フロー情報を対象にした情報検索システム (4)
－ 文書分類 －*

4F－9

廣田 誠　上田隆也　柴田昇吾　伊藤史朗　池田裕治　藤田 稔
キヤノン(株) 情報メディア研究所

1　はじめに

コンピュータネットワークの発達がもたらす“情報洪水”という問題への対処として，我々は，フロー情報からの情報収集・整理を支援する情報検索システムの研究開発を進めている[1]．その中で，大量の情報の中からユーザにとって必要な情報だけを選別して取り込む情報フィルタリングの技術や，取り込んだ文書情報を効率的に利用・整理するために，ユーザの設定したカテゴリに分類する文書分類の技術を検討した．

情報フィルタリングは，文書情報を「必要なもの」と「不要なもの」に「分類」することと言える．従って，情報フィルタリング，文書分類はいずれも文書を「分類」する技術が基盤となる．本研究では，この「分類」する技術について検討し，これを基盤として，情報フィルタリングの機能，カテゴリへの文書分類の機能を実現した．本稿では，その内容を述べる．

2　フロー情報の「分類」

フロー情報に対する情報収集・整理を支援するシステムにとって我々が重要と考える要件として「ユーザへの適応」「複数情報源への対応」がある．文書を「分類」する機能がこれらの要件を満たすには，主に次の２つの処理が必要である．

文書の内容に基づいた処理

「複数情報源へ対応」するには，ヘッダー情報を頼りに電子メールを分類するというような，情報源特有のフォーマットに依存した手法は適切ではない．必要なのは文書の内容に基づいて処理することである．本システムでは，文書内容の効果的な表現方法であるベクトル空間モデル[2]に基づいて「分類」処理を行なう．

事例からの学習

情報に対する興味や分類の仕方は，ユーザによって異なる．個々のユーザに適応するために，ユーザが「必要/不要」の判断を下した文書，カテゴリに分類した文書を事例としてそのユーザがどのような情報を必要としているのか，情報をどのように分類しようとしているのかを学習する．

3　「分類」のプロセス

「分類」のプロセスは，学習フェーズと実行フェーズからなる．その内容を順に説明する．

3.1　学習フェーズ(1)–有効語辞書の作成–

文書のベクトル表現は，文書中の単語の出現パターンに基づく．目的は「分類」であるから，「分類」に有効な情報を担った単語(**有効語**と呼ぶ)に注目したほうがよい．統計的観点からすれば，有効語は，特定のカテゴリに偏って出現する単語と考えられる．そこで，学習文書に現れる各単語について，その偏りの度合に基づく評価値 E を計算し，この評価値の高い順に N 個を有効語として抽出する．

次に，有効語どうしの意味的な距離を考慮するため，有効語間の共起確率に基づいて，有効語の意味を多次元空間のベクトルで表現する(共起ベクトル[3][4])．手順は次の通りである．

1. **共起確率の算出**
   有効語どうしの共起確率を，学習文書から求める．
2. **基底語選択による次元圧縮**
   有効語の意味をなるべく低い次元のベクトルで表すことで，保持するデータのサイズ，処理コストの軽減を図る．そのために，$L(\ll N)$ 個の**基底語**を選択し，これらをベクトル空間の軸とする．

このようにして，有効語 W_i の意味ベクトルは，有効語 W_i と基底語 W_k^B の共起確率を $c_{i,k}$ として，

$$W_i = (c_{i,1}, c_{i,2}, c_{i,3}, \cdots, c_{i,L})$$

のような L 次元ベクトルで定義される．以上によって求められた有効語とその意味ベクトルを**有効語辞書**として保持する．

---

*An Information Retrieval System for Flow-type Information (4) – Document Categorization –

HIROTA Makoto, UEDA Takaya, SHIBATA Shogo, ITOH Fumiaki, IKEDA Yuji and FUJITA Minoru (Media Technology Laboratory, Canon Inc.)

3.2 学習フェーズ(2)–重み学習–

文書内容は，その中に含まれる有効語の意味ベクトルの重み付き平均によって表現する(以降，文書ベクトルと呼ぶ)．文書内容を適切にベクトル表現するには，適切な有効語選択，有効語の適切なベクトル表現の他に，有効語の適切な重み付けが重要な要素となる．本研究では，次のようにして重みを求めた．

まず，文書中での有効語での重要性に関連のありそうな，

- その有効語の出現位置
- その有効語の格役割，修飾タイプなどの言語的役割

に注目して，評価項目をあらかじめ作成しておき，有効語が各評価項目の条件を満たした場合に与える重みの値を，学習文書から学習する．

3.3 学習フェーズ(3)–代表ベクトルの算出–

学習文書すべての文書ベクトルを求め，各カテゴリの平均ベクトルを算出し，これを各カテゴリの代表ベクトルとする．

3.4 実行フェーズ

学習の結果として作成された有効語辞書，重みデータ，代表ベクトルデータに基づき，入力文書を分類する．まず，入力文書の文書ベクトルを求める．この文書ベクトルを Dとしカテゴリ C_kの代表ベクトルを V_kとして，カテゴリ C_kへの帰属度 d_kを次のように計算し，帰属度の高い順に分類カテゴリの候補とする．

$$d_k = \frac{D \cdot V_k}{|D||V_k|} \quad (\cdot\text{は内積}, |\cdot|\text{はノルム})$$

4 性能評価

以上に説明した「分類する」技術を基盤として，情報フィルタリング，文書分類の各機能をインプリメントし，その性能を評価するための実験を行なった．実験データとして，オンラインデータベースからダウンロードした新聞記事(2紙)のうち，経済関連の記事を用いた．

まず，情報フィルタリングについては，上記の経済関連記事のうち，我々の興味の対象である情報処理技術関連の記事を必要記事として取り出すように学習した．学習文書は，約10000の経済関連記事を人手で必要/不要と判定して作成した．

一方の文書分類については，必要記事約600記事をさらに「マルチメディア」「人工知能」…といった8つのカテゴリのいずれか(複数可)に人手で分類し，これを学習文書とした．

学習後，学習に用いなかった記事をテスト用記事として情報フィルタリング，文書分類をそれぞれ実行した．結果を表1に示す．

表 1: 実験結果

情報フィルタリング

テスト用記事数	2077記事(必要記事111，不要記事1966)
再現率	93%(必要記事111記事中103記事の取り出しに成功)
適合率	35%(取り出した298記事中必要記事は103記事)
記事数の削減率	86%(トータル2077記事のうち1779記事を削減)

文書分類

テスト用記事数	124記事	
精度	第一候補の正解率	77%(96/124)
	第二候補までの正解率	91%(113/124)
	第三候補までの正解率	94%(116/124)

5 おわりに

“情報洪水”の問題に対処する手法として，情報フィルタリングの技術，文書分類の技術について検討し，その内容を述べた．今後は，「ユーザの興味の変化への対応」，「分類カテゴリの変化への対応」，「複数分類軸への対応」，「新聞記事以外の文書情報に対する性能検証」などの課題を解決して，情報フィルタリング，文書分類の機能向上を図る予定である．

参考文献

[1] 上田他：フロー情報を対象にした情報検索システム(1) −概要−，情報処理学会第50回全国大会4F-6，1995.

[2] G.Salton, M.J.McGill : *Introduction to Modern Information Retrieval*,McGraw-Hill,1983 .

[3] 湯浅，上田，外川：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類，自然言語処理研究会98-11,1993.

[4] 丹羽，新田：単語ベクトルを用いた多義語の意味推定–共起ベクトルと定義距離ベクトルの比較，自然言語処理研究会102-7,1994.