Webディレクトリの階層構造を利用した検索対象文書の分野推定に基づいた言語横断情報検索

木村文則[†1] 前田亮[†2] 波多野賢治[†1]
宮崎純[†3] 植村俊亮[†4]

本論文では，Web 文書の言語横断情報検索において，Web ディレクトリの階層構造を利用して問合せの検索対象分野の推定に基づいた検索手法を提案する．提案手法では，Yahoo! カテゴリのような複数の言語版を持つ Web ディレクトリを問合せ翻訳における訳語の曖昧性解消のための言語資源として利用し，Web ディレクトリの下層のカテゴリを上位のカテゴリに統合したうえで，利用者が入力した問合せ語群から検索対象分野の範囲を推定することで，問合せ語群の最適な訳語に翻訳することにより，言語横断情報検索を行う．評価実験では，Web 文書の言語横断情報検索に適切なカテゴリ統合度がどの程度であるのか検証し，提案した検索対象分野の推定の有効性を検証した．

Cross-Language Information Retrieval by Estimation of Domain Using Web Directory Structure

Fuminori Kimura,[†1] Akira Maeda,[†2] Kenji Hatano,[†1]
Jun Miyazaki[†3] and Shunsuke Uemura[†4]

In this paper, we propose a cross-language information retrieval (CLIR) method based on an estimate of query domain related with search results using hierarchic structures of Web directories. To get the most appropriate translation of the queries, we utilize the Web directories written in many different languages as multilingual corpus for disambiguating translation of the query and estimate a domain of search results using hierarchical structures of Web directories. Experimental evaluations showed that we could have an advantage in retrieval accuracy using our proposal for disambiguating translation in CLIR system.

1. はじめに

1990 年代以降インターネットが急速に普及し，膨大な情報がインターネット上で発信されるようになった．また，インターネットの量的な普及だけでなく，国際的な普及も急速に進んでいる．インターネットを利用しているユーザの使用言語の調査結果[⋆1]によると，1990 年代に過半数を超えていた英語を母国語とするユーザの割合が，現在では 30%を切っている状態であり，欧州やアジア各国の言語の台頭が目立っている．しかしながら，すべてのユーザが容易に情報にアクセスし，多くの恩恵を受けるわけではない．なぜなら，多くのユーザは母国語以外の言語に精通していないからである．このようなユーザにとって，母国語以外の言語で記述された情報を理解することを始め，その中からそのユーザが必要とする情報を探し出すことも困難である．たとえば，台湾の人工衛星「ST1」についての記述がある Web ページ（NTCIR-3 CLIR タスク日本語検索課題 9）を英語版 Yahoo! で検索するといくつもの適合するページが検索されるが，日本語版 Yahoo! で検索すると適合文書は検索されない（2008 年 1 月現在）．このような場合，海外の情報または直接現地の情報に当たる必要がある．

多国語に精通しているわけではない一般のユーザが

†1 同志社大学文化情報学部
Faculty of Culture and Information Science, Doshisha University
†2 立命館大学情報理工学部
College of Information Science and Engineering, Ritsumeikan University
†3 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†4 奈良産業大学情報学部
Faculty of Informatics, Nara Sangyo University

⋆1 http://www.internetworldstats.com/stats7.htm 2006 年 12 月現在.

こういったインターネット上の情報を有効に活用するためには，いわゆる「言語の壁」を乗り越える必要がある．つまり，インターネット上で情報を発信する側が各国語版の Web ページを作成するなど，他国語を母国語とするユーザを想定した多言語化が必要となるが，多言語化に必要となる労力は膨大であり，また同時に情報を発信する側が各国語に精通していることが要求されるため，情報の多言語化は容易に実現できるものではない．

このような要求から，ある言語で書かれた文書群を別の言語による問合せで検索することを可能とする言語横断情報検索（Cross-Language Information Retrieval: CLIR）に関する研究が 1990 年代後半から活発に行われるようになった．言語横断情報検索に関する研究では，問合せの翻訳や訳語の曖昧性解消などにコーパスを利用する手法が提案されており[1),2)]，検索精度の向上において一定の成果が得られている．しかしコーパスを利用した手法では，学習に用いるコーパスが扱う分野に依存してしまうため，それ以外の分野の検索には精度が低くなる可能性がある．一般的に Web 文書の言語横断検索では，扱われる文書内容の話題は広範囲であるため，こうした問題の解決は必要不可欠である．こうした問題を解決するために，我々は Yahoo! カテゴリ[⋆1]のような複数の言語で類似の構造を持つ Web ディレクトリを利用する手法を提案してきたが[3)]，Web ディレクトリの最上位の構造だけを利用しているにすぎなかったため，訳語の曖昧性解消を十分に行うことができず，結果的に言語横断情報検索の性能を大幅に改善できたとはいえなかった．

そこで本論文では，Web ディレクトリの最上位の構造だけでなく，その下層にある構造を利用して，あるカテゴリで扱われている内容を複数のサブカテゴリに細分化したうえで，ユーザが入力した問合せ語群から検索対象となる Web 文書のサブカテゴリを推定し，問合せ語群の最適な訳語を得ることで，精度の高い言語横断情報検索の実現を目指す．

2. 関 連 研 究

言語横断情報検索では，問合せと検索対象文書との間の言語的な相違を吸収し，単言語検索へと帰着させることが主要な課題である．言語横断情報検索に用いられる手法は大きく分けて 3 つあり，検索対象の文書群を翻訳する方式，言語に依存しない中間言語を用いる方式，そして問合せを翻訳する方式であるといわれている．

まず，検索対象の文書群を翻訳する方式[4)] は，既存の機械翻訳システムを用いることができ，文脈を考慮できることにより訳語の曖昧性も低くなることから，一般に問合せを翻訳する方式より高い検索精度が得られるとされている．しかしながら，大規模な文書群をすべてあらかじめ翻訳しておくことは現実的ではなく，対応言語の拡張も困難であるため，Web のように多言語が混在し，かつ大規模で更新が頻繁な文書群の検索には不向きである．この問題を改善するために，文書群を翻訳する方式と問合せを翻訳する方式を組み合わせた手法[5)] が提案されている．

また，言語に依存しない中間言語を用いる方式[6)–8)] では，シソーラスの意味クラスや Latent Semantic Index（LSI）などを用いることが多い．この方式では，言語の違いを意識することなく処理することが可能であるが，学習に用いるコーパスの規模が大きくなると計算量が膨大となるため，大規模な文書群に対しては実現は困難である．

一方，問合せを翻訳する方式は，文字どおりユーザによって入力される検索キーワードを翻訳する方法である．しかし，Web 検索エンジンのログを分析すると分かるように，一般ユーザが入力する問合せキーワード数は平均 2 語程度と少なく[9)]，また単語の羅列である場合が多いため，訳語の曖昧性の解消をいかに正確に行うかが問題となってくる．訳語の曖昧性解消は，一般にまず対訳辞書を用いて問合せを翻訳し，訳語の共起情報などを利用して行われる．この方式は，翻訳された問合せを既存の単言語検索エンジンでそのまま用いることができるという利点があるため，他の手法と比べて実現が比較的容易であることから，言語横断情報検索ではこの方式が用いられることが多い．その流れを受け，訳語の曖昧性解消を効果的に行うためにコーパスを用いる手法の提案[2),10),11)] が行われたり，コーパスを用いる手法の問題の 1 つであるコーパスの扱う分野の問題を解決するための提案[1),12)] がなされたりしている．また，対訳辞書やコーパスなどの言語資源が十分にそろっていない言語間においては，仲介言語を介する手法[13),14)] などが提案されている．

訳語の曖昧性解消を効果的に行うためには，大規模かつ扱う分野が一致した複数のコーパスが必要であるといわれている．しかし，本論文が対象とする言語横断情報検索の技術を利用した Web 文書検索のようなアプリケーションを想定した場合，さまざまな分野のコーパスを複数個用意することは現実的ではない．そのため本論文では，Yahoo! カテゴリのような複数の

⋆1 http://dir.yahoo.co.jp/

言語で作成された Web ディレクトリに登録されている文書群をコーパスとして用いる．Web ディレクトリには多種多様な分野の Web 文書が登録されているため，Web ディレクトリをコーパスとして用いることは，現存するほとんどの分野に対応したコーパスを利用することに等しいといえる．また，Web ディレクトリによって問合せが対象としている分野を限定することができるので，問合せが対象とする分野と一致したコーパスを利用でき，訳語の曖昧性解消を効果的に行うことができる．以上により，問合せが対象とする分野に依存しない言語横断情報検索システムの実現が可能となる．

3. Web ディレクトリのカテゴリを利用した訳語の曖昧性解消

本論文で提案する手法は，問合せが対象としている分野を Web ディレクトリを用いることで推定し，その推定された分野において適切な訳語を選択することで訳語の曖昧性解消を行うというものである．問合せが対象としている分野の推定は，問合せ語の共起情報を基に，対象としていると推定される Web ディレクトリのカテゴリを 1 つあるいは複数選択することで行う．問合せ語と各カテゴリとの比較は問合せを翻訳する前の言語において行っている．一方，訳語の選択では，対訳辞書から得られた問合せ語の訳語候補から，それぞれのカテゴリから抽出された統計情報を用いてその分野に適切である訳語を絞り込んでいく．そのため，それぞれのカテゴリにおいてその分野に適切な統計情報を，カテゴリに属する Web 文書から抽出する必要がある．前者の場合はカテゴリの選択が，後者の場合は訳語の選択がうまく働かなければ効果的に訳語の曖昧性解消を行うことはできない．

本章では曖昧性解消を効果的に行ううえで，従来研究がかかえている問題を明確にし，その問題点を解消するための提案手法について述べる．

3.1 提案手法の概略

図 1 は提案手法を実装した言語横断情報検索システムの概略である．本システムは，問合せおよび検索対象のそれぞれと同じ言語版の Web ディレクトリ，それぞれの言語の特徴語データベース，対訳辞書，検索対象となる文書群から構成されている．図 1 において点線で囲まれている部分は，問合せの翻訳処理を行うモジュール部分である．

本システムの問合せ処理は，大きく 2 つに分けることができ，1 つは Web ディレクトリの各カテゴリから特徴語を抽出してそれを特徴語データベースに事前に格納しておく前処理部分と，与えられた問合せを翻訳して実際に検索を行う検索処理部分である．

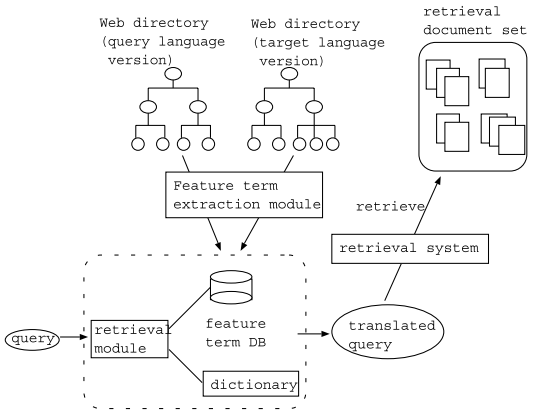

図 1 言語横断情報検索システムの概要
Fig. 1 Outline of our CLIR system.

3.2 先行研究の問題点

文献 3) で提案されている方法は，Yahoo! カテゴリで構築されている Web ディレクトリの最上位の構造，すなわち 13 のカテゴリから問合せに関連のあるカテゴリを 1 つ選択し，そのカテゴリ以下に含まれているすべての Web 文書から抽出された統計情報を用いて，問合せに適切な訳語を選択している．この場合，Web ディレクトリに登録されている Web 文書数は膨大であるため，それらを 13 のカテゴリに分類したとしても，1 つのカテゴリに属している Web 文書数は大量となる．すなわち，問合せに適合すると判断されたカテゴリの選択をうまく行えたとしても，大量の Web 文書がそのカテゴリに属しているため，訳語の選択の際に利用される分野情報の特定がうまく行えず，問合せに適切な訳語を選択できない可能性がある．

3.3 提 案 手 法

3.2 節に述べたように，先行研究の問題は 1 つのカテゴリが対象とする分野が広く，適切な訳語を選択できなかったためである．したがって，Web ディレクトリの構造を最上位だけではなくより低い階層まで利用し，検索対象となる Web 文書の分野を限定することで，問題を解決することにした．

Web ディレクトリの低い階層のカテゴリを利用することで，検索対象となる Web 文書の分野を限定することができる．しかし，低い階層のカテゴリが対象とする Web 文書も少なくなり，そのカテゴリから抽出される統計情報も少なくなるという問題が生じる．これは，そういったカテゴリでは得られる特徴語も少なくなることを意味する．カテゴリの特徴語数が少ないと，問合せ語の訳語候補がそのカテゴリに存在しな

いなどの理由により，適切な訳語を抽出できないという問題が起こりうる．この問題への対処のために，サブカテゴリに属する Web 文書から抽出される特徴語数が n 語に満たないものは，分野推定の対象から除外している．

3.4 提案手法の構成

3.4.1 前 処 理

本システムでは，異なる言語で構築された複数の Web ディレクトリの利用を想定しており，利用者の問合せ時に利用する Web ディレクトリを「問合せ言語版」，利用者が検索対象としている他言語の Web ディレクトリを「検索対象言語版」と呼ぶことにする．

図 2 は，実装した言語横断情報検索システムの前処理のフローを示したものである．前処理では，Web ディレクトリのカテゴリにおいて，事前に特徴語の抽出と異言語のカテゴリとの対応付けを行っている．以下に，前処理の手順を示す．

(1) 特徴語の抽出
問合せ言語版および検索対象言語版の Web ディレクトリのすべてのカテゴリに対して以下の処理を行う．
- (a) あるカテゴリに属する Web 文書から単語を抽出し，各単語に対し重み付けを行う．
- (b) 重みの高い上位 n 語の単語をそのカテゴリの特徴語として抽出する．
- (c) 抽出された特徴語を特徴語データベースに格納する．

(2) それぞれの Web ディレクトリ間でカテゴリの対応付け
それぞれの Web ディレクトリに属するすべてのカテゴリに対して，互いに対応するカテゴリを推定し，対応付ける．

図 2 を用いて説明すると，図中 (1)(a) において問合せ言語版のカテゴリ a に対する対応付けを行うために，まずカテゴリ a に属する文書群から単語を抽出し，それらのカテゴリ a における重みを計算し，次に (1)(b) で抽出された単語のうちから重みの高いものから n 語を特徴語として抽出し，特徴語集合 f_a を得ている．こうして得られた特徴語集合 f_a を (1)(c) で特徴語データベースに格納しているわけである．最終的に，(2) で得られた特徴語集合 f_a に最も類似していると思われる検索対象言語版のカテゴリを推定し，カテゴリ a と推定されたカテゴリ間に対応付けを行う．

なお，対応付けの方法にはさまざまな方法があり，たとえば，カテゴリの特徴語を比較することにより対応付けや人が直接対応付けを行うことができる．こうして得られたカテゴリ間の関連は，実際に他の言語で書かれた文書の検索を行うときに利用される．

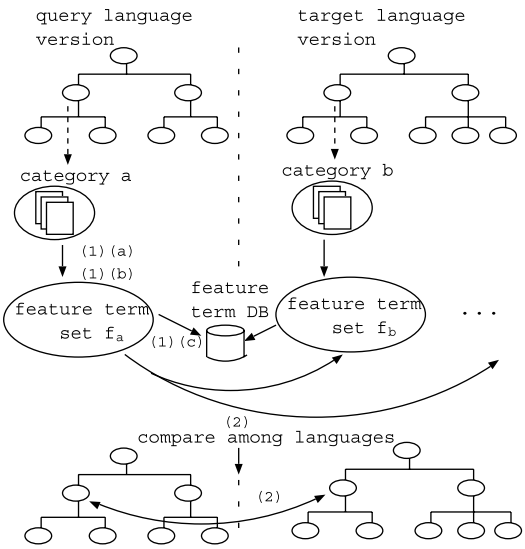

図 2 前処理における処理の流れ
Fig. 2 Flow of preprocessing.

ここで，(1)(a) において行われる特徴語の抽出について，その処理手順の詳細を説明する．

Web ディレクトリの各カテゴリの特徴は，そのカテゴリに属している Web 文書から抽出される単語集合で表現することができる．したがって，特徴語を抽出するために，まず各カテゴリに属する Web 文書から単語を抽出し，抽出した単語をカテゴリごとに集計し重みを計算している．本論文では，Web 文書から抽出された単語のうち，重みが大きいものをそのカテゴリの特徴語としている．

Web 文書から抽出された単語の重み付けには，Term Frequency · Inverse Category Frequency (TF·ICF)[15] を用いて計算する．これは，文書検索の研究でしばしば用いられる単語の重み付けの手法である Term Frequency · Inverse Document Frequency (TF·IDF) を応用したものである．TF·IDF は単語の出現頻度 (TF) と文書頻度の逆数 (IDF) との積により求められるため，TF·IDF は網羅性と特定性がともに高い単語の重みが大きくなるようになっている[⋆1]が，TF·ICF では文書単位ではなくカテゴリ単位で特定性の計算を行うため，本論文で扱うようなカテゴリの重み付けの計算には有用であるとされている．

3.4.2 検 索 処 理

本システムにおける検索処理の流れを**図 3** に示す．

⋆1 TF は単語の網羅性を表し，IDF は単語の特定性を表している．

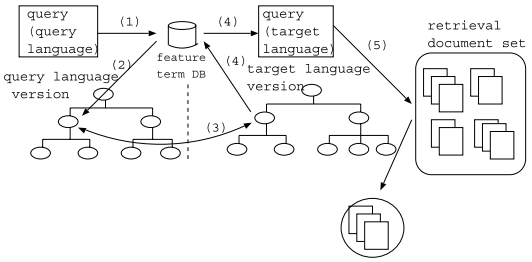

図 3 検索処理の流れ
Fig. 3 Flow of retrieval.

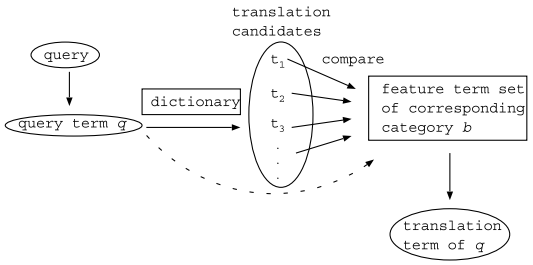

図 4 問合せの訳語の決定
Fig. 4 Translation of feature terms.

まず，問合せの適合カテゴリを選択し，続いて適合カテゴリに対応づけられている異言語のカテゴリを選択する．さらに，そのカテゴリの特徴語集合を利用して問合せの翻訳を行い，最後に翻訳された問合せを用いて検索処理が行われるのである．

検索処理の手順は次のようになる．

(1) 問合せ言語版のすべてのカテゴリに対して，問合せとカテゴリに関連づけられた特徴語集合との適合度を計算する．
(2) 最も適合度の高いカテゴリを問合せの適合カテゴリと決定する．
(3) 検索対象言語版のカテゴリから，適合カテゴリに対応づけられているカテゴリを選択する．
(4) 選択された対応カテゴリの特徴語集合を利用して問合せを翻訳する．
(5) 翻訳された問合せにより，検索対象の文書群を検索する．

3.4.3 問合せの適合カテゴリの選択

本システムにおける問合せは自然言語で表現されるのではなく，数語の単語から構成されていることを前提としている．ここで，単語 $t_1, t_2, \ldots, t_n$ の単語から構成される問合せ q に対する問合せベクトル $\vec{q}$ を次のように定義する．

$$\vec{q} = (q_1, q_2, \ldots, q_n) \quad (1)$$

なお，q_k は問合せの k 番目の単語 t_k に対応しており，その値は 1 である．

与えられた問合せ $\vec{q}$ に対して，まず問合せ言語版を用いて，問合せと各カテゴリとの適合度を計算し（図 3 (1) 参照），そのうち最も適合度が高くなるカテゴリを，問合せが適合するカテゴリと決定する（図 3 (2) 参照）．問合せとカテゴリの適合度は，問合せベクトルとカテゴリの特徴語集合のベクトルの内積にこの 2 ベクトルのコサイン距離を掛けることにより計算することができる．本論文では，カテゴリ c の特徴語集合のベクトル $\vec{c}$ を，単語 t_k のカテゴリ c における特徴語の重み w_k を用いて，

$$\vec{c} = (w_1, w_2, \ldots, w_n) \quad (2)$$

と定義しているため，問合せとカテゴリの適合度 $rel(q, c)$ は次のように表される．

$$rel(q, c) = \frac{\vec{q} \cdot \vec{c}}{|\vec{q}| \cdot |\vec{c}|} \quad (3)$$

この適合度 $rel(q, c)$ は，問合せベクトルとカテゴリの特徴語集合のベクトルの 2 つのベクトルのコサイン距離である．こうして求めた適合度が最も高いカテゴリを，問合せに対する適合カテゴリとする．本論文では適合カテゴリを 1 つだけ選択したが，適合度が閾値以上となるカテゴリを適合カテゴリとする方法も考えられる．このとき適合度が閾値以上となるカテゴリが複数ある場合は，これらをすべて適合カテゴリとして選択する．

次に，前処理で得られた対応付けからそのカテゴリに対応する異言語のカテゴリが決まる（図 3 (3) 参照）．こうして得られた異言語のカテゴリの特徴語集合および対訳辞書を利用して，次節で述べる方法により問合せを翻訳する（図 3 (4) 参照）．こうして得られた問合せを用いて検索対象文書に対して検索を行う．以上の処理を経て得られた文書群が検索結果となる（図 3 (5) 参照）．

3.4.4 問合せの翻訳

問合せの翻訳の流れを**図 4** に示す．まず，問合せ q 中の各単語に対する対訳辞書のすべての訳語 $t_1, t_2, t_3, \ldots$ を，訳語の候補として抽出する．抽出されたすべての訳語候補について，適合カテゴリに対応づけられている異言語のカテゴリ（以下，対応カテゴリと呼ぶ）b の特徴語に含まれているかを調べる．適合カテゴリの決定およびその対応カテゴリの決定方法については，3.4.3 項において述べた方法で行う．含まれていた訳語のうち，特徴語の重みが最も大きい訳語を，その問合せ語の訳語と決定する．このとき，対応カテゴリの特徴語集合の中にいずれの訳語候補も存在しない場合，その問合せ語は使用しない．しかし，たとえば，日本語で書かれた Web 文書中において英単語が使われる

といったことも頻繁にあるため，翻訳を行わないほうが良い場合もある．そこで，いずれの訳語候補も比較している対応カテゴリの特徴語に含まれていない場合，翻訳する前の問合せの単語そのものが，比較している対応カテゴリの特徴語に含まれているかを調べる（図 4 中の点線部参照）．もし含まれていれば，翻訳前の単語そのものをこの問合せ語の訳語と見なす．たとえば，英語のカテゴリ “Computers and Internet” が問合せの適合カテゴリであるときに英語の “system” という単語の訳語を決定する場合を考える．“system” の訳語の候補として，「宇宙」，「方法」，「組織」，「器官」，「システム」，「系統」などが得られる．この訳語の候補のすべてに対して，適合カテゴリの対応カテゴリである日本語のカテゴリ「コンピュータとインターネット」の特徴語集合に存在するどうかを調べる．そのうち重みが最も高いもの，今回は「システム」を英単語 “system” の訳語と決定する．もし，“system” のいずれの訳語候補も対応カテゴリの特徴語集合に存在しない場合は，“system” という単語そのものが対応カテゴリの特徴語集合に存在するか調べ，存在していれば “system” という単語そのものを訳語と見なす．

3.4.5 文書の検索

前項で述べた手法により翻訳された問合せを用いて検索対象文書群に対して検索を行う．検索対象文書群は，必ずしも Web ディレクトリに登録されている文書でなくてもよい．検索システムは既存のシステムを使用することができる．こうして得られた文書群が，問合せに対する検索結果となる．

4. 評価実験

本章では，3.3 節で提案した Web ディレクトリの階層構造を利用した訳語の曖昧性解消が，先行研究と比較して有用であることを示すための評価実験について述べる．

4.1 実験環境

本実験では，第 3 回 NTCIR ワークショップ[⋆1] の言語横断検索タスク（以下 NTCIR-3 CLIR と呼ぶ）で用いられた文書群と検索課題を用いて検索の実験を行った．このテストコレクションのうち，1998〜1999 年に台湾で発行された英字新聞各種からなる EIRB010，および同年に日本で発行された英字新聞である毎日デイリー 1998–1999 の 2 つを検索対象として用いた．また，このテストコレクションの日本語の検索課題を本実験における問合せとして用いた．NTCIR-3 CLIR の日本語検索課題は 50 の問合せが用意されており，このすべての問合せを用いて実験を行った．問合せには日本語問合せの TITLE フィールドを用い，これを提案手法により英語に翻訳し，英語の文書群に対して検索を行った．検索対象となった文書は 22,927 文書である．

提案手法は一般の Web 文書を検索対象とした言語横断情報検索であるため，NTCIR-3 の Web タスクをテストコレクションとして使用することも考えられる．しかし，Web タスクには言語横断情報検索用の正解集合が用意されていないため，言語横断情報検索の検索実験には使用できない．それゆえ NTCIR-3 CLIR テストコレクションを Web 文書群にみたてて検索実験を行った．

また，訳語の曖昧性解消のために用いる Web ディレクトリとして，Yahoo! カテゴリの英語版と日本語版を用いた．本実験では，英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ「地域情報」以下のカテゴリを除いたすべてのカテゴリから Web 文書を収集し，曖昧性解消に用いた．英語のトップレベルカテゴリ “Regional”，および日本語のトップレベルカテゴリ「地域情報」以下のカテゴリを除いたのは，これらのカテゴリには世界各地の地域に関する文書が属しているため，英語および日本語の翻訳に用いるのには適さないからである．たとえば日本語版の「地域情報」のカテゴリには，日本の各市町村ごとのカテゴリがあるが，これらのカテゴリが英語版のカテゴリに存在しないため，言語間でカテゴリの対応付けをとることはできない．それゆえ，本手法においてはこれらのカテゴリは利用することができない．

Web 文書から単語を抽出する際に，英語版では単語の活用形を原形にしたのち，ストップワードを取り除いた．ストップワードのリストは，文献 16) に掲載されている “A Stoplist for General Text” を用いた．また日本語では，英語のように単語の区切りが明確でないため，形態素解析ツールを用いてわかち書き処理を行う必要がある．本実験では茶筌[⋆2]を用いて，わかち書き処理を行った後，単語に分割し，名詞，動詞，形容詞，未知語を特徴語の候補として抽出した．さらに，問合せの翻訳のための対訳辞書には，EDR 電子化辞書の日英対訳辞書[⋆3]を用いた．単純に対訳辞書で翻訳した場合，1 単語に対して平均で 5.17 語の訳語

⋆1 http://research.nii.ac.jp/ntcir/ntcir-ws3/

⋆2 http://chasen.naist.jp/hiki/ChaSen/

⋆3 http://www2.crl.go.jp/kk/e416/EDR/J_index.html

表 1 各階層のカテゴリ数
Table 1 The number of categories in each level.

		1 階層	2 階層	3 階層	4 階層
英語	除外前	13	397	4066	8672
	除外後	13	255	644	292
日本語	除外前	13	391	2953	3259
	除外後	13	154	153	42

候補が得られた．なお，カテゴリの特徴語の抽出において，各カテゴリの特徴語数は 10,000 語とした．

表 1 は，本実験において用いた各階層のカテゴリ数を示している．英語，日本語とも 1 階層では 13 カテゴリであり，階層が深くなるに従って，カテゴリ数も増加している．また，階層が深くなるほどカテゴリは細分化されるため，1 つのカテゴリから得られる特徴語数は減少する．そのため，本システムにおいて利用するために十分な特徴語数が得られないカテゴリも存在する．こういったカテゴリが適合カテゴリとして選択されると問合せの翻訳が適切に行えないため，十分な特徴語数が得られないカテゴリは除外した．本実験では，特徴語が 10,000 語未満であるカテゴリを除外した．利用するカテゴリの特徴語数の下限を 10,000 語とした理由については 4.2 節において述べる．

言語間におけるカテゴリの対応付けは人手により行った．カテゴリの対応付けは，すべてのカテゴリに対して適合する他言語のカテゴリを決定する．そのため，各階層のカテゴリ数が多くなるほど対応付けのコストも大きくなる．また Yahoo! カテゴリでは，1 階層はどの言語においてもカテゴリ構造は同じである．しかし，2 階層以下は各言語版ごとに独自に構築しているため，階層が深くなるほどカテゴリの構造の差異も大きくなる．そのため，階層が深くなるほど対応付けが困難となる．カテゴリの対応付けのコストを見積もるために，試験的にカテゴリの対応付けを行ったところ，3 階層以上では 1 カテゴリを対応付けるのにかかる時間はおよそ 5 分程度であるが，4 階層になるとその倍程度の時間を要した．

しかし，Web ディレクトリのカテゴリ構造は固定的であるため，1 度カテゴリの対応付けを作成するとしばらくの間は使用可能である．また，特徴語数が 10,000 語以上のカテゴリ数は 1,000 カテゴリ以下であり，対応付けが不可能なカテゴリ量というわけではない．これらのことを考慮すると，人手による対応付けを行うことは実現可能である．

問合せを翻訳したあとで行う検索処理においては，独自でシステムを構築した．検索対象文書群の索引付けは，Augmented TF·IDF [17] と呼ばれる米国 Cornell 大学で開発された SMART Retrieval System[⋆1] の重み付け法を用いて行っている．

表 2 4 階層における特徴語数に対するカテゴリ数
Table 2 The number of categories in 4 level for feature terms.

特徴語数	3,000	5,000	10,000	除外前
英語	1185	674	292	8672
日本語	233	115	42	3259

表 3 4 階層における特徴語数の下限と平均適合率
Table 3 Average precision for the under bounce of feature terms in each 4 level categories.

特徴語数	3,000	5,000	10,000
平均適合率	0.0301	0.0278	0.0361

4.2 利用するカテゴリの特徴語数の下限

本実験では 4.1 節で述べたように，特徴語数が 10,000 語に満たないカテゴリは除外している．カテゴリを除外することにより，そのカテゴリが対象としている分野についての情報が失われてしまうため，本来であればカテゴリは除外しないほうがよい．しかし，十分に統計情報が得られないカテゴリによる悪影響も考慮しなければならない．よって，カテゴリを除外することと，特徴語が少ないこととのどちらの要因がより大きな影響を及ぼすかについて検討する必要がある．

表 2 は，特徴語数の下限に対して，それを満たす 4 階層のカテゴリ数について示している．表 2 に示した特徴語数の下限が 3,000 語，5,000 語，10,000 語の場合について，4.1 節で述べた検索実験を予備実験として行った．**表 3** はその結果を補間なし平均適合率によって示したものである．10,000 語の場合が 0.0361 と最も良い結果であり，3,000 語，5,000 語の場合よりも大きく上回る結果となった．このことは，十分に統計情報が得られる特徴語数を下限とすることが，対象となるカテゴリの数よりも重要であることを示している．以上より，本実験においても特徴語数の下限を 10,000 語とした．

4.3 実 験 結 果

表 4 は，提案手法において利用する Web ディレクトリのカテゴリの階層を 1 階層から 4 階層のいずれかを選択した場合について，検索実験を行った場合の補間なし平均適合率である．1 階層の場合は，Yahoo! カテゴリのトップページにリンクされているカテゴリを用い，2 階層の場合はさらにもう 1 階層下のカテゴリまでを用いている．また，1 階層，2 階層のいずれの場合においても，そのカテゴリの下位に属しているサ

⋆1 ftp://ftp.cs.cornell.edu/pub/smart/

表 4 各問合せごとの平均適合率
Table 4 Average precision about each query.

課題番号	1 階層	2 階層	3 階層	4 階層	2 + 3 階層	対訳辞書	Web 翻訳（WT）
2	0.0971	0.0870	0.1042	0.1048	0.0971	0.0270	0.1195
5	0.0027	0.0027	0.0029	0.0034	0.0034	0.0022	0.0020
9	0.0084	0.0086	0.0193	0.0000	0.0206	0.0146	0.0452
12	0.0001	0.0001	0.0009	0.0017	0.0009	0.0009	0.0075
13	0.0222	0.0222	0.0102	0.0222	0.0222	0.0067	0.0070
14	0.0059	0.0059	0.0075	0.0097	0.0075	0.0059	0.0166
18	0.1321	0.0961	0.1523	0.1523	0.1082	0.0951	0.0000
19	0.0056	0.0053	0.0052	0.0053	0.0053	0.0016	0.0017
20	0.1627	0.2048	0.2390	0.2390	0.2168	0.1313	0.2321
21	0.0245	0.0245	0.0281	0.0002	0.0281	0.0189	0.0495
23	0.1962	0.2059	0.2059	0.0001	0.0001	0.0000	0.0003
24	0.0003	0.0005	0.0008	0.0000	0.0000	0.0002	0.0080
26	0.0031	0.0004	0.0005	0.0023	0.0005	0.0018	0.0000
27	0.1640	0.2200	0.1640	0.2200	0.1640	0.0160	0.1870
28	0.0098	0.0126	0.0005	0.0005	0.0005	0.0001	0.0008
29	0.1596	0.1596	0.2332	0.2332	0.2332	0.2332	0.1988
31	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015
32	0.0087	0.0087	0.0165	0.0099	0.0099	0.0090	0.0359
33	0.0158	0.0158	0.0158	0.0158	0.0158	0.0084	0.0158
34	0.0040	0.0036	0.0016	0.0045	0.0040	0.0036	0.0037
35	0.0052	0.0057	0.0055	0.0123	0.0055	0.0044	0.0248
36	0.1078	0.1078	0.1078	0.0050	0.0050	0.0050	0.0059
37	0.0086	0.0228	0.0355	0.0111	0.0084	0.0084	0.0227
38	0.0057	0.0057	0.0072	0.0093	0.0057	0.0057	0.0082
39	0.0121	0.0141	0.0040	0.0040	0.0040	0.0035	0.0000
42	0.0016	0.0016	0.0008	0.0004	0.0005	0.0005	0.1200
43	0.0012	0.0012	0.0005	0.0004	0.0004	0.0001	0.0035
45	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0021
46	0.0179	0.0198	0.0000	0.0003	0.0002	0.0024	0.0000
50	0.0151	0.0212	0.0151	0.0151	0.0137	0.0137	0.0011
平均	0.0400	0.0429	0.0462	0.0361	0.0328	0.0203	0.0377

ブカテゴリはすべてそのカテゴリに属しているものとした．以下，3 階層，4 階層についても同様である．また，“2 + 3 階層” は，2 階層と 3 階層の両方を利用した場合である．さらに “対訳辞書” は，対訳辞書から得られる訳語候補をすべてその問合せ語の訳語とした場合についての結果である．“Web 翻訳” は，問合せを Web 翻訳エンジンにより翻訳して得られた訳語を用いて検索を行った結果である．本実験では，“Excite 翻訳[⋆1]” を Web 翻訳エンジンとして用いた．提案手法と “Excite 翻訳” では用いている辞書が異なっているため，単純に比較することはできないが，参考として “Web 翻訳” の結果も提示する．

4.4 考　　察

4.4.1 提案手法の有効性

表 4 から分かるように，対訳辞書のみの場合に対していずれの階層においても平均適合率は上回っている．また，Web 翻訳に対しても，3 階層以上では平均適合率は上回っていることからも，提案手法の有用性が示されている．ただし，4 階層においては Web 翻訳よりも平均適合率が下回る結果となった．

4.4.2 利用する階層と問合せ語の訳語

提案手法で Web ディレクトリの階層構造を利用して検索対象となる Web 文書の分野を限定することで，平均適合率に変化が見られる問合せが 30 個の問合せ中 11 個存在する．このうち，提案手法を適用することで平均適合率が高くなった問合せは 9 個，逆に低くなった問合せは 2 個であった．

表 5 は，1 階層と 2 階層で差異の見られた問合せにおいて，それぞれの階層の場合で選択した訳語のリストを示している．また，1 階層と 2 階層で選択された適合カテゴリについても示している．

平均適合率が高くなった問合せの訳語数に着目すると，訳語数が多くなった問合せは 3 個，少なくなった問合せは 6 個であった．訳語数が多くなった問合せでは，問 34（import → import，importation）や問 50（fashion → fashion，fashinable closes，vogue）

⋆1 http://odn.excite.co.jp/world/text

表 5 各問合せごとの訳語候補
Table 5 Translation list for each query.

課題番号	階層	訳語
9	(問合せ)	人工 衛星 ST1
	1 階層	human labor human skill artificial artificial heart artificial satellite satellite moon secondary dependency ST1 (適合カテゴリ) Health
	2 階層	human labor human skill artificial artificial heart artificial satellite satellite moon secondary ST1 (適合カテゴリ) Science/Space
19	(問合せ)	欧州 通貨 統合 経済的 影響
	1 階層	Europe currency money synthesis integration economic influence effect consequence (適合カテゴリ) Government
	2 階層	Europe currency money synthesis integration economic economical influence effect consequence (適合カテゴリ) Business_and_Economy /Finance_and_Investment
20	(問合せ)	日産 ルノー 資本 提携
	1 階層	Nissan Renault funds capital fund investment money joint business cooperation (適合カテゴリ) Computers_and_Internet
	2 階層	Nissan Renault capital fund investment money cooperation (適合カテゴリ) News/Business
28	(問合せ)	日本 北朝鮮 訪問 派遣
	1 階層	Japan North Korea visit call dispatch send (適合カテゴリ) Health
	2 階層	Japan North Korea visit call send (適合カテゴリ) News/Journalism
34	(問合せ)	米 輸入
	1 階層	rice meter American America import introduction (適合カテゴリ) Computers_and_Internet
	2 階層	rice meter American America import importation introduction (適合カテゴリ) Government/Military
50	(問合せ)	ティーンエイジャー ファッション
	1 階層	fashion mode style (適合カテゴリ) Government
	2 階層	fashionable clothes vogue fashion mode style (適合カテゴリ) Arts/Design_Arts/Fashion

のように派生語や類義語が多く得られる傾向があった．また，その逆の訳語数が少なくなった問合せでは，問 9 に見られるように誤訳（「衛星」の訳語として “dependency”）がなくなったり，問 20（joint business，cooperation → cooperation）のように，問合せが対象としている分野にふさわしい訳語を選択できるなどの効果が見てとれる．問 9 では，「人工衛星」について調べる質問であることから，「衛星」に対する適切な訳語は “satellite” である．それ以外にも，「衛星」という単語は「衛星国」という使われ方があるように，「ある国の属国」という意味も持っている．しかし問 9 においてはこの意味である “dependency” は訳語としてはふさわしくない．1 階層では “dependency” が訳語として選択されているが，2 階層では選択されなくなっており，対象を限定する効果が現れている．したがって，提案手法の Web ディレクトリの階層構造を利用することは，訳語数の変化につながり，結果的に平均適合率の向上につながったと考えることができる．

これに対し，平均適合率が低くなった 2 つの問合せの訳語に着目すると，問 19 の「経済的」という訳語が “economic” から，“economic” と “economical” の 2 語に増えていたり，問 28 では「派遣」という訳語 “dispatch” が欠落したりという現象が起こっている．前者は Web ディレクトリの階層構造を利用することによる検索対象 Web 文書の分野の限定がうまく働いた例，後者は分野の限定がうまく機能せず訳語を抽出できなかった例であるということができる．前者は正確な訳語を得ることとしては意図した結果となったのであるが，検索語としてはあまり適切でない訳語が選択される結果となった．従来手法では特徴語集合に含まれない訳語であるが，分野を限定することにより，その訳語の重要度が相対的に上がり，提案手法では訳語として選択されることとなったと考えられる．

一方，平均適合率に変化が見られなかった問合せも 20 存在していた．これは，Web ディレクトリの階層構造を利用しても訳語に違いが見られなかったためである．これは，2 階層目のカテゴリであっても対象とする分野が十分に限定できていないため，問合せが対象としている分野に適切な訳語を選択する効果が得られていないからであると考えられる．このことはさらに低い階層のカテゴリを利用することにより，検索対象とする Web 文書の分野をより限定し，適切な訳語を選択できる可能性が残されている．この場合，訳語の曖昧性解消に Web ディレクトリのどの階層のカテゴリまでを利用するのかさらに精査が必要となる．

表 6 t 検定による階層間の p 値
Table 6 Probability value of t-test among proposed method.

階層	1，2 階層	2，3 階層	3，1 階層	3，4 階層	2 階層，対訳辞書	3 階層，対訳辞書
p 値	0.2987	0.4103	0.1069	0.2156	0.0462	0.0296

4.4.3 最適なカテゴリ階層

表 6 は，各階層における検索結果に有意差の有無について，t 検定により検定を行った結果である．t 検定は，母集団が正規分布であることを仮定した場合の 2 組の標本について平均に有意差があるかどうかの検定を行う，パラメトリック検定法である．比較する両群 X，Y の標本サイズがそれぞれ m，n であるとき，検定統計量 t_0 を，

$$t_0 = \frac{|\bar{X} - \bar{Y}|}{\sqrt{U_e(\frac{1}{m} + \frac{1}{n})}}$$

により算出する．

このとき，$\bar{X}$，$\bar{Y}$ は両群の標本平均，U_e は両群を合わせた分散の推定値を表す．両群の平均が等しい場合には “統計量 t_0 は自由度 $m + n - 2$ の t 分布に従う” ことより，これを帰無仮説として両側検定を行う．この t 分布における t_0 の上側の p 値を求め，有意水準 α と比較することで有意差の有無を検定する．p 値が有意水準 α を下回れば，帰無仮説が棄却され，明確に有意差があることがいえる．一般に有意水準 α を 0.05 または 0.10 として検定することが多い[18]．

表 4 が示すように，1 階層，2 階層，3 階層とより下層の階層を利用するに従って，平均適合率は徐々に高くなっている．階層間の平均適合率の有意差については表 6 に示している．隣接する階層間では有意差は十分にあるとはいえないが，1 階層と 3 階層の間では p 値が 0.1069 であり，ある程度の有意差がみられる．さらに下層の階層である 4 階層では，t 検定による検定では有意差が十分にあるとはいえないものの，上位の階層に比べて平均適合率が低下している．これらの結果は，深い階層を利用することで検索精度は向上するが，あまり深すぎると検索精度が低下していくことを示唆している．

深い階層ほど対象となる分野をより特定できるため，深い階層を利用することは適切な訳語に絞り込むことに対して有効である．ただし，あまりにも深い階層を利用すると，悪影響のほうが大きくなる．悪影響の要因として，以下の 2 点があげられる．

1 点目は，過度の絞り込みにより訳語が得られない可能性が生じることである．分野を絞り込むことにより適切な訳語を選択するのであるが，絞り込んだ分野が必ずしもすべての問合せ語に対して適切であるとは限らない．問合せのうちのいくつかの問合せ語に対しては適切であるが，残りの問合せ語に対してはそうではない場合もある．このときその残りの問合せ語に対する訳語候補がそのカテゴリの特徴語集合に存在しない可能性が高くなる．これにより必要な訳語が十分に得られず，検索精度が低下する原因となる．

表 7 適合カテゴリの選択結果
Table 7 Result of relevant category selection.

問 9	（問合せ）人工 衛星 ST1
階層	選択した適合カテゴリ
2	Science/Space
3	Recreation/Travel/Transportation/Planes
4	Social_Science/Psychology /Branches/Neuropsychology

2 点目は，分野を絞り込みすぎることにより，適切な分野が選択したカテゴリから漏れてしまうことである．絞り込みがある程度の段階までであるときには適切な分野が選択したカテゴリに含まれていたとする．これをさらに絞り込むためにはより下層のカテゴリを選択することになるが，これらの下層のカテゴリはより絞り込む前のカテゴリの部分集合である．適切な分野を含んだ部分集合を選択できればよいが，場合によっては適切な分野を含まないカテゴリを選択してしまう可能性もある．その場合適切な訳語が得られない可能性が非常に高くなる．そうなると検索精度が低下する結果となる．**表 7** は，問 9 に対する各階層で選択した適合カテゴリである．2，3 階層では宇宙や航空機に関するカテゴリが選択されている．それに対して 4 階層では神経心理学に関するカテゴリが選択されており，問 9 に対して適切でないカテゴリが選択されている．この結果は，階層が深すぎると適切な適合カテゴリが選択できない可能性が高くなることを示唆している．

4.4.4 ま と め

4.4.1，4.4.2，4.4.3 項における考察より，深い階層を利用することにより対象となる分野を特定する必要があるが，悪影響が大きくなりすぎないようにある程度の階層までで絞り込みをとどめておくことが重要である．本システムにおいて言語資源として利用する Web ディレクトリに Yahoo! カテゴリを用いた場合，2 階層または 3 階層を利用することが適切であるとい

える．平均適合率だけを見ると 3 階層を利用することが最も良い．しかし事前の処理コストに着目した場合，上位の階層のほうが処理コストは小さい．よって検索精度が同等であるならば，上位の階層で済ますほうが良い．本実験においては 2 階層または 3 階層を利用することが適切であるが，それぞれのシステムにおいて適切な階層は異なるため，どの場合においても同じように 2 階層または 3 階層を利用することが適切であるとは限らない．しかし，上の階層であるほど分野特定の効果が薄く，あまり下の階層となると十分な統計情報が得られないことの悪影響が大きくなることは共通している．それゆえ，事前の処理コストと検索精度を考慮して最適な階層を決定する必要がある．

5. おわりに

本論文では，Yahoo! カテゴリに代表されるような各国の言語で構築された Web ディレクトリを，言語横断情報検索における訳語の曖昧性解消と検索精度の向上に用い，より適切に曖昧性解消を行うための問合せ翻訳手法について提案した．また，本手法の有効性を検証するために，NTCIR-3 CLIR テストコレクションを用いて検索の実験を行い，本手法が言語横断情報検索の曖昧性解消において有効であることを示した．さらに，カテゴリの統合の度合いをどの程度まで行えばよいかについて調査し，1 つのカテゴリが対象とする範囲をある程度限定することが有効であることが分かった．本システムにおいて言語資源として利用する Web ディレクトリに Yahoo! カテゴリを用いた場合，2 階層または 3 階層を利用することが適切であった．

本手法は，Yahoo! カテゴリなどの複数の言語で用意されている Web ディレクトリに登録されている文書群をコーパスとして用いることにより，分野に対する依存性が生じることはない．この特徴は Web 文書のようにさまざまな分野が対象となる検索において有効であると思われる．また，本手法で用意すべき言語資源は対訳辞書および Web ディレクトリのみであり，Web ディレクトリのカテゴリの対応付けができればそれ以外に特別に必要となる言語資源はない．さらに，Web ディレクトリには多数の言語版があるが（たとえば Yahoo! カテゴリは 2007 年 9 月の時点で 30 カ国以上の言語版が存在），対訳辞書さえあればそれらの言語のすべての組合せに対して本手法は適用できるため，対応言語の拡大が容易である．

今後の課題としては，4.4 節でも述べたように，Web ディレクトリのカテゴリ情報を詳細に利用することによる平均適合率低下の原因を特定すること，およびどの階層までのカテゴリ情報を利用することが訳語の曖昧性解消に有用であるかを精査することがあげられる．

謝辞 本研究の一部は，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」の公募研究（課題番号 19024058）の支援による．ここに記して謝意を表します．

参考文献

1) Lin, C.-J., Lin, W.-C., Bian, G.-W. and Chen, H.-H.: Description of the NTU Japanese-English Cross-Lingual Information Retrieval System Used for NTCIR Workshop, *Proc. 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.145–148 (1999).
2) 前田 亮，吉川正俊，植村俊亮：言語横断情報検索における Web 文書群による訳語曖昧性解消，情報処理学会論文誌：データベース，Vol.41, No.SIG6(TOD7), pp.12–21 (2000).
3) 木村文則，前田 亮，宮崎 純，吉川正俊，植村俊亮：Web ディレクトリを言語資源として利用した言語横断情報検索，情報処理学会論文誌：データベース，Vol.45, No.SIG7(TOD22), pp.208–217 (2004).
4) 酒井哲也，梶浦正浩，住田一男，Jones, G., Collier, N.：機械翻訳を用いた英日・日英言語横断検索に関する一考察，情報処理学会論文誌，Vol.40, No.11, pp.4075–4086 (1999).
5) Kishida, K. and Kando, N.: A Hybrid Approach to Query and Document Translation Using a Pivot Language for Cross-Language Information Retrieval, *Working Notes for the CLEF 2005 Workshop*, pp.93–101 (2005).
6) Gonzalo, J., Verdejo, F., Peters, C. and Calzolari, N.: Applying EuroWordNet to Cross-Language Text Retrieval, *Computers and the Humanities*, Vol.32, No.2–3, pp.185–207 (1998).
7) Rehder, B., Littman, M.L., Dumais, S. and Landauer, T.K.: Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing, *Proc. 6th Text Retrieval Conference* (*TREC-6*), pp.233–239 (1997).
8) 国分智晴，田中 崇，森 辰則：空間分割型 CL-LSI による大規模言語横断情報検索，情報処理学会論文誌：データベース，Vol.43, No.SIG2(TOD13), pp.27–36 (2002).
9) Jansen, B.J., Spink, A. and Saracevic, T.: Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web, *Information Processing and Management*, Vol.36, No.2, pp.207–227 (2000).
10) Resnik, P. and Smith, N.A.: The Web as

a Parallel Corpus, *Computational Linguistics*, Vol.29, No.3, pp.349–380 (2003).
11) Seo, H.-C., Kim, S.-B., Rim, H.-C. and Myaeng, S.-H.: Improving query translation in English-Korean cross-language information retrieval, *Information Processing and Management*, Vol.41, No.3, pp.507–522 (2005).
12) 奥村明俊，石川　開，佐藤研治：コンパラブルコーパスと対訳辞書による日英クロス言語検索，自然言語処理，Vol.5, No.4, pp.77–98 (1998).
13) Kishida, K., Kando, N. and Chen, K.-H.: Two-Stage Refinement of Transitive Query Translation with English Disambiguation for Cross-Language Information Retrieval: A Trial at CLEF 2004, *Working Notes for the CLEF 2004 Workshop*, pp.135–142 (2004).
14) Sakai, T., Manabe, T., Kumano, A., Koyama, M. and Kokubu, T.: Toshiba BRIDJE at NTCIR-5 CLIR: Evaluation using Geometric Means, *Proc. NTCIR-5 Workshop Meeting*, pp.56–63 (2005).
15) Cho, K. and Kim, J.: Automatic Text Categorization on Hierarchical Category Structure by using ICF (Inverted Category Frequency) Weighting, *Proc. KISS Conference*, pp.507–510 (1997).
16) Frakes, W.B. and Baeza-Yates, R. (Eds.): *Information Retrieval: Data Structures and Algorithms*, Chapter7, Prentice Hall (1992).
17) Salton, G. and Buckley, C.: Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, Vol.24, No.5, pp.513–523 (1988).
18) 金　明哲，中村永友，山田智也：データ解析の基礎，ムイスリ出版 (2003).

(平成 19 年 9 月 20 日受付)
(平成 20 年 1 月 8 日採録)

(担当編集委員　酒井 哲也)

木村　文則（正会員）

1999 年大阪教育大学教育学部教養学科情報科学専攻卒業．2001 年同大学大学院教育学研究科総合基礎科学専攻修士課程修了．2005 年奈良先端科学技術大学院大学情報科学研究科情報システム学専攻博士後期課程研究指導認定退学．博士（工学）．2005 年より同志社大学文化情報学部実習助手，現在に至る．言語横断情報検索の研究に従事．

前田　　亮（正会員）

1995 年図書館情報大学図書館情報学部卒業．1997 年同大学大学院図書館情報学研究科修士課程修了．2000 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（工学）．日本学術振興会特別研究員，科学技術振興事業団 CREST 研究員を経て，2002 年立命館大学理工学部助教授，2004 年同大学情報理工学部助教授，2007 年より同准教授，現在に至る．2000〜2001 年米国バージニア工科大学客員研究員．ディジタル図書館，情報アクセス技術に関する研究に従事．平成 10 年度情報処理学会論文賞受賞．ACM，IEEE CS，電子情報通信学会，日本データベース学会各会員．

波多野賢治（正会員）

1995 年神戸大学工学部計測工学科卒業．1999 年同大学大学院自然科学研究科博士後期課程修了．博士（工学）．同年日本学術振興会未来開拓学術研究事業研究員，同年奈良先端科学技術大学院大学情報科学研究科助手を経て，2006 年より同志社大学文化情報学部専任講師．2005〜2006 年米国 AT&T Labs-Research 客員研究員．XML データベース，情報検索に関する研究に従事．平成 18 年度電子情報通信学会論文賞受賞．ACM，IEEE Computer Society，電子情報通信学会，日本データベース学会各会員．

宮崎　　純（正会員）

奈良先端科学技術大学院大学情報科学研究科准教授．1992 年東京工業大学工学部情報工学科卒業．1997 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了．博士（情報科学）．同大学助手を経て，2003 年より現職．2003〜2007 年科学技術振興機構さきがけ研究員（兼務）．2000〜2001 年テキサス大学アーリントン校客員研究員．高性能・高機能データ工学システムの研究に従事．電子情報通信学会，日本データベース学会，IEEE CS，ACM SIGMOD 各会員．

植村 俊亮（フェロー）

奈良産業大学情報学部情報学科教授．1964 年京都大学大学院工学研究科修士課程修了．同年電気試験所（現産業技術総合研究所）．マサチューセッツ工科大学電子システム研究所客員研究員，東京農工大学教授，奈良先端科学技術大学院大学教授を経て，2007 年から現職．データ工学，データベースシステムの研究に従事．工学博士．IEEE Fellow，電子情報通信学会フェロー．