

UT-Kiwi: 検索支援としてのテキストマイニングシステム

吉田 稔[†] 中川 裕志[†]

東京大学情報基盤センター[†]

1. はじめに

近年、WWW 上や組織内に蓄積される電子的文書の量は増大の一途を辿り、それらの文書を人間が把握することが困難となっている。WWW 全体における文書量の増大のみならず、その中でのトピックの限定された部分集合（特定サイト内の Web 文書集合、Wikipedia、さらには企業内文書集合等）のサイズも増大し、WWW 全体と同様に、把握が困難となりつつある。

検索エンジンは、文書集合を効率的に利用する有力な手段の一つであるが、その機能は基本的には「入力された文字列（クエリ）を含む文書を返す」という範囲に留まっており、「どのようなクエリを入力すべきか」という問題に対しては、研究の余地が多く残されている。

本研究では、このような「クエリ入力支援」に対する新しい提案として、テキストマイニングによる検索支援システム UT-Kiwi を提案する。

テキストマイニングとは、与えられたテキスト集合の中での、「言葉の使われ方」（主に、言葉に関する統計的情報）について分析するタスクである。UT-Kiwi では、入力されたクエリに対し、「用例抽出」「同義語抽出」という二種類のテキストマイニングをリアルタイムに行い、マイニング結果を提示する。

UT-Kiwi は、前述の「トピックの限定された文書集合」を主に対象としており、それらのテキストに対する検索を支援する。基本アイデアは、「文書集合をオンメモリで配置し、Suffix Array による高速検索を実装する。さらに、Suffix Array による検索を応用することで、高速なテキストマイニングを行う」というものである。これにより、テキストマイニングを検索支援として用いることが可能となる。

2. システム概要

UT-Kiwi^{*}は、現在東京大学の Web サイト内の HTML 文書を文書集合として、サービスをを行っている。図 1 に、UT-Kiwi の画面を提示。図中、A. に示されているのが検索窓であり、ユーザがここにクエリを入力すると、システムは入力内容に応じて用例や同義語を提示する。ここで用例とは、クエリ前方、後方にどのような文字列が頻繁に接続しているかを示したものであり（図中「B. 後方接続」「C. 前方接続」）、この場合、「シンポジウム」の前方に「記念」や「国際」、後方に「日本」や「講演」といった言葉が続き易いことがわかる。同様に、シンポジウムの同義語として、「講演会」や「ワークショップ」という文字列が用いられやすいということも提示する（図中、「D. 類義語・同義語」）。ユーザは、提示された用例等をクリックすることで、クエリの補完を行える。補完された結果は、そのまま Google へのクエリとして、検索に用いることができる。

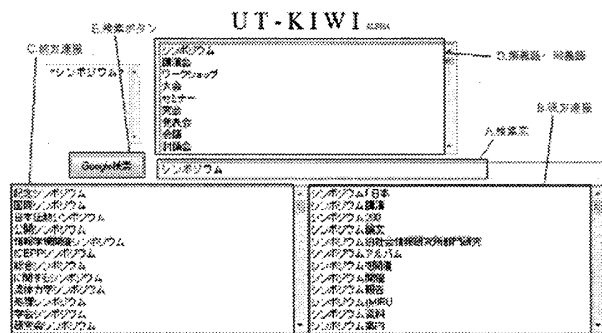


図 1: UT-Kiwi^{*}

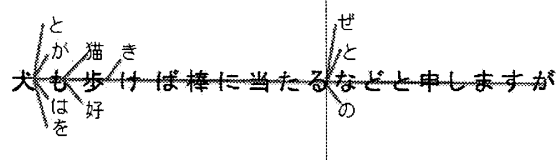


図 2: Kiwi アルゴリズム概要

3. テキストマイニングの詳細

3.1. 用例検索

用例検索は、Kiwi というアルゴリズムを用いて行う。Kiwi についての詳細は、文献[1]を参照されたい。Kiwi では、クエリを検索エンジンで検索し、その結果をマイニングに用いていたが、UT-Kiwi では、これを Suffix Array によるオンメモリの検索に置き換えることで、応答時間を大幅に短縮し、クエリ補完に利用可能な速度を実現している。

図 2 に、Kiwi アルゴリズムの概要を示す。Kiwi アルゴリズムでは、ある文字列 s が与えられたとき、「 s に接続する文字の種類数」（以下、「分岐数」）を計測する。例えば、図 2 では、「犬も」の後続文字として、「猫」「歩」「好」の 3 種類の文字があり、分岐数は 3 となる。文字列 s にどのような文字列が接続するかを、木構造の探索により列挙していくが、そのさい、分岐数が増加している場合のみ、用例候補として考慮する。（図 2 の例では、例えば「犬も歩けば棒に当たる」が、分岐数 1 から 4 へ変化する点として、候補になる。）これは、「分岐数の増加は、意味的な切れ目とよく合致する」という経験則に基づいたルールである。また、UT-Kiwi では、用例候補が見つかったノードに対しては、それ以上深く探索を行わないという制限を加えることにより、探索の高速化を行っている。

最終的に、このようにして得られた用例候補を、スコア関数

$$\text{score}(s) = \text{freq}(s) \cdot \log(1 + \text{length}(s))$$

により順位付けし、提示する。

3.2. 同義語抽出

同義語抽出において問題となるのは、ユーザーによって入力される文字列の多様性である。同義語抽出や言い換えに関しては従来より多くの研究が行われているが、

UT-Kiwi: A Text-Mining-Based Query Support System

[†] Minoru Yoshida and Hiroshi Nakagawa, Information Technology Center, the University of Tokyo.

^{*} <http://kiwi.r.dl.itc.u-tokyo.ac.jp/ut-kiwi/>

