Web検索とクエリログを併用した同位語発見手法

Discovering Coordinate Terms Using both Web Search and its Query Logs

大島 裕明[1] 山口 雅史[2]
小山 聡[3] 田中 克己[3]

Hiroaki OHSHIMA Masashi YAMAGUCHI
Satoshi OYAMA Katsumi TANAKA

本稿では，同位語を発見する手法について提案する．我々はこれまで，Web検索エンジンが保有している情報を利用して同位語を取得する手法を2つ提案してきた．一方はWeb検索エンジンの検索結果を利用する手法であり，他方はクエリログの情報を利用する手法である．それらは共に1語が与えられたときにいくつかの同位語を返すという手法であるが，得られる同位語の数や適合率に違いがある．本稿では，そのような異なる手法を組み合わせた同位語発見手法について提案する．

We proposed a method for discovering coordinate terms. There had been two methods for finding coordinate terms; one is using Web search, and the other is using query logs in a Web search engine. Both of them can find some coordinate terms of a term which is given by a user. The average number of coordinate terms each method can find and the precision is different. We proposed some combinations of these methods.

1. はじめに

語と語の関係性には様々なものがある．「上位語」「下位語」はそれぞれ，ある語が表す概念の上位概念や下位概念を表す語のことである．「同義語」「同意語」「類義語」は，ある語と同様の意味を持つ語のことである．「同位語」も語と語の関係性を表す語の１つであり，ある語と共通の上位語を持つ別の意味の語のことである．例えば，「テニス」と「ゴルフ」は共通の上位語「スポーツ」を持つことから，同位語であるといえる．同位語に限らず，ある語に対して関連する語を発見するサービスは，様々なアプリケーションにおいて利用可能である．例えば，Web 検索の際にはユーザのクエリとしてキーワードが使われるが，求めるものを発見するために最適な語を考えるには慣れやコツが必要であることが多い．特に子供や初心者の場合や，慣れている人であっても未知の分野について調べる場合には，ある語に対して関連のある語を提示するといった想起支援が検索において助けになると考えられる．特に同位語を利用したアプリケーションも考えられる．例えば，デジタルカメラについて調査したいが「LUMIX」しか知らないような場合，このユーザに「EXLIM」，「FinePix」，「Cyber-Shot」，「IXY」といった同位語を提示して比較対象を思い起こさせるサービスが考えられる．

我々はこれまで同位語を求める手法を 2 つ提案してきた．一方は Web 検索の検索結果を利用する手法[1]で，他方はクエリログの情報を利用する手法[2]である．両手法とも，1 語を与えられた時に，いくつかの同位語を出力するというシステムである．本稿ではまず，それらの手法について説明する．次に，いくつかの同位語発見手法がある時に，それらを組み合わせて，結果として得られる同位語の数を増やしたり，正解率を高めたりするための統合手法について提案する．

2. Web検索やクエリログを利用した同位語発見手法

2.1 Web 検索を利用した同位語発見手法

本節では Web 検索を利用した同位語発見手法について述べる．本手法は，同位語が並列助詞「や」で接続されることが多いことに着目し，与えられた語と並列助詞「や」で接続される語を発見するものである．

まず，ユーザはクエリの語 Q を与える．クエリの語 Q は 1 語であり，単語でも複合語でもかまわない．次に，Web 検索エンジンに対するクエリを 2 つ作成する．並列助詞「や」をクエリの語 Q の前後に付加したものである．例えば，Q が「白鳥の湖」であるとき，Web 検索エンジンに対する 2 つのクエリは「"白鳥の湖や"」と「"や白鳥の湖"」となる．引用符で括っているのは，多くの Web 検索エンジンが実装しているフレーズ検索を表している．フレーズ検索では，引用符で括られた部分がそのまま出現するようなページが検索される．

Web 検索の結果の各アイテムは，タイトル，URL，スニペットからなるのが一般的である．スニペットは，検索された Web ページの中に含まれいくつかの文で，検索語が出現するような文から成る．作成したクエリで得られた 100 件ずつの Web 検索の結果から，タイトルとスニペットを得る．それらが本手法において解析するテキストである．

解析対象のテキストの中から，クエリの語 Q と並列助詞「や」で接続されている語，すなわち，「Q や」の直後と，「や Q」の直前に現れるような語を取得し，各語において「や」の直前と直後における出現回数を求める．そして，並列助詞「や」の両側に出現するような語のみを同位語とみなす．

例えば，クエリの語が「白鳥の湖」の時，以下のような文が検索結果のスニペットに存在している．

- まぁ，この曲自体，白鳥の湖やくるみ割り人形と比較して，こういうシンフォニックな 演奏でも聞き映えするように書かれてるから...
- 選ばれた曲はくるみ割り人形や白鳥の湖，カルメン，惑星…のようなポピュラーなものから，...

この場合，「くるみ割り人形」という語が「白鳥の湖」と並列助詞「や」と接続されて前後両方において出現しているため，「くるみ割り人形」を「白鳥の湖」に対する同位語であると判定する．並列助詞「や」の両側に出現した場合のみ同位語とみなすという条件によって，複合語を正しく取り出すことも可能となっている．

2.2 クエリログを利用した同位語発見手法

本節では Web 検索エンジンのクエリログを用いる同位語発見手法について述べる．クエリログは，ユーザが Web 検索エンジンを利用した際のクエリの履歴である．Web 検索エンジ

[1] 京都大学大学院情報学研究科博士後期課程
ohshima@dl.kuis.kyoto-u.ac.jp
[2] 京都大学大学院情報学研究科博士前期課程
yamaguti@dl.kuis.kyoto-u.ac.jp
[3] 京都大学大学院情報学研究科
{oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

ンの運営主体はクエリログを収集していることがある．本研究においては，クエリログとして Overture による「キーワードアドバイスツール」で得られるデータを利用した．

キーワードアドバイスツールは Web ベースのシステムで，検索フォームに語を入力して問い合わせを行うと，その語を含んだ様々な組み合わせのクエリを提示する．提示される組み合わせは，前月における検索回数が多い順に最大 100 組である．例えば，「金閣寺」という問い合わせに対して，

・金閣寺　　50000 件
・金閣寺 京都　　5000 件
・金閣寺 銀閣寺　　3000 件
・銀閣寺 金閣寺　　500 件
・金閣寺 アクセス　　300 件

といったような結果が返される．結果において，「金閣寺」と「銀閣寺」が両方出現するクエリであっても順序が異なれば別の組み合わせとして扱う．本稿では，「金閣寺 京都」といったような各組み合わせのことをログレコードと呼ぶ．

「金閣寺」という語を含んだログレコード集合の中に，「金閣寺 アクセス」がある時，「金閣寺」の部分を変数 x とした「x アクセス」という型を考えることができる．このような型のことを，「『金閣寺』の共起型」と呼ぶことにする．「x アクセス」という「金閣寺」の共起型に対して，「交通 アクセス」というログレコードがあるとき，「交通」はこの共起型に適合する語と呼ぶ．また，「金閣寺」のような 1 語のみのログレコードに対する共起型は考えないこととする．

本手法は，同位語どうしは共通の共起型に適合するという仮定に基づいている．例えば，「トヨタ」と「ホンダ」は同位語であるが，これらの語が含まれるログレコードには，「トヨタ 自動車」「ホンダ 自動車」などが含まれる．このとき，「x 自動車」という共起型に両方の語が適合している．このような，適切な共起型を発見することが重要となる．

まず，ユーザがクエリの語 Q を与える．以下では「トヨタ」をクエリの語 Q の例として用いて説明する．「トヨタ」を含むログレコード集合を取得すると，「トヨタ」「トヨタ 自動車」「トヨタ レンタカー」「トヨタ クラウン」などが得られる．このとき，「トヨタ」の共起型は，「x 自動車」「x レンタカー」「x クラウン」などである．次に，それらの共起型に現れる語を含むログレコード集合を求める．例えば，「自動車」を含むログレコード集合では，「自動車 趣味」「自動車 メーカー」「トヨタ 自動車」「日産 自動車」「ホンダ 自動車」「旅行 自動車」などが得られる．得られたログレコード集合から，「x 自動車」という共起型に適合する語を発見すると，「トヨタ」「日産」「ホンダ」「旅行」などがあることが分かる．まず，「トヨタ」自身が含まれていることに注目し，この共起型は同位語発見において適切な共起型であるとみなす．

例えば，「トヨタ」を含むログレコードの上位 100 件中には，「トヨタ 壁紙」というログレコードが存在する．しかし，逆に「壁紙」を含むログレコードの上位 100 件中には「ディズニー 壁紙」「浜崎あゆみ 壁紙」などは存在するのだが，「トヨタ 壁紙」は存在しない．これは，「壁紙」といった語が非常に多くの語と共起して使われるからであり，「トヨタ 壁紙」が「壁紙」を含むログレコードの上位 100 件中に含まれていないことから，「トヨタ」の同位語を含むログレコードもこれらの中に含まれるとは考えにくい．同様のことは「画像」という語にもいえる．よって，「x 壁紙」や「x 画像」という共起型は「トヨタ」の同位語の発見には適さないとみなす．そして，「日産」「ホンダ」「旅行」といった語を同位語の候補とする．

全ての共起型に対して同様の操作を行うと，各共起型に適合する同位語の候補が得られる．このとき，いくつもの共起型に適合する同位語の候補があれば，その語は同位語として有力な候補であると考えられる．また，逆に同位語として有力な候補の多くが適合するような共起型があれば，その共起型は同位語発見により有用な共起型であると考えることができる．そのような，良い「共起型」や良い「共起型に適合する語」を求めるために，これらの関係を行列で表現しその行列の特異値分解を行う．まず，行を「共起型」，列を「同位語の候補」とし，行列の成分を，共起型に適合するときに 1，適合しないときに 0 とする．その行列を特異値分解し，最大特異値に対応する左特異ベクトルと右特異ベクトルを求める．そして，右特異ベクトルで大きな値を持つ同位語の候補を始めに与えられた語の同位語として出力する．

2.3　2 手法の性質の違い

これら 2 つの手法は，1 語が与えられたときにいくつかの同位語を出力するというものであるが，得られる同位語の数や適合率は異なっている．性質の違いを比較するためテストセットを作成し，どのような違いがあるか比較を行った[3]．テストセットではクエリとして 50 語用意した．そして，あらかじめ各語に対して上位語を設定し，各手法によって得られた語のうち，設定された上位語の下位語であると判定できる語を正解とした．例えば，「姫路城」というクエリに対しては「日本の史跡」という上位語を設定した．表 1 が実験結果のまとめである．

表 1　2 つの手法の性質の違い
Table 1　Difference of the methods

	Web 検索	クエリログ(30 語出力)
出力語数	9.8 語	28.8 語
正解数	6.6 語	11.9 語
不正解数	3.2 語	16.9 語
適合率	67.6%	41.4%
最低 1 語の正解	98.0%	78.0%

Web 検索を利用した手法では，同位語として正解の語が平均 6.6 語得られ，適合率は 67.6%であった．クエリログを利用した手法は，全ての同位語の候補がランキングされる手法であるため，上位の最大 30 語を出力としたときの結果を示している．この時，同位語として正解の語は 11.9 語得られ，適合率は 41.4%であった．また，出力中に最低 1 語の正解が存在するようなクエリの割合は，Web 検索を用いる手法では 98%と非常に良く，クエリログを用いる手法では 78%であった．まとめると，得られる同位語の数はクエリログを用いる手法の方が多く，適合率や適応範囲は Web 検索を用いる手法の方が良いことが分かった．

今回は正誤判定のためにあらかじめ上位語を設定したが，実際に得られた語で不正解と判定された語の中にもクエリの多義性から正解と判定できる語も存在していた．例えば，Web 検索を用いる手法では，「姫路城」に対して同じ世界遺産である「屋久島」「白神山地」が出力されたが，今回の正誤判定では不正解と判定した．他にも，今回「ピアノ」に対して設定していた上位語は「楽器」であったが，クエリログを用いる手法では，「書道」「英会話」といった習い事が出力されたりした．よって，表 1 に示した適合率は，人間が判定した場合よりも多少悪い適合率となっている．

3. 手法の組み合わせによる同位語発見手法

3.1 同位語の性質と組み合わせる目的

ユーザに与えられた語を t_0 とし，t_0 に対していくつかの同位語があるとする．語の多義性を無視した場合，t_0 の同位語に対する同位語もまた t_0 の同位語であると考えられる．このような同位語の性質もあり，同位語発見手法のいくつかの手法を組み合わせることによって，性能の向上を図ることが考えられる．組み合わせる目的としては，大きく「拡張」と「洗練」という 2 つが考えられる．

拡張とは，1 語を入力としたときに最終的に得られる同位語の数を増やすことである．洗練とは，出力された語が与えられた語の同位語であるという適合率を上げることである．

以下ではいくつかの同位語発見手法があるときに，それらを組み合わせて拡張や洗練を行うことについて考える．

3.2 並列的組み合わせ

いくつかの同位語発見手法があるときに，考えられる 1 つの組み合わせ方は並列的組み合わせである．2 つの手法 M_A と M_B があるとする．また，ユーザに与えられた語を t_0 とする．t_0 を入力として M_A によって得られる同位語を $\{t_1, \cdots, t_p, \cdots, t_q\}$ とし，t_0 を入力として M_B によって得られる同位語を $\{t_p, \cdots, t_q, \cdots, t_r\}$ とする．これらの 2 つの出力を組み合わせて新たな出力とすることを，ここでは並列的組み合わせと呼ぶ．

並列的組み合わせによる出力には拡張と洗練のどちらを目的とするかによって 2 種類考えることができる．拡張を目的とする場合には，出力数の増加が求められるため，$\{t_1, \cdots, t_p, \cdots, t_q, \cdots, t_r\}$ の全てを出力とすればよい．それに対して，洗練を目的とする場合には，適合率の上昇が求められるため，$\{t_p, \cdots, t_q\}$ を出力とすることが考えられる．$\{t_p, \cdots, t_q\}$ は両手法で共に t_0 の同位語であるとされた語であり，同位語である可能性は高いと考えられる．

3.3 直列的組み合わせ

同位語発見手法を直列的に組み合わせることも考えられる．2 つの手法 M_A と M_B があるとし，ユーザに与えられた語を t_0 とする．直列的組み合わせではまず，M_A によって t_0 の同位語を求める．これを $\{t_1, \cdots, t_n\}$ とする．次に，$\{t_0, t_1, \cdots, t_n\}$ のそれぞれに対する同位語を，M_B を用いて求める．M_B を用いて得られた $t_k(0 \leqq k \leqq n)$ の同位語を $\{t_{k,1}, \cdots, t_{k,m}\}$ とする．

並列的組み合わせと同様に，拡張と洗練のどちらを目的とするかによって 2 種類の出力を考えることができる．拡張を重視する場合には，最終的に得られた全ての語を出力とすることが考えられる．もし，出力する語のランキングを行う場合には，下記の洗練の考え方も利用可能である．洗練を目的とする場合には，ある語 x がいくつの $t_k(0 \leqq k \leqq n)$ に対する同位語として出力されたかが重要な指標となる．例えば，M_A によって求められた t_0 の同位語が 4 つであったとして，t_0 から t_4 に対して M_B によって同位語を得たときに，ある語 x がそれら全てにおいて含まれていたとすると，x が t_0 の同位語である可能性は高い．ただ，同位語発見手法が 100%の適合率をもっているわけではないため，現実的に洗練を行う条件として考えられるのは，M_B を用いて求められた同位語集合のうちの 30%以上や，50%以上に含まれる語を正解とする，など，閾値を設定することが実用上必要であると考えられる．

3.4 Web 検索とクエリログを併用した同位語発見手法

2 節において，Web 検索を用いる同位語発見手法と，クエリログを用いる同位語発見手法について述べた．本節ではそれらの様々な組み合わせについて考える．

洗練目的で並列的組み合わせを行うと，非常に精度良く同位語を取得することができる．しかし，クエリログを用いる手法では 2 割程度の質問においては正解が全く含まれていないことがあり，その場合は組み合わせ手法でも 1 語も出力されないようになる．拡張目的で並列的組み合わせを行うと，適合率は従来のクエリログを用いる手法よりも良くなり，正解数は Web 検索を用いる手法よりも良いという，中間的な手法になる．

直列的組み合わせでは，2 つの手法を利用する順序と，拡張と洗練のどちらを目的にするかで，4 つの手法を考えることができる．しかし，特にクエリログを用いる手法において適合率があまり良くなく，拡張を行った場合にはさらに適合率が悪くなることが予想できるため，拡張目的の組み合わせは行わない方が良いと考えられる．

洗練目的の直列的組み合わせについては，先に Web 検索を用いる手法を用いる方法と，先にクエリログを用いる手法で得られた上位 30 位までの語に対して Web 検索を用いる手法で同位語を発見する方法が考えられるが，どちらでもある程度の精度向上を行うことができると考えている．ここでは，後者について行った実験の例を示しておく．

ユーザの入力語としたのは「バッハ」である．まず，クエリログを用いる手法で同位語を 30 語取得し，各語に対して Web 検索を用いる手法で同位語を取得した．30 語のうちの 3 語以上から同位語と判定された語は 14 語あり，それらのうち 86%の 12 語は正解であった．また，30 語のうちの 2 語以上から同位語と判定された語は 38 語あり，それらの 59%にあたる 22 語は正解であった．Web 検索を用いる手法のみの場合は 13 語出力され，うち 77%の 10 語が正解であり，クエリログを用いる手法のみの場合は 30 語を出力としたときに 47%の 14 語が正解であった．これより，組み合わせた手法の出力語数や適合率がある程度向上したと考えられる．

3.5 複数語入力を受ける手法を含む組み合わせ

ここまで述べた手法の組み合わせは，1 語の入力に対していくつかの同位語を取得する手法を組み合わせることであった．複数の語の入力を受けて，それらに対していくつかの同位語を取得する手法について考えると，さらに多くの組み合わせについて考えることができる．

そのうちの 1 つは直列的組み合わせである．2 つの手法 M_A と M_B があり，M_B は複数語入力であるものとする．ユーザに与えられた語を t_0 とする．まず，M_A によって t_0 の同位語を求める．これを $\{t_1, \cdots, t_n\}$ とする．次に，t_0 と $\{t_1, \cdots, t_n\}$ からいくつかを M_B に対する入力とし，M_B の出力を求める．この組み合わせによる出力は M_B の出力そのものであるため，先述した拡張や洗練とは少し異なるものであるが，M_B の手法が 1 入力の場合よりも複数入力の場合の方が何らかのメリットがある場合に有用であると考えられる．

2.2 節で述べたクエリログを利用した同位語発見手法は，少し変化させることによって複数語の入力を受けるようにすることが可能である．クエリログを用いる手法では同位語が適合する共起型を発見することが重要となるが，あらかじめ複数の語が与えられていると，あらかじめそれらの多くに共通する共起型が判明する．それらの共起型は他の同位語にも適合する可能性が高いため，それらのみを利用して同位語を求めることが可能となる．複数語入力にすることによって，適合率が上昇する可能性があり，また，利用するクエリログの量も少なくてすむというメリットが存在している．

この手法を用いた組み合わせについて実験を行った．2.1節で述べたWeb検索を用いる手法によって，ユーザの入力 t_0 に対する同位語を求め，その上位2語と t_0 の計3語を上記手法の入力とした．実験で用いたデータは2.3節のものと同様のものを用いた．

図1が適合率についてのグラフである．上位において適合率が良くなっていることが分かる．

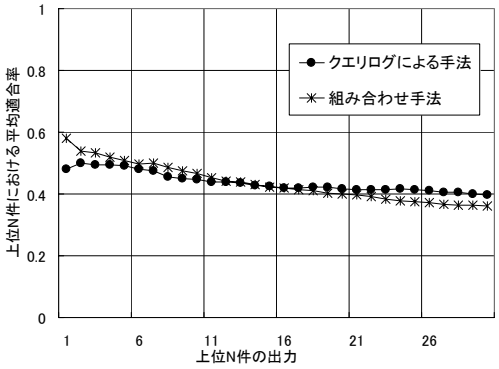

図1　複数語入力の組み合わせ手法

Fig.1　A combination of methods with multiple term input

4. 関連研究

同位語を取得するサービスとしてGoogle Sets[4]が存在する．Google Setsのアルゴリズムは公開されていないが，Googleが収集したWebページに含まれる語に対して大規模なクラスタリングを行い，それによって同位語のクラスタを大量に生成しているようである．現時点では英語に対応している．

同位語の発見に関する研究はいくつか存在している．Churchら[4]は，相互情報量を用いて意味的に関連があるような語を発見する手法について提案した．厳密には同位語の発見を目的としたものではないが，発見される語には同位語が多く含まれることになる．Ghahramaniら[5]によるBayesian SetsはGoogle Setsと同様のシステムを目指したものであり，語の共起テーブルのような大規模なデータに対して，ベイズ推定を用いて同位語のクラスタを発見する．

Linら[6]は類似するような語のクラスタを作成する手法を提案した．係り受け関係を利用して語どうしの類似度を計算するため，係り受け解析が行われている大規模コーパスが必要となる．Shinzatoら[7]はHTML文書から同位語を発見する手法を提案した．HTMLの構造上において同レベルに列挙されているような語を，同位語である可能性がある語として取得している．

5. まとめ

本稿では，同位語発見手法として，Web検索を用いる手法と，Web検索エンジンのクエリログを用いる手法について紹介した．さらに，いくつかの同位語発見手法があるときに，それらの様々な組み合わせについて考案し，Web検索を用いる手法とクエリログを用いる手法の組み合わせについて提案を行った．

[謝辞]

本研究の一部は，文部科学省21世紀COE拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」（リーダー:田中克己，平成14～18年度），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しいIT基盤技術の研究」計画研究「情報爆発に対応する新IT基盤研究支援プラットフォームの構築」(研究代表者:安達淳, Y00-01, 課題番号:18049073）ならびに計画研究「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，文部科学省研究委託事業「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．

[文献]

[1] 大島裕明，小山聡，田中克己，「Web検索エンジンのインデックスを用いた同位語とそのコンテキストの発見」，情報処理学会論文誌（トランザクション）データベース, Vol.47, No.SIG19, TOD32, pp.98-112 (2006).

[2] 山口雅史，大島裕明，小山聡，田中克己，「サーチエンジンのクエリログを利用した同位語の発見」，日本データベース学会Letters, Vol.5, No.2, pp.17-20 (2006).

[3] 大島裕明，山口雅史，小山聡，田中克己，「Web検索エンジンのインデックスとクエリログを用いた同位語発見」，情報処理学会，DBWeb2006 シンポジウム論文集, pp.305-312 (2006).

[4] K. W. Church, P. Hanks: “Word Association Norms, Mutual Information, and Lexicography”, Proc. of the 27th Annual Meeting of the Association for Computational Linguistics, pp.76-83 (1998).

[5] Z. Ghahramani, K. Heller: “Bayesian Sets”, Advances in Neural Information Processing Systems 18, pp.435-442 (2006).

[6] D. Lin: “Automatic Retrieval and Clustering of Similar Words”, Proc. of the 36th annual meeting on Association for Computational Linguistics, pp.768-774 (1998).

[7] K. Shinzato, K. Torisawa: “A Simple WWW-based Method for Semantic Word Class Acquisition”, Proc. of the Recent Advances in Natural Language Processing (RANLP05), pp.493-500 (2005).

大島 裕明　Hiroaki OHSHIMA
京都大学大学院情報学研究科博士後期課程在学中．2004年神戸大学大学院自然科学研究科博士前期課程修了．Web検索やパーソナライゼーションの研究に従事．情報処理学会，電子情報通信学会，日本データベース学会，ACM各学生会員．

山口 雅史　Masashi YAMAGUCHI
京都大学大学院情報学研究科博士前期課程在学中．2005年京都大学工学部情報学科卒業．Web環境におけるパーソナライゼーション，クエリログ活用の研究に従事．日本データベース学会学生会員．

小山 聡　Satoshi OYAMA
京都大学大学院情報学研究科社会情報学専攻助手．2002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）．主に機械学習，データマイニング，情報検索の研究に従事．電子情報通信学会，情報処理学会，人工知能学会，日本データベース学会，IEEE，ACM，AAAI各会員．

田中 克己　Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．

[4] http://labs.google.com/sets