中心キーワードからの周辺キーワード抽出手法の検討

溝渕 正剛 † 坪川 宏 †
† 東京工科大学 バイオ・情報メディア研究科

1 はじめに

Web に情報が増大している現在，Web から情報を探す方法として，検索エンジンを使ったキーワード検索が一般的に行われている．知りたい情報に関係するキーワードを入力して知りたい情報を検索し，必要な情報のあるページを探す方法である．この場合探している内容のページに関係するキーワードを入力できない場合は目的のページにたどり着けない場合がある．

本研究では，一般的に知られていない検索結果を表示し，一般外の情報 (一般的に知られていない情報) を抽出・可視化することとする．

Web 上にある膨大な情報から一般的な検索結果に表示されないもの，一般的な検索結果ではないものを抽出・可視化するという目的としている．

そのための手法として現在思考している行動・物・言葉の情報からその周囲にあるものを抽出し，グループ化することで，思考の中心にあるもの，それ以外のもの，に分類し，一般的な検索結果ではないものにあるものを抽出・可視化する．

また，検索語の共起情報を利用した単語の分類研究として [1] では，検索語の共起情報を利用して，単語クラスタリングを行い，その応用例としてシソーラスの階層構造に基づく類義語検索を行っている．

従来行っていた研究 [2] の課題を踏まえ，本研究では検索語の共起情報を利用し，情報を抽出しその周辺に含まれる情報をグループ化し可視化することで一般外の情報に到達する手法の検討を行う．

2 中心キーワードと周辺キーワード

- 中心キーワード
  ユーザが検索したキーワード　以下中心キーワードを中心語とする．

- 周辺キーワード
  現在思考している行動・物の情報から抽出したその周囲にあるキーワード

A study of method of Extraction of keywords from the center keyword
† Masataka Mizobuchi
† Hiroshi Tsubokawa
Graduate School of Bionics,Computer and Media Sciences
Tokyo University of Technology (†)

3 システム概要

図 1 にシステム概要図を示す．本システムは以下の機能を実装する．

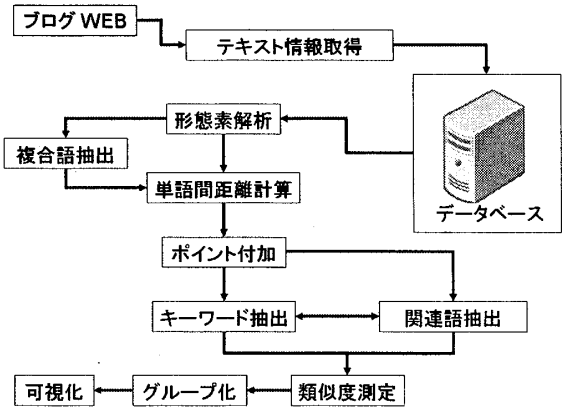

図 1: システム概要図

- ブログ情報・テキスト取得の登録
  テキスト情報を取得するページ (RSS/Ping サーバ) を登録する．システムに URL・サイト名を入力・登録する．取得日時，件名，テキスト情報，URL の登録を行う．

- 形態素解析
  データベース からテキスト情報を文単位で取り出し形態素解析を行う．形態素に分割された単語ごとに，形態素，品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用形，活用型，原形，読み，発音をデータベースに登録する．

- 複合語抽出
  形態素解析された情報を基に複合名詞を抽出し，２語以上の語にわかれていた名詞を複合名詞として登録する

- 単語間距離計算
  形態素解析された情報を基に単語を距離 1 として距離を計算する．

- ポイント付加
  下記の式で中心キーワードと同一文に出現した各

語のポイントPを計算し，関連語抽出を行う．

$$P = \sum_{k=1}^{\text{出現頻度}} \frac{1}{\text{単語間距離}}$$

- 類似性を用いたグループ分け
  中心語と関連語の関係性の上位の語をnとし，上位の語が全体に占める割合を計算し，各中心語通しを比較して関連語の上位に占める割合の似ている語を調べ，下記の式で類似度を判定する．

$$\text{類似度} = \sum_{k=1}^{n} \frac{(100 - |A - B|) \times \frac{(A+B)}{2}}{100}$$

$$A = \frac{\text{中心語と関連語Aのポイント P}}{\text{各関連語のポイントの合計}} \times 100$$

$$B = \frac{\text{中心語と関連語Bのポイント P}}{\text{各関連語のポイントの合計}} \times 100$$

4 検索の可視化

本手法で可視化する手順を示す

1. キーワードを入力する
   ユーザは中心として表示させるキーワードを入力する．（今回は「ラーメン」と入力）
2. 入力したキーワードと周辺キーワードの表示を図2に示す．周辺キーワードはグループごとに分類して，大きな○で示したものはグループの代表する語、小さな○はそれ以外の語として表示している．

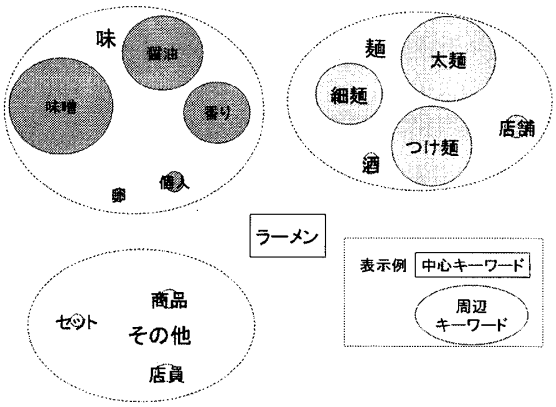

図 2: 入力したキーワードと関連する語

関連する語の表示基準

- さまざまな関係の語を表示
  関係の強さに関係なく、グループを代表する語とそれ以外の語をランダムで表示する
- 一般的な結果の上位を表示する一般的に関係が高いもの（グループを代表する語）を表示する機能

3. ユーザが関連するキーワードを選択したときの図3に示す．（周辺キーワードとして「商品」を選択）

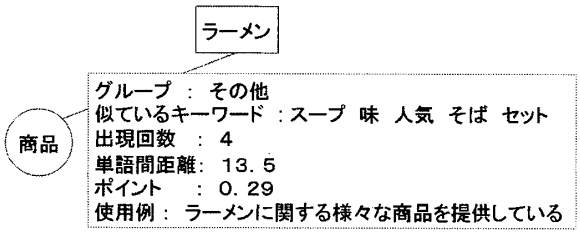

図 3: ユーザが関連するキーワードを選択したとき

- 関連する語と選択されたキーワードを表示
- 選択された語がなぜ関係があるのか表示する
  - 所属するグループ
  - 関係性が似ているキーワードの表示
  - 単語の出現頻度，単語の距離，ポイント数を表示
  - 関係が見られたときの実際の文の表示

5 まとめ

中心キーワードから周辺キーワードを抽出する手法の検討を行った．従来手法[2]から抽出手法・可視化手法を変更した．図3に示したようにキーワードを選択すると似ているキーワードに一般的に関係があるのか判断できないキーワードが抽出されている．これにより従来手法では埋もれていたキーワードを抽出した．

今後の課題としては，中心キーワードとして指定できるキーワードの一般化を行うことや可視化までの時間の短縮，精度向上の改善が必要である．

参考文献

[1] 有田一平, 菊池英明, 白井克彦：”検索語の共起情報を利用した単語クラスタリングとWeb検索への応用”情報処理学会,2007(76) pp.115-120 20070724

[2] 溝渕正剛, 坪川宏：”着目キーワードからの連想検索手法の検討”，第70回情報処理学会全国大会(2009)