審査の結果の要旨

論文提出者氏名　金澤 輝一

本論文は「文書クラスタを用いた情報検索のモデルとその応用に関する研究」と題し、テキスト情報の検索の新しいモデルの提案とテストコレクションによる評価実験によりその有効性を論じたものであり、10章から構成されている。

第１章は、「序論」であり、本研究の背景、問題の所在、研究の目的と論文の構成について述べている。

第２章は「関連研究」と題し、本研究に関連した情報検索技術に付き、従来研究を概観しており、情報検索のモデルと特徴量抽出について説明した後、意味曖昧性を解消するためのQuery Expansion (QE)などの従来研究を紹介し、現在の課題である言語横断検索、クラスタリングなどの諸技術について解説している。

第３章の「関連性の重ね合わせモデル」では、本研究で新しく提案する情報検索手法として「関連性の重ね合わせモデル」(Relevance-based Superimposition、RSモデル)を提案している。これはベクトル空間モデルの検索モデルに立脚し、検索対象の文書の持つ関連性に着目して検索性能を大幅に向上する手法を考案したものである。文書の集合を話題毎に非排他的な文書クラスタに分け、そのクラスタを代表するような特徴量の期待値を推定し、各々の文書の特徴ベクトルに文書の属するクラスタの特徴量を重ね合わせることにより、文書単体の特徴量を補正する。この結果、質問との間での類似性判定の性能が向上することが期待される。代表ベクトルの算定、重ね合わせのための手法、最適パラメータの推定について、テストコレクションに基づき定量的に評価して、実効的な式と値を求めている。

第４章は「評価用検索システムの実装と評価指標」と題し、提案しているRSモデルの評価をするための環境条件について論じている。RSモデル評価用のシステムとして文献検索システムR^2D^2を実装し、その構成と機能について説明している。これはRSモデルをQEなどと比較できるようにするため、広範なパラメータ調整機能と多様な検索パターンの設定ができるようになっている。本論文で用いた様々な検索方法を実行するための機能について紹介している。また、検索評価に用いた日本語および英語のテストコレクションNTCIRおよびTRECの規模と内容について説明している。

第５章は、「単一言語検索特性の評価」と題し、従来手法のtf·idfに比較して、RSモデルが、日本語テストコレクションでは6から9%の性能向上を実現できたことを示している。一方、英語のTRECでは3から4%程度の向上に止まり、文書クラスタ構成についての分析と検討を要するが、第８章にてこの問題を解消する方法について論じている。

第６章は、「多言語検索特性の評価」と題し、言語横断検索にRSモデルを適用する際に、RSモデルによる検索性能への寄与が言語に依存するかどうか、また問い合

わせの翻訳によって生じる意味曖昧性に対するRSモデルの耐性について、実証的に論じている。コーパスからの対訳辞書自動抽出による翻訳とEDRの日英対訳辞書を用いた問い合わせ翻訳による検索性能の違いをNTCIR-2テストコレクションにて実施した。その結果、RSモデルが言語に依存せず良好な性能を発揮すること、問い合わせ翻訳により意味曖昧性が増大した場合にも安定した性能を持っていることを示した。

第７章は、「query expansionとの融合」と題し、RSモデルとQEとの間での特性の違いをテストコレクションの性格と比較しつつ論じている。RSモデルは同概念異表記の問題に対しして効果を持つこと、報道記事データのように概念記述の統制がなされている場合にはQEの性能が安定していることを実験的に調べた。

第８章の「重要語の自動抽出を用いた文書関連性解析」では、あらかじめキーワードが付与されていないデータベースにRSモデルを適用する手法を提案し、良好な成果を得たことを述べている。頻度分析による重要語の抽出、これによるクラスタ構成、クラスタの特性向上のためのSVM (Support Vector Machine) を用いたクラスタ再構成手法の提案などを行った結果、7.6%の性能向上を達成できた。クラスタの様々な特性分析、重要語の自動抽出の一般的特徴を論じ、RSモデルが一般のデータベースに適用できる基盤を確立した。

第９章は「考察」であり、本論文の全体を総括し、今後の展望を述べている。最後に、第１０章は「結論」として、本論文をまとめている。

以上のように、本論文は、テキストデータベースを対象とした情報検索手法として、従来手法に比し有意な性能向上をもたらす「関連性の重ね合わせ(RS)モデル」を提案し、この検索特性と言語非依存性を日英のテストコレクションを使用して実証的に示しつつ、さらに一般のデータベースに適用するための文書クラスタ構成の一般的手法を確立した研究であり、電子情報工学に貢献するところが少なくない。

よって本論文は博士(工学)の学位請求論文として合格と認められる。