情報フィルタリング技術

Information Filtering

住田　一男　K. Sumita
三池　誠司　S. Miike

情報の電子化が進行し，種々の情報への電子的なアクセスが容易になりつつあり，増大するいっぽうの情報の中から，利用者の望む情報だけを的確に入手する技術の実現が求められている。日々発生する新聞記事を対象として，利用者の必要とする記事だけを選別する情報フィルタリングシステムを開発した。検索条件と記事との類似度を求める手段をもち，類似度があるしきい値を超えた記事だけを利用者に提供する。また，記事間で類似度を求め内容の重複する記事を指摘することができる。選別された記事は，電子メールやWWW (World Wide Web) などのインターネットを介して利用者に提供される。

開発した技術およびシステムは，今後当社が事業展開を進める情報提供サービス事業に応用していく。

We have developed an information filtering system for daily newspaper articles published in digital form. The system computes the similarity between a user's information need and an article based on an expanded vector space model, and then selects articles that correspond to the user's need. The system also detects articles having similar contents to a particular article, so that it can indicate a cluster of such similar articles. The selected newspaper articles are provided to users by means of Internet communication tools such as e-mail and the World Wide Web (WWW).

We will apply this newly developed technology to new products and systems in a Toshiba information providing business.

1 まえがき

米国が全米情報基盤計画 (NII : National Information Infrastructure) を提唱して以来，情報ハイウェイやインターネットといった用語が一般的になってきた。わが国においても，2010年までに各家庭に光ファイバ網をはりめぐらすことが計画されている。家庭にいながらにして，テキスト，音声，画像など世界中のさまざまなマルチメディア情報へアクセスすることが可能になってくる。

このようなネットワークインフラの整備とともに，ネットワーク上に存在すると思われる所望の情報を入手する技術 (検索技術)，日々新たに発生する情報から不必要な情報を除き，必要なものだけを取り出してユーザに提供する技術 (情報フィルタリング技術) の重要性が増してきている。

マルチメディア情報の中でも，テキストはものごとの内容を記述するうえで中心的なメディアであり，情報検索やフィルタリングにおいてもテキストを手がかりにすることで，内容に基づいた処理が可能となる。

ここでは，テキスト情報を対象とした情報フィルタリング技術について，技術の概要と，当社の取組みについて概説する。

2 情報フィルタリング技術

新規に発生するデータだけを対象にして，利用者が必要とするデータを選択して継続的に提供するサービスを情報フィルタリングサービスと呼ぶ。このようなサービス形態は，SDI (Selective Dissemination of Information)，Routingとも呼ばれる。利用者が必要とする情報を探し出して利用者に提供するという点は，いわゆる情報検索システムと同様であるが，情報検索システムとの違いは，**表1**に示すようにその応用形態にあると言える。

情報フィルタリングでは，新しい情報が発生した段階で，それらの情報の中から利用者が必要とする情報を，利用者に通知／提供する。検索と比較して，情報の受取りという面からは利用者は受動的な立場にある。このため利用者の満足を

表1．情報検索と情報フィルタリングの比較
Comparison of information retrieval and information filtering

	検索対象	検索条件指定	実行形態
検　索	蓄積データ	逐一	検索指定時
フィルタリング	新規データ	ほぼ一定	定期的／データ発生時

得るためにはより精度の高い処理が求められると予測される。

現在の商用のオンラインデータベースや全文検索システムでは，検索条件で指定する語が，各テキストにあらかじめ付与されたキーワードやテキスト中の語に照合することを判定材料として検索を行う。検索条件には，ANDやORといった論理演算子を用いて指定する語を組み合わせることもできる。このような検索手法をexact matchと呼ぶ。

この手法では，検索条件で指定した語が，あるテキストの主題でない部分で用いられた場合でもそのテキストが検索されることになり，十分な精度が得られない[1]。

精度の向上のため，次のような情報の利用が考えられる。

(1) 統計情報の利用　語の頻度情報で，その語の重要度をある程度推定することができる。例えば，あるテキストで何度も用いられている語は，そのテキストで重要な語だと推定される。しかし，頻度の高い語でも，多くの異なるテキストで用いられているような語は，検索条件としての弁別能力に欠ける。語の頻度などの統計情報を利用することが考えられる[1]。

(2) 言語情報の利用　語の頻度情報だけでは，検索の精度は十分なものとすることはできない。例えば，“コンピューター”と“計算機”のようにほとんど同じ意味をもつ語どうしを“同義語”と呼ぶ。また，“アップル”という言葉は本来なら“りんご”という意味となるが，情報処理分野で使われる場合は，米国のパソコンメーカの名前とみなしたほうが自然である。このような語を“多義語”と呼ぶ。精度を上げるためには，検索語を同義語で展開する必要がある。また，多義語を扱うためには，文脈によって語義を限定するといった処理も必要となってくる。テキストにおいてそれぞれの単語は，個々独立で用いられているわけではなく，意味的な関係や文脈的な関係もその間に存在する。こういった情報の利用も考えられる[2],[3]。

(3) 書式情報の利用　新聞記事や技術論文のようにある程度形式の定まったテキストでは，多くの場合見出しや，節や段落といった書式情報が利用できる。見出しは，そのテキストの内容を知るうえで重要なキーとなる。また，新聞記事では1文目や1段落目で，主要な情報が述べられる。これらの情報も，関連性判定のための手がかりとして用いることができる[4]。

上述の情報は，個々独立の情報として取り扱われるべきではない。検索条件とテキストとの関連性判定にあたっては，これら情報の統合的な利用が必要である。

3 情報フィルタリングシステム

新聞記事を対象にした情報フィルタリングシステムを試作した。システムの構成を図1に示す。各利用者の情報要求に対応する検索条件に従って，関連記事を選別する。

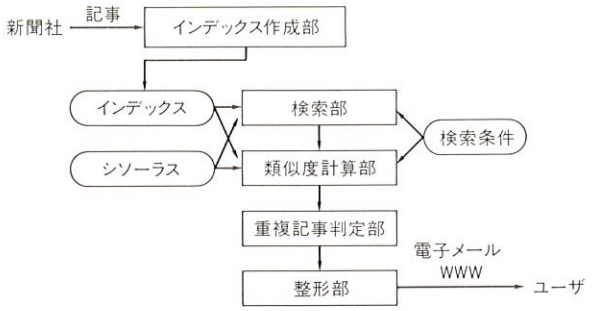

図1．情報フィルタリングシステムの構成　システムは，インデックス作成フェーズとフィルタリングフェーズとから成る。
Configuration of information filtering system

3.1 インデックス作成部

取り込まれた記事情報に対して，文字，単語，共起情報などに関するインデックスを作成する。この際，記事の書式解析により見出しや段落の認識を行うとともに，形態素解析により記事中のすべての語を抽出する。

3.2 検索部

検索条件に従って記事を検索する。検索条件に含まれる語を，シソーラスによって同義語，下位語，上位語などを展開した後，それらの語が含まれる記事をインデックスを用いて高速に検索し，類似度計算を行う記事を選別する。

3.3 類似度計算部

検索された各記事について，検索条件との類似度を求める。類似度は，ベクトル空間法[5]を拡張した手法を考案し用いている。

ベクトル空間法では，テキストは含んでいる語を要素とするベクトルで表現される。ベクトルの各要素の値は，通常，対応する語のテキスト中の頻度やその単語が含まれるテキストの数などによって定められる値が設定される。検索条件側もテキストベクトルと同様にベクトルとして表現される。これらのベクトル間の角度の余弦をそのテキストと検索条件の類似度とする。

開発したシステムでは，語の共起情報，見出しで用いられているか否かなどの出現位置などを個別のベクトルとして表現し，個々のベクトルで求められる類似度の荷重平均をとることで全体の類似度を求めている。

3.4 重複記事判定部

複数のニュース源から記事を受けることになるため，同じ内容の記事が含まれる可能性がある。重複記事を判定し，その情報を利用者に提示することで，異なる新聞社からの記事を比較したり，一社の記事だけを参照するといった利用者の利便性を目ざした機能が実現できる。この判定では，記事間で重複する名詞の数と両記事の全単語数との比率を求める。この値が，あるしきい値を超えた場合，両記事が同一内容で

あると判定する。

新聞記事においては，企業名などの固有名詞が重要な役割を果たす。判定の際には，主語の固有名詞が一致しているかなどの情報も利用している。

3.5 整形部

検索条件との類似度があるしきい値を超える記事だけを利用者に提供する。システムでは，WWWや電子メールでの利用を想定し，情報提供形態に対応する形式に変換する。例えばWWWで提供する場合，検索結果の記事はHTML (HyperText Markup Language) 形式に変換されることになる。

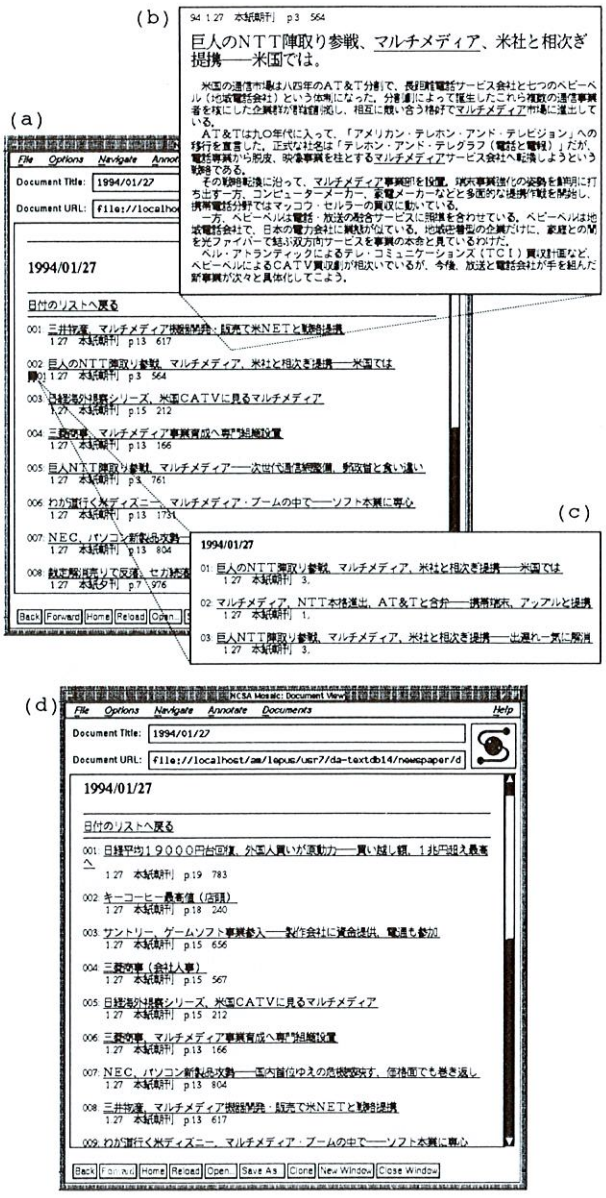

図2．フィルタリング結果例　“マルチメディア”という話題に関してフィルタリングした結果を示す。(a)はシステムの出力する記事一覧，(b)は一記事の全文，(c)は内容が重複すると判定された記事，(d)はexact matchによる処理結果。

Filtering result

WWWにおけるHTMLブラウザでフィルタリング結果を表示した例を**図2**に示す。図2(a)は絞り込まれた記事の一覧，(b)はそのうちの一記事の全文である。記事一覧において，■でマーキングされているものは，内容の類似する記事が存在することを示している。この箇所をマウスでクリックすることにより類似記事の一覧が表示される図2(c)。また，記事は，検索条件と関連性の高いものから順に並べられている。

比較のためexact matchによる記事一覧を図2(d)に示す。記事番号順に表示されており，関連性の大小は判定できない。また，4番目の記事のように会社人事といった関係のない記事も含まれている。これは，“マルチメディア”という語が，たまたまその記事中で用いられていたためである。

4 あとがき

情報フィルタリング技術を概観するとともに，試作した情報フィルタリングシステムの概要を述べた。情報提供サービスを想定して，新聞記事データを用いた試行実験を行い，フィルタリング処理の有効性を確認している。今後さらに精度を高めるためには，テキストの内容に立ち入ったより深い言語処理が必要になってくる。技術の深耕に努めたい。

なお，今回のフィルタリングの実験には，日本経済新聞CD-ROM 94年版を使用した。

文　献

(1) 住田一男，他：文の意味解析に基づく全文検索，第8回人工知能学会全国大会，**24-2**，pp.667-670 (1994)

(2) K. Sumita, et al: Document Structure Extraction for Interactive Document Retrieval System, Proc. ACM SIGDOC'93, pp.301-310 (1993)

(3) S. Miike, et al: A Full-Text Retrieval System with a Dynamic Abstract Generation Function, Proc. ACM SIGIR'94, pp.152-161 (1994)

(4) K. Sumita, et al: Document Structure Extraction for Interactive Full-Text Retrieval, Natural Language Processing Pacific Rim Symposium'93, pp.144-152 (1993)

(5) G. Salton: The Vector Space Model, Automatic Text Processing, Addison-Wesley Publishing Company, pp.312-325 (1989)

住田　一男　Kazuo Sumita

1982年入社。自然言語処理技術，情報検索技術の研究・開発に従事。現在，研究開発センター 情報・通信システム研究所研究主務。

Communication and Information Systems Labs.

三池　誠司　Seiji Miike

1984年入社。自然言語処理技術，情報検索技術の研究・開発に従事。現在，研究開発センター 情報・通信システム研究所研究主務。

Communication and Information Systems Labs.