文書群をクエリとした“似て非なる”文書の検索

Sibling Page Search by Page Examples

大島 裕明[1] 小山 聡[2] 田中 克己[3]

Hiroaki OHSHIMA Satoshi OYAMA
Katsumi TANAKA

本研究では，ユーザがすでに保有している文書のいくつかをクエリとして，それらとカテゴリ的に「兄弟(sibling)」の関係にあるような文書，すなわちクエリの文書群と同一カテゴリに分類される文書で，その内容がクエリの文書群とは異なるような文書を検索する手法を提案する．我々はこのような文書の検索を「似て非なる文書の検索」と呼ぶ．現在，ある事柄について網羅的に調べるためには，目的の文書で使われているような語を考え，その語をクエリとして検索エンジンに与え，検索結果からすでに持っている文書とは異なるような文書をユーザが1つ1つ調べていくというような事が行われている．そのような場面において「似て非なる文書の検索」のニーズがあると考えられる．本稿では，ユーザがすでに持っている文書をクエリとして与えた時にそれらの文書に対して，ある文書が似て非なる文書として適合するものであるかどうかを評価する手法について提案を行う．

We propose methods of searching Web pages that are “semantically” regarded as “siblings” with respect to given page examples. That is, our approach aims to find pages that are similar in theme but have different content from the given sample pages. We called this “sibling page search”. The proposed search methods are different from conventional content-based similarity search for Web pages. Our approach recommends Web pages whose “conceptual” classification category is the same as that of the given sample pages, but whose content is different from the sample pages.

1. はじめに

現在，Web検索は新しい情報を得るための主要な手段となった．有名なものとしてはGoogle[1]，Yahoo![2]，AltaVista[3]などが挙げられる．しかし，Web検索エンジンを利用する際にユーザができることはいくつかの検索キーワードを渡すことのみであり，何をすでに知っているかということなどを伝えることはできない．そのため，ある分野についてユーザが網羅的に調べているような際には，ユーザはWeb検索エンジンに対していくつかの検索キーワードを与え，返された結果のページを1つ1つチェックして自分が知らない情報があるかどうかを調べる，といった作業を行わなくてはならない．

我々は，何らかの事柄について網羅的に調べているようなときに，自分がすでに持っている文書と関連はあるが，内容は異なるような文書を検索するという，「似て非なる文書検索」を行う手法について提案を行う．ユーザは自分がすでに持っている文書のいくつかをクエリとして与え，システムはクエリとして与えられた文書が属する概念的なカテゴリを考え，そのカテゴリに属するが内容的には異なるような文書を結果として返すことを目的とする．

例えば，ワインに興味がある人が，「ボルドーワイン」と「ブルゴーニュワイン」についての文書をそれぞれいくつか持っているときに，それらの文書をクエリとすることによって，ワインには関係するが「ボルドーワイン」と「ブルゴーニュワイン」とは異なるものに関する文書，例えば「ローヌワイン」に関する文書のように，カテゴリ分類的に兄弟関係にあるような文書を結果として返すようなものが，本研究の「似て非なる文書検索」である．

2. クエリの要件

本研究において，ユーザが与えるクエリは，いくつかの文書から成り立つ文書集合の複数の集合によって成り立つ．つまり，P_kが1つ以上の文書で構成される文書集合であるとき，$P_1,\ldots,P_n(n>1)$がクエリとなる．このとき，文書集合は複数である必要がある．

図1はクエリとされる文書集合が1つである場合と，複数である場合を示している．図中の(a)はクエリが1つの文書集合からなる場合である．クエリとされる「ボルドーワイン」に関する文書集合の中には，ボルドーに関する話題とワインに関する話題の両方が含まれていると考えられる．クエリとされる文書集合が1つである場合，目的とする文書がその文書集合の中のどの部分と類似しているべきか，どの部分と非類似であるべきか，という事を判断することはできない．一方，図中の(b)ではクエリが複数の文書集合からなっている．「ボルドーワイン」に関する文書集合と「ブルゴーニュワイン」に関する文書集合という2つの文書集合によってクエリが構成されていれば，両方に含まれている部分である「ワイン」に関する部分が，似て非なる文書においても含まれていることが期待される．また，それぞれの文書で特有の部分である「ボルドー」に関する部分や「ブルゴーニュ」に関する部分は，似て非なる文書にはあまり現れないことが期待される．

以上より，本研究においては，複数の基準となる文書集合によってクエリを表すものとする．

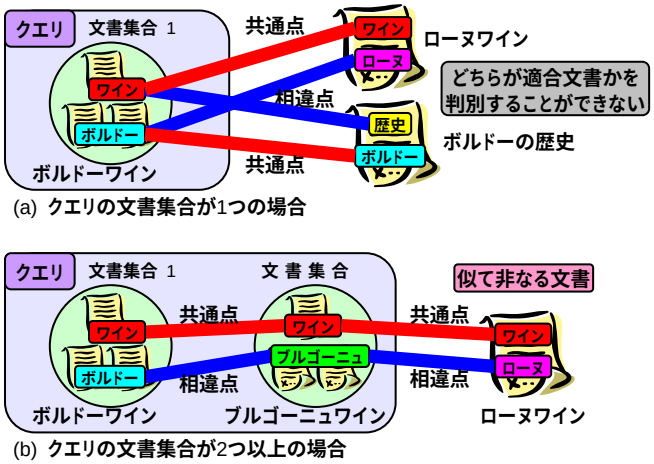

図1 クエリを構成する文書集合の数

Fig.1 The number of page collection in a query

[1] 学生会員 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] 正会員 京都大学大学院情報学研究科助手 oyama@dl.kuis.kyoto-u.ac.jp
[3] 正会員 京都大学大学院情報学研究科教授 tanaka@dl.kuis.kyoto-u.ac.jp

3. 似て非なる文書としての適合性の評価

似て非なる文書として適合する文書は，ユーザがクエリとして与えた文書集合のすべてにおいて共通な部分は含んでおり，クエリの文書集合のそれぞれにおいて特有な部分については含んでいないような文書のことである．

それを判定するために，クエリの文書集合の共通部分と，クエリの文書集合ごとの特有部分を，それぞれ特徴ベクトルで表現する．それらの特徴ベクトルと何らかの方法で得られた文書の類似度を基にして，文書が似て非なる文書としてどの程度適合しているのかということを判定する．

3.1 クエリの文書集合の表現

まず，クエリとして与えられたいくつかの文書集合のそれぞれを特徴ベクトルとして表す．与えられた文書集合を$P_1,\ldots,P_n$とするとき，それぞれの文書集合の特徴ベクトルを$t_1,\ldots,t_n$とする．文書の特徴ベクトルは単語の出現回数であるTerm Frequency (TF)を用いて表現されることが多いが，本研究においても基本的にはTFを用いることとする．しかし，単純なTF以外にも，TFを基礎として様々なベクトル生成手法が考えられている．その中でもしばしば用いられるのが，TFの対数である．そこで，以下の2つの手法について検討する．

(N) 単語の出現回数
(L) 単語の出現回数の対数

なお，本研究において特徴ベクトルの要素として用いるのは，文書を形態素解析システム茶筌[4]で解析して切り分けられた語のうち，名詞，未知語，またはアルファベットを連結したものと判定されるものの中から，独自に用意したストップワードリストにあてはまらない語を対象としている．

ある文書Dにおいて語wが出現する回数を$tf(w,D)$とする．このとき，語w_iに対する文書集合P_kの特徴ベクトルt_kの値は，上記(N)と(L)それぞれの手法において以下のようになる．

(N) $t_k(w_i) = \sum_{D_j \in P_k} tf(w_i, D_j)$

(L) $t_k(w_i) = \log\left(1 + \sum_{D_j \in P_k} tf(w_i, D_j)\right)$

それぞれの文書集合を構成する文書の違いにより，各文書集合の特徴ベクトルの大きさにも差が出るため，このベクトルを以下のように正規化する．ただし，w_{Pk}はt_kにおいて最大の値を持つ要素の語とする．

$$t'_k(w_i) = \frac{t_k(w_i)}{t_k(w_{P_k})}$$

3.2 クエリの文書集合の共通部分の表現

次に，クエリを構成する全ての文書集合に共通するような部分の特徴ベクトルcを生成する手法について述べる．文書集合の共通部分の特徴を表すには，全ての文書集合における特徴を抽出することである．ここでは，以下の3つの手法について検討する．

(M) 文書集合の特徴ベクトルの相乗平均
(A) 文書集合の特徴ベクトルの相加平均
(L) 文書集合の特徴ベクトルのうちの最小値

nを文書集合の総数とするとき，上記それぞれの手法において語w_iに対するcの値は以下のようになる．

(M) $c(w_i) = \sqrt[n]{\prod t'_k(w_i)}$

(A) $c(w_i) = \frac{\sum t'_k(w_i)}{n}$

(L) $c(w_i) = \min(t'_1(w_i), \ldots, t'_n(w_i))$

3.3 クエリの文書集合の特有部分の表現

次に，クエリの文書集合の共通部分の特徴ベクトルを基にして，それぞれの文書集合の特有部分の特徴ベクトルu_kを表現する手法について述べる．文書集合P_kの特有部分は，文書集合そのものの特徴ベクトルから共通部分の特徴ベクトルを差し引くことによって表すことができると考えられる．すなわち，u_kを以下のように表す．

$$u_k(w_i) = \max(t'_k(w_i) - c(w_i), 0)$$

3.4 似て非なる文書の評価

次に，何らかの手法で文書が得られたときに，その文書が与えられたクエリに対する似て非なる文書として適するかどうかを評価する方法について述べる．直感的に，クエリとして渡された文書集合の共通部分をある程度含み，各々の文書集合の特有部分とは似ていない文書が適する文書と考えられる．ベクトル空間でこれを考えると，対象とする文書Dの特徴ベクトルdが，文書集合の共通部分の特徴ベクトルcと類似度が高く，文書集合の特有部分の特徴ベクトルu_kと類似度が低い，すなわち，非類似度が高い場合に，その文書が似て非なる文書として適合していると考えられる．まず，これらの類似度と非類似度を計算する手法について述べる．

はじめに，似て非なる文書の候補となる文書Dを特徴ベクトルとして表す．その際に用いる式は，文書集合の特徴ベクトルを生成した手法に準じて，単語の出現回数，まはた，単語の出現回数の対数のいずれかで表現する．式はそれぞれ，以下のようになる．

(N) $d(w_i) = tf(w_i, D)$

(L) $d(w_i) = \log(1 + tf(w_i, D))$

ベクトルどうしの類似度は，コサイン類似度で測定することとする．2つのベクトルv_1，v_2のコサイン類似度cosは以下の式によって定義される．

$$\cos(v_1, v_2) = \frac{\sum_w (v_1(w) \cdot v_2(w))}{\sqrt{\sum_w v_1(w)^2} \cdot \sqrt{\sum_w v_2(w)^2}}$$

コサイン類似度は2つのベクトルが作る角のコサイン値である．つまり，2つのベクトルの方向が完全に一致するときに最大値1となり，2つのベクトルが直交するときに最小値0となる．このとき，ベクトルの長さは類似度には影響しない．そのため，類似度を測定する前にベクトルの長さの正規化を行う必要はない．

似て非なる文書の候補となる文書の特徴ベクトルと文書集合の共通部分の特徴ベクトルとの類似度$Sim_c(d)$は以下の式で表される．

$$Sim_c(d) = \cos(c, d)$$

似て非なる文書の候補となる文書の特徴ベクトルと各文書集合の特有部分の特徴ベクトルの類似度は，文書集合の数

表1 文書集合の特徴ベクトルの一部
Table 1 The page collection vectors

「競艇」	(N)	(L)	「競輪」	(N)	(L)	「競馬」	(N)	(L)
競艇	1.00	1.00	競輪	1.00	1.00	競馬	1.00	1.00
舟	0.23	0.65	開催	0.36	0.77	記念	0.30	0.77
券	0.23	0.65	選手	0.34	0.75	有馬	0.29	0.76
選手	0.16	0.57	自転車	0.20	0.64	馬	0.28	0.76
ギャンブル	0.14	0.54	グランプリ	0.16	0.59	予想	0.17	0.66
投票	0.13	0.51	競技	0.15	0.58	投票	0.14	0.62
レース	0.13	0.51	投票	0.15	0.58	賞	0.13	0.61
予想	0.13	0.51	開設	0.15	0.58	コラム	0.11	0.58
軍資金	0.13	0.51	記念	0.14	0.56	優勝	0.10	0.57

表2 文書集合「競艇」「競輪」「競馬」の共通部分の特徴ベクトルcの一部
Table 2 The common part vectors for "Kyotei", "Keirin", "Keiba"

	(NM)		(NA)		(NL)		(LM)		(LA)		(LL)
予想	0.136	競馬	0.339	投票	0.125	投票	0.570	投票	0.571	投票	0.510
投票	0.120	競輪	0.333	予想	0.081	予想	0.536	予想	0.543	レース	0.451
レース	0.100	競艇	0.333	レース	0.081	レース	0.508	レース	0.510	予想	0.451
優勝	0.088	選手	0.168	優勝	0.054	優勝	0.472	選手	0.486	優勝	0.343
選手	0.067	記念	0.145	特集	0.027	選手	0.385	優勝	0.483	特集	0.254
電話	0.044	開催	0.143	競走	0.018	電話	0.333	記念	0.442	戦	0.210
募集	0.040	投票	0.137	データ	0.018	募集	0.320	開催	0.421	競争	0.171
会員	0.034	予想	0.125	バンク	0.018	特集	0.295	競馬	0.390	シリーズ	0.171

表3 文書集合「競輪」の特有部分の特徴ベクトルu_kの一部
Table 3 The unique part vectors for "Keirin"

	(NM)		(NA)		(NL)		(LM)		(LA)		(LL)
競輪	1.000	競輪	0.667	競輪	1.000	競輪	1.000	競輪	0.667	競輪	1.000
開催	0.365	開催	0.222	開催	0.365	開催	0.772	自転車	0.428	開催	0.772
選手	0.271	選手	0.170	選手	0.332	自転車	0.642	keirin	0.356	自転車	0.642
自転車	0.203	自転車	0.135	自転車	0.203	グランプリ	0.594	開催	0.351	選手	0.622
グランプリ	0.162	グランプリ	0.099	グランプリ	0.162	競技	0.576	開設	0.339	グランプリ	0.594
競技	0.149	競技	0.097	開設	0.149	開設	0.576	競技	0.339	開設	0.576
開設	0.149	開設	0.097	競技	0.149	記念	0.555	管内	0.339	競技	0.576
記念	0.135	決定	0.082	記念	0.135	keirin	0.533	周年	0.339	記念	0.555

だけ計算される．似て非なる文書の候補となる文書が，クエリの文書集合のどれか1つとでも類似していると判断される場合，似て非なる文書として適していると言うことはできない．そのため，似て非なる文書の候補の特徴ベクトルと特有部分の特徴ベクトルの間に求められた類似度の中で最大のものを評価に用いることが妥当だと考えられる．文書集合の特有部分の特徴ベクトルとの類似度の最大値$Sim_u(d)$は以下のような式となる．

$$Sim_u(d) = \max(\cos(u_1, d), \ldots, \cos(u_1, d))$$

今回利用したいのは非類似度であるため，このコサイン類似度を非類似度に変換する必要がある．コサイン類似度はコサイン値であるため，値の範囲は[0-1]である．この変換にはいくつかの方法が考えられるが，ここでは1からコサイン類似度の値を引くことで非類似度とする．よって，文書集合の特有部分の特徴ベクトルとの非類似度$Dissim_u(d)$は以下のようになる．

$$Dissim_u(d) = 1 - Sim_u(d)$$

似て非なる文書として適している文書とは，$Sim_c(d)$と$Dissim_u(d)$が共に大きくなるような文書である．よって，本稿ではこれらを掛け合わせた値を，似て非なる文書の適合性の評価値とする．式は以下のようになる．

$$R(d) = Sim_c(d) \cdot Dissim_u(d)$$

4. 提案手法の評価

ここまで，文書集合の特徴ベクトルt_kを作成する手法を2種，文書集合の共通部分の特徴ベクトルcを作成する手法を3種，それぞれ挙げた．それらの組み合わせにより6種の手法を挙げたこととなる．以降では，それらの組合せをラベルの組み合わせによって表す．例えば，(L)というラベルはt_kの計算において語彙の出現頻度の対数を用いた事を表し，(LM)というラベルはt_kの計算において(L)を用いた上で，cの計算においては相乗平均を用いたことを表す．

これらの手法の評価のために，Open Directory Project (ODP)[5]を利用してテストセットの作成を行い，各手法の比較評価を行った．ODPは人手によって編集されているウェブページの巨大なディレクトリである．ODPからリンクされているページを取得し，それらから，クエリのセット，正解ページ，不正解ページを設定した．まず，対象とする文書は，

http://dmoz.org/World/Japanese/レクリエーション/

以下の文書である．この際，特徴ベクトルがある程度の大きさになるように，取得したページから内容のテキストを抽出した時のサイズが2Kbyte以上の文書のみを対象とした．

以下の3つのディレクトリのそれぞれからリンクされているページ集合を，クエリを構成する文書集合とした．ただし，これらのサブディレクトリに存在するページは含まない．

/レクリエーション/ギャンブル/競艇/
/レクリエーション/ギャンブル/競輪/
/レクリエーション/ギャンブル/競馬/

表1はこれらの文書集合の特徴ベクトルt_kを各手法で計算したものを表している．各文書集合を特徴づけるような語の値が大きくなっていることが分かる．次に，これらの文書集合の共通部分の特徴ベクトルcを各手法で計算したものを表2に示す．(NA)の手法を除いて，「予想」「投票」「レース」など全ての文書集合において出現しそうな語の値が大きくなっていることが分かる．表3は「競輪」の文書集合の特有部分の特徴ベクトルu_kを計算したものを表している．(NM)(NA)(NL)はt_kと大差なく，この部分だけ見る限りではあまり特有部分を表しているとは思えない．(LM)(LA)(LL)では，「開催」「選手」といった語の値が相対的に小さくなり，逆に「自転車」「keirin」などの値がより大きくなっており，特有部分が強調されていることが分かる．

これらの特徴ベクトルを用いて評価を行う．対象となるページ総数は2630文書である．そのうち，正解となるページは，クエリの文書集合の兄弟カテゴリに分類されるような文書である．つまり，「/レクリエーション/ギャンブル/」ディレクトリ以下にある162文書のうち，「競艇」「競輪」「競馬」以下にある91文書を除いた71文書とした．

6つの手法それぞれにおいて，全ての文書の適合性の評価値Rを計算し，Rが大きい方をより高い順位とするような順序づけを行った．その順序づけにおいて任意のθ番目までの文書をシステムの出力とした時に，適合率と再現率を計算することが可能である．図2が6つの手法それぞれにおいて，θを1から2630まで変化させて描いた適合率-再現率グラフであ

る．適合率がより高い位置を推移するグラフの手法が良い手法と言えるため，6つの中では(NM)と(NL)が同程度に良く，残りの4手法についてはそれらに比べて悪いことが分かる．

一般に，Web検索においては結果として提示されたページの上位10ないしは20ページ程度しか見ない．(NM)と(NL)において上位20件に着目すると，それぞれ70%と65%の文書が正解文書であり，全対象ページにおいて約2.7%のみが正解文書であることを考えると良い結果と言える．以上より，クエリの文書集合ならびに評価対象の文書は語彙の出現回数によって特徴ベクトルの生成を行い，文書集合の共通部分の特徴ベクトルは文書集合の特徴ベクトルからの相乗平均，または文書集合の特徴ベクトルの中で最も小さい値を取る手法によって，似て非なる文書の評価が行えることが言えた．

5. 関連研究

ベクトル空間モデルは，文書とクエリを特徴ベクトルとして表現し，ベクトルどうしの類似度を計算することによって文書検索を行う．システムとしてはSMART[5]が有名である．本研究における特徴ベクトルの生成や類似度の計算などにおける基礎はベクトル空間モデルにおいて研究されたものである．Robertsonら[6]によるOkapi Weightingは，類似度計算を発展させたクエリと文書の適合度計算手法である．これらを基礎とする検索システムは種々あるが，クエリとして与えられたベクトルと類似していれば適合度が高くなり，本研究における似て非なる文書の検索を行うものではない．

Web検索のような大規模な検索システムではキーワード検索が主流である．ユーザは目的とする文書が含んでいると考えられる検索キーワードを推測する必要がある．本研究の「似て非なる文書の検索」におけるクエリは，ユーザが保有するいくつかの文書であり，検索におけるユーザへの負担は少ないといえる．このように，いくつかの例によってクエリを生成する事は特にマルチメディアデータベースなどでは良く行われている．例えば，Ishikawaら[7]によるMindReaderは，いくつかのサンプルと各サンプルに対するユーザが決定した適合度をクエリとする画像検索システムである．

6. まとめ

本稿では，似て非なる文書の検索を行う手法について提案した．クエリは複数の文書集合から構成される．それらの文書集合から，共通部分と特有部分をそれぞれ特徴ベクトルで表現し，それらを利用してある文書がクエリに対する似て非なる文書として適しているかどうかを判定する手法を提案した．テストセットを作成して実験を行い，提案手法が有効であることを示した．今後はさらに様々な手法に対して実験を行い，最適な手法を検討していきたい．

[謝辞]

本研究の一部は，21世紀COEプログラム「知識社会基盤構築のための情報学拠点形成」，文部科学省科学技術振興費プロジェクト「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」（代表：田中克己），平成18年度科研費特定領域研究「情報爆発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」(課題番号:18049041, 代表:田中克己)，および，平成18年度科研費若手研究(B)「参照の同一性判定に基づく複数Webページの検索閲覧方式の研究」（課題番号：16700097，代表：小山聡）によるものです．ここに記して謝意を表すものとします．

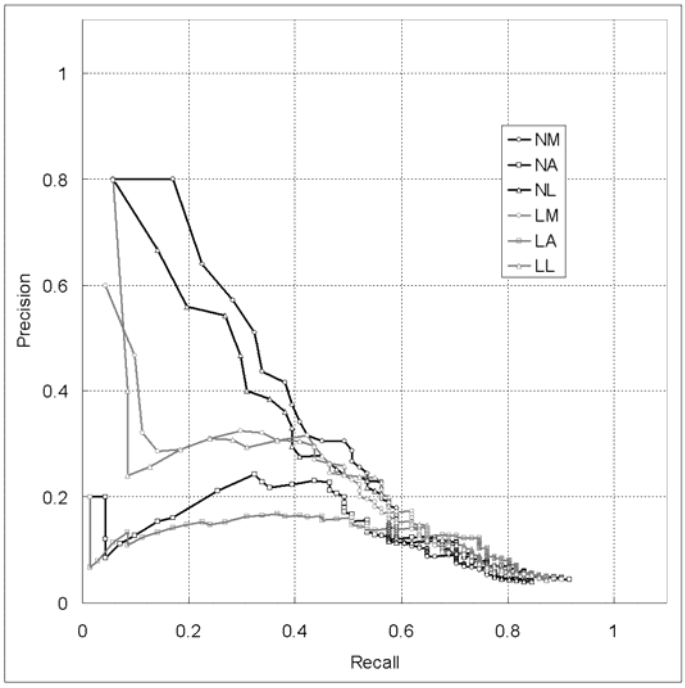

図 2　適合率-再現率グラフ
Fig.2　Recall-precision graphs

[文献]

[1] Google.　http://www.google.com/
[2] Yahoo!　http://www.yahoo.com/
[3] AltaVista.　http://www.altavista.com/
[4] 形態素解析システム「茶筌」. http://chasen.naist.jp/hiki/ChaSen/
[5] Open Directory Project.　http://dmoz.org/
[6] G. Salton and M. McGill: “Introduction to Modern Information Retrieval ”, McGraw-Hill (1983).
[7] M. Beaulieu, M. Gatford, X. Huang, S. Robertson, S. Walker, and P. Williams: “Okapi at TREC-5”, Proceedings of TREC-5, pp. 143-166 (1997).
[8] Y. Ishikawa, R. Subramanya, and C. Faloutsos: “MindReader: Querying databases through multiple examples”, Proceedings 24th Inernational Conference on Very Large Data Vases, pp. 218-227 (1998).

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 環境におけるパーソナライゼーションの研究に従事．情報処理学会，日本データベース学会，ACM 各学生会員．

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．