Vol.2010-ICS-161 No.3
2010/11/22

着目キーワードとつながりのある埋もれた語の抽出手法の検討

溝渕正剛[†1]　坪川宏[†1]

ユーザが着目したキーワードを用いて収集した情報から形態素解析，共起頻度などの情報を利用しキーワードの抽出を行い，それらの関係性を利用してグループを生成する．ユーザは作成されたグループの中に分類されたキーワードから気になる語を選択する．システムはユーザが選択した気になるキーワードを中心に情報の再抽出を行い，着目したキーワードとつながりがあるが埋もれていた語の抽出を行う．

A study of method of extraction of buried keyword from attention keyword

Masataka Mizobuchi[†1] and Hiroshi Tsubokawa[†1]

In this paper, the extraction method of the hidden word connected with the keyword which the user chose is proposed. First, the words which coincide with the word chosen by user from the result of web are collected. Next, a relevant word is extracted from the word which exists around the word which coincided, and a group is formed by word which relevance resembles. The word connected with the keyword which the user chose in the group is extracted. Although the frequency of the extracted word is low, the word has a strong relation with the word in the group which the user chose.

1. はじめに

最近，ブログやSNS，Twitterのような気軽に個人が情報発信する機会の増加に伴い，インターネット上の情報は爆発的に増大している．こうした情報の中から，ユーザが必要な情報を探す方法として，様々な方法が検討されてきたが，現在では一般的にキーワード検索が利用されている．キーワード検索では，検索エンジンが事前に収集したWebSiteにランク付けを行って上位から順に情報が表示されてため，より多くのリンクが張られている主要な情報や出現頻度が高い情報を探すことには優れているが，リンク数が少ないWebSiteや出現頻度の少ない情報は表示されにくくなる．そのため，関連がある情報であっても主要ではない情報,「埋もれた語」にたどり着くことは困難となってしまっている．

†1 東京工科大学
Tokyo University of Technology

2. 関連研究

キーワードを抽出する方法として単語の共起情報を利用した研究も盛んに行われている．検索語の共起情報を利用した単語の分類研究[1]として，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，階層構造を持ったシソーラスの自動構築を行っている．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．この手法では，新語に対応するため定期的なシソーラス更新などが必要になってしまう．このほか共起に基づく研究[2]として，任意の関係にある語のペアを与え，類語関係にある語のペアをWebから取得している．この手法では，任意の関係にあるペアの語を用意する必要があるが，埋もれた語のペアを予め用意することは難しいと考えられる．

3. 目的

そこで本研究では，従来研究[3][4]の課題を踏まえ，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からグループを作成し分類・整理を行い，埋もれた語を抽出する手法の検討を行う．

4. 埋もれた語

本研究において「埋もれた語」とは，ユーザが着目したキーワードと直接共起していないがいくつかの関連語とつながりがあり，出現頻度が低い語と定義する．埋もれた語のイメージを図1に示す． ユーザが着目したキーワードとつながりのあるキーワードを基に単語の関連性・類似性を抽出しグループを作成する．ユーザが気になった語の周辺にあり，気になる語の属するグループとつながりのある語の中から，出現頻度が低い語を選択し埋もれた語と定義する．

Vol.2010-ICS-161 No.3
2010/11/22

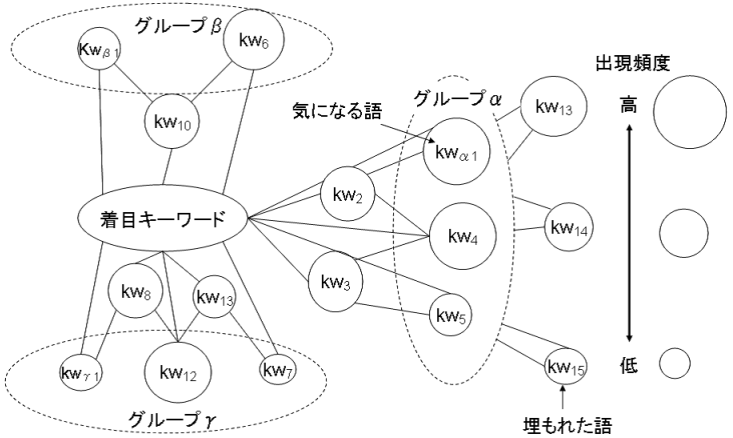

図 1 埋もれた語
Fig. 1 buried keyword

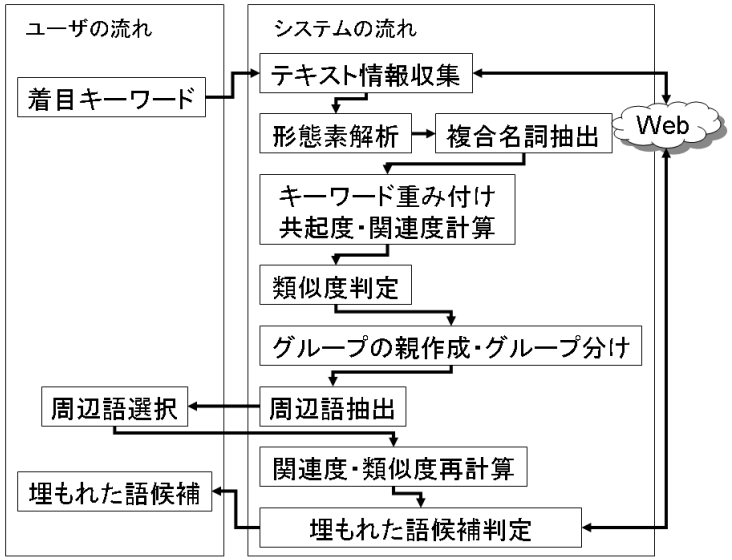

図 2 システム概要図
Fig. 2 flow of system

5. システム提案

5.1 システムの流れ

図 2 にシステム概要図を示す．システムはユーザが着目したキーワードを基に Web から情報を収集し，形態素解析，出現頻度，複合名詞統合，単語間距離の計算を行い，キーワードに重み付けを行い関連語として抽出する．関連語をそれぞれ比較し類似度を判定する．それらの結果を用いてシステムはグループを作成し，周辺語を抽出する．ユーザは周辺語を選択し，システムは着目キーワードと選択された周辺語を用いて関連度，類似度の再計算を行う．その結果から埋もれた語候補を抽出しユーザに提示する．

5.2 キーワード抽出手法

- テキスト情報収集
  ユーザが着目したキーワードを基に Web から情報を取得する．情報取得には Yahoo! 検索 Web API[5] を使用する．
- 形態素解析
  着目キーワードを利用して WEB から収集した情報を用いてテキスト情報を文単位で取り出し形態素解析を行う．また，形態素解析には MeCab[6] を用いる．
- 複合名詞統合
  形態素解析結果を用いて複合名詞を抽出し，複数に分割された形態素を 1 つの形態素として統合する．形態素の統合は「名詞」が連続して出現した場合に統合する．
- 出現頻度計算
  形態素解析結果を用いて各形態素の出現頻度 (wf) を集計する．
- 単語間距離計算
  形態素解析結果から形態素を距離 1 として距離を計算する．例えば，s 番目に存在したキーワード x とキーワード y の距離は $wl_s(x, y)$ とする．
- 単語の間に含まれる関係性
  単語の間に含まれる助詞の関係性を抽出する．これらの情報を取得することで，抽出された関連語の前後関係，上下関係を判定する．
- 共起関係の利用

キーワード x とキーワード y が 1 つの文において同時に出現した場合共起関係にある．この場合の共起出現頻度を $wf(x,y)$ とする．ここで x と y のシンプソン係数を $sim(x\_y)$ と定義し，式 (1) で計算する．

$$sim(x\_y) = \frac{wf(x,y)}{\min\{wf(x), wf(y)\}} \qquad (wf(x), wf(y) \geqq N) \tag{1}$$

式 (1) では全体的な出現頻度は上位ではないキーワードでも，特定のキーワードとよく結びついているキーワードを抽出するために，共起度が高くなるように式を定義している．また，今回は N=5 として計算を行う．

- 関連度
  式 (2) を用いてキーワード x とキーワード y の関連度 (R) を計算し，関連語抽出を行う．

$$R_{(x\_y)} \;=\; sim(x\_y) \times \{ \sum_{s=1}^{wf(x,y)} \frac{1}{wl_s(x\_y)} \} \tag{2}$$

  関連度の計算では，単語間距離が近ければ，関連度は高くなり，遠ければ低くなる．キーワード x とキーワード y が共起出現するたびに，それぞれのキーワードの単語間距離を用いて計算し，累計を行い，最後に共起度の重み付けを行う．同様に他のキーワードの組み合わせの関連度を計算する．
- 類似度判定
  キーワード x の関連語とキーワード y の関連語に含まれるキーワードを比較し，同じキーワードが存在した場合に類似していると考えることができる．キーワード x とキーワード y の共通の関連語として r_1,r_2,r_3,$\cdots$,r_L が存在する場合，式 (2) を用いてキーワード x の関連度 $R_{(x\_r_1)}$,$R_{(x\_r_2)}$,$R_{(x\_r_3)}$,$\cdots$,$R_{(x\_r_L)}$ を求め，キーワード x の関連度として $R_{(y\_r_1)}$,$R_{(y\_r_2)}$,$R_{(y\_r_3)}$,$\cdots$,$R_{(y\_r_L)}$ を求める．L はキーワード x の関連語に含まれるキーワードとキーワード y の関連語に含まれるキーワードに同じキーワードが存在した回数となる．式 (3) を用いて類似度 (S) を判定する．

$$S_{(x\_y)} = \sum_{k=1}^{L} \frac{\left(100 - \mid R_{(x\_r_k)} - R_{(y\_r_k)} \mid\right) \times \frac{\left(R_{(x\_r_k)} + R_{(y\_r_k)}\right)}{2}}{100} \tag{3}$$

  同様に他のキーワードの組み合わせの類似度を計算する．
- グループ作成
  グループの作成以下の手順で行う．
  - グループの親を決定する
    式 (4) を用いてグループの親を決定する．多くのキーワードと似ていると判定したキーワード (G の値が最も大きいキーワード) をグループの親 (a) として抽出する．

$$G_x = \sqrt{R^2_{(x\_y)} + S^2_{(x\_y)}} \tag{4}$$

    $R_{(x\_y)}$ = グループの親 (x) と関連語 (y) の関連度
    $S_{(x\_y)}$ = グループの親 (x) と関連語 (y) の類似度

  - グループに属する関連語を決定する
    グループの親 (a) と似ている関連語をグループ (G_a) に入れる．グループ (G_a) に属するかの判定を行うために，G_a と各関連語の関連度，類似度を計算する．式 (4) を用いて関連語がグループ (G_a) に属するかを判定する．G_a の値が上位に属する語をグループに属すると判定するため，複数のグループに属する場合もある．また，グループに属する基準については後述する．

  同様にグループを他のキーワードが親となるグループを作成する．
- 周辺語表示
  分類された関連語は主要なキーワードを中心に表示されている．これらを周辺語とする．ユーザは気になった周辺語を選択する．システムはユーザが選択した周辺語と着目キーワードに着目して，類似度・関連度を再計算する．
- 類似度・関連度を再計算
  類似度・関連度の再計算では以下のように重みを追加して行う．
  - 助詞の情報に着目する
    キーワード x とキーワード y の間に含まれる助詞の情報に着目して前後関係・上下関係を抽出する．助詞の中で格助詞の「ガ格」「ニ格」「ヲ格」に着目し関係性の抽出を行い，抽出ルールは表 1 に従う．例えば,「キーワード x がキーワード y に××する」のような文が存在した場合キーワード x とキーワード y を抽出して前後関係があり，キーワード x がキーワード y より前にあると判定する．似たような

文の構成でも受動態の場合が考えられる.「キーワード y がキーワード x に××される」のような表現を抽出した場合は「される」を「する」に変化させ，キーワード x とキーワード y を交換することで能動態の文に変形させることができる．このような変換を行い，キーワード x がキーワード y より前にあると判定する． これらの情報を集計し，また，他のキーワードも格助詞抽出ルールに基づき関係性を抽出する．これにより他のキーワードの前に多く出現するキーワードを抽出する．そのようなキーワードは他のキーワードと比べると上位の関係にあると考えられ，単語間距離が＋となるように wl_s を変更する．変更した wl_s 結果を基に式 (2) で再計算を行う．

– 単語間距離の前後関係

関係性を再計算するキーワード x とキーワード y の前後関係・上下関係を抽出するため，単語間距離 (wl_s) を計算する場合に±を追加して計算する．この正負の情報を関連語の式に追加することで前後関係を抽出する．また，キーワードごとに出現数を正の場合は wp，負の場合は wn と定義し，各キーワードの情報ごとに集計する．関連語の式にすべての正負の情報を追加すると，平均化され 0 となる可能性があるため，正負の数どちらか多いほうの情報のみを式 (5) に追加する．式 (5) でキーワード x とキーワード y の関連度 (R) を再計算し，関連語抽出を行う．同様に他のキーワードの組み合わせの関連度を計算する．

$$\begin{cases} R_{(x\_y)} = sim(x\_y) \times \{ \sum_{s=1}^{wf(x,y)} \frac{1}{wl_s(x\_y)} \} & wp > wn \\ R_{(x\_y)} = sim(x\_y) \times \{ \sum_{s=1}^{wf(x,y)} \frac{1}{wl_s(x\_y)} \} \times (-1) & wp < wn \end{cases} \tag{5}$$

表 1 格助詞抽出ルール
Table 1 a case particle rule

能動態	○○が△△に××する
	○○が△△を××する
	○○を△△が××する
受動態	△△が○○に××される

- 埋もれた語候補

選択された周辺語と着目キーワードに着目して再計算した結果から，周辺語選択時には抽出されていない周辺語を埋もれた語候補として抽出する．着目キーワード，選択した語，埋もれた語候補が Web 検索で何件ヒットするかを示し，埋もれた語候補としてユーザに表示する．

6. 結　　果

6.1 共起情報の追加

グループの親を効率的に抽出するために共起情報を強く反映した結果と検証する．共起情報を追加したことによる結果の変化は着目キーワードを「スポーツ」として設定し取得した 3000 件の文 (2010 年 1 月に Web から収集した情報) を利用した．グループの親の候補として抽出されるキーワードの件数の比較，上位に抽出されるキーワード (表 2) の検証した． 共起情報を追加する前は，グループ作成時に親の候補となる語に様々なキーワードと

表 2 上位に抽出されるキーワード
Table 2 High-ranking keyword

野球	サッカー
ニュース	ゴルフ
社会	大会
オリンピック	ワールドカップ
沖縄	オリックス

広く使われグループに属すべきではないキーワードが含まれていた．そこで共起情報を追加し，より結びつきが強いキーワードを抽出できるように抽出手法を変更した．抽出された結果を比較すると，共起情報追加前の 252 件から，共起情報追加後に 148 件へと減少した．上位に抽出されているキーワードにほとんど変化はないが，そのほかのキーワードでは，日本など様々なキーワードと結びつくと考えられる候補は上位に抽出されなくなった．この結果から様々なキーワードと結びつきやすいキーワードが上位に抽出されにくくなり，グループの精度が向上した．

6.2 グループ作成と周辺語表示

6.1 の結果で上位に抽出され生成されたグループの結果を図 3 に示す． 結果の野球，サッカー，ニュースはグループの親となったキーワードを示している．また，グループの親と同

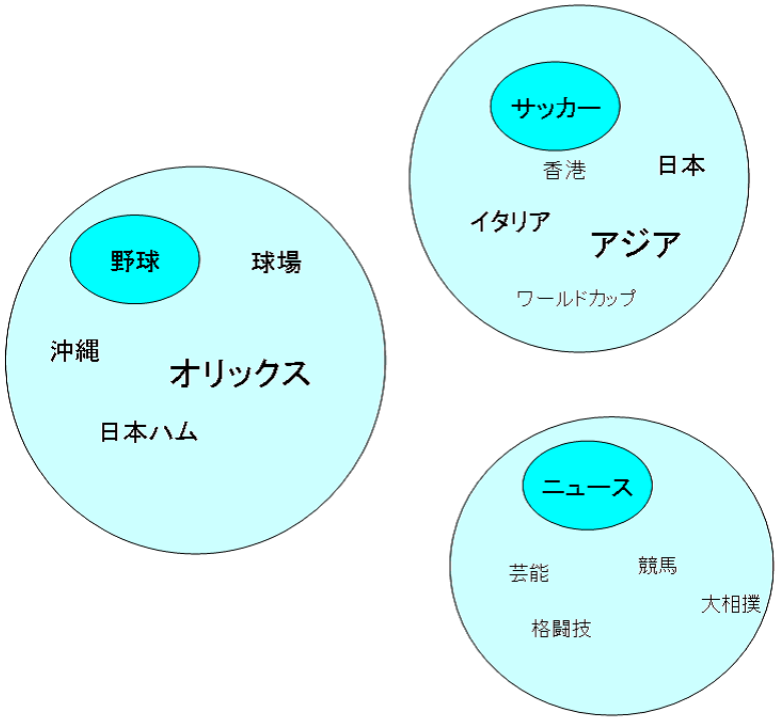

図 3 グループ作成
Fig. 3 group making

一円の中に存在するキーワードはその親のグループに属するキーワードとして抽出されたキーワードである．グループに属するキーワードの文字の大きさはグループの親との結びつきの強さを表現している．グループに属するキーワードの判定の閾値として今回は G_x の値が上位 10 件，上位 5%，上位 10%，を比較し検討を行い 5%の結果を示している．

6.3 埋もれた語候補の抽出

6.2 で作成されたグループの中から「野球」を選択肢，埋もれている語を抽出するため，再計算を行った．その結果を図 4 に示す．6.2 のグループ作成時にはグループとして抽出されていない「小瀬」「宿舎」という埋もれた語候補として抽出された． 埋もれた語候補として抽出されたキーワードを用いて Web 検索を行い検索件数を調査した結果を表 3 に示す．

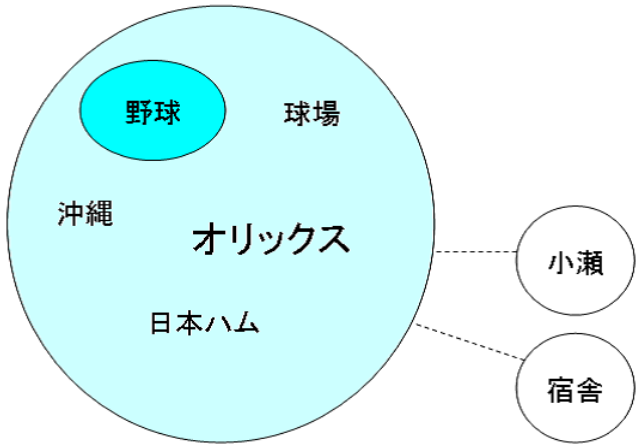

図 4 埋もれた語候補
Fig. 4 buried keyword

検索には Google 検索を用いており，複数のキーワードで検索した場合は AND 検索を行っている． このグループの代表的なキーワードである「オリックス」と「小瀬」のキーワードを「スポーツ」「野球」と組み合わせてヒットした件数を比較するとヒット件数の割合で 8.55%の「小瀬」が抽出されている．また同様に「オリックス」「宿舎」で比較するとヒット件数の割合で 19.7%の「宿舎」が抽出されている．この結果より埋もれた語が抽出されていると考えられる．

表 3 Web 検索結果
Table 3 Retrieval result

検索キーワード	ヒット件数
スポーツ　野球 オリックス	152,000
スポーツ　野球　小瀬	13,000
スポーツ　野球　宿舎	29,900
スポーツ　小瀬	35,300
スポーツ　宿舎	98,900
スポーツ　野球	7,860,000

7. システムの評価

7.1 グループの妥当性

辞書による分類と本システムによる結果を比較しグループの妥当性を適合率，再現率，F値を用いて評価を行う．また，一般的なクラスタリング手法との結果の比較を行う．

7.2 埋もれた語の評価

埋もれた語候補として抽出したキーワードの妥当性を検証する．評価には埋もれた語候補として出現したキーワードがWeb上に何件あれば埋もれた語と考えられるのかという尺度を用いて評価する．

8. ま と め

着目キーワードとつながりのある埋もれた語抽出手法の検討を行った．グループの親を効率的に抽出するために共起情報を強く反映した結果，様々なキーワードと結びつきのあるキーワードが抽出されにくくなり，グループの精度を向上させることができた．

また，周辺語として抽出されたキーワードを選択し，関連度・類似度を再計算した結果から抽出されたキーワードのヒット件数はグループの代表的なキーワードとして抽出された語より格段に少なく有効な埋もれた語として抽出されている．

今後の課題として，システムの評価として提案手法の有効性の評価を行うこと．またユーザインターフェース部の実装などが必要である．

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) 加藤 誠, 大島 裕明, 小山 聡, 田中 克己田：”共起に基づくWebからの類似関係のブートストラップ抽出”, 日本データベース学会論文誌 Vol.8,No.1 2009
3) 溝渕正剛, 坪川宏：”中心キーワードからの周辺キーワード抽出手法の検討”，第72回情報処理学会全国大会 (2010)
4) 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71回情報処理学会全国大会 (2009)
5) Yahoo!デベロッパーネットワーク,http://developer.yahoo.co.jp/
6) MeCab (和布蕪),http://mecab.sourceforge.net/