様式例：課程博士用

１７年　09 月　06 日

電子、情報工学専攻	学籍番号	019305
申請者氏名	徐　盈輝	

指導教員氏名	梅村 恭司 中川 聖一

論　文　要　旨(博士)

論文題目	A Study on Empirical Models for Large Scale Information Retrieval (大規模情報検索のための経験的モデルに関する研究)

（要旨 1,200字程度）

情報検索は文書集合から検索質問に関連している文書を特定するタスクである．本論文では，大規模な情報検索に線形代数を使用した経験的モデルに着目する．

はじめに本論文は，大規模な検索対象における潜在的意味インデキシング(LSI)の計算量の問題を解決することを目的とした狭い質問空間の LSI の分析結果を示す．狭い質問空間の異なる特異値分解(SVD)次元での実験を通じて，低い次元の LSI がベクトル空間モデルに比べはるかに良い精度を，グローバルな LSI に比べ同程度の精度を得たことを示す．この小さな SVD の次元は，ほぼ直線的な平面が狭い質問空間上にあることを示す．最大か最も大きな二つの特異ベクトルは，この直線的な平面を捉える能力を有しており，特定の検索質問にとって有益である．検索質問毎に特異値分解を行う必要はあるが，SVD 次元は驚くほど少ないため，大規模情報検索における LSI の計算量の制限を解決することが可能である．さらにいくつかの関連するサンプル文書が利用可能であるという条件で，独自の局所適合性フィードバックを用い同程度の精度を得ることが可能にする．

World Wide Web(WWW)の出現により，新たな情報検索手法への要求が生まれた．リンク構造に基づき文書が連結される WWW 上のデータは，語-文書行列からなる純粋な枠組みに加え，文書-文書推移行列を伴う拡張を必要とした．テキスト内容と同様に文書間の連結パターンを獲得することが私の次の研究テーマである．私の研究では，より正確なマルコフ変遷モデルによるリンク分析のために，WWW ページのテキスト情報を考慮に入れる方法を提案した．本論文はページ中のハイパーリンクを，報知リンクと参照リンクの二つに分類した．報知リンクは，ページ間の文字上の一致をともなうリンクであり，類似した情報やより詳細な情報，もしくは付加的な情報を示す．参照リンクは，その対象と文字上の一致を持たないページ中のリンクである．報知リンクにおけるリンク連結がそれらに対応する目標との文字上の一致に導かれるのに対して，参照リンクにおけるリンク連結はランダムである．本論文では，この考えに基づいて統一モデルを提案した．このモデルはリンク構造分析に文字上の一致を組み合わせる．本論文は，WWW 情報検索のための文字上のマイニングとリンク解析を統一するための機構を提供する．実験結果として，私のアプローチが情報検索タスクのためのオリジナルの PageRank より良い性能を得たことを示す．

あらゆる線形代数に基づく情報検索アプリケーションにおいて，最も重要な前処理の 1 つは，語-文書行列の各要素に適切な語の重みを割り当てることである．本論文は，語重みとしてよく知られている $tf \cdot idf$ の代わりに，語 w_i の文書中に出現する回数 k の割合として推定した値である $P_i(k)$ を用いた．最終的に本論文では，K mixture model を改良する効果的なアプローチを提案する．提案法における減衰係数は，文書中の語の出現頻度に依存する関数によって補完された二つの可能な減衰係数の組み合わせである．提案モデルによる語の頻度の推定が統計的に有意であることを示す．