文書群をクエリとした“似て非なる”文書の検索

大島　裕明† 　小山　聡† 　田中　克己†

† 京都大学大学院情報学研究科社会情報学専攻
〒 606–8501 京都市左京区吉田本町
E-mail: †{ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本研究では，Web 情報検索を用いてすでに得られた文書のいくつかをクエリとして，それらとカテゴリや主題が似ているが内容が異なるような文書を取得する手法について提案を行う．ある事柄について網羅的に調べているときなどに，自分がすでに持っている Web 文書と関連はあるが異なる文書を取得したいというニーズが存在する．現在，このような目的のためには，既存の文書に共通するような語をクエリとして検索エンジンからページ群を取得し，それら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．本稿では，クエリとして与えられた文書に対して，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．そして，適合する文書の候補を既存の Web 検索エンジンから取得するための手法について提案を行う．

キーワード　Web 検索，文書によるクエリ，似て非なる文書

Search of Web Pages with Similar Categories but Different Contents by Page Examples

Hiroaki OHSHIMA†, Satoshi OYAMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606–8501 Japan
E-mail: †{ohshima,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract　We propose a system called similar but different documents search. A query for similar but different documents search consists of some documents. The system collects candidate documents from Web, evaluates them, and gives a result. When a user is studying about a certain field comprehensively, he or she will want documents that are related to the field and different from the documents he or she already has. The system can be useful in such a situation. We propose a method to evaluate how a document is similar but different from the query documents, and a method to obtain documents using conventional Web search engine that may be suitable for similar but different documents.

Key words　Web Search, Query by Examples, Category-similarity, Content-difference

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り，Google [1]，Yahoo! [2]，AltaVista [3] などが有名である．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，まだ知らない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．

以降，2. 節では似て非なる文書とはいかなるものであるかについて論じ，本研究における似て非なる文書の位置づけを行う．3. 節ではユーザがシステムに渡すクエリに求められる要件と，そのクエリに対してある文書が似て非なる文書として適合するかどうか評価する手法について述べる．4. 節では既存の検索エンジンを用いてクエリに適する文書の候補を取得する手法について述べる．5. 節では関連研究について述べる．6. 節ではまとめと今後の課題について述べる．

2. 似て非なる文書

2.1 種々の似て非なる文書

本研究では似て非なる文書の検索を行うが，一言で似て非なる文書と言ってもいくつもの定義が考えられる．ここではまず，いくつか考えられる似て非なる文書の定義について論じ，その上で，本研究で対象とする似て非なる文書の定義がどのようなものかということについて述べる．

文書の類似性には様々な観点があると考えることができる．似て非なる文書とは，ある観点からは類似性があり，別の観点からは類似性がないような文書どうしが似て非なる文書と言うことができる．ここでは文書の類似性として，

- 内容の類似性
- カテゴリの類似性
- 属性の類似性

の 3 つについて考えてみる．

（1） 内容の類似性

内容の類似性とは，実際に文書に記述されている内容がどの程度似ているかということである．ベクトル空間モデルにおいて，文書は特徴ベクトルとして表現され，文書間の類似度が測定可能であるが，これは文書の内容の類似性を求めていると考えられる．

（2） カテゴリの類似性

カテゴリの類似性とは，例えば，人物のプロフィールや製品に対する評判情報などのように，文書が所属するカテゴリの類似性である．例えば，織田信長の伝記と西郷隆盛の伝記は，伝記であるという分類カテゴリから見れば共通点があり，類似性を持っているといえる．

（3） 属性の類似性

属性の類似性とは，例えば,「対象読者」という属性がいずれも「子ども向け」というように一致 (類似) している場合などである．すなわち，文書の主題・カテゴリそのものではなく，文書が潜在的に持っている属性が似ていることによる類似性を意味する．そのような属性としては，文書が対象とする読者層や，用途などが考えられる．

2.2 似て非なる文書の具体例

ここでいくつかの具体例を挙げて，様々な似て非なる文書の定義が存在することを示す．

阪神タイガースに関する文書と中日ドラゴンズに関する文書は，カテゴリとしては野球のセリーグのチームに関する文書でありその意味では類似しており，それぞれ別々のチームに関する文書という意味では非類似であるということができる．この場合，カテゴリの類似性という観点からは類似していて，内容の類似性という観点からは相違している．

「桃太郎」と「カチカチ山」は，両方とも日本の昔話という観点からは類似しているが，話の内容は異なっている．この場合，属性の類似性という観点からは類似していて，内容の類似性という観点からは相違している．

「1 週間で分かる C 言語」などという書籍と,「プログラミング言語 C (B. W. Kernighan & D. M. Ritchie)」では，両方とも C 言語について述べているという観点からは類似しているが，前者は初心者向けであるのに対して後者は熟練者向けであるという違いがある．この場合，内容の類似性という観点からは類似していて，属性の類似性という観点からは相違している．

2.3 本研究が対象とする似て非なる文書

このように，似て非なる文書の定義がいくつか考えられる中で，本研究において対象とするのは，カテゴリの類似性という観点からは類似しており，内容の類似性という観点からは相違しているというような文書である．

ユーザは複数の文書をクエリとして与え，それらが共通して属するような概念的なカテゴリが考えられるならば，そのカテゴリに属していてクエリの文書とは内容的に異なるような文書を検索する手法について提案を行う．

3. クエリと適合性評価

3.1 クエリの要件

我々が対象とする「似て非なる文書」とは，カテゴリの類似性においては類似しており，内容の類似性においては非類似であるような文書である．そのような文書を検索するために，ユーザはクエリを与える必要がある．

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．例えば，新たに「ローヌワイン」に関する文書が得られたとして，類似している部分と相違している部分が

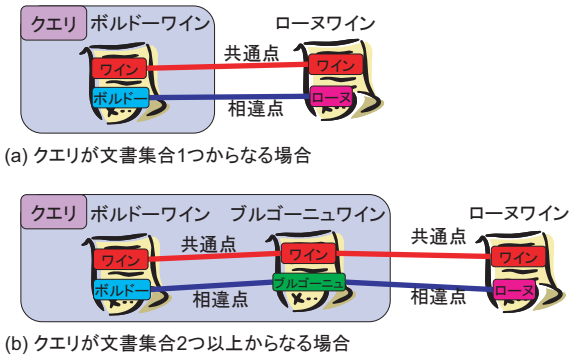

図 1　クエリが文書集合 1 つからなる場合と複数からなる場合

あることを判定できたとしても，どの部分が類似しているべきで，どの部分が相違しているべきかが判定できないため，似て非なる文書として適合かどうかを判断することは不可能である．

図中の (b) はクエリが複数の文書集合からなる場合である．「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．

文書はたいていいくつかの話題を含んでおり，いくつか概念的なカテゴリに属していると考えられる．上記の「ボルドーワイン」の場合以外にも,「京都の寺院」といった場合でも，この文書集合 1 つがクエリとして与えられた場合，京都との兄弟関係を考えて「奈良の寺院」に関する文書を適合とするべきか，寺院との兄弟関係を考えて「京都の神社」に関する文書を適合とするべきかは，判断することができない．

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．

3.2　特徴ベクトルの類似度による似て非なる文書の適合性評価

我々が求める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いくつかの文書集合において共通な部分は含んでおり，クエリのそれぞれの文書集合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である必要がある．

本研究では，まず，クエリとして与えられたいくつかの文書集合の共通部分と，それぞれの文書集合の特有部分を特徴ベクトルとして表現する．次に，何らかの方法で得られた文書の特徴ベクトルと，クエリの文書集合の共通部分や特有部分の特徴ベクトルとの類似度を基にして，その文書の似て非なる文書としての適合度を測定する手法について提案する．

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 $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ とする．次に，それらを用いて複数の文書集合の共通部分を特徴ベクトルとして表し，これを $\boldsymbol{c}$ とする．さらに，それぞれの文書集合の特徴ベクトル $\boldsymbol{t}_k (1 \leqq k \leqq n)$ と共通部分の特徴ベクトル $\boldsymbol{c}$ から，各文書集合における特有部分を特徴ベクトルとして表し，これを $\boldsymbol{u}_k$ とする．

何らかの方法によって似て非なる文書の候補となる文書が得られたとすると，それを特徴ベクトルとして表し $\boldsymbol{d}$ とする．その特徴ベクトルが，$\boldsymbol{c}$ と類似度が高く，全ての $\boldsymbol{u}_k$ との類似度が低いような場合に，得られた文書は与えられたクエリに対する似て非なる文書として適合していると判定することができると考えられる．

以下では，$\boldsymbol{t}_k$，$\boldsymbol{c}$，$\boldsymbol{u}_k$ の作成手法についていくつか提案し，それぞれについて検討を行う．

3.3　クエリの文書集合の表現

始めにクエリとして与えられたいくつかの文書集合それぞれに対して，文書集合そのもの特徴ベクトル $\boldsymbol{t}_k$ を生成する手法について述べる．

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．

また，TF に Inverse Document Frequency (IDF) を乗じて一般的な語の重みを低くするような手法も良く用いられる．しかし，今回は文書集合の共通部分として全体的に良く出現するような語を取得することが必要となるため，IDF は用いない．

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 $\boldsymbol{t}_k$ の値は，上記 (N) ならびに (L) それぞれの手法において下記のように表現される．

$$\text{(N)}\ \boldsymbol{t}_k(w_i) = \sum_{D_j \in P_k} tf(w_i, D_j) \tag{1}$$

$$\text{(L)}\ \boldsymbol{t}_k(w_i) = \log\left(1 + \sum_{D_j \in P_k} tf(w_i, D_j)\right) \tag{2}$$

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，何らかの正規化を行う必要がある．ここでは，ベクトルの最大の値を持つ要素の値で全ての要素の値を割ることによって，ベクトル内の最大の値を持つ要素の値を 1 とする正規化手法を用いる．すなわち，$\boldsymbol{t}_k$ の正規化された特徴ベクトル $\boldsymbol{t'}_k$ は，$\boldsymbol{t}_k$ において最大の値を持つ要素の語を w_{p_k} とすると，以下のように表現される．

$$\boldsymbol{t'}_k(w_i) = \frac{\boldsymbol{t}_k(w_i)}{\boldsymbol{t}_k(w_{p_k})} \tag{3}$$

3.4 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 $\boldsymbol{c}$ 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いてある程度の出現数があるような語を重視し，特定の文書集合のみでしか出現しないような語を軽視する，という事が必要である．ここでは，以下の 3 つの手法について検討する．

(M)　$\boldsymbol{t'}_k(w_i)$ の相乗平均 $(1 \leqq i \leqq n)$
(A)　$\boldsymbol{t'}_k(w_i)$ の相加平均 $(1 \leqq i \leqq n)$
(L)　$\boldsymbol{t'}_1(w_i), \cdots, \boldsymbol{t'}_n(w_i)$ のうちの最小値
n は文書集合の総数とする．

それぞれにおいて，語 w_i に対する文書集合の共通部分の特徴ベクトル $\boldsymbol{c}$ の値は，下記のように表現される．

$$\text{(M) } \boldsymbol{c}(w_i) = \sqrt[n]{\prod \boldsymbol{t'}_k(w_i)} \tag{4}$$

$$\text{(A) } \boldsymbol{c}(w_i) = \frac{\sum \boldsymbol{t'}_k(w_i)}{n} \tag{5}$$

$$\text{(L) } \boldsymbol{c}(w_i) = \min(\boldsymbol{t'}_1(w_i), \ldots, \boldsymbol{t'}_n(w_i)) \tag{6}$$

これらのうち，相乗平均の式 (4) ならびに最小値の式 (6) では，ある要素において文書集合の特徴ベクトル $\boldsymbol{t}_k$ のうちのどれか 1 つでも値が 0 となるような場合には，必ず $\boldsymbol{c}$ においても値が 0 となってしまう．例えば，ある 1 つの文書集合に含まれる文書の量が非常に小さく，ほとんどの要素の値が 0 であるような場合には，$\boldsymbol{c}$ のほとんどの要素の値も 0 になってしまうことが考えられる．そのような場合のために，α を定数として $\boldsymbol{t}_k$ を下記のように補正することが考えられる．

$$\boldsymbol{t^\dagger}_k = \boldsymbol{t}_k + \alpha \tag{7}$$

しかし，本稿では文書集合として与えられる文書群はある程度の大きさを持ったものとし，このような補正については考えない．

3.5 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を特徴ベクトル $\boldsymbol{u}_k$ を表現する手法について述べる．

文書集合 P_k そのものの特徴ベクトルは $\boldsymbol{t'}_k$ として表現されている．そこから共通部分の特徴ベクトル $\boldsymbol{c}$ を差し引くことによって文書集合 P_k に特有の部分が表せる．

文書集合 P_k の特有部分の特徴ベクトル $\boldsymbol{u}_k$ の各要素における値は，$\boldsymbol{t'}_k(w_i)$ から $\boldsymbol{c}(w_i)$ を引き算することによって作成する．ただし，0 以下となった要素は 0 とする．結果として，$\boldsymbol{u}_k$ は以下のような式となる．

$$\boldsymbol{u}_k(w_i) = \max\left(\boldsymbol{t'}_k(w_i) - \boldsymbol{c}(w_i), 0\right) \tag{8}$$

あとで例示するが，共通部分の特徴ベクトル $\boldsymbol{c}$ は手法によっては，文書集合そのものの特徴ベクトル $\boldsymbol{t'}_k$ よりもかなり小さなベクトルになってしまい，上記の式において計算された文書集合の特有部分の特徴ベクトル $\boldsymbol{u}_k$ が $\boldsymbol{t'}_k$ からあまり変化しない場合がある．そのようなときは，$\boldsymbol{c}$ を何らかの手法で正規化するなどして，下記の式のように補正を行ったほうがより文書集合の特有部分を表すことができるかもしれない．

$$\boldsymbol{u^\dagger}_k(w_i) = \max\left(\boldsymbol{t'}_k(w_i) - \beta \cdot \boldsymbol{c}(w_i), 0\right) \tag{9}$$

β は $\boldsymbol{c}$ の正規化手法による値であるが，本稿ではどのように補正すれば良くなるかについての精査は行わず，今後の課題とする．

3.6 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．本稿では，対象とする文書 D の特徴ベクトル $\boldsymbol{d}$ と

- 文書集合の共通部分の特徴ベクトル $\boldsymbol{c}$
- 文書集合の特有部分の特徴ベクトル $\boldsymbol{u}_k$

という 2 種類のベクトルとの類似度で，似て非なる文書としての適合度を定量化する．

まず，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るものとして，

(N)　単語の出現回数
(L)　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，下記のようになる．

$$\text{(N) } \boldsymbol{d}(w_i) = tf(w_i, D) \tag{10}$$

$$\text{(L) } \boldsymbol{d}(w_i) = \log\left(1 + tf(w_i, D)\right) \tag{11}$$

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 $\boldsymbol{v_1}$，$\boldsymbol{v_2}$ のコサイン類似度 Cos は以下の式によって表現される．

$$Cos(\boldsymbol{v_1}, \boldsymbol{v_2}) = \frac{\sum_w (\boldsymbol{v_1}(w) \cdot \boldsymbol{v_2}(w))}{\sqrt{\sum_w \boldsymbol{v_1}(w)^2 \cdot \sum_w \boldsymbol{v_2}(w)^2}} \tag{12}$$

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 $Sim_c(\boldsymbol{d})$ は以下のような式で表される．

$$Sim_c(\boldsymbol{d}) = Cos(\boldsymbol{c}, \boldsymbol{d}) \tag{13}$$

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，文書集合の個数だけ求められるが，似て非なる文書の候補となる文書がいずれかの文書集合と類似していると判断されると似て非なる文書とは考えられないため，求められた類似度の中で最大のものだけを評

表 1　文書集合の特徴ベクトル $\boldsymbol{t'}_k$ で値が大きな要素

「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	0.20	0.63	開催	0.36	0.77	馬	0.30	0.77
券	0.20	0.63	選手	0.34	0.75	記念	0.29	0.77
選手	0.20	0.63	自転車	0.20	0.64	有馬	0.28	0.76
レース	0.20	0.63	グランプリ	0.16	0.59	予想	0.17	0.66
予想	0.17	0.59	開設	0.15	0.58	投票	0.13	0.62
ギャンブル	0.12	0.52	投票	0.15	0.58	賞	0.12	0.60
軍資金	0.11	0.49	keirin	0.15	0.58	コラム	0.10	0.58
投票	0.11	0.49	競技	0.15	0.58	レース	0.10	0.58

価に用いればよい．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 $Sim_u(\boldsymbol{d})$ は以下のような式となる．

$$Sim_u(\boldsymbol{d}) = \max\left(Cos(\boldsymbol{u}_1, \boldsymbol{d}), \cdots, Cos(\boldsymbol{u}_n, \boldsymbol{d})\right) \qquad (14)$$

似て非なる文書として適している文書とは，$Sim_c(\boldsymbol{d})$ はより大きく，$Sim_u(\boldsymbol{d})$ はより小さいような文書であると考えられる．コサイン類似度は最小値が 0，最大値が 1 であることから，下記の式によって文書 D のクエリに対する似て非なる文書としての適合度 $R(\boldsymbol{d})$ を表す．

$$R(\boldsymbol{d}) = Sim_c(\boldsymbol{d}) \cdot (1 - Sim_u(\boldsymbol{d})) \qquad (15)$$

3.7　各手法の評価

ここまで，文書集合の特徴ベクトル $\boldsymbol{t}_k$ 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 $\boldsymbol{c}$ を作成する手法を 3 種，それぞれ挙げたため，それぞれの組み合わせにより 6 種の手法が存在することとなる．以降では，式 (1), 式 (2) と式 (4)〜(6) につけた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 $\boldsymbol{t}_k$ 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 $\boldsymbol{t}_k$ において (L) を用いた上で，$\boldsymbol{c}$ 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) 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，正解ページセット，不正解ページセットを設定した．その際，対象としたのは，http://dmoz.org/World/Japanese/ 以下のディレクトリである．また，特徴ベクトルがある程度の大きさになるように，取得したページの大きさが 5 キロバイト以上のファイルのみを対象とした．

まず，

- /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/

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，サブディレクトリからリンクされているページはクエリには含まなかった．

これらの文書集合の特徴ベクトル $\boldsymbol{t}_k$ を (N) (L) それぞれの手法で計算した時に，値が大きくなる要素の部分を表 1 に表した．各文書集合を表現するような語の値が大きくなっていることが分かる．

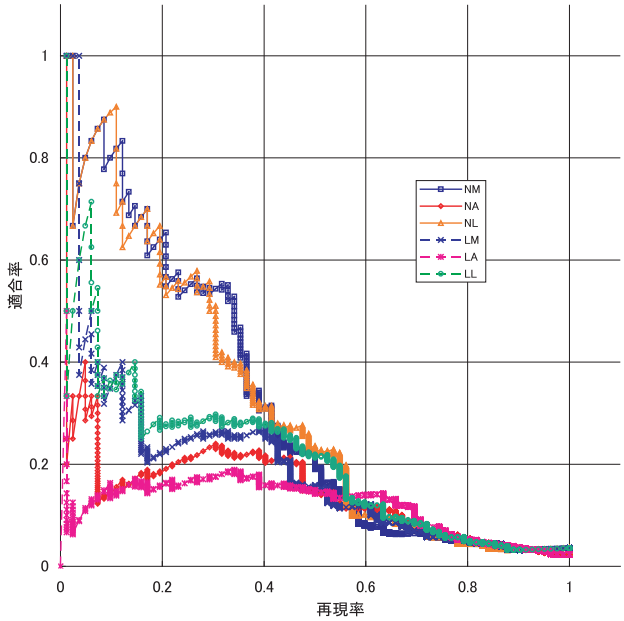

図 2　適合率–再現率グラフ

次に，これらの文書集合の共通部分の特徴ベクトル $\boldsymbol{c}$ を各手法で計算した時に，値が大きくなる要素の部分を表 2 に表した．(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．また，3.5 節で少し述べたように，ベクトルの大きさが小さくなる場合があることが分かる．

表 3 は，「競輪」の文書集合の特有部分の特徴ベクトル $\boldsymbol{u}_k$ を計算したときに，値が大きくなる要素の部分を表している．(NM) (NA) (NL) (LL) は，$\boldsymbol{t}_k$ 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) 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」などがより重要となっており，特有部分が強調されていることが明らかである．

これらの特徴ベクトルを用いて評価を行うのだが，評価において対象とするページは ODP において，

- /レクリエーション/

以下にある全てのページで，3638 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．よって，

- /レクリエーション/ギャンブル/

以下にある 201 文書のうち，

- /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/

以下にある 119 文書を除いた，82 文書とした．

6 つの手法それぞれにおいて，全ての文書の似て非なる文書としての適合度 R を計算して 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

表 2　文書集合「競艇」「競輪」「競馬」の共通部分の特徴ベクトル $\boldsymbol{c}$ で値が大きな要素

	(NM)		(NA)		(NL)		(LM)		(LA)		(LL)
予想	0.131	競馬	0.343	投票	0.106	予想	0.561	予想	0.569	投票	0.495
投票	0.137	競輪	0.338	予想	0.081	投票	0.561	投票	0.563	レース	0.451
レース	0.118	競艇	0.333	レース	0.081	レース	0.547	レース	0.552	予想	0.451
開催	0.088	選手	0.180	優勝	0.045	優勝	0.464	開催	0.507	優勝	0.330
優勝	0.082	開催	0.153	開催	0.030	開催	0.461	選手	0.505	開催	0.261
選手	0.070	記念	0.142	特集	0.027	選手	0.397	優勝	0.477	特集	0.254
年度	0.043	予想	0.138	バンク	0.015	年度	0.332	記念	0.441	戦	0.209
電話	0.041	投票	0.128	シリーズ	0.015	電話	0.327	競馬	0.420	競走	0.165

表 3　文書集合「競輪」の特有部分の特徴ベクトル $\boldsymbol{u}_k$ で値が大きな要素

	(NM)		(NA)		(NL)		(LM)		(LA)		(LL)
競輪	1.000	競輪	0.661	競輪	1.000	競輪	1.000	競輪	0.611	競輪	1.000
開催	0.293	開催	0.217	開催	0.349	自転車	0.642	自転車	0.428	自転車	0.642
選手	0.272	選手	0.170	選手	0.333	グランプリ	0.594	keirin	0.356	選手	0.623
自転車	0.202	自転車	0.135	自転車	0.203	競技	0.576	競技	0.340	開催	0.604
グランプリ	0.162	グランプリ	0.099	グランプリ	0.162	開設	0.576	開設	0.340	グランプリ	0.594
開設	0.149	競技	0.097	開設	0.149	記念	0.555	管内	0.340	開設	0.576
競技	0.149	開設	0.097	競技	0.149	keirin	0.533	周年	0.340	競技	0.576
記念	0.135	決定	0.083	記念	0.135	周年	0.509	選抜	0.300	記念	0.555

番目までの文書をシステムの出力とした時に，適合度と再現率を計算することが可能である．θ を 1 から 3638 まで変化させて得られる適合率と再現率の組を利用して，適合率–再現率グラフを描くことができる．

図 2 が 6 つの手法それぞれにおける適合率–再現率グラフである．このグラフにおいて，適合率がより高い位置を推移する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 2 つに比べると良いとは言えないことが分かる．

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，70%にあたる 14 件が正解文書であり，全対象ページにおいて約 2.3%のみが適合文書であることを考えると，良い結果と言うことができる．

以上より，クエリの文書集合並びに評価対象の文書は単純な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ような手法で作成すれば，ユーザのクエリに対して，ある文書が似て非なる文書としてどの程度適合しているかを測定することが可能である．

4. Web 検索エンジンを用いた適合候補文書の取得

4.1 概　　要

前章において，与えられたクエリに対してある文書が適合文書かどうかを判定する手法について述べた．我々が行おうとしているのは似て非なる文書を Web から得ることであるため，何らかの方法で似て非なる文書の候補となるような文書を取得する必要がある．本章ではその手法について述べる．

例えば，クエリに「阪神タイガース」に関する文書によって構成される文書集合が含まれているとすると，その文書集合を簡潔に説明する語としては,「阪神」や「タイガース」などが考えられる．また，同時にクエリに「中日ドラゴンズ」の文書集合が含まれているとすると，それらの文書集合をまとめ，さらに似て非なる文書として求める文書の内容も包含して説明するような語としては「野球」や「セリーグ」などが考えられる．また，似て非なる文書として適する文書の内容を簡潔に説明するような語としては,「広島」「巨人」「ベイスターズ」などが考えられる．

このような語をいくつか求めることができれば，それらを用いて既存の検索エンジンから Web ページを取得し，似て非なる文書の候補とすることができると考えられる．

ここではそのような，検索に利用できる語を役割に応じて以下の 3 つに分け，それぞれを求める方法を考える．

- **特有語**

クエリとして与えられた各文書集合を特徴づけるような語．上記の例における「阪神」「タイガース」など．

- **広域語**

クエリにおける全ての文書集合や似て非なる文書を包含して説明するような語．上記の例における「野球」「セリーグ」など．

- **対象語**

似て非なる文書の内容を簡潔に説明するような語．上記の例における「広島」「巨人」「ベイスターズ」など．

4.2 特有語の取得

各文書集合は 3.5 節で述べたように文書集合の特有部分の特徴ベクトルという形で表現されている．それらの特徴ベクトルにおいて大きい値を取るような語が文書集合を簡潔に説明する語であることが考えられる．

表 3 で「競輪」文書集合の例を示したが，いずれの手法でも

最も値が大きな要素は「競輪」という語であり，また，(LM) (LA) (LL) の 3 つの手法においては 2 番目に値が大きな要素が「自転車」になっており，文書集合を説明する語として適する語が得られているといえる.「競艇」「競馬」の文書集合でも同様であり，それぞれ,「競艇」「舟」と,「競馬」「馬」が上位 2 つの要素となっており，適した語が得られているといえる．

4.3 広域語の取得

ODP において,「競艇」「競輪」「競馬」の上位のディレクトリは「ギャンブル」であり,「ギャンブル」こそが広域語として適していると考えられる．表 2 の文書集合の共通部分の特徴ベクトルを見ると,「予想」や「投票」が大きな値を持っている．これらの語は「ギャンブル」という話題においてある程度共通して使われる語であると考えることも可能である．少なくとも，これらを検索キーワードとして用いて検索を行えば，結果にギャンブルに関するさまざまな文書が得られる事が予想される．以上より，文書集合の共通部分の特徴ベクトルにおいて高い値を持つ語は広域語として利用できると考えられる．

しかし,「ギャンブル」のように，今回の文書集合の共通部分の特徴ベクトルにおいて高い値を持たないような語でも広域語として適している語が存在している可能性がある．そのような語を，すでに得られている特有語を用いて取得する手法について以下で提案する．

小山ら [5] は任意の語が与えられたときに，その語に対してより詳細な語を Web から発見する手法を提案している．提案手法では，ある語 A に対して

$$O(A,B) = \frac{DF(intitle(A) \wedge B)}{DF(intitle(A))} - \frac{DF(A \wedge B)}{DF(A)} > 0 \quad (16)$$

が有意に成り立つような語 B が見つかれば，語 B は語 A の詳細語である言えるとしている．ただし，$DF(A)$ は Web において語 A を含むページの総数であり，$intitle(A)$ は語 A を文書のタイトルに含むという条件である．これらは検索エンジンに条件を与えて求められる検索総数を用いている．小山らはこの手法によって任意の語（以後，被説明語とする）の詳細語を求めているが，我々はこの手法を，いくつかの詳細語が与えられたときにそれらを詳細語とするような語を求めるために用いることにする．

広域語に対する特有語の関係は，小山らによる手法の被説明後に対する詳細語の関係であると考えられる．まず，特有語に対する広域語の候補を求める必要がある．式**??**fml:oyama) の定義より，広域語はタイトルに頻出する語であると考えられる．そこで，特有語をクエリとして検索エンジンの結果を求め，タイトルに頻出する語を取得する．このとき，同一文書集合の特有語は OR 条件とし，各文書集合ごとは AND 条件とする．すなわち,「競艇」「競輪」「競馬」の例では，各文書集合における特有語を 2 語用いるなら「(競艇 ∨ 舟) ∧ (競輪 ∨ 自転車) ∧ (競馬 ∨ 馬)」などが，1 語のみを用いるなら「競艇 ∧ 競輪 ∧ 競馬」がクエリとなる．これらをクエリとして検索エンジンから結果を求める．ここでは 1 語のみを用いることとする．

そこで得られた広域語の候補に対して式 (16) を用いて広域語として適しているかどうかを測定する．式において A は広域語の候補となる語である．B は広域語の候補を得るために用いた特有語のクエリそのものとする．すなわち,「競艇 ∧ 競輪 ∧ 競馬」である．式によって得られた値が大きい方がより広域語としてふさわしい可能性が高いといえる．本来は統計的に有意かどうかを調べるのだが，簡易的には値が 0 以上となったときに広域語として適していると考えることができる．

表 4　Google の検索結果から得られた広域語の候補

広域語の候補	タイトルでの出現回数	式 (16) の値
ギャンブル	21	0.199
カテゴリ	6	-0.00294
ソフト	5	-0.00335
趣味	7	-0.00936
公営	4	-0.0111
スロット	4	-0.0129
宝くじ	6	-0.0167
パチンコ	15	-0.0295
オートレース	6	-0.309

Google で行った場合の上位 100 件の検索結果のタイトルに頻出する語とそれらの語に対する式 (16) の値を表したのが表 4 である．ただし，検索に用いた語は評価の対象から除外している．タイトルに含まれている回数においても「ギャンブル」という語が最頻出語であったが，小山らの手法を利用すればよりはっきりと「ギャンブル」が広域語として適していることがわかる．

この手法で広域語が求められない場合もある．典型的な失敗例の 1 つは，与えられた文書集合の内容が近いような場合で,「ラグビーチーム」「学生ラグビー」「ラグビー協会・団体」というような文書集合が与えられたときには特有語の抽出がうまく行えず，広域語が得られなかった．

別の失敗例としては，広域語として適するような端的な語が存在しない場合である．例えば,「農業」「林業」「漁業」というような文書集合に対しては適する広域語はなかなか思いつかず，すなわち Web ページにおいてもこれらの語を含むページのタイトルに頻出するような語は考えづらく，実際にも存在しない．

しかし，これらの場合にも文書集合の共通部分の特徴ベクトルには広域的な語が含まれているか，検索エンジンに対してクエリとして用いる際にいくつも組み合わせる事もできるため，最低限の広域語は得ることができると言える．

以上が，文書集合の共通部分の特徴ベクトルや，小山らの手法を用いることによる，広域語を求め方である．

4.4 対象語の取得

似て非なる文書を直接説明するような語を求めることができれば，その語を検索のクエリとして用いることで求める文書の候補が簡単に求められる．前節の広域語の取得において特有語のクエリによって検索を行ったが，対象語はその際に特有語の周辺に現れる可能性が高いと考えられる．検索エンジンが出力する結果には検索語周辺の文書が含まれており，それらに頻出する語が対象語の候補となる．対象語の候補となる語が実際に対象語として適しているかどうかは，広域語が求められていれば式 (16) によって評価することが可能である．

表 5　Google の検索結果から得られた対象語の候補

対象語の候補	検索語周辺での出現回数	式 (16) の値
オートレース	21	0.169
レース	14	0.0682
予想	50	-0.0490
宝くじ	14	-0.0558
パチンコ	26	-0.0852

「ギャンブル」の例の場合，Google の検索結果の Snippet における出現回数と，式 (16) の値は表 5 のようになる．対象語として適切な語としては「オートレース」「レース」が得られた．この結果は，決してクエリに対する全ての似て非なる文書を網羅して説明する語ではないが，一部に対して適している語であるのは確かである．そのため，似て非なる文書の候補を Web 検索エンジンを用いて取得する際に，何回か別のクエリを用いることとするならば，いくつかではこれらの対象語が利用できると考えられる．

4.5　検索エンジンへのクエリの生成

特有語，広域語，対象語を得ることができれば，それらから検索エンジンに対するクエリを生成することができる．似て非なる文書として適する文書では，特有語はなるべく出現せず，広域語と対象語が出現すると考えられる．対象語はいずれもあまり出現しない方が良いため，各対象語が出現しないという条件は AND で結合するのが良いと考えられる．すなわち,「¬ 競艇 ∧ ¬ 競輪 ∧ ¬ 競馬」といった条件になる．広域語はいずれもなるべく出現した方が良いと考えられるが，出現しなくても構わないので各広域語どうしは OR 条件となる．すなわち,「予想 ∨ 投票 ∨ ギャンブル」といった条件となる．対象語もやはりいずれかが出現すれば良いので,「オートレース ∨ レース」といった条件になる．これらのうちいくつかを AND で結合したものを既存の検索エンジンにクエリとして渡して結果を得ることによって，似て非なる文書の候補とすることができる．それらを 3.6 節で述べた評価方法で評価することで，ユーザが与えたクエリに対する似て非なる文書を返すことが可能となる．

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 [6] 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 [7] 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適合していると判断されるような検索キーワードを考えるという作業が必要になっている．本研究ではシステム内では既存の検索エンジンを用いているが，ユーザがクエリとするのは保有するいくつかの文書であり，より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おり，例えば，Ishikawa ら [8] による MindReader は，いくつかの例とそれぞれの例の適合度をクエリとする画像検索システムである．

6.　まとめと今後の課題

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．また，似て非なる文書の候補となる文書を既存の Web 検索エンジンを用いて取得するためにクエリを生成する手法についても提案を行った．それぞれにおいて実験を行い，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 17 年度科研費特定領域研究 (2) 「Web の意味構造発見に基づく新しい Web 検索サービス方式に関する研究」（課題番号：16016247，代表：田中克己），および，平成 17 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．

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] S. Oyama and K. Tanaka: “Query modification by discovering topic from web page structures.”, Proceedings of APWeb 2004, pp. 553–564 (2004).
[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”, pp. 218–227 (1998).