検索要求顕在化のための「喩え」の利用

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

現在のWWW検索システムが有効に機能するためにはキーワードの入力が必須である．WWW検索の能力を享受しようとした場合，ユーザは自身が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．よって，ユーザがキーワードを提示できない場合，WWW検索の恩恵に預れないことになる．

この場合，ユーザは自らが求める適合文書をどのようにして探し当てればいいのであろうか．旧態依然とした手作業によって，文書の山を漁らなければならないのだろうか．

我々は，上記問題を解決するための手段として，比喩の利用が有力候補であると考えている．情報要求について明確なキーワードを提示できないとき，通常は「攻撃側と守備側に別れ，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だろう．その中でも多用されるのが後者に挙げたような比喩表現である[1].

上記情報要求に対する人間同士の対話を考えると，だいたい図1のような流れであろうと想像できる．

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 …
　　クリケットとかラウンダースじゃない？」
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

上記対話では，話者Aが比喩表現として質問を発している．このとき，対象の実態は「競技」のインスタンスである．次に，話者Bは「競技」の特徴を確認して候補範囲の絞り込みをしようとしている．そして，話者Bは，絞り込まれた候補を列挙し，最終的に、話者Aが回答として受け入れている．

情報検索においても，上記のような処理が実現できれば，ユーザの情報アクセス効率は大きく向上するはずである．先に述べたような，キーワードが顕在化できないような状況においても，検索システムの能力を発揮させることが可能となる．

そこで本論文では，比喩表現の機能，比喩的関係の特性を利用して,「喩え」による検索要求を顕在化のための手法を提案する．提案手法では，対話的な応答を設置し，検索範囲の絞り込みを行うことで，図1で示したような過程をインタラクティブな処理として実現する．次に，提案手法の有効性と検索範囲の絞り込み効果を検証するため，提案手法を構成する評価実験を行った．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く評価実験，および，実験結果について述べる．4章で本提案手法の考察を述べる．

2　提案手法

本章では，提案手法について説明する．3つの処理過程を試作し，図2に示すように提案する手法は3段階のステップから構成されている．以下，各ステップ毎に詳述する．

＜ステップ1＞

まず，ユーザが入力するクエリ語（以降，疑似クエリ語と称す）は検索要求そのものではない．疑似クエ

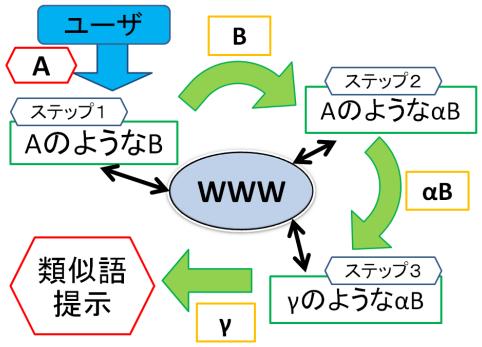

図 2: 提案手法概略図

リ語は真のクエリ語と何らかの点で類似した語であり，疑似クエリ語についての連体修飾を考えると，比喩表現を生成することができる．例えば,「野球のようなスポーツ」,「野球のような競技」,「ぶどうのような果物」,「ハーレーのようなバイク」などが生成できる．このとき，比喩表現の主辞（スポーツ，競技，果物，バイク）は真のクエリ語のカテゴリや属性を意味する手がかり（以下，カテゴリ語と称す）である．

しかし，WWW 検索では「野球のようなスポーツ」と「野球のような競技」は全く異なる表現として認識される．そこで，ステップ 1 の処理として，検索要求を「(疑似クエリ語）のような（カテゴリ語)」という比喩形式で表現できるカテゴリ語の候補を WWW 検索より抽出する．例えば，ユーザが疑似クエリ語「野球」を入力すると，カテゴリ語として「スポーツ」,「競技」,「ゲーム」,「遊び」などが抽出できる．

＜ステップ 2 ＞

ステップ 1 で得られたカテゴリ語を修飾する語を WWW 検索により抽出する．例えば,「野球のようなスポーツ」については「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」などが挙げられる．このとき,「団体」,「チーム」,「新しい」はカテゴリ語の意味を限定する語（以下，特徴語と称す）である．単に「野球のようなスポーツ」という表現を WWW 検索するよりも,「野球のようなチームスポーツ」という拡張された比喩表現を用いることで検索範囲を絞り込むことができると考えられる．

＜ステップ 3 ＞

ここまでに抽出したカテゴリ語と特徴語を利用すると，これらを共通点とした比喩形式を考えることができる．例えば,「のような団体スポーツ」,「のようなチームスポーツ」,「のような新しいスポーツ」である．

さらに，上記比喩形式に基づいた比喩表現を生成することができる．例えば,「ソフトボール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」といった表現が生成できる．このとき,「ソフトボール」,「サッカー」,「クリケット」は「特徴語＋カテゴリ語」を共通項とする疑似クエリ語の類語であり，真のクエリ語候補（以下，類似語候補と称す）となる．

その結果，共通項を用いた比喩表現を WWW 検索することで類似語候補が抽出される．

3 評価実験

3.1 実験環境

前章で述べた提案手法の有効性を検討するために評価実験を実施した．以下に，実験手順を示す．

ステップ 1 の入力として，疑似クエリ語「A」として「野球」の他,「ぶどう」や「ハーレー」など 28 単語を用いた．ただし，A は名詞句とする．図 3 に全疑似クエリ語を示す．

あじさい，イチョウ，イチロー，梅，オランダ，
カーリング，カップヌードル，クッキー，
コロッケ，ゴリラ，サソリ，サッカー，
サボテン，シカ，柔道，スキー，チューリップ，
テニス，ハーレー，バッタ，パスタ，帽子，
ホタテ，マラソン，メロン，野球

図 3: 疑似クエリ語「A」として用いた単語

出力としてカテゴリ語「B」を抽出する．ただし，B は名詞句とし，抽出限度を頻度上位 20 件とした．

ステップ 2 の入力として，ステップ 1 で得られたカテゴリ語「B」を用いた．出力として特徴語「α」が得られる．

ステップ 3 の入力として，カテゴリ語「B」と特徴語「α」を用いた．出力として類似語候補「γ」が得られる．

最後に，類似語候補「γ」について調べた．

3.2 実験結果

前章で述べた実験結果について，各ステップの実験結果毎に述べる．

＜ステップ１＞

全ての疑似クエリ語について 20 個のカテゴリ語が抽出できた．カテゴリ語のほとんどが疑似クエリ語についての上位語，もしくは属性を表現する語であった．表 1 に抽出例を示す．

表 1: カテゴリ語として得られた単語の例

クエリ語	カテゴリ語
野球	スポーツ，ゲーム，競技，ボール，…
梅	香り，酸味，風味，桃，桜，ピンク，…
シカ	動物，角，茶色，足，立派，蹄，…
ぶどう	房，色，香り，実，酸味，紫，爽やか …
ハーレー	バイク，アメリカン，エンジン，排気音 …

＜ステップ２＞

全 560 語のカテゴリ語から特徴語が抽出できたのは 212 語だった．特徴語は名詞，動詞，形容詞，もしくは，これら品詞を組み合わせた語句が抽出された．表 2 に抽出例を示す．

表 2: 特徴語として得られた単語の例

クエリ	カテゴリ	特徴語
野球	スポーツ	チーム，人気，マイナー，…
	ゲーム	ボール，熱い，コールド，…
	競技	団体，集団，戦略
ぶどう	房	長い，黒い
	色	美しい，深い，オレンジ，…
	香り	強い，甘い，甘酸っぱい，…
ハーレー	バイク	アメリカン，大型，高級
	エンジン	空冷，パワフル，OHV，…
	重量	Φ

＜ステップ３＞

類似語候補を含んでいる文書が全 18,342 件抽出された．つまり，１つの疑似クエリ語からは平均 655 件の類似語候補が抽出されたことになる．

また，属性語とカテゴリ語の組合せ１組からは平均 31 件の類似語候補が抽出できた．図 4 と図 5 に類似語候補を人手によって解析した例を示す．

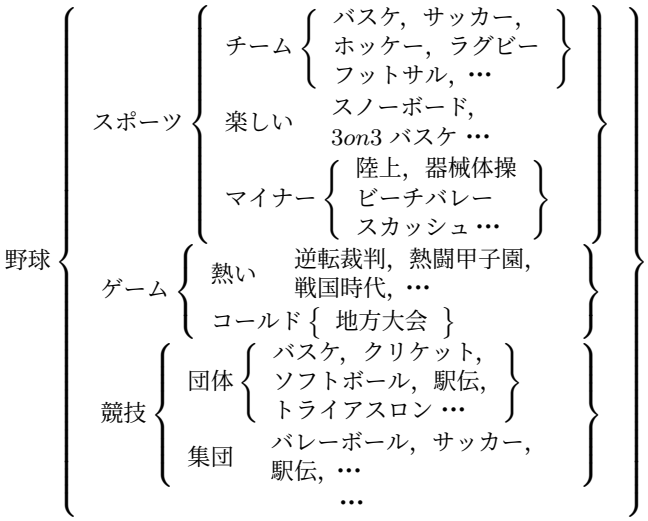

図 4: 「野球」に対する各要素抽出結果の例

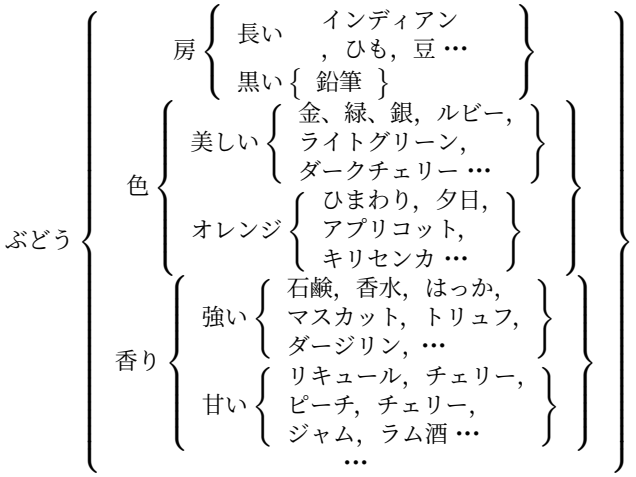

図 5: 「ぶどう」に対する各要素抽出結果の例

4 考察

本章では，本提案手法とステップ毎に抽出された要素群について考察する．

図 4 と図 5 より，抽出された類似語候補の内容が異なることから，「喩え」による比喩表現を利用することが有効であると考えられる．よって，本提案手法によってユーザの検索要求を顕在化するための候補を抽出できることを確認した．

ステップ１では，今回用意した全ての疑似クエリ語についての上位語や特徴を表現する語が抽出できた．例えば，疑似クエリ語「カーリング」については「スポーツ，ゲーム，作戦，ルール，… 」，「サボテン」では「多肉植物，花，乾燥，トゲトゲ，… 」などが得ら

れている．これらは，桝井ら [3] や川村ら [4] の知見と一致していることから妥当であると判断できる．

ステップ２では，図 2 のように，抽出した全ての特徴語について，表現意図が不明となる語はなかった．これは，WWW 上で使用される特徴語が単語に限定されているからであると考えられる．したがって，ステップ３での処理に用いても支障のない語が抽出できたと言える．

ステップ３では，属性語とカテゴリ語の組合せから平均 31 件の類似語候補が抽出できた．例えば，図 4 より「野球のような団体競技」からは「クリケット」,「ソフトボール」,「ラグビー」「カーリング」などが得られた．このように，疑似クエリ語「野球」に似ている「クリケット」や「ソフトボール」が抽出できた．その他のクエリからも類似語候補が抽出できており，本提案手法によって類似語候補を抽出できることを確認した．

以下，今後の課題について考察する．

ステップ２では，特徴語が抽出されなかったカテゴリ語を実験の対象外としている．つまり，カテゴリ語 560 語のうち 348 語が対象外となっている．しかし，実験の対象外としたカテゴリ語の中にも真のクエリ語が含まれている可能性があるため，これらの項目についての検討が必要である．

さらに，抽出された特徴語の対義語を用いることも検討している．対義語を用いることによって特徴語と類似語候補の抽出において網羅性向上が期待できるからである．例えば，特徴語として「大きな」が抽出されたならば「小さな」,「メジャー」ならば「マイナー」などである．

ステップ３において，特徴語とカテゴリ語の組合せによっては抽象的な表現となり，抽出結果に大きな差が出た．例えば，図 4 と図 5 中の，拡張部が「個人競技」の場合と「オレンジ色」の場合である．前者の表現では，全ての「競技」が含まれている文書の中から「個人競技」についての文書を絞り込む成果が見られた．しかし，後者の「オレンジ色」は抽象的な語のため，オレンジ色に関連する多量の文書が抽出された．例えば,「ひまわり」,「夕日」,「アプリコット」などである．これらに対応するため，特徴語と抽象語に分類し，抽象語については別処理が必要であると考えている．

次に，抽出された類似語候補において多数の名詞並列が含まれていた．例えば,「野球やサッカー」,「野球，サッカー，バスケット」などの名詞並列をそれぞれ１組としている．現時点では対応していないが，ユーザがわかりやすい形式で類似語候補を提示するためにも，名詞並列を解析する処理が必須である．

最後に，特徴語によってはユーザの所望する情報から遠ざかってしまう場合があった．例えば,「野球のような団体競技」からは「クリケット」が抽出できたのだが,「野球のようなチームスポーツ」からは抽出できなかった．これは，今回の調査で使用した比喩指標「のような」だけでは情報を網羅することができないためだと思われる．その他の比喩指標「みたいな」や「に似た」などの比喩指標を複数使用することで特徴語の網羅性を向上できる可能性がある．

5　おわりに

本論文では，ユーザの曖昧な検索要求の顕在化を支援することを目的とした手法を提案した．また，提案手法を基にしたシステムの段階的実装を行い，各ステップにおける抽出要素についての調査，および，分析を行った．

結果として，抽出要素の絞り込み方法にさらなる工夫が必要だが，本手法がユーザの情報要求を顕在化するための候補を抽出することができる可能性があることを確かめた．

今後の課題として，名詞並列や同義語などの語句を統合・分類する手法の提案，および，抽象的な表現や曖昧な語句についての調査を行う予定である．

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] 中村明：”比喩表現の理論と分類”. 共立出版,1977.

[2] 桝井文人, 久保真哉, 福本淳一：”比喩表現による検索手法の構想”. 人工知能学会情報編纂研究会第 3 回研究会資料,2010.

[3] 桝井文人, 福本淳一, 荒木健治：”比喩解釈を目的とする World Wide Web を利用した特徴値の適合性判定とそのフィードバック”. 電子情報通信学会論文誌,Vol.J89-D,No.9,pp.860-870,2006.

[4] 川村佳史, 桝井文人, 河合敦夫, 井須尚紀：”WWW から Descriptive 知識を抽出・掲示するシステム Murasaki の試作”. 言語処理学会第 12 回年次大会,P8-10(2006.3)