Web ディレクトリの階層構造を利用した言語横断情報検索

Cross-Language Information Retrieval Using Web Directory Structure

木村 文則♥ 前田 亮♦
吉川 正俊† 植村 俊亮♠

Fuminori KIMURA Akira MAEDA
Masatoshi YOSHIKAWA Shunsuke UEMURA

Web には様々な分野の文書が存在するため，Web に対する言語横断情報検索には特定の分野に依存しない手法を用いることが望ましい．そこで本論文は，Web 情報の言語横断検索において，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する．カテゴリごとに属する Web 文書から特徴語を抽出し比較することにより，対応する異言語のカテゴリを決定する．検索範囲を，問合せが適合する同言語のカテゴリおよび対応する異言語のカテゴリに限定することにより，訳語の曖昧性解消と検索性能の向上を図る．また Web ディレクトリの階層構造を利用して下位のカテゴリを上位のカテゴリにマージすることによりさらなる性能の向上を図る．

Since the Web consists of documents in various domains or genres, the method for Cross-Language Information Retrieval (CLIR) of Web documents should be independent of a particular domain. In this paper, we propose a CLIR method which employs Web directories provided in multiple language versions (such as Yahoo!). In the proposed method, feature terms are first extracted from Web documents for each category in the source and the target languages. Then, one or more corresponding categories in another language are determined beforehand by comparing similarities between categories across languages. Using these category pairs, we intend to resolve ambiguities of simple dictionary translation by narrowing the categories to be retrieved in the target language. Moreover, we consider merging child categories into the parent at certain level in category hierarchy in order to further improve the effectiveness of the proposed method.

1. はじめに

世界的なインターネットの発展に伴い，外国語文書を電子的に入手することが容易となった．しかし従来のWeb検索エンジンは，問合せと同一言語の文書群が検索対象であるため，外国語文書に対する検索は効率的とは言い難い．そこで，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）に関する研究が近年盛んになってきている．言語横断情報検索に関する従来の研究では，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などが提案され，検索性能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のドメインに対する依存が大きいため，それ以外のドメインに対しては検索性能が低くなる可能性がある．Web文書の言語横断検索では文書内容の分野は広範囲に渡っているため，ドメイン依存の問題を改善しなければならない．

そこで我々は，Web情報の言語横断情報検索において，例えばYahoo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いる[1]，[2]．まず，カテゴリごとに属するWeb文書から特徴語を抽出し，これを比較することにより対応する異言語のカテゴリを決定する．検索を行うときは，まず問合せと適合するカテゴリを同一言語間で選択し，次にそのカテゴリに対応する異言語のカテゴリも選択する．選択された異言語のカテゴリの文書に対して再度検索を行う．こうして検索範囲を限定することにより，検索性能の向上を図る．

提案手法による対応付けの実験を以前の研究において行ったが，カテゴリ数が多かったことなどから，充分な対応付けの精度が得られなかった[2]. 本論文ではこの問題を解決するため，Webディレクトリの構造を利用して，カテゴリの統合をおこなった．また，統合したカテゴリに対して，言語間でのカテゴリの対応付けの実験もおこなった．

2. 関連研究

言語横断情報検索において，問合せを翻訳する場合，対訳辞書を用いて問合せを翻訳するが，このとき訳語の曖昧性解消が問題となる．その方法として，コーパスを用いる手法が研究されている．しかし訳語曖昧性解消にコーパスを用いる手法では，検索要求とコーパス間のドメインの相違による検索性能への影響が指摘されている． Hull[3]および奥村ら[4]は，並列コーパスや類似コーパスを用いる手法において，検索要求とコーパス間のドメインの相違が検索性能に悪影響を及ぼす可能性があることを指摘している．またLinら[5]は，単言語コーパスとしてドメインや規模の異なる三つのコーパスを用いて比較実験を行った結果，有用な共起情報を得るには大規模でドメインの一致したコーパスが必要であると結論付けている．

本研究で対象とするWeb検索では，多様な分野の検索要求に対応することが要求される．しかし，そのそれぞれのドメインについて，対応するコーパスをあらかじめ用意することは現実的ではない．本研究では，Yahooなどの複数の言語版が用意されている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，言語間のカテゴリの比較によって言語横断情報検索における検索性能の向上を目標とする．

本論文では，言語横断情報検索にWebディレクトリの階層構造を利用することを提案する．Dumaisら[6]は，Web文書の分類において，カテゴリの階層構造を利用し比較対象と

♥ 学生会員 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 fumino-k@is.aist-nara.ac.jp
♦ 立命館大学理工学部情報学科 amaeda@cs.ritsumei.ac.jp
† 正会員 名古屋大学情報連携基盤センター yoshikawa@itc.nagoya.ac.jp
♠ 正会員 奈良先端科学技術大学院大学情報科学研究科 uemura@is.aist-nara.ac.jp

なるカテゴリを絞り込むことにより，分類精度がある程度向上することを指摘している．本論文では，カテゴリの階層構造をカテゴリの統合に利用することにより，提案したシステムの改良を試みる．

3. 提案するシステム

3.1 システムの概略

本システムでは, Web ディレクトリの複数の言語版を利用する．一つは問合せと同じ言語版(図 1 language A)であり，残りは検索対象となる一つ以上の言語版(同 language B)である．前処理として事前にこれらのそれぞれのカテゴリにおいて，異言語のカテゴリとの対応付けをおこなう．

図 1 は前処理の流れを示したものである．前処理では(1)文書からの単語の抽出，(2)カテゴリの特徴語の抽出，(3)特徴語の翻訳，異言語間でのカテゴリの対応付け，がおこなわれる．

例えば図 1 の language A のカテゴリ *a* に対する対応付けでは, まずカテゴリ *a* に属する文書群から単語を抽出し(1)，次にそれらのカテゴリ *a* における重みを計算して特徴語を抽出し，特徴語集合 *fa* を得る(2)．さらに特徴語集合 *fa* を検索対象となる language B に翻訳する(3)．これを language B の全てのカテゴリの特徴語集合と比較し，language B のカテゴリの中からカテゴリ *a* に適合するものを推定し，対応付けをおこなう(4)．

この対応付けを利用することにより, Web 文書の検索をおこなう．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付けられている異言語のカテゴリを選択し，最後に選択されたカテゴリの文書に対して検索がおこなわれる．

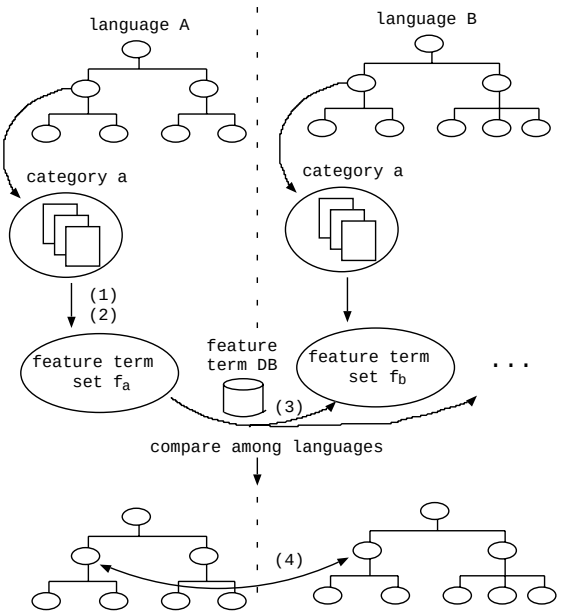

図 1 前処理
Fig.1 Preprocessing

3.2 前処理

3.2.1 特徴語の抽出

各カテゴリは特徴語集合によりその特徴を表現される．特徴語集合は，そのカテゴリの特徴を表現していると思われる単語の集合である．

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する．次に，抽出された単語をカテゴリごとに集計し，その単語がカテゴリの内容を表現する程度を表す重みを計算する．抽出された単語のうち，重みが大きいものから上位n語をそのカテゴリの特徴語として抽出する．また，ある閾値以上となるものを特徴語とすることも考えられる．

Web文書から抽出された単語の重みは，TF・ICF (term frequency ・ inverse category frequency)により計算する．これは．一般的に良く知られた単語の重み付けの手法の1つであるTF・IDF (term frequency ・ inverse document frequency)を発展させたものである．TF・IDFは単語の出現頻度(TF)と文書頻度の逆数との積により求められる．TFは単語の網羅性を表し，IDFは単語の特定性を表しており，これらの積である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．TF・IDFでは文書を単位として重みを計算するが，文書のかわりにカテゴリを単位として重みを計算したのがTF・ICFである．TF・ICFにより重みを計算することで, 文書単位で計算するTF・IDFより，カテゴリの内容をより反映した重み付けをおこなうことができる．

3.2.2 異言語間カテゴリの対応付け

3.2.1 で抽出した各カテゴリの特徴語を比較することによりカテゴリ間の適合度を求め，異言語間での対応付けをおこなう．

異言語間でカテゴリを比較するには，特徴語を翻訳する必要がある．特徴語の翻訳の流れを図 2 に示す．まず，翻訳したい特徴語に対する対訳辞書の全ての訳語を，訳語の候補として抽出する．抽出された全ての訳語候補について，比較している異言語カテゴリの特徴語に含まれているかを調べる．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，そのカテゴリにおけるその特徴語の訳語と決定する．このとき，比較している異言語カテゴリの特徴語の中にいずれの訳語候補も存在しない場合，その特徴語はこの二つのカテゴリ間の対応付けにおいては使用しない．

しかし，例えば，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，翻訳をおこな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異言語カテゴリの特徴語に含まれていない場合，翻訳する前の特徴語そのものを，比較している異言語カテゴリの特徴語に含まれているか調べる．もし含まれていれば，翻訳前の単語そのものをこのカテゴリにおける訳語とみなす．

例えば，英語の“ system ”という単語の，日本語のカテゴリ“ コンピュータとインターネット > ソフトウェア > セキュリティ ”(以下 「カテゴリ“ セキュリティ ”」)における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．この訳語の候補の全てに対して，カテゴリ“ セキュリティ ”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 ”のカテゴリ“ セキュリティ ”における訳語と決定する．もし，“ system ”のいずれの訳語候補もカテゴリ“ セキュリティ ”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をこのカテゴリにおける訳語とみなす．

言語 A におけるカテゴリ *a* に対して, 比較対象である言語 B のカテゴリのうちで最も *a* と適合度の高いものを，適合カ

テゴリとして対応付ける．カテゴリ *a* に対する異言語のカテゴリ *b* の適合度は，*a* の特徴語の訳語が *b* にあるならば互いの重みを掛けることを *a* の全ての特徴語に対しておこない，その値を全て足し合わせることによって計算する．カテゴリ *a* のカテゴリ *b* に対する適合度を *sim*(*a,b*)とすると，

$$sim(a,b) = \sum_{f \in f_a} w(f,a) \cdot w(t,b)$$

であり，*f* は特徴語，*fa* はカテゴリ *a* の特徴語集合，*w*(*f, a*)は特徴語 *f* のカテゴリ *a* における重み，*t* は *f* のカテゴリ *b* における訳語を表す．

例として，英語のカテゴリ“Computers and Internet > Security and Encryption”(以下 「カテゴリ“Security”」)を先程のカテゴリ“セキュリティ”と比較する場合について説明する．「カテゴリ“Security”」の特徴語集合として，“privacy”,“system”,…,があるとする．“privacy”,“system”の重みはそれぞれ 0.007110, 0.006327 とする．これらの単語の訳語が“プライバシー”,“システム”と決まり，重みはそれぞれ 0.023999, 0.047117 であったとする．このとき，カテゴリ“security”(*s1*)のカテゴリ“セキュリティ”(*s2*)に対する適合度 *sim*(*s1, s2*)は，

$$sim(s_1, s_2) = 0.007110 \times 0.023999$$
$$+ 0.006327 \times 0.047117$$
$$+ \mathrm{L}$$

のように計算される．

3.2.3 検索

検索の流れを図 2 に示す．与えられた問合せについて，まず同言語間において，問合せと各カテゴリとの適合度を計算し，問合せが適合する同言語のカテゴリを決定する(1). 問合せとカテゴリの適合度は，問合せから抽出された語群とカテゴリの特徴語集合間の内積を求めることにより計算する．こうして求めた適合度がある閾値以上となるカテゴリを，問合せに対する適合カテゴリとする．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，これらを全て適合カテゴリとして選択する．なぜなら，問合せが対象としている分野の文書がいくつかのカテゴリに分割されていたり，同じ分野の文書が全く違うカテゴリに属しているといった場合もあるため，問合せに適合するカテゴリは必ずしも一つであるとは限らないからである．

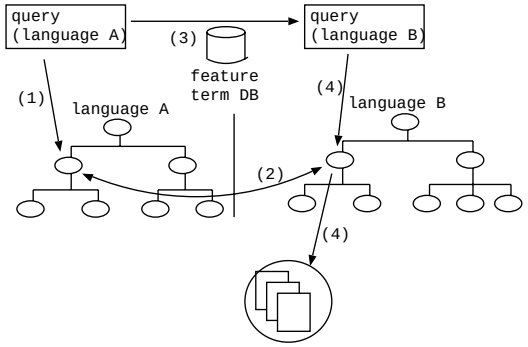

図 2 検索

Fig.2 Retrieval

次に，3.2.2 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(2). こうして選択されたそれぞれのカテゴリに属する個別文書に対して検索をおこなう．問合せとは異なる言語のカテゴリに対する検索は，対訳辞書とそのカテゴリの特徴語集合を利用して問合せを翻訳(3)したのち，そのカテゴリに属する Web 文書の検索をおこなう．以上の処理を経て得られた文書群が，検索結果となる(4).

4. カテゴリの統合

4.1 以前の結果

これまでの我々の研究[1],[2]において，提案手法を用いて，Yahoo におけるカテゴリについて言語間での対応付けをする実験をおこなったが，正確な対応付けをすることができなかった．このときの実験対象は，英語版 Yahoo におけるカテゴリ「Computers and Internet」以下の 559 カテゴリと，日本語版 Yahoo におけるカテゴリ「コンピュータとインターネット」以下の 654 カテゴリであった．HTML タグ除去後の各カテゴリにおける Web ページのバイト数の総計は，英語版では平均 45,905 バイト，最小 476 バイト，最大 1,084,676 バイトであり，日本語版では平均 22,770 バイト，最小 467 バイト，最大 409,576 バイトであった．正確に対応付けができなかった原因として，文書不足のカテゴリが存在したこと，カテゴリが細分化されすぎていて数が多すぎることなどが考えられた．

そこで，文書不足のカテゴリを対象から除外し，Web 文書の総バイト数が 30KB 以上の英語版 210 カテゴリ，日本語版 138 カテゴリに対して同様の実験をおこなった．しかし，対応付けの結果は改善されなかった．

4.2 カテゴリの統合方法

4.1 節の実験結果から，カテゴリの対応付けを正確におこなうには，カテゴリが細分化されすぎているという問題点を解消する必要があると考えられる．

その理由としてまず，カテゴリが細分化されすぎていることにより，類似のカテゴリが複数存在し，それにより対応付けの精度が低下することが考えられる．また，同一のカテゴリに属するべき Web 文書群が複数のカテゴリに分散するため，文書数が不足するカテゴリが生じ，十分な統計量が得られないことも，対応付けの精度を低下させていると考えられる．以上の点から，カテゴリ数を減らすこと，各カテゴリに属する文書を増やすことにより，カテゴリの対応付け結果が改善されることが予想される．

そこで，Web ディレクトリの構造を利用し，下位の階層のカテゴリを上位の階層に統合することにより，カテゴリの細分化の問題を解決する．下位の階層のカテゴリは上位のカテゴリの内容を特化したものであることから，これらが対象とする分野は同じとみなせることが多い．また，直接繋がっていないカテゴリ間でも，それほど階層の離れていない共通の上位のカテゴリを持つならば，互いが対象としている分野に大きな差異はないといえる．よって，このように下位のカテゴリを上位に統合し，一つのカテゴリとして扱うことを考える．カテゴリを統合すると，一つのカテゴリに属する総文書数も増加する．

このようにカテゴリを統合することにより，カテゴリ数を減少することができ，言語間でのカテゴリの対応付けの精度が向上することが期待される．さらに，一つのカテゴリに属する Web 文書数が増加し，その総バイト数も増加するため，有意な統計量が得られると考えられる．

しかし，どの程度の深さ以下の階層にあるカテゴリを統合すればよいかという点については明確になっていないため，今後検討をする必要がある．

5. 実験

今回の実験では，4.1 節の以前の実験で使用した実験対象に対してカテゴリの統合をおこない，そのうえで言語間でカテゴリの対応付けをおこなうことによって，カテゴリの階層構造を利用することの有効性を検証した．

今回カテゴリの統合は，英語版は “ Computers and Internet ” から 1 階層下のカテゴリまで，日本語版は “ コンピュータとインターネット ”から 1 階層下のカテゴリまでに統合した．統合後のカテゴリ数は，英語版は 16 カテゴリ，日本語版は 18 カテゴリである．カテゴリの統合をおこなった後，3 節で提案した本手法により言語間でのカテゴリの対応付けをおこなった．今回，各カテゴリの特徴語数は，特徴語の重みの上位 100 語とした．

実験の結果，英語から日本語，日本語から英語のどちらの対応付けもほとんど正確におこなうことができなかった．

その原因として，カテゴリを統合しすぎたことが考えられる．今回の実験では，最上位のカテゴリから 1 階層下のカテゴリまでに統合したが，これによりそれぞれのカテゴリが対象とする分野が広範囲になったため，各カテゴリの特色があいまいになってしまったと思われる．そのため，特定のカテゴリの特徴を表す単語より，いくつものカテゴリに現れる単語の影響が大きくなり，対象とする分野の範囲が広いカテゴリほど対応付けられやすい結果となった．

また，カテゴリを統合しすぎた結果，カテゴリ数が少なくなったため，特徴語の重み付けが適切におこなえなかったことも考えられる．全てのカテゴリに出現する単語と，ただ 1 つのカテゴリにしか出現しない単語の間でも，IDF の値の差があまり大きくなくなったため，TF の値の影響ばかりが強く現れた．そのため，特徴語の重みにカテゴリの特定性を反映することができなくなった．

以上のことから，言語間におけるカテゴリの対応付けの精度を向上するには，カテゴリの統合をある程度の階層までにとどめ，あまり上位の階層にまで統合をおこなわないようにする必要がある．また，ほとんどのカテゴリに出現するような単語は特徴語から除外することなどにより，特徴語のカテゴリを特定性をより高くする重み付けをおこなう必要がある．

6. おわりに

本論文では，Yahoo に代表されるような，複数言語版が存在する 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性能の向上に用いる手法を提案した．本手法は，複数の言語版が用意されている 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 文書のように様々な分野が対象となる場合において有効であると思われる．また，本手法で必要とされるのは対訳辞書のみであり，それ以外に特別に必要となる言語資源はない．さらに，Web ディレクトリにはいくつもの言語版があるが(例えば Yahoo は 2003 年 2 月の時点で 23 カ国版が存在)，対訳辞書さえあればそれらの言語の全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．

本手法の有効性を検証するために，カテゴリの対応付けの実験をおこなったが，以前の研究では十分な結果が得られなかった．そこで，カテゴリの対応付けの精度を向上するために，Web ディレクトリの構造を利用し，下位の階層のカテゴリを上位の階層に統合することを提案した．現状では言語横断情報検索における有効性を示せる結果が得られていないが，Web のような雑多な分野の文書に対する言語横断情報検索に対して，本論文で提案した手法はある程度の有効性を示せるものと考えている．

本研究の今後の課題として，カテゴリの統合をどの程度までおこなうか検討することが挙げられる．またそれ以外にも，言語間でのカテゴリの対応付け手法のより詳細な検討，検索の評価実験などが挙げられる．

[文献]

[1] 木村文則，前田亮，吉川正俊，植村俊亮．ディレクトリ型検索エンジンのカテゴリ間対応付けによる言語横断検索．第 13 回データ工学ワークショップ論文集 (2002). http://www.ieice.org/iss/de/DEWS/proc/2002/papers/C4-4.pdf

[2] 木村文則，前田亮，吉川正俊，植村俊亮．ディレクトリ型検索エンジンを利用した言語横断情報検索．第 1 回情報科学技術フォーラム論文集，第 2 分冊 pp.69-70 (2002).

[3] David A. Hull. Using structured queries for disambiguation in cross-language information retrieval. In *Electronic Working Notes of the AAAI Symposium on Cross-Language Text and Speech Retrieval* (1997).

[4] 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.77-93 (1998).

[5] Chuan-Jie Lin, Wen-Cheng Lin, Guo-Wei Bian, and Hsin-Hsi Chen. Description of the NTU japanese-english cross-lingual information retrieval system used for NTCIR workshop. In *Proceedings of First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp.145-148 (1999).

[6] Susan Dumais and Hao Chen. Hierarchical classification of Web content. In *Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval* (*SIGIR2000*), pp.256-263, Athens, Greece (2000).

木村 文則 Fuminori KIMURA
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在学中．言語横断情報検索の研究に従事．日本データベース学会学生会員．

前田 亮 Akira MAEDA
立命館大学理工学部情報学科助教授．ディジタル図書館，情報検索，多言語情報処理の研究に従事．ACM，情報処理学会，電子情報通信学会，情報メディア学会各会員．

吉川 正俊 Masatoshi YOSHIKAWA
名古屋大学情報連携基盤センター教授．データベースシステムの研究に従事．情報処理学会正会員．電子情報通信学会正会員．日本データベース学会理事．

植村 俊亮 Shunsuke UEMURA
奈良先端科学技術大学院大学情報科学研究科教授．データベースシステムの研究に従事．情報処理学会，電子情報通信学会フェロー． IEEE フェロー． 日本データベース学会正会員．