E-018

比喩表現を使用して検索を行うシステムの開発
Development of an Information Retrieval System Using Metaphorical Expression

松原 悠子 †
Yuko Matsubara

小林 一郎 ‡
Ichiro Kobayashi

1. 研究背景と目的

現在の情報検索システムには,google など優れた検索エンジンが数多く存在するが, ユーザの意図を明確に反映し, 検索できるシステムは未だ少ない. その問題点として, 検索をする際にユーザは検索対象に直接関連するキーワードを入力する必要があるが, ユーザがそのようなキーワードを知らない場合も多いという点が挙げられる.

本研究では, ユーザが検索対象の名称などを知らない場合に, 比喩などを使い別の表現を用いても, その意図を汲み取り検索ができるシステムの開発を目指す.

システムの具体例として, 今回は検索対象を MS ワードのヘルプ文書に特化して, 初心者ユーザが MS ワードの操作に困った際に, 比喩などの表現を用いて操作方法を検索できるシステムを開発した.

2. 検索エンジン作成

2.1 インデックス作成

本研究では, インデックスを用いた検索システムを基礎となるシステムとして, 比喩検索機能を追加し開発する. インデックスを作成するには, まず前処理として文書中の語の形態素解析を行う. 形態素解析には, 奈良先端科学技術大学院大学松本研究室にて開発された形態素解析器, 茶筅 (ChaSen)[2] を使用し,HTML 形式で保存されている MS ワードのヘルプ文書の形態素解析を行った. 次に, 形態素解析処理された文書中の単語を昇順に並べてインデックスを生成した (図 1 参照).

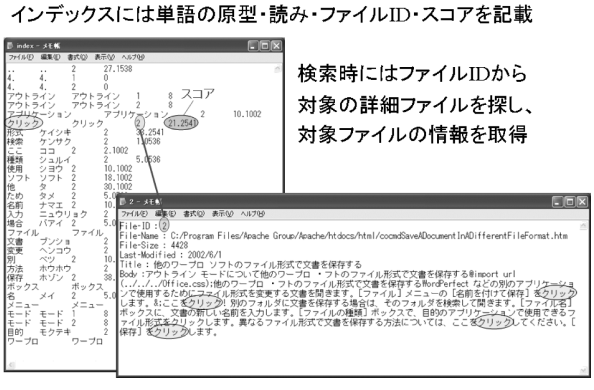

図 1: インデックス例

インデックス内の単語の重要度を決定するスコア計算は,TF.IDF 法に HTML タグによる重み付けを加えた計算方法を採用した.

†お茶の水女子大学大学院人間文化研究科　数理・情報科学専攻, Graduate Division of Mathematics and Computer Science, Graduate School of Humanities and Sciences,Ochanomizu University

‡お茶の水女子大学理学部情報科学科,Dept. of Information Sciences,Faculty of Science,Ochanomizu University

スコア= TF・log(N/DF) + HTML タグの重みづけ

例:“ ＜H1＞ワープロ＜/H1＞ ”があれば, 単語“ ワープロ ”に HTML タグのポイント (H1 タグの場合 8 ポイント) を加算する.

“ ワープロ ”のスコア= TF.IDF 法でのポイント＋8 ポイント

3. 比喩検索機能の追加

検索者が検索対象の正式名称を知らない際に, 比喩などを用いて, 他の表現でそれを説明する. そこで使用された語彙を, 検索者の意図にかなう索引語に変換して検索をかけるための辞書を作成し, システムに比喩検索の機能を追加した.

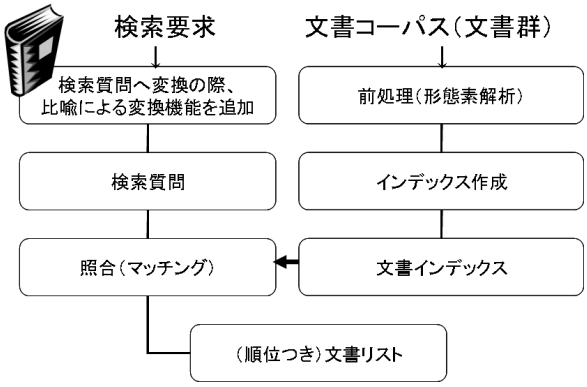

図 2: 比喩検索システム

3.1 理研コーパス分析

実際にユーザが MS ワードを使用していて, 操作方法や名称が分からず, インストラクタに, その操作方法や名称を尋ねる時のコーパスデータ (理研コーパス [1]) を約 3000 行分析し, ユーザがどの様な表現を用いて自分の必要とする情報を入手しているかを調べた. 表にその一例を示す. 以上の対話から,“ 区切る ”,“ 区切り ”のキーワードから“ スラッシュ ”が連想されることが分かるので「区切る」,「区切り」と「スラッシュ」を比喩検索用辞書に登録する. これにより, 検索質問文中に「区切り」があれば,「スラッシュ」に変換して検索を行えるようにした.

表 1: 対話例:スラッシュ

U:	「あの二千一年とかで、区切る、あれってどうやって入れるんでしたっけ」(スラッシュのこと, 「2001/1/31」を例に挙げている)
I:	「あ、区切りですか？」

(以下, 上記対話例中の「区切り」のような単語を『入力

キーワード』,「スラッシュ」のような単語を『出力キーワード』の仮称で説明する.)

入力キーワードは複数の単語の組み合わせを用いることも可能である. 例えば「ウィンドウ上部の青い部分」は,「タイトルバー」のことだと推測できることから,「ウィンドウ」「上部」「青」の組み合わせを入力キーワードとし,「タイトルバー」を出力キーワードとする (表 2 参照). 出力キーワードは全部そろわないと入力キーワードに結びつかないつくりになっている　入力キーワードの組み

表 2: 比喩検索用辞書 (一部)

入力キーワード				出力キーワード
区切り				スラッシュ
斜め	棒			スラッシュ
ウィンドウ	上部	青		タイトルバー
いちばん	上	ファイル	名	タイトルバー

合わせについては, どのような単語を登録するかの判断は人の手で行っている. 例えば, 青・ブルーなど入力キーワードに同義語が存在する場合は, 青を入力キーワードに含めたものとブルーを入力キーワードに含めたもの, 二種類の入力キーワード群を作る. 入力キーワード群 A と, 入力キーワード群 A ＋他の単語が同じ出力キーワードを導き出す場合は, 入力キーワード群 A の登録みで後者も変換されるので, 後者を辞書に加えていない. つまり, 出力キーワードを判別できる必要最小限の入力キーワードが登録されている.

3.2　比喩検索アルゴリズム

以下に, 上記した比喩検索用辞書を用いて検索を行うためのアルゴリズムを示す.

step1 辞書中の入力キーワード群をリスト化する.

step2 検索質問文を形態素解析して名詞と形容詞のみ抜き出し, 入力キーワードリストと照合.

step3 検索質問文中に入力キーワードリストに存在する語があれば, 今度は入力キーワードリスト中の他の単語が検索質問文中に存在するか照合する.

step4 リスト中の全単語が検索質問文中に存在すれば, 対応する出力キーワードを索引語として, インデックスを用いた検索を行う.

3.3　実行結果

図 3 に通常の検索結果を, 図 4 に比喩検索結果を示す.

図 3: 通常検索結果画面

図 4: 比喩検索結果画面

3.4　関連研究との比較

徳永ら [3] は, 概念のプロトタイプを意味ベクトルで表現し, 比喩の source 概念と target 概念を一致させる手法や語の共起関係を用いて,target 概念の検索を可能にする手法を提案している. また清田ら [4] は, ユーザが質問文中に使用する換喩表現に対して, 換喩解釈表現とペアをなす辞書を用意して, ユーザの入力表現の柔軟性を提供している. これらに対して, 我々の提案する手法は, 実際のコーパスから比喩表現を収集し, 比喩表現の対象となっている表現と比喩表現を一致させる辞書を構築することにより, 比喩表現解釈を実現している. これは, 比喩表現解釈に対して, 徳永らのアプローチとはまったく異なる方法を提案し, かつ, 清田らのアプローチより汎用性があるといえる. 本研究の長所としては, 実際に使用された表現なので, より確実なマッチングが行える点が挙げられる.

4.　まとめと今後の課題

本研究では、比喩で表現された検索質問を、本来のユーザの意図にかなう (と推測される) 索引語に変換して検索を行うシステムを開発した。
今回作成した比喩検索用辞書は、実際のコーパスを分析して、それに基づき作成したものである、収集された比喩表現の数は十分であるとは言えない。今後の課題として、比喩検索用辞書をより実用的なものとするため、より多様な比喩表現を集め、辞書を充実させていくつもりである。

参考文献

[1] 理研模擬対話音声コーパス (仮称), http://www.brain.riken.go.jp/labs/lbis/corpus/Notice040524.htm

[2] 奈良先端大学院大学松本研究室　茶筅 (ChaSen),http://chasen.naist.jp/hiki/ChaSen/

[3] 徳永健伸, 田中穂積, 岩山真.“概念・属性間の確率的関係を用いた比喩理解の計算モデル－比喩検索を前提として－”,Proc. of Symposium on Large-Scale Knowledge Resources, 東京工業大学,pp.61-64,March,2005.

[4] 清田陽司, 黒橋禎夫, 木戸冬子.“自動抽出した換喩表現を用いた係り受け関係のずれの解消”, 言語処理学会 第 10 回年次大会 発表論文集, pp. 305-308, March, 2004.