新聞記事自動分類システム構築の検討と評価

3T－9

森本　由起子　　間瀬　久雄　　辻　洋　　絹川　博之
(株)日立製作所 システム開発研究所

1. はじめに

大量の記事情報を扱う新聞社では、複数の情報源から電子化されたデータをリアルタイムで受信すると同時に、社外への配信を行なっている。これらの記事情報を一つの窓口で受信し、内容別にカテゴリに自動分類して一括管理し、必要な情報を効率良く検索／抽出／配信したいという要求が高まっている。

我々は、テキストを既存のカテゴリに自動分類するシステムの構築を支援するテキスト分類支援ツールFLUTE (Filtering Lens software system for Unclassified TExts) [1][2][3] を開発しており、本ツールを適用した新聞記事自動分類システムの構築を検討した。更にシステム実現可能性を検証すべく、新聞記事１年分を用いて、２２の大分類カテゴリに自動分類する実験を行なった。本稿では、新聞記事自動分類システムの概要と分類精度向上を目的とした実験、実運用時における分類用知識ベースの保守に関連した実験結果について報告する。

2. FLUTEの新聞記事自動分類への適用

2.1 新聞記事自動分類システムの概要

本システムは、以下の機能を有する。

(1) 受信した大量の記事情報を予め定義した分類体系にリアルタイムで自動分類して一括管理する。よって効率の良い検索／情報抽出が可能である。

(2) 顧客に関心のある記事情報を自動抽出して配信する。その際に、分類／検索の付加価値を付け、顧客がさらにパーソナルな分類体系で階層的に分類することを可能とする。

2.2 FLUTEの概要

FLUTEは、以下の機能を有する（図1参照）。

(1) 分類済み文書からカテゴリを特徴付けるキーワードを自動抽出し、分類用知識ベースを自動生成。

(2) 新規未分類文書からキーワードを自動抽出し、分類用知識ベースを参照して予め定義した分類体系に自動分類。

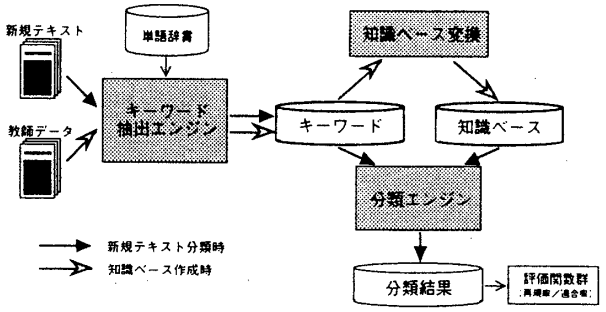

図１　FLUTEの構成

(3) 分類結果を評価するための履歴を出力し、分類精度を算出。

3. 分類実験方法及び結果

FLUTEを新聞記事自動分類に適用し、実現可能性を検証するために、分類実験を行なった。

3.1 実験方法

実験一覧を表１に、各実験の目的を以下に示す。

(1) 共通語・不要語の除去の有効性検証実験
　共通語とは、多くのカテゴリに共通して出現する単語を指し、不要語とは、文書の内容に依らずキーワードとなり得ない一般的な単語を指す。キーワードとして不要なこれらの単語を予め選定、除去することによりノイズを減らす。

(2) キーワードの重みの正規化の有効性検証実験
　キーワードが持つ重みの価値をカテゴリ間で同一になるように正規化しその影響を検証する。

(3) キーワード抽出範囲の違いの影響検証実験
　重要な内容は冒頭に記述するという新聞記事の特徴を活かし、キーワードの抽出範囲を第１段落＋見出しとした場合の分類精度を把握する。

(4) データ量多少の違いの影響検証実験
　実運用時の保守の観点から分類精度を安定させるために必要な教師用データの量を検証する。

(5) データの年度の違いの影響検証実験
　上記実験と同様に分類精度を安定させるために必要な知識ベースの更新時期を検証する。

3.2 実験で使用する記事データ

実験では、日経四紙の新聞記事１年分を使用した。実験で使用したデータの規模を表２に示す。

---

*Evaluation of News Articles Categorization System.

†Yukiko MORIMOTO, Hisao MASE, Hiroshi TSUJI, and Hiroshi KINUKAWA
‡Systems Development Laboratory, Hitachi Ltd.

表 1 実験一覧

項番	評価項目	教師用データ	評価用データ	不要語除去	重みの正規化	解析範囲
1	初期知識ベースによる精度測定	'94/1-3	'94/4-6	しない	しない	見出し＋本文全部
2	共通語・不要語の除去の有効性検証	'94/1-3	'94/4-6	する	しない	見出し＋本文全部
3	キーワード重み正規化の有効性検証	'94/1-3	'94/4-6	しない	する	見出し＋本文全部
4	実験項目 2 + 3	'94/1-3	'94/4-6	する	する	見出し＋本文全部
5	キーワード抽出範囲の違い影響検証	'94/1-3	'94/4-6	する	する	見出し＋第 1 段落
6	データ量多少の違いの影響検証	1 週間 ('94/3) 2 週間 ('94/3) 1 か月 ('94/3) 3 か月 ('94/1-3) 6 か月 ('93/10-3)	'94/4-6	する	する	見出し＋本文全部
7	データ作成年度の違いの影響検証	'88/1-3	'94/4-6	する	する	見出し＋本文全部

表 2 実験で使用したデータ

評価用データ	'94/4-6	37462 件
教師用データ	'94/1-3	34556 件
	'93/10-12	36472 件
	'88/1-3	32276 件
一記事あたりの平均カテゴリ数		3.5
知識ベースのキーワード種類数（キーワード数）		
	見出し＋第 1 段落	94028
	見出し＋本文全部	140923
知識ベースの規模（レコード数）		
	見出し＋第 1 段落	620719
	見出し＋本文全部	1059721

3.3 実験結果

分類結果の評価指標には再現率を用いた。再現率は、あるしきい値よりも高い確信度（分類結果の正しさを表す値）をもつカテゴリを付与した場合の、

$$\text{再現率} = \frac{\text{機械が付与した正解カテゴリ数}}{\text{正解カテゴリ数の総数}}$$

とした。

分類結果を表 3に示す。

表 3 分類結果

項番	評価項目		再現率
1	初期知識ベース		61.34 %
2	共通語・不要語の除去		66.83 %
3	キーワードの重み正規化		68.13 %
4	実験項目 2 + 3		68.95 %
5	キーワード抽出範囲の違い		69.74 %
6	データ量多少の違い	1 週間	63.60 %
		2 週間	67.91 %
		1 か月	68.72 %
		3 か月	68.95 %
		6 か月	68.50 %
7	データ年度の違い	'94/1-3	68.95 %
		'88/1-3	65.78 %

実験項番 1、2、3 の結果から、知識ベース作成時において、共通語・不要語除去及びキーワードの重みを正規化した場合、それぞれ分類精度が数%ずつ向上することがわかった。また、これらの方法を組み合わせると、さらに分類精度が向上することがわかった。

実験項番 1、5 の結果から、キーワード抽出範囲は、見出し＋第 1 段落とした場合の方が良かった。このことから、解析範囲を減らすことで、分類精度を保持しつつ知識ベース作成のコストを削減させることが可能であることがわかった。

実験項番 1、6 の結果から、約 2 週間の記事データの量があれば、分類精度を安定させることが可能であることがわかった。実験項番 7 の結果から、新しい記事データの方が古い方よりも分類精度が良く、知識ベースを定期的に更新していく必要があることがわかった。

4. 研究開発のまとめ

我々が開発したテキスト分類支援ツール FLUTE を新聞記事自動分類システム構築に適用し、その実現可能性を検証すべく、新聞記事 1 年分を用いて、大分類 2 2 カテゴリに自動分類する実験を行なった。

謝辞

本研究の機会と、本研究への貴重な御意見、御助言を頂いた (株) 日本経済新聞社システム局、データバンク局の関係者の方々に感謝致します。また、本実験に御協力頂いた (株) 日立製作所情報システム事業部の関係者の方々に感謝致します。

参考文献

[1] 辻他 3 名: テキスト自動分類エキスパートシステムの一構成法、情報処理学会第 49 回全国大会講演論文集 (3)3-93,('94.9)

[2] 間瀬他 3 名: テキスト分類支援ツール FLUTE の開発 (1) - 機能と構成 -、情報処理学会第 52 回全国大会講演論文集 (3)3-303,('96.3)

[3] 森本他 3 名: テキスト分類支援ツール FLUTE の開発 (2) - 障害事例分類への適用 -、情報処理学会第 52 回全国大会講演論文集 (3)3-305,('96.3)