審 査 の 結 果 の 要 旨

論文提出者氏名　　クー　キュウ　ブン

本論文は「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)」と題し，7 章から成り，英文で記されている．

第 1 章「Introduction」では，社会におけるグローバルな情報流通，蓄積，共有の基盤に成長した WWW(Web)では，サーチエンジンが情報探索の中心的役割を果たしているが，Web 上に新規に出現する情報が高い価値をもつことから，それらを自動的に検知し，主要なトレンドを要約して提示するシステムの必要性があることを記し，本研究を行った動機になっていることを述べている．同時に本論文の構成を示している．

第 2 章「Web Intelligence and Data Mining」では，関連のある研究領域として Web インテリジェンスとデータマイニングを挙げて，動向を記している．本論文の内容に特に関係する Web 上の変化監視システムの研究開発例として，7 種のシステムを挙げて説明している．これらのうち，WebBeholder は著者の研究グループにより開発されたものであり，本研究の先行研究に当たることになる．また，本研究と関係する新種の Web crawler(Web robot, spider とも称される)の研究開発例についても示し，特定のトピックに注目して Web ページを探索するシステムなどについて紹介している．

第 3 章「Topic Trends Detection and Tracking」では，Web 上のトピック検知とトラッキングに関係する基礎技術を挙げ，システム化する際に必要な他の要素技術と構成について記し，考察している．テキスト変化分を分類するに際し重要な役割を果たすのが，出現単語によるベクトル空間モデルと，TF*IDF (Term Frequency*Inverse Document Frequency)を代表とする単語重み付けである．他所での関連したシステム開発例も挙げている．

第 4 章の「Automatic Online Journalism」では，これまで社会に起きた出来事や情報は新聞，雑誌といったジャーナリズムが分類，整理して多数の人々に届けてきたように，Web の世界でもそのような機能の必要性があり，またその機能はこれまでの主に人手によるジャーナリズムとは異なり，大半がコンピュータによって自動化される形態になることを述べ，本研究もそのような形態へ向けての研究であると位置付けている．そのような Web 上の自動オンライン・ジャーナリズムに向けて必要な技術要素として，テキスト文書のクラスタリングと識別，複数文書の要約についての検討を示している．また，このような自動オンライン・ジャーナリズムを指向するシステム例として，Google News 等を挙げている．

第 5 章と第 6 章では Web 上の情報をフロー型情報（典型的にはオンライン・ニュースなど）とストック型情報とに分け，フロー型情報，ストック型情報，それぞれの新規情報の検知と要約を行う研究開発したシステムについて記している．

第 5 章「Flow Type Information Topic Detection and Summarization」では，Web のフロー型情報を対象にして開発した，主要トピック検知・要約システムについて記している．システムは複数の Web オンライン・ニュースを情報ソースとし，主要なトピックは多数のニュース文に出現すると仮定し，その検知と関連事項の要約を作成する構成になっている．複数情報ソースのニ

ュース文から上記のような主要トピックを検知するに際し，新たに導入した TF*PDF (TF*Proportional Document Frequency, IDF とは逆に該当の単語を含む文書数の指数に比例する量）による出現単語の重み付けが有効であることを示している．要約文は，これにより重み付けされた単語により文の重みを求め，重み上位文を単語ベクトルによりクラスタリングし，それに基づき生成する構成となっている．作成したシステムの主要トピック検知と要約の効果を 4 種のオンライン・ニュースソース(Associate Press, The New York Times, Reuters, USA Today)を用いて評価，検証している．

第 6 章「Emerging Topic Tracking System(ETTS)」では，関心を持つ領域を表す指定したキーワードに関係する Web ページを対象とする，ストック型 Web 情報の新規更新情報を整理し要約を生成する，ETTS と名付けたシステムについて記している．この場合，関連する Web 情報はキーワードが出現している Web ページだけでなく，リンクで結ばれた子ページ，孫サブページ，兄弟ページにも掲載されることが多いことを考慮し，これらの Web ページも更新検査の対象にするようにしている．差分によって検知された新規更新情報の整理と要約は， TF*PDF を利用する前章の方法と同様にして作成される．作成した ETTS の効果の評価と検証を，幾つかのキーワード（“アジア経済”，“原子兵器”，“e コマース”）で表される領域について示している．関連システムとの差異についても言及している．

第 7 章は「Conclusion」であり，本論文の成果をまとめている．

以上を要するに，本論文は社会における膨大な情報の流通，蓄積，共有の基盤となってきた WWW(Web)において，情報価値が高い新規情報を検知し，主要なトピックを抽出，その要約を自動生成する手法とシステム化の技術を，フロー型情報(オンライン・ニュースなど)，ストック型情報それぞれについて提示し，具体例を通じてその効果を実証したものであり，電子情報学上貢献するところが少なくない．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる．