早稲田大学大学院理工学研究科

博 士 論 文 概 要

論 文 題 目

情報検索および情報フィルタリングの
高精度化に関する研究
(A Study on Performance Improvements for
Information Retrieval and Filtering Systems)

申 請 者
酒井 哲也
Tetsuya Sakai

1999年11月

近年，インターネットに代表される計算機ネットワークおよびデジタル情報の普及に伴い，世界中のテキスト・画像・音声など様々なメディア情報に誰もがどこからでも高速にアクセスできる環境が整いつつある．なかでもテキスト情報は，ＷＷＷ（Ｗｏｒｌｄ－Ｗｉｄｅ　Ｗｅｂ）ページや電子メールに見られるように，マルチメディア情報の中で依然として主要な役割を果している．

高度情報化社会はその一方で，膨大な情報の中に自分の所望する情報が埋もれてしまうという情報過負荷の現象を引き起こしている．この問題を解決するために，大量の情報の中から所望の情報をいかに効率的に，洩れなく，かつ誤りなく探し出すかを研究する分野に情報検索の分野がある．とくに，テキスト情報を対象とした検索は１９５０年代から欧米の言語を対象に続けられており，歴史的に情報検索という言葉は通常テキスト情報の検索を意味する．本研究でも，テキスト情報の検索のことを単に情報検索と呼ぶことにする．

１９８０年代から，情報が絶え間なく発生する時代を反映して，情報検索に加えて情報フィルタリングという言葉が使われ始めた．情報検索には，固定的なテキストデータベースからユーザの短期的な検索要求を満たす情報を選出するという意図があるのに対し，情報フィルタリングには，刻々と発生する新しい情報からユーザの長期的な興味に合った情報を選出するという意図がある．さらに，情報検索におけるユーザは能動的に情報を探し，情報フィルタリングにおけるユーザは一旦自分の興味のある分野の話題を指定した後は受動的に情報を得るという違いがある．しかし，両者の本質は検索要求と検索対象とのマッチングにより大量の情報からユーザにとって有用な情報を選出することであり，実際に情報フィルタリングは情報検索の技術の多くを受け継いでいる．

ユーザの本当に求めている情報を計算機により検索するには，究極的には自然言語で書かれた検索要求および検索対象の内容を理解する高度な自然言語理解技術が不可欠であると考えられる．しかし，現状の自然言語理解技術はそのレベルに到達しているとは言い難い．そこで，情報検索・情報フィルタリングでは，検索要求および検索対象の双方を検索語の集合あるいは構造体として表現し，検索語の頻度情報などの統計的手がかりを用いることにより，検索結果（ユーザにとって有用であると推定された情報のリスト）の精度を高めようとする．本研究はこのような背景から，情報検索・情報フィルタリングの高精度化のための具体的なアルゴリズムを提案し，実験的に検証したものである．

申請者は情報検索・情報フィルタリングシステムの開発に携わり，実際に情報フィルタリングサービス会社Ｎ社およびＦ社が設立され，開発の成果が生かされている．１９９６年に設立されたＮ社は，日本経済新聞などを含む国内の約５０の新聞社・通信社から電子化された記事の提供を受け，有用そうな記事のみを毎日電子メールなどによりユーザに配信している．一方，１９９８年に設立されたＦ社はＷＷＷの日本語ページ検索サイトを実現しており，新

規に作成されたＷＷＷページや頻繁に更新されるＷＷＷページの中から特定の話題にマッチするものを選出し，ユーザに紹介している．上記のようなサービスにおけるユーザ数の拡大には，情報フィルタリングの高精度化が不可欠である．本研究の背景には，このシステムの検索精度を向上させるという実用的なニーズがあったと言える．

以下に本研究の目的を記す．

第１は，日本語の検索要求（ユーザの求める情報の内容を自然言語文で端的に表現したもの）を検索条件（計算機が理解可能な形式の検索語の集合あるいは構造体）に変換するための有効な手法を提案し構築することである．応用としては前述の新聞記事フィルタリングサービスを考え，見出しや段落といった記事の文書構造情報を利用した有効な検索条件の構成方法を求める．また，このように自動生成された検索条件と，人手で作成した検索条件との性能比較を行うことにより，人間の検索要求に対する計算機による解釈の可能性について考察する．

第２は，上記の日本語文から自動生成された検索条件をスタート地点とし，最初の検索結果に対してユーザが行った評価の情報を検索条件にフィードバックすることにより，さらなる検索精度の向上を実現することである．応用としては，前述の新聞記事フィルタリングサービスにおいて，過去に配信された記事に対するユーザの評価をもとに，今後そのユーザに配信する記事の選出方法を修正することを考える．また，ユーザから検索結果に対する評価情報が与えられない場合にも，教師なしフィードバックのアプローチにより自動的に精度を向上させる．

第３は，上記フィードバック技術が言語横断検索（検索要求と検索対象の言語が異なる場合の情報検索）においても有効であることを示すことである．この応用としては多言語文書が混在するＷＷＷ環境などが考えられるが，ここではとくに日本語・英語間の双方向の情報検索を扱う．

以下に本論文の構成を示す．

第１章では，本研究の序論として，研究の背景・目的などを述べる．

第２章では，情報検索における従来の研究を概観し，本研究の位置づけを行う．まず情報検索研究の歴史的な推移について様々な角度から明らかにし，次に情報検索システムを客観的に比較評価するためのデータであるテストコレクションを用いた検索精度評価手法について述べる．さらに，検索結果のテキストを重要度順に並べてユーザに提示するランキング検索における従来の研究についてより詳しく議論する．最後に，情報検索研究における本研究の位置づけと，本研究の第４～６章の間の関係を示す．

第３章では，本研究の第４～６章における検索精度評価実験の実験環境について述べる．まず検索の題材に用いた３つの日本語テストコレクションについ

て，次に申請者が情報フィルタリングサービスのために開発した情報フィルタリングシステムＮＥＡＴ（Ｎｅｗｓ　Ｅｘｔｒａｃｔｏｒ　ｗｉｔｈ　Ａｃｃｕｒａｔｅｌｙ　Ｔａｉｌｏｒｅｄ　ｐｒｏｆｉｌｅｓ）について述べる．

第４章では，情報フィルタリングのための文書構造を利用した初期検索条件生成について述べる．第１の実験では，実際の情報フィルタリングサービスにおいて検索条件の作成経験をもつ専門家が，テストコレクションの検索要求および補足説明文を読んで検索条件を作成する．第２の実験では，人手による検索条件作成の負荷を軽減するために，検索要求から検索語を抽出し検索条件の自動生成を行う．自動生成実験では，計算効率重視の観点からブール式を用いる場合と，精度重視の観点からブール式を用いない場合の二通りを検討し，いずれの場合にも，文書構造を利用した検索条件が文書構造を利用しない従来の検索条件に比べて有効であることを示す．また，人手と自動で検索される正解記事集合が互いに相補的であることを示す．

第５章では，情報フィルタリングのためのフィードバックによる検索条件展開について述べる．ここでは，英語検索において有効性が示されている確率検索モデルを情報フィルタリングシステムに実装し，検索結果に対するユーザの評価情報を用いたレレバンスフィードバック技術およびユーザの評価情報を用いないローカルフィードバック技術により初期検索条件を自動修正して再検索を行うことにより検索精度を向上させる．レレバンスフィードバックでは初期検索の１８％程度，ローカルフィードバックでは初期検索の５％程度の精度向上を実現する．

第６章では，第５章で確立したローカルフィードバック技術の適用事例として，機械翻訳を用いた英日・日英言語横断検索について述べる．英語検索要求による日本語文書の検索実験では，文書の日英機械翻訳と検索要求の英日機械翻訳のアプローチを比較し，前者のアプローチにより単言語検索の精度の９０％以上，後者のアプローチにより８０％以上が実現できることを示す．さらに，日本語テストコレクションの検索要求を人手で英訳して英日言語横断検索の模擬実験を行う場合に二人の翻訳者を起用することにより，検索精度が人手による翻訳結果に大きく左右されることを示す．一方，日本語検索要求による擬似英語文書の検索実験では，検索要求の日英機械翻訳の前後にローカルフィードバックを行うことにより，単言語検索の場合と同程度の精度を実現する．これは実際の英語文書を検索対象とした場合の精度の上限を示すものである．最後に，機械翻訳精度と検索精度との関係について考察を行う．

第７章では，本研究のまとめを行い，得られた成果を要約し，今後の課題とこの分野の展望を述べる．