D-021

単語クラスタリングを用いた検索キーワードの提示

小西　隆太†　　上原子　正利‡　　小柳　滋‡

†立命館大学大学院理工学研究科　　‡立命館大学

1 はじめに

インターネットの発展と優れた検索エンジンの登場によって，我々はさまざまな情報を容易に手に入れられるようになった．しかし同時に，急激に増加し膨大な量となったテキスト情報は目的とする情報の発見を困難にした．

検索エンジンの利用には検索キーワードの入力が不可欠であり，検索結果は検索キーワードによって大きく変化するため，適切なキーワードの選択が重要になる．検索したい対象が明確に決まっている場合は比較的容易にキーワードを決定できるが，検索したい対象が曖昧である場合，キーワードの決定は難しい．

現在，このような問題を解決するためにさまざまな技術が開発されている．その1つとして，ユーザがキーワードを入力すると，他のユーザが過去の検索で利用したキーワードの組を提示するものがある．この技術は入力されたキーワードに関連した複数の検索キーワード候補を表示することにより，ユーザの検索キーワード決定を補助する．また，キーワードの組を提示することにより文書の絞込みが可能となり，結果として目的の文書にたどりつきやすくなる．この技術は多くのユーザにとって有益であろう．

しかし，この技術は入力されたキーワードを含む検索キーワード候補しか提示しないため，キーワードについての異なった表現には対応できない．また，他のユーザが過去に行なった検索データを用いるため，検索にあまり利用されない単語や具体性の高い単語については候補が提示されにくい．

本稿では，過去に用いられた検索キーワードに依存するのではなく，入力されたキーワードで検索された文書に含まれる単語をクラスタリングすることにより，キーワードの関連語をユーザに提示するシステムを提案する．

2 提案方式

図1に処理の流れを示す．本システムではまず，ユーザが入力したキーワードを既存の検索エンジンに問い合わせ，検索結果を得る．得られた検索結果に含まれる複数の文書に対し単語抽出を行い，各文書に含まれる単語の一覧とその出現回数を取得する．つぎに，入力されたキーワードの関連語を決定するために，得られた単語一覧とその出現回数を用いて単語のクラスタリングを行なう．その後，関連語として検出された語に対し，同位語・上位語・下位語のいずれであるかの判別を行なう．ユーザが入力したキーワードの言い換えとして同位語，より広い話題を検索するために上位語，より話題を絞り込むために下位語をユーザに提案する．この際，ユーザが単語の選択を容易に行なえるようにするため，単語をクラスタ化して提示する．

以上の説明をまとめると次のようになる．

(1) 既存の検索エンジンを用いて文書を取得
(2) (1) の文書から単語を抽出
(3) 単語のクラスタリング
(4) 単語の判別
(5) 関連キーワードの提示

主に絞込みのための関連語だけを提示する従来技術と異なり，関連語を判別し幅広い柔軟なキーワードの提示を行なう点が本研究の特徴である．

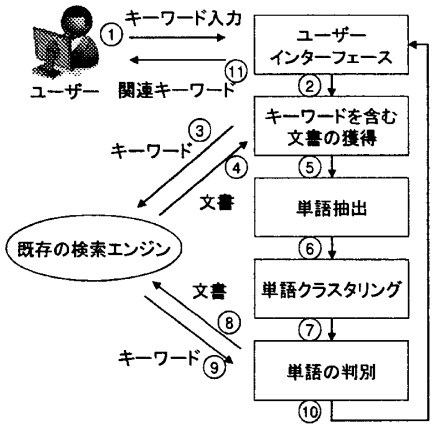

図1: 処理の流れ

3 実現方法

3.1 文書の取得

ある単語と他の単語の関係を調べるために，ある単語を含む複数の文書が必要である．この複数の文書を得るためにYahoo!検索 Web サービス*を用いる．Yahoo!検索 Web サービスはウェブページ検索 Yahoo!の検索結果を XML 文書形式で取得できるサービスである．Yahoo!の検索結果にはページのタイトルや URL，ページサマリーなどが含まれており，このページサマリーを用いることで効率的に文書を獲得する．

ここで，大量の文書を得ればその分精度が向上すると考えられるが，一方で処理に要する時間が増加する．そのため，文書数を適切に設定する必要がある．

3.2 単語の抽出

単語には名詞，形容詞，動詞などさまざまな品詞があるが，検索に頻繁に用いられる名詞と形容詞のみを抽出する．これらの抽出には MeCab‡を用いる．MeCab は高速な形態素解析ツールであり，単語の切り出しと同時に品詞判定が行なえる．

3.3 単語クラスタリング

MeCab によって得られた単語の一覧から文書ごとに各単語の出現回数を調べる．そして，単語を行，文書を列とするような行列を作成し，単語の出現回数を要素の値とする．ここで，行と行の余弦を計算することで，単語の出現回数と共起性を考慮した単語間の関連度を数値化する．共起性とはある単語とある単語が複数の文書にどの程度の割合でともに出現するかをいう．つまり，多くの文書にともに含まれるような単語同士は何らかの関係があるという考えである．最後に関連度の高い単語をまとめ，複数のクラスタを生成する．

*http://developer.yahoo.co.jp/search/
‡http://mecab.sourceforge.net/

Query Suggestion Using Word Clustering
†Ryuta KONISHI ‡Masatoshi KAMIHARAKO ‡Shigeru OYANAGI
†Graduate School of Science and Engneering, Ritsumeikan University
‡Ritsumeikan University

3.4 単語の判別

ユーザが入力したキーワードと関連度の高いいくつかの単語について単語の判別を行う．まず，判別対象の単語を一つずつ再度検索エンジンに問い合わせることで文書を獲得し，行列を拡大する．この拡大によって得られた部分と初期行列とを比較することで，拡大に用いた単語がユーザの入力したキーワードとどのような関係を持つかを判別する．

拡大によって得られた部分に含まれる単語が初期行列に含まれる単語と高い関連度を持てば，ユーザが入力したキーワードと類似の話題を取得できたことから同位語であると判断する．また，初期行列に含まれる一部の単語とのみ高い関連度を持てば，ユーザが入力したキーワードによって得られた話題の一部とのみ関係していることから下位語であると判断する．もし，関連度が低ければ，拡大に用いた単語は異なる話題を取得したことから上位語であると判断する．図 2 にそれぞれの場合の拡大された行列の例を示す*．

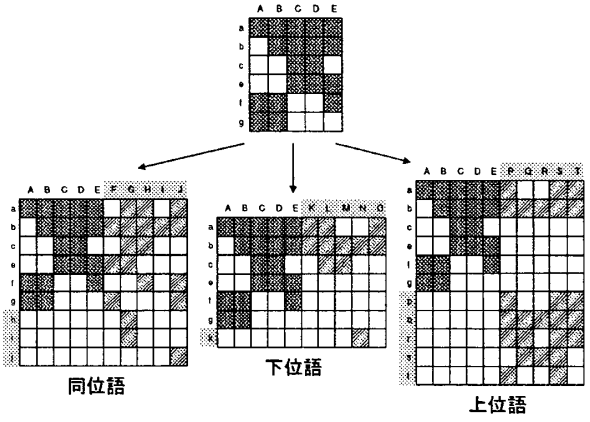

図 2: 拡大された行列の例

3.5 関連キーワードの提示

ユーザが入力したキーワードと関係の判別を行なった単語を判別結果ごとに提示する．同時に，これらの単語を含むクラスタ内の単語をクラスタごとにまとめて提示することで，判別を行なわなかった単語についても提示を行なう．

また，ユーザがこれらの提示された単語を自由に組み合わせて検索を行なえるような提示の方法をとる．

4 関連研究との比較

4.1 キーワード候補の提案を行なう検索エンジン

ユーザにキーワード候補を提示する代表的な検索エンジンとして Google が提供する Google Suggest† と Ask.com が提供する Ask‡ が挙げられる．

Google Suggest では他のユーザが過去に行なった検索データをもとにユーザにキーワード候補を提示する§．ユーザがフォームに文字を入力するとリアルタイムでその文字に続くキーワード候補が提示されることにより，キーワード決定だけでなく，文字入力の助けともなる．また，過去に他の人が用いた検索キーワードを提示するため高い精度が得られる．本研究と比較すると，ユーザの検索キーワード決定を助けるという目的は同じであるが，検索キーワードの単語間の関係については考慮していない点で異なる．

Ask では単語間の関係を考慮した関連語の提示を行なっている．これは検索キーワードを入力し検索を実行すると，結果とともに検索を絞り込むための関連語や，検索を広げるための関連語を提示する．ただし，本研究と異なり，言い換えのための関連語の提示は行なわれない．多くの点で本研究と類似しているが，Ask が関連語の組み合わせを固定で提示しているのに対し，本研究ではユーザがキーワードを自由に組み合わせられるような提示方法を提案している．

4.2 クラスタリングを行なう検索エンジン

クラスタリング処理を行なう検索エンジンとして，米 Vivisimo 社の Clusty¶ が挙げられる．これはユーザが入力した検索キーワードを他の複数の検索エンジンに問い合わせるメタ検索エンジンである．Clusty では検索結果の文書に対してクラスタリング処理を行なうことによって結果を分類し，ユーザの結果閲覧作業を効率化している．

検索結果の文書をクラスタリングしている点で Clusty は本研究と類似している．しかし，Clusty では入力された検索キーワードによる結果を分類するためにクラスタリングを行なっており，キーワードの提示については行なわれていない．

4.3 関連語の抽出を行なうもの

文献 [1] や [2] では日本語の連体助詞による修飾関係を用いることで関連語を抽出している．文献 [1] では親概念と話題語という 2 つの補助概念をキーワードに付帯することによってキーワードが指す曖昧性を回避し，意味内容に即したクラスタリングを行なうことを目標としている．関連語を抽出するためにクラスタリングを行なう本研究に対し，文書クラスタリングのために関連語の抽出を行なっているという点で本研究と異なる．

文献 [3] では単語の共起関係に基づき文書のトピックと関係の強い単語を抽出する手法を提案している．また，抽出された単語に対してクラスタリングを適用し，トピックごとにわけて単語を提示することで，ユーザが求めるトピックに対応する検索キーワードの発見を支援している．

5 まとめと今後の課題

本稿ではユーザが入力したキーワードに対して単語クラスタリングを用いることで関連のあるキーワードを提示するシステムを提案した．

今後の課題として，1 つの単語についてどれだけの文書を用いると妥当な精度が得られるのかを検証する必要がある．また，上位いくつの単語を行列の拡大に用いると適当であるかを決定する必要がある．そのため，実際にシステムを実装し定量的な実験を行なう必要がある．

参考文献

[1] 野田武史, 大島裕明, 手塚太郎, 小山聡, 田中克己. Web 検索結果のクラスタリングに用いる話題語の質問キーワードからの自動抽出. In *DEWS2006*, 2006.

[2] 佐々木稔, 新納浩幸. 単語クラスタリングの語義判別問題への応用. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2003, No. 23, pp. 145–152, 20030306.

[3] 若木裕美, 正田備也, 高須淳宏, 安達淳. 検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング. 情報処理学会論文誌. データベース, Vol. 47, No. 19, pp. 72–85, 20061215.

*図 2 では便宜上，単語の出現回数ではなく単語が出現しているかどうかについてのみを表している．

†http://www.google.co.jp/webhp?complete=1&hl=ja

‡http://ask.com/

§http://labs.google.com//intl/ja/suggestfaq.html

¶http://clusty.jp/