様式例：課程博士用

17年　1月　11日

電子・情報工学専攻	学籍番号	983425
申請者氏名	武　田　善　行	

指導教員氏名	梅　村　恭　司 中　川　聖　一

論　文　要　旨(博士)

論文題目	全文字列の統計量に基づく索引語分析に関する研究

（要旨 1,200字程度）

本研究は，形態素解析などの分かち書きを行わず，また辞書を用いずに任意文字列の中から興味深い文字列の特徴を分析したものである．過去にも多くの研究が任意文字列を分析の対象としているが，たとえば自然言語のようにその構造がよく知られているデータの中にでさえ多くの未知の要素があり，それらを捉えることが重要な応用があることを示唆している．任意文字列を扱うことの利点は，こういった未知の要素をも一般的なケースとして分析できることにある．

本稿では大まかに三つの話題について取りあげる．はじめに，全文字列の反復度分析を行う．音声認識の分野に，語を参照するたびにその後に続く語の推定値を調節するモデル（Adaptive Language Model）があり，このような調節において用いられる語の特徴量として反復度（Adaptation）がある．反復度は，語が1回以上出現する条件で2回以上出現する事後確率により計算される特徴量である．反復度は自立語と付属語とでその値が明確に異なることが報告されている．自立語と付属語の性質を捕らえることは興味深い問題であり，本研究ではとくに任意文字列による反復度の値に興味を持ち分析を行った．分析の結果，日本語や中国語のような言語の違いや新聞記事や論文抄録のようなコーパスの違いを問わず，自立語とそれ以外の文字列がもつ反復度には明確な差異が確認された．また，今まで明らかでなかった平滑化されていない反復度の値や，任意文字列における反復度の値を確認し，反復度は他の文字列と自立語と分離しうるものであることが明らかにした．

次に，反復度を用いた索引語抽出法の提案を行う．本論文で提案する索引語抽出法は，分析により得られた反復度の特徴を元に任意文字列中の索引語らしさを定義し，ビタビアルゴリズムを用いて索引語らしさを最大化する分割を見つけることで索引語の分かち書きを行う．本論文で提案する索引語抽出法は，辞書や既存の言語知識を一切必要とせず，言語やコーパスを問わずに実行される．日本語や中国語での索引語抽出実験によりその実証を行う．また提案法により抽出された索引語が情報検索性能を有意に向上させることを確認することで，提案法がもつ有用性を定量的に示した．

最後に，任意文字列対を分析の対象とする共起文書頻度を計数するために有用である，任意文字列を代表的な文字列に区分する方法を紹介する．出現の交わりの度合いに基づく類似尺度は，どんなものであれ共起文書頻度に基づく．任意文字列の頻度や文書頻度の計数が既に実現された現状で任意文字列対の相関を分析する問題は，任意文字列対の共起文書頻度を計数する問題に帰着する．任意文字列対を分析することは任意文字列を分析することに比べ難しい．長さNの文字列において，全部分文字列は$N(N+1)/2$であるのに対し，その対は$N^2(N+1)^2/4$である．本研究では，山本ら(2001)の提案した等頻度クラスを改良し，よりコンパクトで文書出現の等しい文字列のクラスへの分割を実現することで任意文字列対の分析を容易にする．日本語論文抄録集を用い，実証を行った．結果として任意文字列対は，高々$0.01N^2$の両方向等頻度クラスに減少された．