論文の内容の要旨

論文題目　　Topic Trend Detection and Mining in World Wide Web
（WWW 上でのトピックトレンドの探知とマイニング）

氏名　　クー　キュウ　ブン

WWW(Web)では膨大な量の情報が常にダイナミックに変わるため、そのオンライン・トピック検知およびトラッキングの研究は重要で挑戦的になって来ている。本研究の目的は Web 上の変化を捕らえて分析することができる技術を提案することである。言い換れば、本研究のゴールは豊富な量の情報源を持つがダイナミックに変わる Web 上の変化の主要なトピックを検知して、追跡する問題に取り組むことである。

Web は大量の情報が流通・蓄積・共有される、最も重要なチャンネルとして出現しており、Web 自体も世界で最も大きなネットワークにつながれた情報記憶機構になっている。しかしながら、Web の高度成長は止まらず、その情報は膨張し続けている。従って、膨大な量の新しい情報あるいは変化は、Web にダイナミックに付け加えられている。情報化時代での競争力を持つために、これらの新しい情報は不可欠であり、遅滞なく入手することが重要になる。しかしながら、手動でブラウズすることにより変化を見つけることは非効率かつ非現実的である。したがって、多くの変化の中に埋まれてる価値のある情報を収集、処理してユーザへ伝達する知的な情報システムが不可欠となる。Web 上これらの情報変化は 2 つのタイプに分類することができる：即ち「フロー」タイプと「ストック」タイプ情報である。我々の研究は、Web 上の変化を検知し追跡するためのフレームワークを提示することであり、本論文は、Web 上の新しい情報(変化)の自動ジャーナリズムに関するアプローチを示している。「フロー」タイプ情報(例えばニュース)は、定期的で高頻度に Web に現れる。これに対して、我々は与えられた幾つかのニュース・チャンネルのニュース・アーカイブから重要なトピックを検知し要約するためのシステム News Topics Summarizer を提案している。このシステムは新しい TF*PDF (Term Frequency * Proportional Document Frequency) アルゴリズムを使用して、トピックの単語の重みを計算して、これらの単語の重みを分析することにより、重要なトピックを検知する。この TF*PDF アルゴリズムは、多くのニュース・チャンネルでの多くのドキュメントの中でトピックについて説明する単語へ高い重みを与える。異なるトピックからの単語グループでは、それらの単語の重みは異なる特性を示すことがある。一時的な話題のトピックについて説明する単語は、ある時間枠中の正の値の系列が続いた後に負の値の系列を示す。これらのトピック単語およびその出現の時間枠を認識した後に、トピック時間枠に現われる重要な文を用いて文ベクトル・クラスタリングを行うことにより、トピックの要約を生成する。このようにして、本システムは各トピックをカバーするよい要約を作成でき、ユーザに主要なトピックに関する要約の報告を定期的に提供することが出来る。この問題の領域において、このアルゴリズムは従来の TF*IDF アルゴリズムより効果的で、過去分の Web コーパスを必要とせず、かつトピックの検知およびトラッキングの軌跡を失う危険性が少ない。その上、我々のシステムは高い柔軟性を持ちながら所有計算量は少ない。このシステムは、Web を周回（クロール）し、更新情報を集め、ユーザに新たに出現したトピックの要約を記事として提供する。これは Web 上の個別化された電子ジャーナリスト（e-journalist）となり、定期的に新しい出来事の収集とその電子出版化（e-publication）を可能にするものとなろう。

「ストック」タイプ情報（主に静的 Web ページ）は予測できずに変わる。したがって、モニタリング・システムはユーザが興味のあるページあるいは情報領域を常にチェックし、変化を報告することが要請される。従来の Web モニタリング・システムは変化が発生した Web ページの URL を単に知らせるものが多いが、あまり無意味の変化を知らされるのは良くないので、我々はある特定分野をトラッキングし、価値のある変化を報告する知的なシステム ETTS（Emerging Topic Tracking System）を提案し、構築している。ユーザの入力キーワードに対して、ETTS がキーワードを表わす Web 上の情報エリアを見い出して、定期的にこのエリアを周回し変化を収集する。その後、TF*PDF アルゴリズムを用いて新規トピックを表す単語を抽出して、これらトピック単語を含んでいる重要な文に基づいてで要約を生成する。簡潔に述べると、ETTS システムは Web 上の知的なエジェントとして、ユーザの興味のある情報分野の変化を検知し、変化の要約を生成する。この変更の要約は、その特定な分野におけるホットな話題を提示することによって、その情報分野の新規に出現しつつあるトピックを明らかにする。このシステムを用いることにより、我々は WWW 情報空間の最新の傾向について常に知ることが出来る。