するのではなく，コーパス中から発見する必要がある．

そこで我々は語の共起を用いて，類似関係にあるオブジェクト名の検索を提案する．まず，語 A，B がともに出現する文書においてのみ，著しく出現確率が高くなる語を t とする．このときに，語 t と語 C がともに出現する文書において，$\mathrm{Sim}(\mathrm{Relation}(A, B), \mathrm{Relation}(C, D))$ が高くなるような語 D の出現確率が高くなることを仮定する．この仮定を用いた我々の提案手法は以下のとおりである．

4.1 入力 A と B を結び付ける語の発見

最初に，入力 A と入力 B を結び付けるような語集合 T を Web 検索を用いることによって取得する．ここで，A と B を結び付ける語は，A と B をともに含む文書において出現確率が高くなると仮定している．

(1) 入力 A および入力 B に対して，A を含み B を含まない文書を検索するクエリと，B を含み A を含まない文書を検索するクエリで Web 検索を行い，上位 n 件の検索結果のタイトルとスニペットを取得する．

(2)　入力 A と入力 B を含む文書を検索するクエリで Web 検索を行い，検索結果のタイトルとスニペットを取得する．

(3)　タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不要な語を除去する．

(4)　品詞が「名詞」であると推定された各語 t_i に対して，“語 A を含み語 B を含まない文書と語 A と B をともに含む文書において，語 t_i の出現確率が等しい” という帰無仮説に対して χ^2 検定を行う．同様に，語 B を含み語 A を含まない文書と語 A と B をともに含む文書に対しても検定を行う．

(5)　有意水準 α の検定において棄却された語 t_i のうち，語 A，B が出現したときに出現確率が高くなるものを，入力された語 A と B を結び付ける語として採用し，これを語集合 T とする．

各プロセスの詳細は以下のとおりである．

(1)，(2) 入力 A を含み入力 B を含まない文書の検索は，多くの Web 検索エンジンが採用している検索オプションを利用することによって，検索することができる．たとえば，この場合，“A AND -B” というクエリによって検索することが可能である．同様に，検索オプションを利用することによって，入力 A と入力 B を含む文書は “A AND B” というクエリで検索する．なお，実装では，Yahoo!検索 Web サービス[⋆1]を利用している．

本稿ではタイトル，スニペットのみを解析しているが，検索結果中の文書に直接アクセスし，文書全体を解析することも可能である．しかし，これを行うには多数の Web アクセスが必要であるため，実用的な速度を達成できないと考えられる．100 件の検索結果を用いる場合には，タイトルおよびスニペットだけを用いる場合の 50 倍ほどの時間がかかる．また，A と B を結び付ける語は，語 A および B の周辺のテキストに出現することが期待される．そのため，本手法ではクエリの周辺テキスト，すなわちスニペットの利用だけで十分であると考えられる．

(3) 形態素解析器には MeCab[⋆2]を用いた．また，ストップワードで除去するのは，全角記号や半角記号などである．また，複合名詞に対応するために，連続する名詞列をまとめて複合名詞としている．

取得する語を名詞だけに限定するのは，名詞以外の品詞はそれ単体だけでは，名詞に比べて具体的な意味を持たないためである．たとえば，A = 愛媛，B = みかん，C = 青森という入力で検索した場合，A と B を結び付ける語として，名詞以外では「有名な」,「買う」,「良い」などが得られる．しかし，これらは「産地」や「直送」といった名詞と比べて，一般的な意味を持った語が得られるため精度が低下する．よって，経験的ではあるが本稿では名詞のみに着目をする．

(4) 語 A と語 B をともに含む文書における語 t_i の出現確率 $P(t_i|A,B)$ は，語 t_i が出現するかしないかが母数 $P(t_i|A,B)$ の二項分布に従うものと考えた場合，最尤推定により以下のように求めることができる．

$$P(t_i|A,B) = \frac{N_{A,B}(t_i)}{N_{A,B}} \tag{4}$$

ここで，$N_{A,B}$ は検索によって得られた語 A，B を含む文書数であり，$N_{A,B}(t_i)$ は $N_{A,B}$ 文書のうち，語 t_i を含む文書数である．

“語 A を含み語 B を含まない文書と語 A と B をともに含む文書において，語 t_i の出現確率が等しい” という帰無仮説が正しいかどうかを判断することは，語 A を含み語 B を含まない文書においても語 t_i が，語 A，B を含む文書において推定された出現確率 $P(t_i|A,B)$ で出現しているかどうかを，χ^2 検定によって検定することで可能である．この場合，χ^2 検定値は以下のようにして求めることができる．

$$\chi^2 = \frac{(N_{A,\overline{B}}(t_i) - N_{A,\overline{B}}P(t_i|A,B))^2}{N_{A,\overline{B}}P(t_i|A,B)} + \frac{(\overline{N_{A,\overline{B}}(t_i)} - N_{A,\overline{B}}\overline{P(t_i|A,B)})^2}{N_{A,\overline{B}}\overline{P(t_i|A,B)}} \tag{5}$$

ここで，$N_{A,\overline{B}}$ は検索によって得られた語 A を含み B を含まない文書数であり，$N_{A,\overline{B}}(t_i)$ は $N_{A,\overline{B}}$ 文書のうち，語 t_i を含む文書数である．$\overline{N_{A,\overline{B}}(t_i)}$ は $N_{A,\overline{B}}$ 文書のうち，語 t_i を含まない文書数である．

(5) この場合の χ^2 検定値は自由度 1 の χ^2 分布に従う．有意水準 α で 2 つの仮説が棄却され，かつ，語 A，B が出現する文書での語 t_i の出現確率の方が，語 A を含み語 B を含まない文書と，語 A を含まず語 B を含む文書よりも高い場合，語 t_i を入力 A と入力 B を結び付けるような語として語集合 T に含める．

以上が，入力 A と入力 B を結び付けるような語集合 T を Web 検索を用いて取得する方法である．

4.2　入力 C に対する語 D の発見

次に，語集合 T を用いて，入力 A と B の関係と類似するような，入力 C に対する語 D

⋆1 http://developer.yahoo.co.jp/search/
⋆2 http://mecab.sourceforge.net/

を発見する．ここで，A と B の関係と類似する語は，A と B を結び付ける語 t_i と C をともに含む文書において出現確率が高くなると仮定している．

(6) 語集合 T のすべての語 t_i に対して，入力 C を含み語 t_i を含まない文書を検索するクエリと，t_i を含み C を含まない文書を検索するクエリで Web 検索を行い，上位 n 件の検索結果のタイトルとスニペットを取得する．

(7) 語集合 T のすべての語 t_i に対して，入力 C と語 t_i を含む文書を検索するクエリで Web 検索を行い，検索結果のタイトルとスニペットを取得する．

(8) タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，ストップワードリストを用いて不必要な語を除去する．

(9) 品詞が「名詞」であると推定された各語 d_j に対して，“語 C を含み語 t_i を含まない文書と語 C と t_i をともに含む文書において，語 d_j の出現確率が等しい” という帰無仮説に対して χ^2 検定を行う．同様に，語 t_i を含み語 C を含まない文書と語 C と t_i をともに含む文書に対しても検定を行う．両者の検定の結果，帰無仮説が発生する確率をそれぞれ $P_C(d_j)$，$P_{t_i}(d_j)$ とする．

(10) 有意水準 β の検定において棄却された語 d_j のうち，語 C，t_i が出現したときに出現確率が高くなるものに対して，$P_C(d_j)$，$P_{t_i}(d_j)$ の積を $P_{C,t_i}(d_j)$ とする．

(11) 語集合 T のすべての語 t_i に対する，$P_{C,t_i}(d_j)$ のすべての積を語 d_j のスコアとする．

各プロセスの詳細は以下のとおりである．

(6)，(7)，(8) この手法は，(1)，(2)，(3) において，A を C，B を t_i に置き換えたものと同等である．ただし，語集合 T に含まれるすべての語 t_i に対して行う．

(9) “語 C を含み語 t_i を含まない文書と語 C と t_i をともに含む文書において，語 d_j の出現確率が等しい” という帰無仮説に対して，その帰無仮説が発生する確率 $P_C(d_j)$ は，(5) で示した χ^2 検定値を用いると以下のようになる．

$$P_C(d_j) = \frac{1}{\sqrt{2}\Gamma(1/2)} \int_{\chi^2}^{\infty} x^{-1/2} e^{-x/2} dx \tag{6}$$

また，$P_{t_i}(d_j)$ も同様にして求めることができる．

(10) この場合の χ^2 検定値は自由度 1 の χ^2 分布に従う．有意水準 β で 2 つの仮説が棄却され，かつ，語 C，t_i が出現する文書での語 d_j の出現確率の方が，語 C を含み語 t_i を含まない文書と，語 C を含まず語 t_i を含む文書よりも高い場合，$P_{C,t_i}(d_j) = P_C(d_j)P_{t_i}(d_j)$，そうでない場合，最小値を与える．すなわち，$P_{C,t_i}(d_j) = 1$ とする．

(11) 最終的に語 d_j のスコア $\mathrm{Score}(d_j)$ は語集合 T のすべての語 t_i に対する $P_{C,t_i}(d_j)$ の積としている．

以上が語集合 T を用いて，入力 A と B の関係と類似するような，入力 C に対する語 D を発見する方法である．$\mathrm{Score}(d_j)$ の値が小さいほど，条件を満たすような語 D である可能性が高い．ただし，$\mathrm{Score}(d_j)$ は計算機で扱ううえでは非常に小さな値となるため，実際には $-\log_{10}\mathrm{Score}(d_j)$ をスコアとして用いるものとする．このスコアを用いて，クエリ $q_i = (a_i, b_i, c_i)$ が与えられたとき，本手法における関係の類似度関数 Sim およびランク関数 Rank は以下のように定める．

$$\mathrm{Rank}(q_i, d_j) = \mathrm{Sim}(\mathrm{Relation}(a_i, b_i), \mathrm{Relation}(c_i, d_j)) = -\log_{10}\mathrm{Score}(d_j) \tag{7}$$

5. 実　　験

提案手法の有効性を実証するために，我々は 854 の正解セットを用いて提案手法の性能評価を行った．性能評価を行う前に，854 の正解セットの各関係クラスから 2 つずつランダムに抽出して 66 の正解セットを作成し，これを用いてパラメータの決定および取得件数の増加にともなう精度の変化について事前実験を行った．決定すべきパラメータは，提案手法で用いられた χ^2 検定の有意水準 α および β である．実験では入力 A，B，C を与えた場合に，正解出力 D をどれくらいの精度で発見できるかを評価した．

5.1 正解セット

我々は，入力 A，B，C と期待される出力 D の組で構成された 854 の正解セットを人手により作成した．すべての正解セットクラスを**表 1** に示す．表のカラムはそれぞれ，クラス ID，X，Y の種類，X と Y の関係の説明，1 位に正解が得られた例，そのクラスに含まれる正解セット数を表している．各クラスには X と Y のペアの集合が含まれており，我々は同じクラスに属している任意の 2 組を選び，一方を A と B，もう一方を C と D として割り振り，これを正解セットとしている．同じクラスから取り出すということは，A と C，また B と D は同じ種類のものに限定されている．たとえば，クラス 3 はワインとブドウ，ヨーグルトと牛乳，パンと小麦粉などのペアが含まれており，これらから正解セット A = ワイン，B = ブドウ，C = ヨーグルト，D = 牛乳や，A = ヨーグルト，B = 牛乳，C = パン，D = 小麦粉などの組合せを作ることができる．そして，その組合せの総数が 854 である．正解セットの一部を**表 2** に示す．ただし，正解セットには答えが一意に定まるという制約を設けている．

表 1　正解セットクラス（∗：1 位に正解が得られなかった）
Table 1　Test set classes (∗: means that a correct answer was not obtained at the top).

ID	X の種類	Y の種類	関係の説明	1 位に正解が得られた例 ($A:B::C:D$)	#
1	地域	特産物	Y は X の特産物	愛媛:みかん::静岡:お茶	20
2	国	国技	Y は X の国技	タイ:ムエタイ::日本:相撲	90
3	製品	原材料	Y は X の原材料	チョコレート:カカオ::パン::小麦粉	82
4	地域	お菓子	Y は X で有名なお菓子	フランス:クグロフ::トルコ:ロクム	88
5	地域	名物	Y は X の名物	広島:お好み焼き::香川:うどん	54
6	国	原住民	Y は X の原住民	オーストラリア:アボリジニ::アメリカ:インディアン	12
7	国	首都	Y は X の首都	ドイツ:ベルリン::イギリス:ロンドン	56
8	スポーツ	スポーツ用品	Y は X で使われる	∗	12
9	成虫	幼虫	Y は X の幼虫	カエル:オタマジャクシ::蚊:ボウフラ	12
10	動物	綱	X は Y に属する	クジラ:哺乳類::ヘビ:爬虫類	10
11	日本文学	著者	X は Y の作品	舞姫:森鴎外::こころ:夏目漱石	12
12	寺	建立者	X は Y に建てられた	久遠寺:日蓮::飛鳥寺:蘇我馬子	30
13	衛星	惑星	X は Y の衛星	フォボス:火星::タイタン:土星	12
14	魚の卵	魚	X は Y の卵	キャビア:チョウザメ::数の子:ニシン	12
15	宗教	神	Y は X の神	∗	12
16	本	社会学者	Y は X の著者	資本論:マルクス::法の精神:モンテスキュー	20
17	僧	仏教の宗派	X は Y の開祖	最澄:天台宗::空海:真言宗	30
18	地域	温泉	Y は X で有名な温泉	∗	30
19	世界遺産	国	X は Y の世界遺産	ボロブドゥール:インドネシア::ピラミッド:エジプト	30
20	県	花	Y は X の県の花	兵庫:ノジギク::熊本:リンドウ	30
21	幕府	将軍	Y は X を開いた	鎌倉幕府:源頼朝::江戸幕府:徳川家康	6
22	世界の作家	文学	X の作品は Y に属する	∗	12
23	ディズニーキャラクター	モチーフ	X は Y がモチーフ	∗	20
24	世界の島	国	X は Y の島	∗	30
25	パスタ	ソース	X は Y のソース	アラビアータ:トマト::ペペロンチーノ:オイル	12
26	神社	日本の神	Y は X に祀られている	八坂神社:スサノオ::伊勢神宮:天照大神	12
27	地域	織物	Y は X で有名な織物	秩父:秩父銘仙::京都:西陣織	20
28	花	花言葉	X の花言葉は Y	スミレ:忠実::バラ:愛	12
29	国	通貨	Y は X の通貨	韓国:ウォン::中国:元	20
30	魚	魚	X は成長すると Y になる (出世魚)	ワカシ:ブリ::コッパ:スズキ	12
31	楽器	楽器のクラス	X は Y に属する	∗	20
32	スープ	主原料	X の主原料は Y	∗	12
33	会社	社長	X の社長は Y	Mixi:笠原健治::KDDI:小野寺正	12

この正解セットは，第 1 著者とコンピュータサイエンスを専門としていない者 1 名，合計 2 名によって製作された．関係クラス ID1 から 14 が第 1 著者，ID15 から 33 までがもう 1 名によって構築されている．正解セットを構築するにあたって，我々は以下の点を前もって被験者に伝えてある．

(1)　同じ関係クラスに属するすべてのペア (X, Y) が類似関係であり，またそれらが一般的な関係であること．
(2)　類似関係を考慮した場合，語 X に対して語 Y が一意に定まるようなものであること．
(3)　各関係クラスにおいて 2 個から 10 個の類似関係ペアをあげること．

表 2　正解セットの例
Table 2　Examples of test sets.

A（入力）	B（入力）	C（入力）	D（正解出力）
千葉	落花生	静岡	茶
日本	相撲	韓国	テコンドー
チョコレート	カカオ	かまぼこ	魚
長崎	カステラ	京都	八ツ橋
名古屋	手羽先	広島	お好み焼き
フランス	パリ	カナダ	オタワ
野球	バット	ゴルフ	クラブ
チョウ	イモムシ	トンボ	ヤゴ
クジラ	哺乳類	カエル	両生類
舞姫	森鴎外	人間失格	太宰治
金閣寺	足利義満	銀閣寺	足利義政
江戸幕府	徳川家康	鎌倉幕府	源頼朝
月	地球	エウロパ	木星
キャビア	チョウザメ	とびこ	トビウオ

（1）は特殊すぎる関係であった場合，評価が困難であり，一般性に欠け，実験の正確さを欠くためである．（2）は正解が複数あるような正解セットを用いた場合，正解をすべてを列挙することが困難であるためである．また，（3）に関しては，1 個の場合は正解セットが生成できず，10 個より多い場合は組合せ数が多くなりすぎるためである．

5.2　パラメータ決定

提案手法の性能を評価する前に，正解セットの各関係クラスから 2 つずつランダムに抽出した 66 組の正解セットを用いて，パラメータの決定を行った．パラメータを決定するための評価指標として，順位付きの検索結果の評価に対して用いられる平均逆順位（MRR）を採用した．これは各課題において最初に正解が現れた順位の逆数を，すべての課題で平均したものであり，式 (8) で表される．

$$\mathrm{MRR} = \frac{1}{N}\sum_{k=1}^{N} rr_k \tag{8}$$

ただし，rr_k は k 番目の課題において最初に正解が現れた順位の逆数であり，正解がなければ 0 である．また，N は総課題数であり，評価値 MRR の最大値は 1，最小値は 0 となっている．我々が上位数件の正解率を主な評価指標として用いず，MRR を用いたのは，出現順位に重みをつけて評価することで，より上位に正解が現れることを重視したためである．

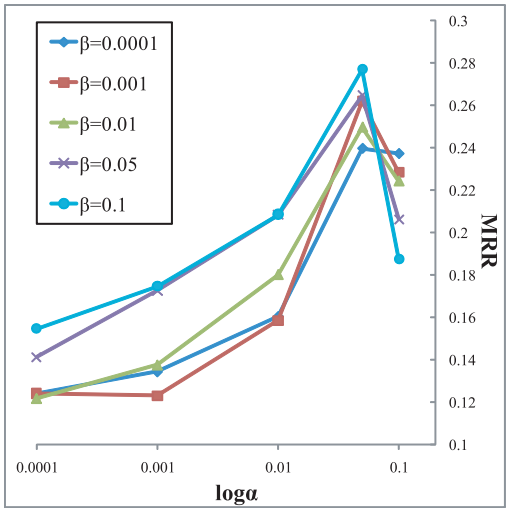

図 1　各パラメータごとの MRR
Fig. 1　MRR scores for each parameter.

実験で用いられるパラメータは χ^2 検定に用いられる有意水準 α および β であり，これらに対して，0.0001，0.001，0.01，0.05，0.1 の値を割り当てて，それぞれの組合せで実験を行った．正解セットのうち，入力 A，B，C を用いて出力 D に順位をつけ，66 セットの正解セットで MRR をとったものを**図 1** に，また，各パラメータごとの，上位 20 件の検索結果に正解を含む割合を**図 2** に，有意水準 α ごとの実行時間を**図 3** に示す．

図 1 の横軸は仮説検定で使用するパラメータ α の値を，縦軸は MRR の値を示しており，また，β の値ごとにグラフを表示している．平均逆順位のスコアでは，$\alpha = 0.05$ かつ $\beta = 0.1$ のときが最大の値を示している．すべての各 β の値において，$\alpha = 0.05$ のときに最大となっており，$\alpha = 0.1$ ではそのスコアを低下させている．また，$\alpha = 0.1$ 以外の各 β の値において，$\beta = 0.1$ のときに最大となっている．

上位 20 件の正解率を考えた場合，最大の正解率（54.5%）を示すのは，$\alpha = 0.05$ かつ $\beta = 0.1$ および $\alpha = 0.1$ かつ $\beta = 0.01$，$\beta = 0.05$ の 3 種類のパラメータである．α の値が増加するにつれて正解率は向上するが，$\alpha = 0.1$，$\beta = 0.1$ の場合だけは正解率が低下している．α の値が大きくなれば得られる語集合 T の大きさが増加するため，様々なクエリによる Web 文書検索が行われそれらが集約される．そのため，得られる答えの種類が増え，

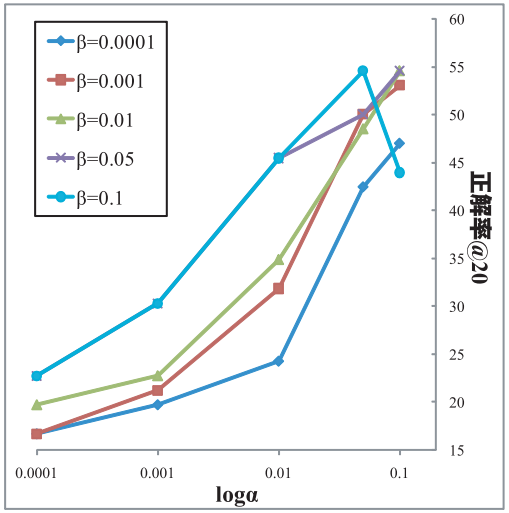

図 2　各パラメータごとの上位 20 件正解率
Fig. 2　Percentage of relevant search results in top 20 for each parameter.

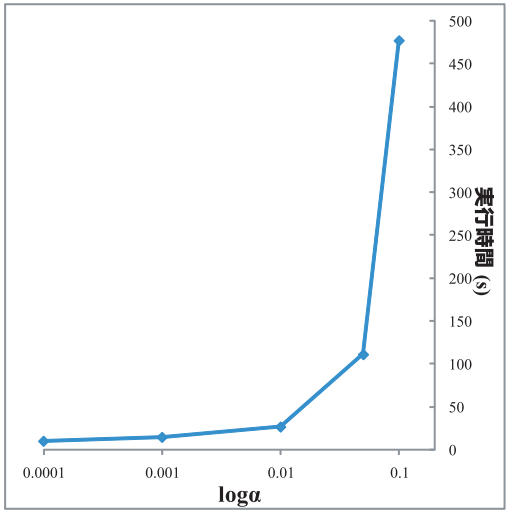

図 3　各パラメータごとの実行時間の平均（単位は秒）
Fig. 3　Average runtime for each α.

上位 20 件以内に正解を含められる割合が高まると推測される．しかし，$\alpha = 0.1$，$\beta = 0.1$ ではノイズの増加によって，正解が下位に順位付けされ，MRR のスコアが低下していると考えられる．

精度のみを考えた場合，最適なパラメータの組合せは $\alpha = 0.05$ および $\beta = 0.1$ である．しかし，実時間で実行できるという条件を与えた場合，最適なパラメータは異なってくる．本手法は語集合 T の数の 3 倍の回数，Web 検索エンジンにアクセスする必要がある．パラメータ α を大きくすることで，語集合 T の数は増加していく．そのため，図 3 に示すようにパラメータ α の値が大きくなるにつれ，実行時間は指数的に増加する．実行時間は環境によって異なるが，60 秒を超えるようなパラメータは実時間で実行できるとはいい難い．したがって，採用するパラメータ α は 0.0001，0.001，0.01 のいずれかである．

精度および速度の点を考慮し，我々は $\alpha = 0.01$ かつ $\beta = 0.1$ を最適なパラメータとして採用し，以下の実験ではこの値を用いるものとする．

5.3　比 較 手 法

我々の提案手法を相対的に評価するために，提案手法と同等の入出力をとる 2 つの手法をここで紹介する．1 つ目は tf-idf 法を用いたベースライン手法であり，以下のようなものである．

(1)　入力 A と入力 B を含む文書と入力 C を含む文書を Web 検索エンジンを用いて検索し，上位 n 件の検索結果のタイトルとスニペットを取得する．ここで，前者で得られる文書集合を D，後者を D' とする．

(2)　タイトルとスニペットに対して形態素解析を行い，形態素ごとに分割する．また，形態素は名詞だけに限定し，ストップワードリストを用いて不要な語を除去する．連続する名詞列は複合名詞として扱っている．

(3)　文書集合 D に含まれるそれぞれの文書に対して tf-idf ベクトルを作成する．ここで，$tf_{i,j}$ は文書 d_j に含まれる単語 t_i の数であり，df_i は単語 t_i を含む文書数，そして，$idf_i = \log(|D|/df_i)$ である．同様にして，D' に対しても tf-idf ベクトルを作成する．

(4)　文書集合 D' に含まれるそれぞれの文書に対して，文書集合 D に含まれる文書との類似度を計算し，またその数を足し合わせ，D' の中で最大値をとるものを d' とする．類似度にはベクトル間のコサイン類似度を用いている．

(5)　文書 d' に含まれる語を tf-idf のスコアに基づいてランキングし，出力する．すなわち，$Rank(q, d_i) = tf_{i,j} idf_i$ である．ただしここでは，クエリ $q = (A, B, C)$ であり，単語 d_i は文書 d' に含まれる単語である．

この基準となる手法は関係の類似性を計るために tf-idf 法を用いている．文書集合 D との類似度の総和が関係の類似度を表している．

2 つ目は言語パターンを用いた手法である．これには 2 章で述べた大島らの手法を採用した．

(1) 「$A \wedge B$」をクエリとして Web 検索を行い，1000 件の検索結果を取得する．
(2) 得られた文書群から B の直後に現れる文字列を取得し，それらの出現回数を数える．出現回数と文字列評価により得られた文字列のスコア付けを行い，最も高いスコアを得た文字列を $\text{Patterns}^{\text{Pre}}$ とする．
(3) 同様にして，B の直前に現れる文字列のうち最も高いスコアを得たものを $\text{Patterns}^{\text{Post}}$ とする．
(4) $\text{Patterns}^{\text{Pre}}$ に文字列 A を含めば，これを C に置き換えた文字列を $\text{webQuery}^{\text{Pre}}$ とする．A を含まなければ，「$C \wedge \text{Patterns}^{\text{Pre}}$」を $\text{webQuery}^{\text{Pre}}$ とする．
(5) 同様に $\text{webQuery}^{\text{Post}}$ を決定する．
(6) $\text{webQuery}^{\text{Pre}}$ で Web 検索を行い，上位 n 件のタイトルとスニペットから $\text{Patterns}^{\text{Pre}}$ の直前に現れる文字列を取得する．
(7) 同様にして $\text{webQuery}^{\text{Post}}$ を用いて検索し，$\text{Patterns}^{\text{Post}}$ の直後に現れる文字列を取得する．
(8) 得られた文字列を出現回数に応じてスコアリングし，得られた各文字列を順位付けを行って出力する．

スコアリング手法，細かなパラメータなどについては，大島らの論文に基づいて設定している．

5.4　最適な Web 検索結果取得件数の決定

次に提案手法で取得する Web 検索結果の取得件数を変化させ，その精度と実行時間の変化を調べる．また，比較のためにベースライン手法と言語パターンを用いた手法でも取得件数を変化させ，最適な Web 検索結果数を決定する．利用した正解セットはパラメータ決定で用いたもので，それぞれの手法で変化させる取得件数パラメータ n は以下のとおりである．

tf-idf を用いたベースライン手法

クエリ「$A \wedge B$」と「C」の Web 検索検索結果件数

言語パターンを用いた手法

クエリ $\text{webQuery}^{\text{Pre}}$ と $\text{webQuery}^{\text{Post}}$ の Web 検索結果件数

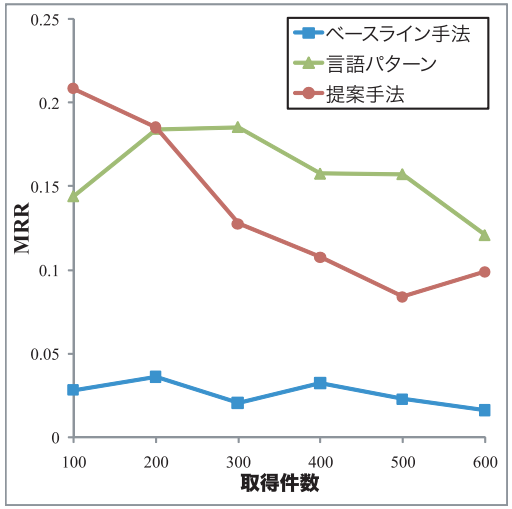

図 4　各手法ごとの MRR
Fig. 4　MRR scores for each method.

提案手法

入力 A と B を結び付ける語の発見，および，入力 C に対する語 D の発見での Web 検索結果件数

取得する Web 検索結果の件数 n を 100 から 600 まで変化させたときの，それぞれの MRR を**図 4** に，上位 20 件の正解率を**図 5**，実行時間を**図 6** に示す．

MRR と上位 20 件に正解が得られた正解セットの割合は，ある程度の取得件数を境にして減少していくのが分かる．これは取得件数の増加とともに得られる答えにノイズが増え，本来正解であるものが下位に順位付けされていくためであると思われる．

ベースライン手法は取得結果数 $n = 200$ において MRR が最大，$n = 100$ において上位 20 件に正解が得られた正解セットの割合が最大になっている．しかし，両方を考慮した結果，どちらの値も最大値に近い $n = 400$ を採用した．

言語パターンを用いた手法は取得結果数 $n = 300$ において MRR と上位 20 件に正解が得られた正解セットの割合が最大となっている．

得られる件数が増えるにつれて，提案手法は統計的に精度が改善されると考えられるが，実際は取得結果数 $n = 100$ 以降は低下し続けている．また，実行時間は取得件数が増える

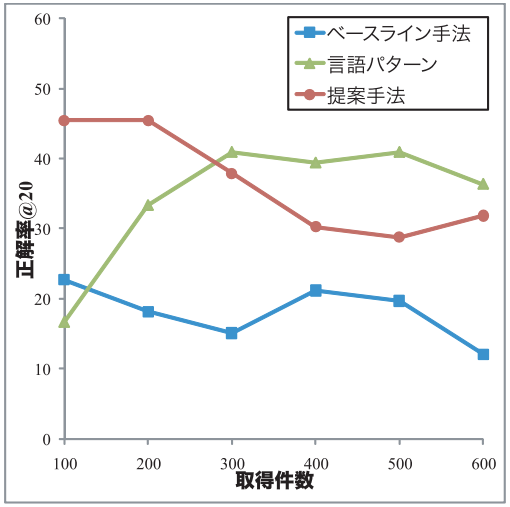

図 5　各手法ごとの上位 20 件正解率
Fig. 5　Percentage of relevant search results in top 20 for each method.

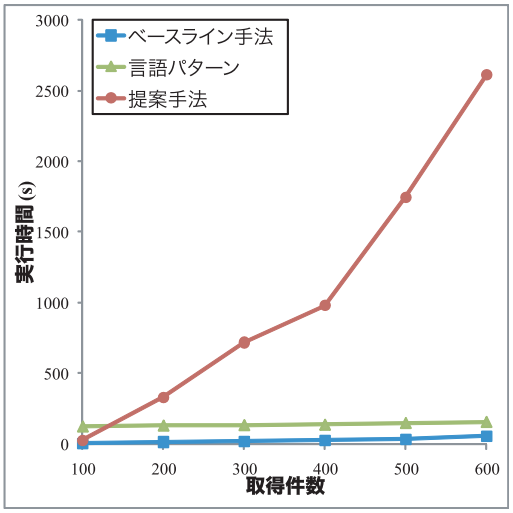

図 6　各手法ごとの実行時間の平均（単位は秒）
Fig. 6　Average runtime for each method.

表 3　クラス平均値
Table 3　Average scores for each test class.

	MRR	@5	@10	@20	実行時間 (s)
ベースライン	0.045	8.3	10.8	18.7	31.9
言語パターン	0.150	24.5	30.9	33.8	94.2
提案手法	**0.249**	**34.4**	**42.1**	**49.8**	**79.8**

につれて急激に増加していくことが図 6 から見てとれる．提案手法で取得件数を増加させた場合，語集合 T が大きくなっていくため Web 検索の回数が増大する．そのため，単純に実行時間が取得件数に比例しないのである．また，この実行時間の急激な増加および精度の低下は，パラメータ α を増加させたときと似たような傾向を示している．

提案手法が取得結果数 $n = 100$ 以降低下し続けるのは，$n = 100$ で最適なパラメータを決定したことが 1 つの原因であると考えられる．Web 検索結果の取得件数を増加させて，提案手法の精度を向上させるためには，有意水準のパラメータ α および β を取得件数に応じて変化させる必要がある．少ない検索結果数であった場合，これらのパラメータを大きくしなければ，語の出現頻度に有意な差が発見されず，正解が得られない場合がある．しかし，一方で大量の検索結果数を用いた場合，有意水準を小さくしてノイズを除かなければ，正解とノイズのスコア間に差がなくなり精度が低下しまうと考えられる．

最終的な比較実験では，ベースライン手法は $n = 400$，言語パターンを用いた手法は $n = 300$，そして提案手法は $n = 100$ として行うことにする．

5.5　各手法との比較実験

決定されたパラメータ，および，取得件数によって，より大きな正解セットでの実験を行い，各評価値および各手法，各関係クラスごとの特性を明らかにする．全 854 の正解セットを用いた実験結果を**表 3** に，各クラスごとの MRR を**図 7** に示す．

@5，@10，@20 はそれぞれ 5，10，20 位に正解を含めることができたテストの割合を，実行時間は 1 つの課題にかかる時間を表している．これらの値を各クラスごとに平均し，全クラスでの平均を示している．提案手法と言語パターン，ベースライン手法を比較したとき，MRR，@5，@10，@20 において顕著な差が出ている．出力として期待される D を上位 20 件以内に含めることができた正解セットは全体の中の 49.8%で，これはベースライン手法よりもはるかに優れているが，まだ精度的に不十分であると思われる．我々の手法は，約 50 回の Web アクセスを必要としており，並列処理を行うことである程度，実用的な速度で処理をすることができると考えられる．また，言語パターンと比較した場合もパターン

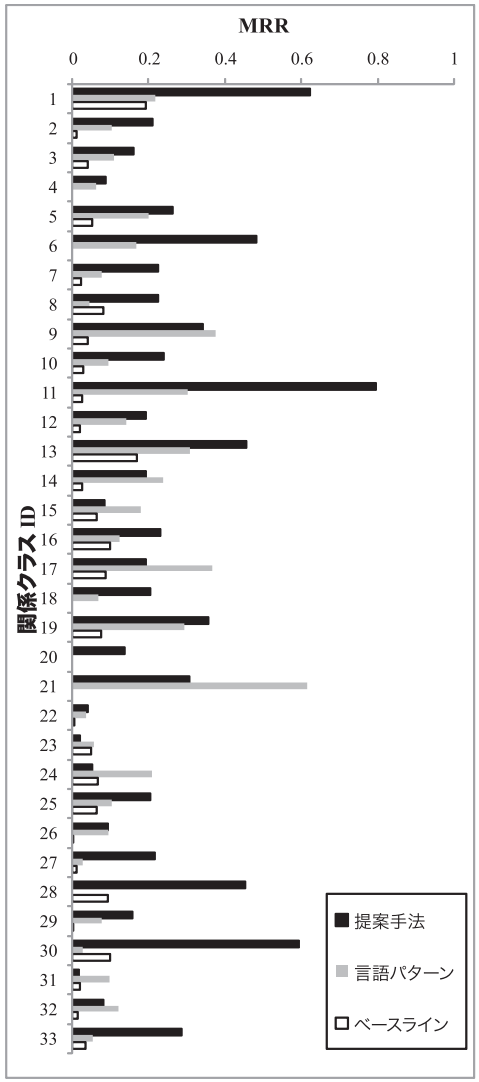

図 7　各クラスごとの MRR
Fig. 7　MRR score for each class.

抽出に時間が必要となる分，我々の提案手法の方が高速であることが分かる．

各クラスごとの MRR を示した図 7 において，いくつかのクラスは高い MRR を示しているが，あまり良い結果が得られていないものもある．MRR の値が高いクラス，1，11，30，すなわち，日本の特産物，日本文学，出世魚のクラスは，すべて日本に特有の話題である．これらの話題は日本語 Web ページに現れやすく，比較的詳しい情報が得られたため，成功しているのではないかと考えられる．

一方，クラス 22，23，24，31，32 などは提案手法において低い MRR スコアを示している．クラス ID 24，31，すなわち，世界の島と楽器は，Web ページ作成者が言及するまでもないような，あまりにも一般的すぎる関係であるため，Web 文書中にそれらに関する記述が少なく，期待される答えを発見できなかったのではないかと考えられる．一方，言語パターンはこのような関係に対してもある程度取得できている．

また，クラス 32，すなわち，スープとその原料では，入力 A と B を結び付けるような語集合 T の要素を名詞に限定してしまったことが，精度低下の原因であると考えられる．これらの関係は名詞で表現するのが困難であり，そのために，十分な結果が得られていないと思われる．

ベースライン手法と比べた場合，我々の手法は 6，9，28 などのクラス，原住民，成虫と幼虫，花言葉において大きく上回ったスコアを得ている．ベースライン手法は，D に相当する語を C で検索した結果の中からしか探すことができない．つまり，期待される出力 D が語 C と共起確率が低い場合には，ベースライン手法は有効に働かないと考えられる．そこで，共起度を測るために，クラス ID 6，9，28 における C と D の Jaccard 係数を Web 検索エンジンのヒットカウントを利用して計算した．その結果，これらの Jaccard 係数は 0.015 以下であり，非常に低い共起度であることが分かった．このように，C と D の共起度が低く，取得する文書数が限られている場合は，C による検索結果に D が含まれず，性能が大きく低下しているのではないかと思われる．一方で，提案手法は C のみではなく，語集合 T 中の語を含む文書を対象としているため，C と D の共起度が低い場合でもうまく働くのである．

提案手法と言語パターンの MRR 値で差が生じているのは，クラス ID 1，28，30，および 21，24 である．クラス 1，28，30，すなわち，特産物，花言葉，出世魚は提案手法が高い MRR を得ているが，言語パターンでは低い MRR を得ているクラスである．これらのクラスの関係は言語によって明示的に表されにくいクラスであるため，このような差が生まれているのだと考えられる．一方，クラス 21，24，すなわち，幕府と将軍，世界の島は言

表 4　入力：秋田，きりたんぽ，山梨に対する語集合 T
Table 4　Term set T for input Akita, Kiritampo, Yamanashi.

語	A	B	A, B
販売	4	5	32
稲庭うどん	0	1	18
郷土	3	9	23
特産品	1	0	11
あきたこまち	1	2	10
名産品	0	1	8
当店	0	0	7
本物	0	0	7
厳選	0	0	7
通販	0	0	7

表 5　入力：秋田，きりたんぽ，山梨に対する検索結果
Table 5　Search result for input Akita, Kiritampo, Yamanashi.

語	スコア
ほうとう	**123.1**
桃	97.6
ぶどう	95.3
直送	67.7
群馬	64.1

語パターンの方が優れていたクラスである．2 つのクラスに明確な共通点はないが，クラス 24 は「フィリピンのルソン島」のように従属関係が暗に助詞によって表現されることが多い．そのため，提案手法では検索に失敗し，言語パターンでは良い結果が得られているのだと考えられる．

5.6 結　果　例

クラス 5，名物，から生成されたテストの 1 つである，入力 A = 秋田，B = きりたんぽ，C = 山梨，期待される出力 D = ほうとうの結果について考察を行う．**表 4** はこれらの入力が与えたときに得られた語集合 T であり，項目 A は “秋田” を含み “きりたんぽ” を含まない文書での語の出現回数で，項目 B はその逆の条件であり，項目 A，B は “秋田” と “きりたんぽ” を含む文書での出現回数である．また，**表 5** は得られた結果上位 5 件とそのスコアである．

この例は成功例であり，“秋田” と “きりたんぽ” の関係と類似した関係を持つ “山梨” に対するものとして，1 位に “ほうとう” があげられている．これらの関係は，前者の地域特有の郷土料理が後者であるという関係である．“秋田” と “きりたんぽ” を結び付けるような語として，表 4 の語が得られており，特に “販売”，“特産品”，“郷土” などの単語が “山梨” と “ほうとう” をつないでいると考えられる．しかし一方で，“あきたこまち” や “稲庭うどん” という “山梨” と “ほうとう” を結ばないような単語が含まれている．これは，“秋田” と “きりたんぽ” が出現する文書に高確率で含まれるような単語であるが，“秋田” と “きりたんぽ” に固有の単語であるため，今回提案した手法ではノイズとなってしまっている．

表 6　入力：クジラ，哺乳類，カエルに対する語集合 T
Table 6　Term set T for input whale, mammal, frog.

語	A	B	A, B
イルカ	9	1	38
ジュゴン	0	1	10
アザラシ	0	0	9
呼吸	0	0	8
鯨類	1	0	8
水中	0	1	8
適応	0	0	7
肺	0	0	7
昔	0	0	7

表 7　入力：クジラ，哺乳類，カエルに対する検索結果
Table 7　Search result for input whale, mammal, frog.

語	スコア
皮膚	127.0
呼吸	94.2
両生類	**73.4**
肺	72.0
水中	66.8

表 6 は入力 A = クジラ，B = 哺乳類，C = カエルを与えたときに得られた語集合 T であり，項目 A は “クジラ” を含み “哺乳類” を含まない文書での語の出現回数で，項目 B はその逆の条件であり，項目 A，B は “クジラ” と “哺乳類” を含む文書での出現回数である．また，**表 7** は得られた結果上位 5 件とそのスコアである．このテストはクラス 10 から生成されたものの一例である．

“クジラ” と “哺乳類”，“カエル” と “両生類” の関係は，後者が前者の上位概念であって，綱という生物の分類における階級に属している，という関係である．“クジラ” と “哺乳類” を結び付ける語集合のうちに上位概念を明確に示唆するような語は見られないが，“呼吸”，

表 8　入力：こころ，夏目漱石，人間失格に対する語集合 T
Table 8　Term set T for input *Kokoro*, *Soseki Natsume*, *No Longer Human*.

語	A	B	A, B
先生	1	3	24
まなざし	0	0	11
私	5	7	18
新潮文庫	0	1	9
特別展	0	0	8
教科書	0	0	7
遺書	0	0	7
江戸東京博物館	0	0	6

表 9　入力：こころ，夏目漱石，人間失格に対する検索結果
Table 9　Search result for input *Kokoro*, *Soseki Natsume*, *No Longer Human*.

語	スコア
メロス	70.4
愛欲	59.0
斜陽	52.9
炎	45.5
自伝	40.5

“適応”，“肺” などといった生物学的な語が出現する文書に “両生類” という言葉が頻出したため，成功していると考えられる．

適切な出力 D が得られなかった例として，作品とその著者の関係である，“こころ” と “夏目漱石”，“人間失格” と “太宰治” という関係があげられる．**表 8** は得られた語集合 T であり，項目 A は “こころ” を含み “夏目漱石” を含まない文書での語の出現回数で，項目 B はその逆の条件であり，項目 A, B は “こころ” と “夏目漱石” を含む文書での出現回数である．また，**表 9** は得られた結果上位 5 件とそのスコアである．この例はクラス 11 から生成されたテストの 1 つである．

“こころ” と “夏目漱石” を結び付ける語集合に含まれる “先生” と “私” という単語は，“こころ” の登場人物である．これらの単語は “こころ” に固有のものであるため，“人間失格” とこれらの語が出現するような文書には，“太宰治” が出現する確率が高くならないと考えられる．“新潮文庫”，“教科書” などの語は “人間失格” から “太宰治” を発見するのに有効な語であると考えられるが，この例の場合，“人間失格” という語が現れる文書にはほとんど “太宰治” という語が出現するため，共起の差を用いた本手法では求めることができない．

我々は，関係に基づく情報検索手法の 1 つの実装例として，今回述べた手法に関する評価を行ったが，実際に用いるためには精度の点において大きな改善が必要である．本稿では，Web 検索エンジンの検索結果のうち上位 100 件しか用いていなかったが，統計に基づく本手法において，使用する検索結果件数を増やし適切なパラメータを用いれば，速度の低下と引き替えにその精度を改善できると考えている．

6. まとめと今後の課題

本稿では，入力として A，B，C が与えられた場合，A と B の関係 R と，C と D の関係 R' が類似するような D を検索する，関係に基づく情報検索手法を提案した．また，それを実現するための 1 つの手法として，Web 文書中の単語の共起を用いるものについて述べ，各種パラメータごとにその精度を評価した．共起を用いる手法は，入力 A，B をつなぐような語集合 T を求め，それと入力 C を用いたときに出現確率が高くなるような語に，その出現確率に応じてスコアを付けるものである．そして，スコアに基づいて順位付けを行い，A と B の関係 R と，C と D の関係 R' が類似する可能性が高い順に，D を出力する．

実験では，最も良いパラメータを設定した場合，854 のテストセットのうち，49.8%に対して上位 20 件以内に正解を得ることに成功した．しかし，共起による関係に基づく情報検索の多くの問題点が明らかとなる結果となった．1 つ目は，語に注目しただけでは，多様かつ複雑な関係を十分に表現することができないという点である．2 つ目は，入力 A と B を結び付けるような語集合 T が，入力 C に対する D を結び付けるような語であるとは限らないという点である．そして，3 つ目は，共起の差を用いた本手法では，語 C と求めたい語 D が頻繁に共起する場合，D を発見することが困難であるという点である．これらの点は，関係に基づく情報検索手法を実現するうえでの課題として，取り組んでいきたいと考えている．

また，我々はマルチメディアへのアナロジの適用にも関心がある．Heidorn[9] は画像について説明をするうえで，アナロジが重要な役割を果たしていることについて言及しており，また，画像をアナロジを用いて検索することを可能にする ABI[5] が提案されている．今回は，語を入力として語を検索するものであったが，関係に基づく情報検索は，様々な入力と出力に対して適用可能である．たとえば，語を入力として画像を検索したり，画像を入力として画像を検索したりすることも可能であると考えられる．

謝辞　本研究の一部は，京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」，文科省科研費（特定領域研究）計画研究「情報爆発時代に対応するコ

ンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041），および，文科省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築：異メディア・アーカイブの横断的検索・統合ソフトウェア開発」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

参 考 文 献

1) Baroni, M. and Bisi, S.: Using cooccurrence statistics and the web to discover synonyms in a technical language, *Proc. 4th International Conference on Language Resources and Evaluation* (*LREC 2004*), pp.1725–1728 (2004).
2) Bollegala, D., Matsuo, Y. and Ishizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engines, *Proc. 16th International World Wide Web Conference* (*WWW 2007*), pp.757–766 (2007).
3) Bollegala, D., Matsuo, Y. and Ishizuka, M.: WWW sits the SAT-Measuring Relational Similarity on the Web, *ECAI*, pp.333–337 (2008).
4) Church, K.W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol.16, No.1, pp.22–29 (1990).
5) Cibelli, M., Nappi, M. and Tucci, M.: ABI: analogy-based indexing for content image retrieval, *Image Vision Computing*, Vol.22, No.1, pp.23–34 (2004).
6) Falkenhainer, B., Forbus, K.D. and Gentner, D.: The Structure-Mapping Engine: Algorithm and Examples, *Artificial Intelligence*, Vol.41, No.1, pp.1–63 (1989).
7) Ghahramani, Z. and Heller, K.A.: Bayesian Sets, *Proc. 19th Annual Conference on Neural Information Processing Systems* (*NIPS 2005*), pp.435–442 (2005).
8) Hearst, M.: Automatic Acquisition of Hyponyms om Large Text Corpora, *Proc.14th International Conference on Computational Linguistics* (*COLING 1992*), pp.539–545 (1992).
9) Heidorns, P.: The Identification of Index Terms in Natural Language Object Description, *Proc. American Society for Information Science Conference*, Vol.36, pp.472–481 (1999).
10) Hokama, T. and Kitagawa, H.: Extracting Mnemonic Names of People from the Web, *Proc. 9th International Conference on Asian Digital Libraries* (*ICADL 2006*), pp.121–130 (2006).
11) Lin, D.: Automatic Retrieval and Clustering of Similar Words, *Proc. 36th Annual Meeting of the Association for Computational Linguistics* (*ACL 1998*), pp.768–774 (1998).
12) Luo, G., Tang, C. and li Tian, Y.: Answering relationship queries on the web, *WWW*, pp.561–570 (2007).
13) Miller, G.: WordNet: A Lexical Database for English, *Comm. ACM*, Vol.38, No.11, pp.39–41 (1995).
14) Oyama, S. and Tanaka, K.: Query Modification by Discovering Topics from Web Page Structures, *Proc. 6th Asia-Pacific Web Conference* (*APWeb 2004*), pp.553–564 (2004).
15) Turney, P.D.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, *Proc. 12th European Conference on Machine Learning* (*ECML 2001*), pp.491–502 (2001).
16) Turney, P.D.: Measuring Semantic Similarity by Latent Relational Analysis, *IJCAI*, pp.1136–1141 (2005).
17) Turney, P.D.: Expressing Implicit Semantic Relations without Supervision, *ACL*, pp.313–320 (2006).
18) Turney, P.D.: Similarity of Semantic Relations, *Computational Linguistics*, Vol.32, No.3, pp.379–416 (2006).
19) Turney, P.D. and Littman, M.L.: Corpus-based Learning of Analogies and Semantic Relations, *Machine Learning*, Vol.60, No.1-3, pp.251–278 (2005).
20) アリストテレス（著），高田三郎（訳）：ニコマコス倫理学，岩波文庫 (1971).
21) 大島裕明，田中克己：両方向構文パターンを用いた Web 検索エンジンからの高速関連語発見手法，情報処理学会研究報告，Vol.2008, No.88, pp.37–42 (2008).
22) 中島伸介，田中克己：相対的マッピング処理に基づく相対的情報検索手法，情報処理学会論文誌：データベース，Vol.45, No.4, pp.63–75 (2004).

(平成 20 年 12 月 20 日受付)
(平成 21 年 4 月 7 日採録)

（担当編集委員　戸田 浩之）

加藤　誠

京都大学大学院情報学研究科修士課程在学中．2008 年京都大学工学部情報学科卒業．情報検索の研究に従事．

大島　裕明（正会員）

京都大学大学院情報学研究科社会情報学専攻特定助教．2007 年京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．主に Web，情報検索，データベースの研究に従事．電子情報通信学会，日本データベース学会，ACM 各会員．

小山　　聡（正会員）

京都大学大学院情報学研究科社会情報学専攻助教．1994 年京都大学工学部数理工学科卒業．1996 年同大学大学院工学研究科数理工学専攻修士課程修了．1996〜1998 年日本電信電話株式会社．2001〜2002 年日本学術振興会特別研究員（DC2）．2002 年京都大学大学院情報学研究科社会情報学専攻博士後期課程修了．博士（情報学）．同年より同大学大学院情報学研究科社会情報学専攻助手．2003〜2004 年スタンフォード大学 Visiting Assistant Professor．2005 年度人工知能学会論文賞．機械学習，データマイニング，情報検索等に興味を持つ．

田中　克己（正会員）

1974 年京都大学工学部情報工学科卒業．1976 年同大学大学院修士課程修了．1979 年神戸大学教養部助手．1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学専攻）．1995 年同大学大学院自然科学研究科情報メディア科学専攻専任教授．2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．工学博士．主にデータベースとマルチメディア情報システムの研究に従事．人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．