氏　　名	竹田　隆治	
学位（専攻分野）	博士（情報学）	
学位記番号	総研大甲第1240号	
学位授与の日付	平成２１年３月２４日	
学位授与の要件	複合科学研究科　情報学専攻 学位規則第６条第１項該当	
学位論文題目	局所的類似情報に基づいたテキストマイニングに関する研究	
論文審査委員	主　査　教授	高須　淳宏
	教授	大山　敬三
	准教授	相原　健郎
	准教授	市瀬　龍太郎
	客員教授	相澤　彰子
	准教授	影浦　峡（東京大学）

論文内容の要旨

現代は高度情報化社会といわれ、さまざまな情報が電子化されている。こうした電子化された大量の文書が存在する一方、その中から必要とされる情報を効率よく見つけたいというユーザの情報取得要求も強くなっている。たとえば、最新ニュースなどに関する記事は、リアルタイム性の高いストリーム型の文書で、その中から、ユーザの興味のある記事をタイムリーに提示することが望まれる。

また、Web上に形成されるコミュニティには、ユーザの意見、要望、苦情等、企業側が想定しない潜在的なニーズやリスクに関する情報が含まれている。Web上に蓄積された文書の分析には、各種のテキスト解析ツールを備え、大量テキストを効率よく処理できるシステムが望まれる。

このようなニーズに答えるために、電子化された情報から類似文書をまとめるクラスタリング、特定の話題に関する文書を選別するフィルタリング、類似文書クラスタの内容をまとめる多文書要約といった高度なテキスト処理技術が必要になる。

電子化された文書の特徴のひとつとして、部分的に類似した表現が複数の文書に現れることがあげられる。本研究では、この特徴に着目し、大量の文書から部分的に類似した表現を効率良く列挙することによって、効果的な文書クラスタリング、文書フィルタリング、文書要約を行う手法を提案する。

1 章では本研究の背景と目的を述べる。また、本研究で扱う局所的類似表現を事例を示しながら定義する。

2 章では、本研究の基礎となる文書類似度、近似文字列マッチングについての既存研究を述べる。さらに、本研究で提案手法の応用を試みるスプログフィルタリングおよび多文書要約の研究をサーベイする。

第 3 章で本論文の核となる局所的類似性を抽出する方法を述べる。本論文では、特に大規模なテキストに対して効率よく局所的類似表現を抽出することに力点をおいており、

（1）問題に応じて定義された部分文字列の類似度に基づいて接尾辞配列を作成し、

（2）接尾辞配列を一回スキャンすることによって局所的に類似する部分文字群を列挙する方法を提案する。

第4章では、提案手法をスパムブログ（スプログ）のフィルタリング問題に適用し、その有効性を示す。スプログは、他のブログやWebの文書に現れる語、フレーズ、文を組み合わせることによって自動的に生成されることが多い。そのため、スプログは他の文書からのコピーコンテンツを多く含む。本研究では、スプログが持つこの特性に着目し、ブログやWebの文書から構成される文書データベースを構築し、各ブログと文書データベース中のテキストとの局所的類似性を抽出することによって、スプログのフィルタリングを行うシステムを構築した。そして、システムのフィルタリング性能を測定するために、およそ25,000件のブログよりなる評価用のコーパスを作成し、フィルタリングシステムを評価した。この評価実験により、提案手法がフィルタリング性能を示す代表的な尺度の一つであるF値において0.76程度の比較的高い性能を得られること、また、実用的な観点からは、このフィルターを未知のタイプのスプログに対する検知器として用い、スプログのタイプ

ごとに専用のフィルターを構築することによって、より精度の高いスプログフィルタを構成することが望ましいことを示す。

第5章では、本研究で提案した局所類似性検出法を複数文書要約の問題に適用し、その有効性を示す。Web上に公開されるオンラインニュースでは、複数の新聞社より同種のニュースが発信されており、また、同一の新聞社からも続報といった形で類似情報を含む記事が発信される。そのため、ニュース記事には、他の記事と部分的に同種の情報が含まれていることが多い。本研究では、この特徴に着目し、（1）これらのニュース記事集合より局所的類似性を検出し、（2）抽出した局所的類似性に基づいて記事間の類似性を求めることによって、同一トピックに関する記事のクラスタを作成し、（3）各クラスタから頻出する局所類似性を含む文を組み合わせて、記事クラスタの要約を作るシステムを構築した。さらに、このシステムを複数文書要約システムの性能評価用コーパスの一つであるNTCIR4-TSC3を用いて評価し、冗長性の少ない要約を作るのに優れた手法であることを示す。

最後に第6章で本論文の成果をまとめる。

論文の審査結果の要旨

近年、多くの文書がインターネット上に公開されているが、その中には、他の文書からのコピーされたものが多く見受けられる。竹田隆治君の研究は、これらのコピーコンテンツを検出し、取り除くことによって、利用者にコンパクトな情報を提示し、利用者が効率的に情報を取得できる情報活用システムの構築技術を作ることを目的としている。本論文では、複数の文書に共通して現れる局所的に類似する記述を効率的に抽出する方法を提案し、オンラインニュースの要約システムおよびブログからスパムをフィルタリングするシステムに応用し、提案手法の有用性を示している。

本論文では、特に大規模なテキストに対して効率よく局所的に類似した表現を抽出することに力点が置かれており、（1）問題に応じて定義された部分文字列の類似度に基づいて接尾辞配列を作成し、（2）接尾辞配列を一回スキャンすることによって局所的に類似する部分文字群を列挙する方法を提案している。接尾辞配列の効率的な構築や接尾辞配列を用いた効率の良い全文検索アルゴリズムについては多くの研究があるが、本研究では、局所類似表現を効率よく列挙するために、接尾辞配列を利用した部分表現のクラスタリング法を提案した点に新規性がある。

局所的類似表現は文書の重要な特徴として用いることができる。本研究では、提案手法をスパムブログ（スプログ）のフィルタリング問題に適用し、その有効性を示している。スプログは、他のブログやWebの文書に現れる語、フレーズ、文を組み合わせることによって自動的に生成されることが多い。そのため、スプログは他の文書からのコピーコンテンツを多く含む。本研究では、スプログが持つこの特性に着目し、ブログやWebの文書から構成される文書データベースを構築し、各ブログと文書データベース中のテキストとの局所的類似性を抽出することによって、スプログのフィルタリングを行うシステムを構築している。さらに、スプログ検出性能を評価するためのコーパスを作成し、提案手法が既存の他の手法と比較し高い性能を持つことが実験的に示されている。

また、本研究では、局所類似性検出法を複数文書要約の問題に適用し、その有効性を示している。Web上に公開されるオンラインニュースでは、複数の新聞社より同種のニュースが発信されており、また、同一の新聞社からも続報といった形で類似情報を含む記事が発信される。そのため、ニュース記事には、他の記事と部分的に同種の情報が含まれていることが多い。本研究では、この特徴に着目し、局所類似表現を記事の類似度、および、記事中の文の重要度を図る尺度として用い、新聞記事要約システムを試作するとともに、複数文書要約システムの性能評価用コーパスの一つであるNTCIR4-TSC3を用いて評価し、冗長性の少ない要約を作るのに優れた手法であることを示している。

以上、竹田隆治君の研究は、大規模なテキストデータから類似情報を効率よく抽出することを可能にし、特に、コピーコンテンツや重複した情報を含むテキストからコンパクトな情報を抽出するのに適したテキストマイニング手法を実現した。また、本研究のスプログフィルタリングの研究を発展させた手法が実システムに組み込まれて使用されており、実用性の高い研究である。本研究の成果の一部は、学術雑誌論文1篇および国際会議論文

４篇にまとめられており研究業績も学位取得の基準を満たしている。よって、竹田隆治君の論文は学位（情報学）を授与するに値すると審査委員会委員全員一致で判断した。