論文の内容の要旨

論文題目Query Refinement based on Comprehensive Representation of Multiple Topics
（複数トピックの包括的提示による検索支援に関する研究）

氏名　　若木　裕美

本研究では，特定のトピックに強く関係する単語を抽出するための単語の重み付け手法として単語共起の統計に基づく定式化を提案する．すなわち，『特定の単語とのみ頻繁に共起する単語ほど個別のトピックを持ちやすい』という仮説を立て，その定式化を試みた．さらに，この手法によって抽出された単語を用いて単語クラスタリングを行う．その結果，検索結果の中に混在していた幾つかのトピックに分けて単語を質問者に提示することが可能となる．そして，質問者は求めるトピックを検索するのに有効な検索語を発見できると考える．本研究では，提案手法を用いた単語抽出および単語クラスタリングに関して様々な実験を行い，提案手法の有効性を確認している．

現在の検索エンジンでは，ユーザによって入力された検索語に関連する文書の中で，より検索語と関係が深いと思われる文書が一次元的にランキングされる．ユーザ側もこの検索エンジンの特性に合わせて，既に必要とするものがはっきりと分かっているときにブックマーク的に使うことが多い．しかしWeb上には様々な内容の文書が存在し，検索結果としてトップページに表示される中には多様なトピックが混在している．このような背景をふまえ近年では，文書クラスタリング型の検索エンジンが幾つか登場している．例えばClustyではメタサーチを行って複数の検索エンジンの結果を文書クラスタリングして，各クラスタに名前を付ける．そして，文書クラスタとその名前をユーザに提示することにより，多くの検索結果を整理することを目的としている．

このように従来の検索エンジンでは，検索語に関連する話題を幅広く調べにくいという問題がある．また，検索するのに適切な言葉が分らないときには，様々なページを閲覧してトピックが絞りこめるような単語をユーザ自身が発見して追加する必要がある．そこで本研究では，多様な内容を含む検索結果の中から，含まれる複数のトピックを分けるのに効果的な検索語を提示し，検索語の曖昧性を解消する手法を提案した．一般的に多義性解消で求められるのは辞書的な複数の意味に分けることである．しかし，検索語は１～2語であることが多いため，その検索語が複数の概念や対象を示しうるために生じる多義性の問題がある．このような多義性の解消は，個々のトピックを分離することで解決できると考えている．一方，検索語の示すものが一意に決まる場合であっても，同一のものを異なる視点から見ることで異なった話題が考えられるという問題がある．これもまた，個々のトピックを分離することで解決できると考えられる．

本研究では，頻繁に同じ文書に出現する一定の単語群がトピックの現れであると考え，ある単語が一定の単語群と頻繁に共起する場合，その単語は特定のトピックに強く関係しているとみなす．ここで，『特定の単語群とのみ頻繁に共起する』という単語の性質をTangibilityと呼ぶ．また，Tangibilityをもつ単語を選ぶための単語への重み付けとして，本研究ではTNGという定式化を提案した.そして,Tangibilityの高い単語,すなわち特定のトピックに強く関わる単語のみを抽出する．ただし，本研究では，ある単語ペアが何回であれ同一文書中に出現することを1回共起したと数えることにする．こうして特定のトピックに強く関係する単語を抽出することで，トピックを際立たせることが出来る．そして，抽出された単語を用いて，Distributional Clustering アルゴリズムに

基づく単語クラスタリングを行う．生成された単語クラスタによって，検索結果の中に混在していた複数のトピックを分けて質問者に提示することができ，質問者は自分が求めるトピックに対応する検索質問拡張(Query Expansion) 用の単語を発見しやすくなる．さらに，提示された単語群の中に検索対象分野に詳しくない質問者にとって未知の単語を含む場合，質問者の学習支援も期待できる．

提案手法では，複数のトピックが混在した文書集合の中で，トピックにのみ強く関わる単語が抽出できていることが期待される．しかし，各単語のトピックへの偏り具合を測る方法はなく，また各単語がいずれのトピックに関わりがあるかの正解は存在しない．そこで，文書分類の正解データを使って，各単語が強く関わりのある分類とその関連度を測る方法を提案した．この方法では，単語が関連するトピックとそのトピックへの関連度についての正解を与えることができる．また，被験者による単語へのラベル付け実験を通して，その妥当性を確認した．評価方法として妥当性があると確認できたため，このトピックへの関連度とそのトピックを推定する方法を用いて，抽出された単語の性能を評価することとした．本実験では，MI(相互情報量)やKLD(カルバックライプラー情報量)などの単語重み付け手法をTNGに対する比較対象とした．使用したデータセットはNTCIR3, NTCIR4, NTCIR-CLIR, Web上の産経スポーツニュースの記事，Dmoz，Reuters，Newsgroup20の7種類で，それぞれのデータセットについて実験を行った．その結果，TNGがもっとも各トピックに強く関係する単語を抽出することができていた．また，文書データに含まれる複数のトピックに関連のある単語が網羅的に抽出されていた．さらに，単語クラスタを生成した後においても，TNGがもっとも各トピックに強く関係し，また，網羅的な単語クラスタを生成していることが分った．

次に，実験用のデータセットを用いるのではなく，実際の検索エンジンの検索結果の上位に対して提案手法を適用した実験も行った．TNGを用いたシステムの出力する単語クラスタに対する比較対象として，検索結果の整理を行うClustyの提示する単語セットを利用した．また，主観的評価実験を通じて各々のシステムの生成する単語クラスタを評価した．その結果，提案手法を用いたシステムでは，Clustyと同程度に検索語自身の多義性の発見に役立つことが分った．また，Clustyに比べて個別具体的な単語を提示するため，提案手法によるシステムの出力した単語クラスタを用いることで検索語に関連する新しい話題の発見につながることを評価実験から明らかにした．さらに，Clustyに比べて積極的に単語クラスタを生成するため，Clustyでは複合語や言い換えになるような単語が多く提示されるのに対し，提案手法によるシステムでは複数の異なる見方を表す単語セットが提示されやすいことが分った．