D-35 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

Cross-Language Information Retrieval Using Web Directories

木村 文則† 前田 亮‡ 吉川 正俊*† 植村 俊亮†
Fuminori Kimura Akira Maeda Masatoshi Yoshikawa Shunsuke Uemura

1. まえがき

WWW において，母国語以外の文書も電子的に入手することが容易となったことにより，それらの文書を検索したいという要求は高まっていると思われる．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．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．

そこで本論文では，Web 情報の言語横断情報検索において，例えば Yahoo のようなディレクトリ型検索エンジンを利用する手法を提案する．あらかじめそれぞれのカテゴリにおいて，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推定する．検索は問合せと適合する同一言語のカテゴリおよびそのカテゴリに対応する異言語のカテゴリに対して行う．こうして検索範囲を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．

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の解消が問題となる．訳語曖昧性解消の手法として，コーパスを用いる手法が研究されている．しかしこの手法では，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．奥村ら[1]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．

本研究で対象とする Web 検索では多様な分野の検索要求への対応が要求されるが，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 などのディレクトリ型 Web 検索エンジンから複数の言語版を利用する．これら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間の対応をとり，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を目標とする．

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，ディレクトリ型検索エンジン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 言語 A)であり，残りは検索対象となる一つ以上の言語版(同言語 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行う．

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，(4)異言語間でのカテゴリの対応付け，が行われる．例えば図 1 の言語 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 a を得る(2)．さらに特徴語集合 a を検索対象となる言語 B に翻訳する(3)．これを言語 B の全ての特徴語集合と比較し，言語 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)．

前処理で行われた対応付けを利用することにより，Web 文書の検索を行う．検索は，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することで行われる．

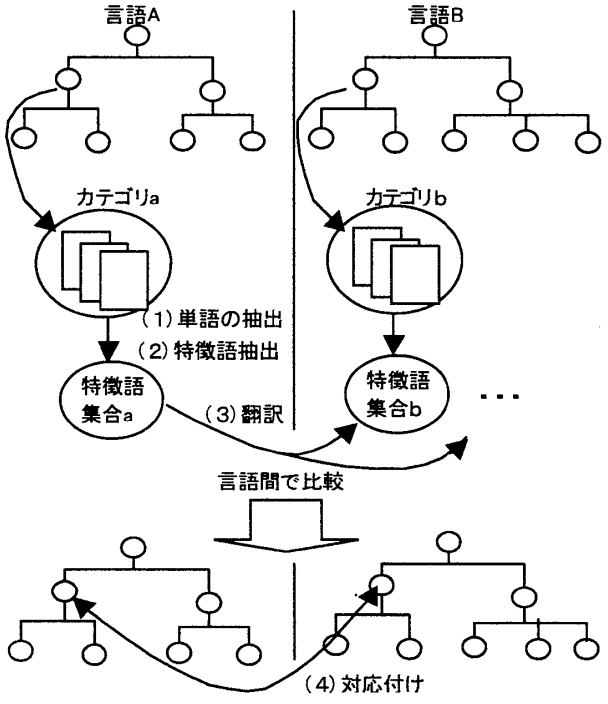

図 1．前処理

3.2 前処理

3.2.1 特徴語の抽出

各カテゴリにおいて特徴語を抽出するために，まずそれに属する Web 文書から単語を抽出する．各単語の重みは，

†奈良先端科学技術大学院大学 情報科学研究科
‡立命館大学 理工学部 情報学科
*名古屋大学 情報連携基盤センター

その出現単語数で正規化することにより計算する．文書 d における単語 t の重み $w(t,d)$ は

$$w(t,d)=\frac{f(t,d)}{\sum_{s\in d} f(s,d)}$$

であり，$f(t,d)$ は Web 文書 d における単語 t の出現頻度を表す．

こうして得られた単語群からカテゴリの特徴語を抽出する．同一のカテゴリには内容の類似した文書が分類されることから，そのカテゴリに属する文書の多くに出現している単語ほどそのカテゴリの特徴を表していると言える．よって，特徴語の重みの計算には DF(document frequency)を用いて計算する．カテゴリ c における単語 t すなわち特徴語 t の重み $df(t,c)$ は

$$df(t,c)=\frac{\sum_{d\in c} w(t,d)}{N}$$

であり，N はそのカテゴリに属する文書数である．こうして計算された特徴語の重みの大きいものから n 語をそのカテゴリの特徴語とする．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．

3.2.2 異言語間カテゴリの対応付け

3.2.1 で抽出した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，異言語間でのカテゴリの対応付けを行う．

異言語間で特徴語を比較するには，特徴語を翻訳する必要がある．まず，特徴語の訳語の候補を対訳辞書から全て抽出する．抽出された全ての訳語について，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調べる．含まれていた訳語のうち，特徴語の重みがもっとも大きい訳語を，そのカテゴリにおけるその特徴語の訳語と決定する．

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行い，その値を全て足し合わせることによって計算する．

3.3 検索

検索は，まず問合せが適合するカテゴリを，同言語において選択し，次に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．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全てのカテゴリを，問合せに対する適合カテゴリとする．さらに，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を決定する．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行う．

4．実験

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行った．英語版 Yahoo におけるカテゴリ「Computers and Internet」以下のカテゴリと，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」以下のカテゴリについて対応付けした．英語版のカテゴリ数は 559，日本語版のカテゴリ数は 654 である．HTML タグ除去後の各カテゴリでの Web ページのバイト数の総計は，英語版は平均 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．

英語版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．日本語版は「茶筌」(http://chasen.aist-nara.ac.jp/)を用いて名詞，動詞，形容詞，未知語を抽出した．特徴語の翻訳には「EDR 電子化辞書」を用いた．今回の実験では，各カテゴリの特徴語数を 100 語とした．

英語から日本語への対応付けでは，調査した 60 カテゴリ中，正しいと言えるのは 5 件であった．また日本語から英語の対応付けでは，63 カテゴリに対して，正しいと言えるのは 6 件であった．また，カテゴリが細分化されているため，完全には正確ではないが，適切な対応付けであると思われるものもあった．さらに特定のカテゴリばかりが対応付けられるなど，うまく対応付けできなかった．

その原因としてまず，カテゴリの文書数の不足が考えられる．日本語版の"/Science/Computer_Science/Courses" の文書数は 1 文書であり，抽出された特徴語は 11 語であった．このカテゴリの特徴語の重みの最大値は 0.166667 であり，最小値は 0.083333 であった．特徴語が 100 語存在するあるカテゴリの場合，最大値 0.043468，最小値 0.002155 であった．以上より，特徴語が少ないとその重みが大きくなり，結果的にカテゴリの適合度も高くなったと考えられる．実際"/Science/Computer_Science/Courses"カテゴリは 63 件中 13 回対応付けられている．

また，実験対象を Yahoo の特定の分野に限定したため，あるカテゴリに対する異言語の適合カテゴリが実験対象中に存在しないこともあった．このような場合そのカテゴリに適合しない別のカテゴリが対応付けられたため，対応付けが不正確となった一因となった．

5．おわりに

本論文では，Yahoo に代表されるような，ディレクトリ型検索エンジン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．また本手法の有効性を検証するための実験として，カテゴリの対応付けを行った．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は，特徴語の抽出およびカテゴリ間対応付け手法の再検討や文書数不足への対応が挙げられる．また，今回はカテゴリの木構造を考慮せずフラットなものとして扱ったが，木構造を活用することで[2]，より高精度なカテゴリの言語間対応付けが可能になると考えられる．さらに情報検索テストコレクションを用いた評価実験を行う必要がある．

参考文献

[1]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, October 1998.

[2]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, July 2000.