着目キーワードからの連想検索手法の検討

溝渕 正剛 † 坪川 宏 †
† 東京工科大学コンピュータサイエンス学部コンピュータサイエンス学科

1 はじめに

近年のインターネットの普及により，インターネット上に多くの Web ページが作成・更新されている．現在 Web ページを探し出す手段として Web 検索（キーワード検索）という手法が一般的に使われている．この手法ではユーザが 1 つのキーワードで検索を行った時，検索エンジンがインデックスしているページ数が増大している今日では目的のページにすぐにたどり着くことが難しくなっている．検索に慣れないユーザが検索を行った場合に適当な第二，第三キーワードが分からず，目的のページに到達するまでに多くの時間を要する場合がある．

そこで，キーワード検索を使用時に第二，第三のキーワードをユーザに提供し，ユーザの利便性を高めようという関連検索という手法がキーワード検索と併用されている．現在の関連検索では他の多くのユーザが組み合わせた語について関連検索として提示しているため，ユーザが探したい内容のページが専門的な内容で知識が必要な場合に関連検索として提示されないという問題点がある．

また、キーワードの組み合わせではなく Blog の内容を解析し，Blog を推薦する MineBlog[1] では記事の内容を関連性，相違性，話題性を考慮し，文章解析を行い興味発見につながる Blog 記事推薦を行っており，キーワード入力時の支援と異なっている．

2 連想検索手法

本研究では関連検索のように多くのユーザが入力した語 (キーワード) の組み合わせではなく，インターネット上（ブログ）から日本語テキスト情報を定期的に収集して各ブログ記事の単語情報からキーワードの共起数，単語間距離，構造解析を行い，各解析結果を組み合わせて数値化して，主要な単語のつながりを可視化する．それによって Web 検索のキーワード入力 (選択) 時にユーザが気づきにくい，第二，第三キーワードを抽出し，新たな視点 (キーワード) からの検索を支援するシステムを検討する．

A study of method of associative searching from attention keyword.
† Masataka Mizobuchi
† Hiroshi Tsubokawa
School of Computer Science, 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．

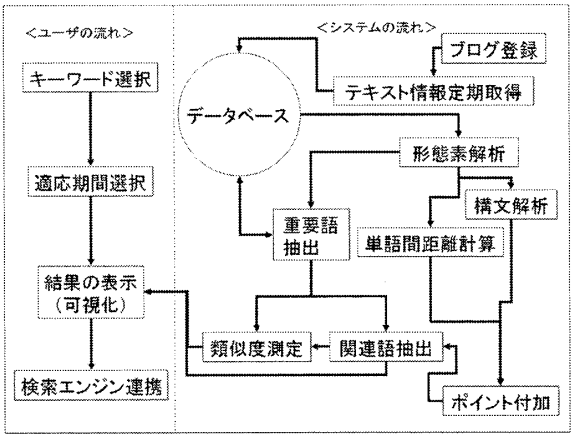

図 1: システム概要図

- 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．登録されている RSS に対して更新が行われているかを確認し，更新が行われている場合は取得日時，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．

- 単語間距離計算

  形態素解析済みのテーブルから単語を距離 1 として距離を計算する．計算済みの距離のデータから名詞，動詞のデータを抜き出し，形態素，品詞，原形，距離情報をデータベースに登録する．

- 係り受け計算

  データベースからテキスト情報を取り出し，構文解析を行う．解析結果から係り受けがあった場合

にはどの単語（文節）ごとにどの単語（文節）と係り受けがあるのかをデータベースに登録する．

- 重要語・関連語の抽出
  集計された形態素の中で頻度が上位にある名詞，動詞を重要語として抽出する．抽出された重要語は重要語リストとしてデータベースに登録する．また，重要語リストに登録されている語から単語間ポイントの付加と関連語抽出を行う．システムは重要語 A と重要語 B が含まれる記事をデータベースから抽出し，記事ごとの重要語の共起回数，単語間距離，係り受け結果に基づき，下記の式で重要語抽出と関連語判定に基づき総合ポイント P を計算し，関連語抽出を行う．

$$P = \left(\sum_{k=1}^{\text{総記事数}} \frac{\text{係り受け結果}}{\text{単語間距離}} \times 100\right) \times \frac{\text{総共起数}}{\text{総記事数}}$$

- 類似性を用いたグループ分け
  データベースに登録されている重要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，各重要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．

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

$$A = \frac{\text{重要語と関連語 A の総合ポイント P}}{\text{各関連語のポイントの合計}} \times 100$$

$$B = \frac{\text{重要語と関連語 B の総合ポイント P}}{\text{各関連語のポイントの合計}} \times 100$$

4 検索の可視化

検索の可視化には以下の機能を実装した．

- 着目キーワードを中心とした関連語の可視化
- 検索期間による表示
- 類似性を用いたグループ分け支援表示
- 着目キーワードの追加と再表示

5 実装結果

本システムを用いて，自動的に収集されたデータから「ラーメン」というキーワードの連想検索を行った結果，図 2 のように連想される抽出し，キーワードを可視化することができた．単純に各記事の中に含まれるキーワードの出現回数で可視化をした場合と本システムで用いた，重要語抽出と関連語判定に基づき総合ポイント P を計算し，可視化をした場合と結果を比較すると上位のキーワードの順位に変動があった．上位のキーワードに変動があった理由として上位に浮上した「味噌」というキーワードに着目すると，出現回数は５８回で共起回数５位であるが，平均単語間距離が１８となっている．共起回数で１位であった「スープ」に着目すると出現回数は１１５回であるが，平均単語間距離が２７となっている．したがってこの結果に重要語抽出と関連語判定に基づき総合ポイント P を計算することで，順位が変動している．

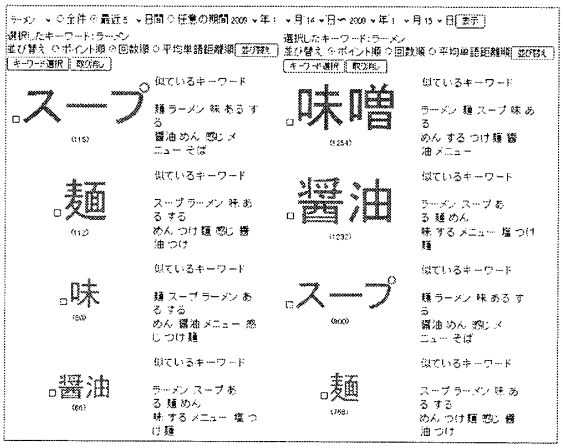

図 2: 検索の可視化

また，検索対象のデータの期間を変更することで浮かび上がったキーワードの変化を確認することができた．このことからブログの流行などの情報をこの手法を用いることで読み取ることが可能と考えられる．

6 まとめ

本システムでは単純に記事の中に含まれるキーワードの出現回数で単語の重要度を測定するのではなく，単語間距離などの情報を付加し，可視化を行っている．この手法を用いることで，出現回数では上位に出現しなかったキーワードが上位に出現することや，出現回数が多く上位にあったキーワードにおいても順位の変化を確認することができた．

参考文献

[1] 森本和伸, 林貴宏, 尾内理紀夫：”MineBlog:興味発見を支援する blog 記事推薦システム”情報処理学会誌，Vol.47，No4，1171-1180（2006）

[2] 奥村学：”blog マイニング――インターネット上のトレンド, 意見分析を目指して――”人工知能誌，vol.21，No.4，pp.424-429（2006）