比喩的関係を利用した検索クエリ拡張手法

久保 真哉 † 桝井 文人 ‡ 福本 淳一 ††
† 北見工業大学大学院工学研究科
‡ 北見工業大学工学部情報システム工学科
†† 立命館大学情報理工学部メディア情報学科

†shinku-@ialab.cs.kitami-it.ac.jp
‡f-masui@mail.kitami-it.ac.jp
††fukumoto@media.ritumei.ac.jp

1　はじめに

既存のWWW検索システムを利用するためにはクエリの入力が必須であることは周知の事実であり，ユーザは自身の指向する検索要求をキーワードとして顕在化させなければならない．それゆえ，ユーザがうろ覚えや無知などの理由によりクエリを明示できない場合は，WWW検索システムを利用することが極めて困難な状況になってしまう．この問題を解決するためには，ユーザの情報要求を顕在化する柔軟なクエリ拡張手法が必要である．

クエリ拡張に関する研究はこれまでにも多数行われている [1] [2]．大石ら [3] はクエリと関連する語を追加するための関連単語抽出アルゴリズムを提案し，検索精度が向上することを確認している．また，大塚ら [4] はYahoo!知恵袋[1] の質問記事を用いて，ユーザが入力したキーワードに関連するカテゴリを特定し，適合フィードバックによって質問記事を提示するシステムについて報告している．

しかしながら，これらの研究ではクエリに関連する語や類似している質問記事の提示に留まっており，ユーザの検索要求を顕在化したキーワード提示までには至っていない．

そこで，我々は比喩的関係を利用することで情報要求を顕在化する立場を考える [5]．情報要求に関する明確なキーワードを提示できない場合，「攻撃側と守備側に分かれ，ボールを打って得点を競うスポーツ」や「野球のようなスポーツ」「野球によく似た競技」といった表現を用いることで対話者に情報伝達できる．その中でも多用されるのが後者の比喩表現である [6]．

図1に示す人間同士のコミュニケーションの例では，比喩表現から「クリケット」や「ソフトボール」などの語を導出している．このように，人間ならではの柔軟かつインタラクティブな処理が施されることによって情報要求を伝えることができる．WWW検索システムにおいて同様の処理を実現できれば，ユーザの情報アクセス効率は大きく向上するはずであり，前述したようにキーワードが顕在化できない状況においても検索システムの能力を発揮させることが可能であろう．

A：「あの競技何だっけ？
　　ほら、野球みたいなスポーツ。」
B：「マイナーなスポーツで？」
A：「そう。マイナーなスポーツ。」
B：「だったら …
　　クリケットとかソフトボールじゃない？」
A：「そうそう、クリケット。」

図 1: うろ覚えの情報要求に対する会話例

本稿では，上述した比喩表現と適合フィードバックを利用した検索要求顕在化手法 [5] を対象として，その性能評価と結果について述べる．具体的には，抽出した語が比喩的関係に合致した語であるか心理学実験によって評価した結果を報告する．

以下，2章で提案手法について説明し，3章で提案手法に基づいた実験準備と心理学実験について述べる．そして，4章で提案手法の性能評価についてまとめる．

2　提案手法

本章では，比喩表現が示す語の推論手法について説明する．

提案手法は，比喩表現の属性を利用してWWW検索を複数回繰り返し行うことでユーザが入力したクエリ（以下，疑似クエリ語と称す）からユーザの情報要求（以下，類似語候補と称す）を推論する仕組みであり，1. カテゴリ語を抽出するステップ，2. 特徴語を抽出するステップ，3. 類似語候補を抽出するステップの3ステップから構成される．各ステップのWWW検

[1] Yahoo!知恵袋 http://chiebukuro.yahoo.co.jp/

索結果を次のステップに引き継ぐことにより検索範囲の絞り込みを実現する．また，語の抽出には形態素解析[2]を用いる．

図 2 に本手法の概要図を示す．以下，それぞれのステップについてを詳述する．

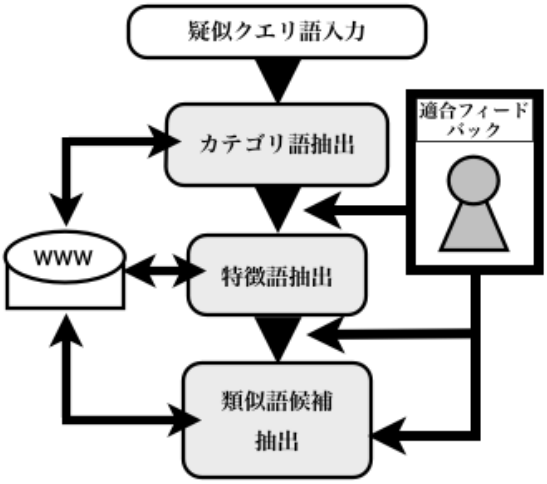

図 2: 処理の流れ

2.1 カテゴリ語抽出

ユーザが入力する疑似クエリ語 a のカテゴリや属性を表現する語（以下，カテゴリ語と称す）を抽出する．

a はユーザの情報要求を満たすクエリのカテゴリや属性を意味する手がかりであるため，検索要求を比喩表現 $X = \{a,$ のような, $b_i\}$ に対して適切なカテゴリ語集合 $B = \{b_1, b_2, b_3, \cdots, b_i\}$ を WWW 検索を介して得る．ここで，b_i は頻度情報を基にランキングする．

例えば，$a =$ { 野球 } の場合，$X =$ {(野球, のような, スポーツ), (野球, のような, 競技), (野球, のような, ゲーム)} などが抽出でき，$B =$ { スポーツ, 競技, ゲーム} となる．

2.2 特徴語抽出

カテゴリ語抽出で得られたカテゴリ語 B を修飾する語（以下，特徴語と称す）を抽出する．

単に比喩表現 X を WWW 検索するよりも，比喩表現 $Y = \{a,$ のような, $c_{i,j}, b_i\}$ として拡張表現を用いることで検索範囲を絞り込むことができるはずである．したがって，検索要求は Y に適切な特徴語集合 $C = \{(c_{1,1}), (c_{1,2}), (c_{1,3}), \cdots, (c_{i,j})\}$ を WWW 検索より得る．ここで，特徴語 $c_{i,j}$ はカテゴリ語 b_i を修飾する名詞，動詞，形容詞，もしくは，これら品詞を組み合わせた語句である．

例えば，$a =$ { 野球 }，$b_i =$ { スポーツ } の場合，$Y =$ {(野球, のような, 団体, スポーツ), (野球, のような, チーム, スポーツ), (野球, のような, 新しい, スポーツ)} などが抽出でき，$C =$ { 団体, チーム, 新しい} となる．

ここで，C に該当する語が抽出できないならば処理を終了する．

2.3 類似語候補抽出

2.1 および 2.2 で抽出したカテゴリ語 b_i と特徴語 $c_{i,j}$ の組み合わせを $w_{i,j} = \{c_{i,j} + b_i\}$ とする（以降，共通項と称す）．疑似クエリ語 a と共通項 $w_{i,j}$ が似通っている類似語候補集合 $D = \{d_1, d_2, d_3, \cdots, d_k\}$ を得る．よって，D を抽出するための検索要求を $Z = \{d_k,$ のような, $w_{i,j}\}$ とする．

例えば，$w_{i,j} =$ { 団体スポーツ, チームスポーツ, 新しいスポーツ } の場合，$Z =$ {(ラウンダース, のような, 団体スポーツ), (サッカー, のような, チームスポーツ), (クリケット, のような, 新しいスポーツ)} などが抽出でき，$D =$ { ラウンダース, サッカー, クリケット } となる．

3 心理学実験

提案手法の性能を検証するため，提案手法によって抽出した類似語候補を被験者に評価してもらった．本章では，心理学実験における評価者間の相関について分析する．

3.1 実験準備

まず，実験環境について述べる．システムに与える疑似クエリ語集合として $A =$ { あじさい, カーリング, カップヌードル, コロッケ, コオロギ, サッカー, サボテン, シカ, 野球 } を用意した．これらの疑似クエリ語は EDR 辞書 [7] に登録されている単語を無作為に選出した．

提案手法の適用にあたりカテゴリ語 b_i の抽出数をランキング上位 20 件（$i \leqq 20$）に制限し，特徴語と類似語候補の抽出に関しては無制限（$j, k \leqq \infty$）とした．また，$b_i =$ { もの, こと, 感じ } の場合は意味カテゴリの絞り込み効果が期待できないため，不要語として取り除いた．

提案手法によって得られた類似語候補集合 D を 3 人の被験者に評価してもらった．具体的には，d_k を下記の判定基準に従って分類してもらった．

[2]ChaSen 形態素解析器 http://chasen-legacy.sourceforge.jp/　評価×　：比喩表現 Y の説明に合致せず不適格である

評価△ ：比喩表現 Y だけでは不十分だが適格である

評価◯ ：比喩表現 Y の説明が適格である

ここで，評価者による判定の差を認識する尺度として Cohen の一致係数 κ（kappa cofficient） [8] を利用し，算出する際に下記の基準を設けた．

基準 I ：評価◯のみを正解とした場合の κ 係数

基準 II ：評価◯と評価△を正解とした場合の κ 係数

例えば，d_k = { クリケット } で Y = { 野球, のような, チームスポーツ } の場合，的確な説明であるため評価◯を選択する．

3.2 実験結果

まず，有効性判定の結果を表 1 に示す．全体の κ 係数は基準 I の場合が 0.59 となり，基準 II の場合は 0.68 となった．また，5 つの疑似クエリ語が基準 I の場合に 0.5 以上となり，基準 II の場合は 0.6 以上となった．

表 1: 評価者間における κ 係数

a	基準 I	基準 II	抽出数
コオロギ	0.44	0.44	96
あじさい	0.42	0.48	239
カップヌードル	0.47	0.52	55
コロッケ	0.45	0.53	99
カーリング	0.45	0.59	118
バッタ	0.55	0.63	126
サボテン	0.56	0.63	200
野球	0.51	0.70	126
シカ	0.67	0.74	222
サッカー	0.74	0.80	145
全体	0.59	0.68	1,426

この結果から基準 I における全体の κ 係数は 0.59 であり，判定結果に相関があることを示しているが，半数の疑似クエリ語が 0.5 以下のため相関が弱い可能性も考えられる．一方，基準 II の場合では全体の κ 係数が 0.68 と相関が強く，8 つの疑似クエリ語が 0.6 以上の値のため相関が基準 I よりも強いと考えられる．また，比喩表現に関する文書の場合は読み手によって解釈が異なることから，基準 I による厳格な判定結果よりも基準 II による評価結果を用いることが望ましいと考えられる．

4 提案手法の評価

提案手法の性能を評価するため，前章の実験結果を用いて分析を行った．

4.1 評価環境

心理学実験による結果を考慮し，評価者 3 人中 2 人以上の評価が一致している類似語候補を正解として適合率を算出した．

また，カテゴリ語の抽出数を制限することによって性能が向上するかどうかを検証するため，ベースラインとしてランダムにカテゴリ語を抽出する手法を用意し，提案手法のランキングによって抽出した場合の適合率と比較した．

4.2 評価結果

全体の平均適合率はカテゴリ語の頻度上位 10 件で 18.4%，頻度上位 20 件では 14.7%となった．このことから抽出した類似語候補全体にはノイズが多く含まれていると考えられる．例えば，代名詞や形態素解析ミスによって意味の理解できない語が抽出される場合などである．さらに，判定者によって類似語候補に対するイメージが異なる場合も考えられ，意味的に広範囲な類似語を抽出していることが考えられる．

そのため，性能向上のためにはノイズを除去する手法を考案する必要がある．

次に，カテゴリ語のランキングの有効性を検証するためにベースラインと比較した結果についてまとめる．ランキングに対する平均適合率の推移を図 3 に記載し，図 4 に疑似クエリ語毎の平均適合率をランキング別に示す．

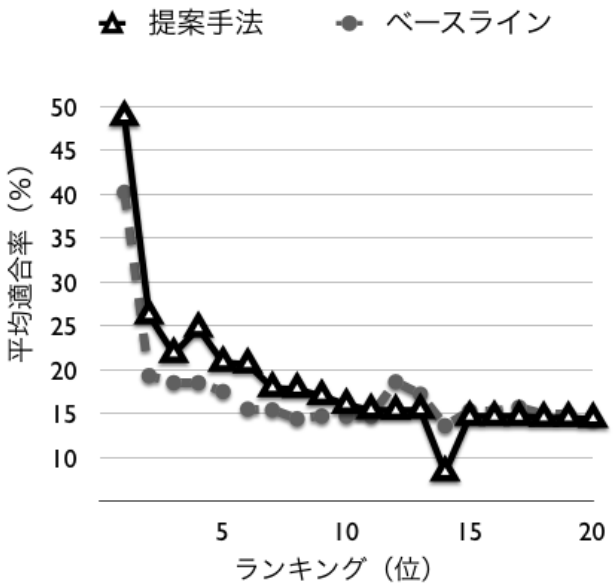

図 3: カテゴリ語の頻度における平均適合率の推移

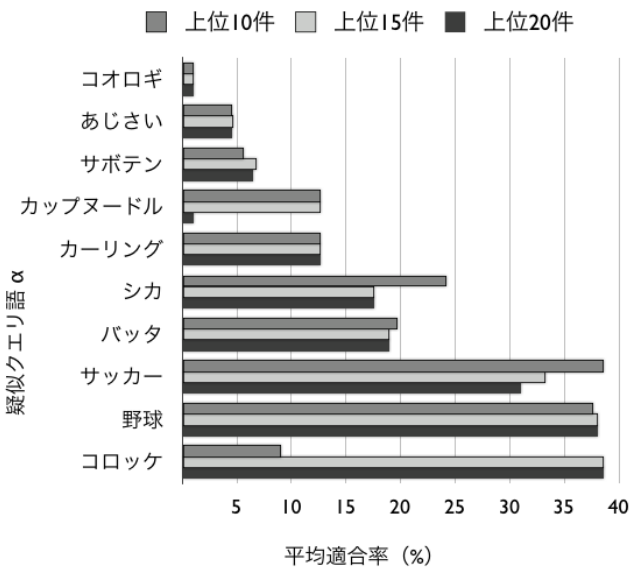

図 4: 疑似クエリ語毎の平均適合率 (%)

図 3 よりカテゴリ語上位 11 件目までは提案手法による平均適合率が高く，カテゴリ語が増えると共に平均適合率が減少している．

さらに，wilcoxon の順位和検定 [9] を行うことで提案手法に頻度情報を用いることが精度向上に関連するのか評価した．２つの結果に対して wilcoxon の順位和検定を行った結果，上位 9 位まで有意水準 5%（$p < 0.05$）において有意差があることを確認した．

このことから，適合フィードバックを適用する場合にはカテゴリ語の上位 10 件程度までをユーザへ提示すればよいことがわかる．

また，図 4 より疑似クエリ語 a によって平均適合率が大きく異なる傾向がある．例えば，$a = \{$ 野球, サッカー $\}$ の平均適合率は 30%を上回っているのに対し，$a = \{$ あじさい, コオロギ, コロッケ $\}$ などは有効な性能が得られていない．その原因として，比喩表現に適用するクエリとして不適当な語が存在することや，生成した比喩表現の意味が曖昧である可能性が挙げられる．特に，$a = \{$ コオロギ $\}$ の場合は顕著にこの傾向が現れていることから，本手法に適用するには日常会話に頻出する語である必要がある．

5　おわりに

本稿では，入力クエリを比喩によって表現することで同カテゴリに含まれるであろう類似語候補の抽出を行い，心理学実験を実施することによって提案手法の評価を行った．

まず，評価による結果は κ 係数を考慮することによって信憑性があることを確認した．そして，全抽出性能は平均適合率 36%と低い値となることから本手法が広範囲に渡って意味的に類似した語を抽出している可能性が高く，提案手法のままでは類似語候補の抽出精度が低いことを確認した．同時に，カテゴリ語の頻度情報を用いて適合率の推移を調査した結果，頻度上位の平均適合率が高いことと wilcoxon の順位和検定 ($p < 0.05$) により，頻度上位 10 件程度において本手法に有効性があることを確認した．

今後の課題として，評価実験規模の拡大と入力クエリが類似語候補に与える影響について調査する．また，類似語候補として適切な語を選択するための手法を考案する．

謝辞

本研究は，科学研究費補助金（基板研究 (C)20500833）の助成を受けている．

参考文献

[1] 吉岡真治, 原口誠：”検索語の網羅性に注目した汎化概念により検索語選択支援を行う情報検索システムの研究”. 人工知能学会論文誌, Vol.20,No.4,pp.270-280, 2005.

[2] 松生 泰典, 是津 耕司, 小山 聡, 田中 克己：”検索結果の概要を表すキーワード式生成による質問修正支援”. 第 16 回データ工学ワークショップ 2005(DEWS2005), 1C-i9, 2005.

[3] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸：”関連単語抽出アルゴリズムを用いた Web 検索クエリの生成”. 電子情報通信学会論文誌 D, Vol.J92-D, No.3, pp.281-292, 2009.

[4] 大塚淳史, 関洋平, 神門典子, 佐藤哲司：”情報要求の言語化を支援するクエリ拡張型 Web 検索システム”. 第 3 回データ工学と情報マネジメントに関するフォーラム 2011(DEIM2011), F6-3, 2011.

[5] 久保真哉, 桝井文人, 福本淳一：”喩える関係を利用した検索クエリ拡張に関する一考察”. 人工知能学会情報編纂研究会第 5 回研究会資料, 2011.

[6] 中村明：”比喩表現の理論と分類”. 共立出版, 1977.

[7] 日本語電子化辞書研究所,EDR 概念辞書, 日本電子化辞書研究所,1995

[8] 丹後俊郎：”統計学のセンス―デザインする視点・データを見る目”. 朝倉書店, 1998.

[9] 舟尾 暢男：”The R Tips―データ解析環境 R の基本技・グラフィックス活用集”. オーム社, 2009.