審査の結果の要旨

氏名　ヤトフト　アダム　ヴァディスラヴ

本論文は「A Study on Temporal Summarization of Web Pages（Webページのテンポラル要約の研究）」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)は，情報が随時更新され，また新情報が随時出現するダイナミックな情報空間であり，その膨大な量の情報の活用を可能にするため，本研究は特に時間軸に沿った指定の話題に関する情報の集約と要約に焦点を当てて行ったものであるとの位置付けを述べている．

第２章「Automatic Document Summarization」では，関連研究として，文書要約及び時系列に沿うニュース文等を対象とする要約に関する研究についてまとめている．要約手法はいくつかの観点から分類することが出来るが，ここではまず，意味内容を把握して要約を作成する抄録的要約(abstractive summarization)と，重要度が高いとみなせる要点文を抜粋し編成する抜粋的要約(extractive summarization)の観点から，各々の技術と具体的システムについて記している．前者は人間による要約に近く好ましいが技術的にまだ難しく，現状での要約には後者の方法が採られることが多い．更に，複数文書要約についても，その技術と具体的システムについて記している．時系列情報の要約に関しては，ニュース文要約を中心に，最近の Time Mines, Google News, Newsblaster や，トピック検知とトラッキング，トレンド検知等のシステムを挙げ，その技術内容についてまとめている．

第３章「Change Detection and Summarization in the Web」では，本研究に直接的に関係する Web 情報の変化の検知と要約に関する既存の研究についてまとめている．Web 情報の要約については，通常の文書要約と異なり，Web テキスト情報が往々にして非構造的であり，単一ページに多様な内容を含むことがあることを，記述の充足性，一貫性，論理性が必ずしも保たれていないこと，関係する内容が複数ページに分散されることがあること等を考慮する必要があることを述べている．そして，現在の Web 情報の要約には，抽出された単一又は少数の限定された Web ページを対象とするコンテント・ベース・アプローチと，ハイパーリンクされた関連 Web ページも含めて要約を作成するコンテクスト・ベース・アプローチが存在し，両アプローチによる具体的システム例の技術的内容を記している．

第４章「Temporal Summarization of Web Pages」では，Web 情報空間のダイナミック性に着目し，指定の話題に関する Web 情報のテンポラル（経時的）要約を作成する，著者が作成したシステムの概要を記している．このシステムでは，ある話題に関係する Web ページを所定の時間間隔で所定の期間に亘り収集し，この Web ページ集合からテンポラル要約を作成する．時間的に静止した Web 情報要約や Web ニュース文に限定した要約と比較すると，時間軸も含むより範囲の広い Web 情報の要約機能を開発していることになる．基本的なアプローチは現時点で品質の高い要約が可能な抜粋的要約であり，主な課題は収集された Web ページ集合から，いかにして情報量の大きいテキスト文を抽出し編成することにある．このため，発生した重要な出来事や示された重要な概念等を発掘するデータマイニング的手法の採用が必要になる．また話題に関する Web コミュニティを発見し，そこから十分な量の関連 Web ページの収集も必要であり，そしてこれらの Web ページは重要なテキスト情報を含むことに加えて，新規のものである必要がある．Web 文書に現れるイベントや出来事の時間的ダイナミック性を測る指標として，ある期間での変化量(change volume)と活性度(activity ratio)を導入している．一般にある期間に集中して発生する変

化量，活性度の高い情報内容が注目に値するものとなる．以上の考えの下，オンライン型テンポラル要約と追想型(retrospect)テンポラル要約の２種のシステムを作成し，評価を行っている．

第５章「Online Temporal Summarization」では，指定話題に関する直近の変化の要約を所定の期間間隔（1 日から 1 週間程）毎にスナップショット的に作成して提示する，オンライン型要約の ChangeSummarizer と名付けたシステムについて記している．検索エンジンにより指定話題に関係する Web の上位 200 ページを取り出しベース集合とし，前サイクルとの有意な単文単位の差分をとり，新規情報中の頻出用語（単語と連結単語，すなわち 1,2-grams）とそのスコア付け，この用語重み及び下記の Web ページ・ランクに基づく文のスコア計算により，要点文を抽出する．用語スコアはそれらの複数 Web における差分での出現の共通性(popularity)と，以前の情報からの新奇性(unpopularity)の両者を考慮して決定している．またここでの Web ページ・ランクは以前のサイクル時点を考慮して，共通的な差分新規用語を含んでいた割合と最近での更新頻度により決定しており，これを文のスコア計算に用い，同様な要点文が複数抽出される場合にランクの高い Web ページの文を優先するようにしている．このランクの高い Web を起点として Web コミュニティマイニング手法より調査の元になるベース集合の拡大を行い，これによって時間が経つにつれて関連 Web ページ集合の充実を図っている．最終的な要約文は重みの高い文を可読性向上のためその前後の文も含めて抜粋する．作成した ChangeSummarizer システムにより，“terrorist attacks”, “Iraq war”, “Harry Potter”, “latest movies”等の話題に関する実験結果を示し，考察と評価を行っている．

第６章「Retrospective Temporal Summarization」では，指定話題に関してオンライン型と同様に関連 Web ページを収集し，ある期間（2 ヶ月など）に亘る特徴的な情報を見出して要約する 2 種の追想型要約システムについて記している．第１のスライディング時間窓を用いる追想型要約では，まず全期間の文単位の変化に含まれる用語の長期間スコアを算出し，時間窓をスライドさせながら新出用語と消失用語について用語の窓区間スコアを計算し，これらの用語スコアに基づき要点文を抜粋し，時間順序に編成して要約を生成する．第 2 の追想型要約では，変化のない用語も含めて全期間に亘る用語の出現の平均傾向と分散から長期的用語スコアを求め，この平均的傾向から大きくずれる時点の用語についての指標も加味して要点文を抜粋し，時間順序とある Web ページにおける前後関係も考慮して要約を編成する．以上より，追想型要約ではある短期間に集中的に出現する出来事や概念は高いスコアを得て，要約文に現れることになる．2 種の追想型要約システムを作成し，“artificial intelligence”, “chicken flu”, “EU enlargement”, “soccer”等の話題に関する実験結果を示し，考察と評価を行っている．

第７章は「Conclusions」であり，本論文の成果をまとめ，その用途を展望している．

以上を要するに，本論文は社会におけるグローバルな情報流通，蓄積，共有の基盤になってきた Web 情報空間のダイナミック性に焦点を当て，その膨大な量の情報の活用を可能にするため，時間軸に沿う Web 情報の要約法について研究したものであり，時間差分情報に含まれる用語のスコアリングとそれによる文のスコアリングに基づく注目すべき要点文の抜粋を基本とし，スナップショット的な要約を生成するオンライン型要約と，ある期間（2 ヶ月など）に亘る特徴的な情報を見出して要約する追想型要約の効果的な方法を示し，システムとして実現し，効果を実験的に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．