論文

テキスト情報を対象とした キーワード抽出と関連情報収集システム[†]

片岡 充照*¹ 今中 武*² 水谷 研治*³ 若見 昇*⁴

近年、放送や通信の発展によって、一般ユーザに提供される情報の量が急増している。特にマルチメディア情報を扱えるようになってきた現代においても、インターネット、文字多重放送などではテキスト情報が内容の主要部分を占めている。テキスト情報の量が多くなれば、ユーザは大量の文章を読まなければならない、情報利用が非常に困難になる。このような状況においては、大量のテキスト情報から要点を抽出し、興味のある情報だけを選択的に取り出す情報要約・検索手法が求められる。この要求に対して、従来より、自然言語処理の分野において、いくつかの手法が提案されている。これらの手法では、文章の意味的側面まで捉えることができ、より高度な要約・検索が実現できる。しかしながら、形態素解析や文脈理解を行うために、類似語辞書、文法知識、意味辞書など、大規模な辞書的情報を必要とするため、固有名詞など新たな表現の出現に対して柔軟に対応できない。一方、これらの辞書的情報を用いずに、単語の出現頻度などから要約・検索を実現する手法がある。この手法は適用範囲が広く、固有名詞など新たな単語の出現に柔軟に対応できる反面、例えば、頻出単語が要約結果として単に提示されるなど、結果がユーザにとって分かりにくい。また、検索が単純なキーワード文字列の照合で行うため、要約・検索能力が低いといった問題がある。著者らは実用面から考えた適用範囲の広さに重点を置くために後者の立場をとり、大量の日本語テキスト情報からホットな話題を抽出して要約するシステム KEIFIS(Keyword Extract and Information Filtering System)を開発した。KEIFIS では、より高い要約・検索能力を実現する以下の特徴を持っている。

- (1) ホットな話題を特徴付ける関連単語を抽出し、関連単語を組み合わせることにより要約結果を表現。
- (2) ユーザの興味のある情報を話題単位で検索し、新たな関連情報の到着を監視。

また、KEIFIS の実問題への適用例として、実際に放送されている文字多重放送のニュース番組に適用し、適切にホットな話題を抽出できることを確かめた。

キーワード：情報要約、情報フィルタリング、キーワード抽出、ファジィ情報処理

1. はじめに

21 世紀のマルチメディア時代に向けて、一般ユーザに提供される情報の量は急増し、情報の洪水ともいえる状況になりつつある。この状況では、提供される情報についての概要を把握することすら困難である。このような時代においても、電子新聞、電子図書館など、テキストを中心とした情報が現在以上に多く用いられると考えられる。テキスト情報が大量になれば、ユーザは大量の文章を読まなければならない、情報の

利用が非常に困難になる。また、この時代には、一般家庭でも情報を活用する機会が急速に拡大と考えられ、誰でも簡単に情報を利用できる操作性の実現が求められる。このため、以下の特徴をもつ情報要約検索システムの実現手法確立が求められる。

(要求 1) 大量のテキスト情報から要点を抽出し、わかりやすく提示する。

(要求 2) テレビのリモコン程度の簡単な操作によって、各ユーザの興味のある情報を選択的に取り出す。

従来、(要求 1) に対して、予め準備した単語間の概念構造などの辞書的情報を用いて重要単語を抽出する手法が提案されている[4-8]。しかし「J リーグ」など、新たな概念、単語が出現する度に、辞書的情報の更新が必要となる。また、(要求 2) に対して、SQL の検索式や検索キーワードによる手法がある。しかし、例えば、「何か新しい情報が無いだろうか」と新聞を見るように情報を眺めるユーザにとって、検索目的を検索式で表現したり、情報の概要がわからないまま適切な検索キーワードを指定することは簡単ではない。

著者らは以下の特徴を持つ情報要約・検索手法の提案と、この手法を実装したシステム KEIFIS(Keyword Extract and Information Filtering System)の開発を行なった。

[†] A System of Automatic Keywords Extraction and Related Data Collection for Japanese Text
Mitsuteru KATAOKA, Takeshi IMANAKA, Kenji MIZUTANI and Noboru WAKAMI

*¹ 松下電器産業(株) AVC 商品開発研究所
AVC products development Laboratory, Matsushita Electric Industrial Co., Ltd.

*² 松下電器産業(株) 中央研究所
Central Research Laboratories, Matsushita Electric Industrial Co., Ltd.

*³ 松下電器産業(株) BIU
Business Incubation Unit, Matsushita Electric Industrial Co., Ltd.

*⁴ 松下電器産業(株) 海外 R&D 推進センター
Overseas R&D Promotion Center, Matsushita Electric Industrial Co., Ltd.

(特徴1) 異なる情報ソースにおいて数多く現れる話題(ホット情報と呼ぶ)を特徴付ける単語(ホットキーワードと呼ぶ)を要点として自動抽出し、ホット情報を表現するホットキーワードの組み合わせといったシンプルな形式で提示する。また、辞書的情報を用いることなく、ファジィ情報処理によって類似語処理、関連語処理を実現している。

(特徴2) ユーザは、提示されたホット情報の中から相対的に興味のあるものを選択するだけで情報ソースから与えられたテキスト情報を検索できる。更に、選択したホット情報に関連する新たな情報の到着を監視できる。KEIFISでは、提示されたホット情報の中にユーザにとって興味のあるものが含まれてない場合があるものの、新しい情報を眺めるユーザに、キーボードからのキーワード入力など、一般家庭のユーザにとって困難な操作を要求しないという利点がある。

以下、本論文では、KEIFISの概要と、その機能を実現するための手法について述べる。

2. 機能の概要

2.1 処理対象とするデータ

KEIFISが処理の対象とするデータ(対象データと呼ぶ)は、複数の情報ソースから与えられる一定単位(情報単位と呼ぶ)に分割された日本語テキストであり、かつ、時間に伴って更新されるものである。例えば、テレビ文字多重放送のニュースは、複数の放送局から記事ごとに分割されて提供され、およそ1時間毎に更新される。他の例としては、電子メール、パソコン通信のフォーラム、特許情報などがある。

2.2 ホット情報の自動抽出

ある時点の対象データに含まれる情報単位を分類し、分類毎に共通の内容を話題とする。各話題について、対応する情報単位を含む情報ソースの数が相対的に多

い話題をホット情報と定義する。このようにホット情報を定めたのは、例えば阪神大震災、地下鉄サリン事件など、その時々ホットな話題は数多くの放送局や新聞社など異なる情報ソースで共通に扱われることが多いからである。

KEIFISでは、ホット情報の情報単位に含まれ、かつ、ホット情報を特徴付ける単語をホットキーワードと呼び、ホットキーワードの組によってホット情報をユーザに提示する(図1)。ホットキーワードの満たすべき要件は、ユーザに提示することで話題を連想させ、かつ、興味のある話題に到達するためのインデックスとして用いられることである。

KEIFISでは、話題を表現する際に、個々の単語の重要性よりも単語間の関連性に着目し、より関連性が高い単語の組み合わせをホットキーワードの組としている。例えば、総選挙が行われて○×党が大勝利を収めたことと同時に、□△党の不正発覚が話題になった場合を考える。このとき、出現頻度に基づいて重要単語を求めるだけなら、「総選挙」が両方の話題に共通して含まれるため重要単語となる[9]。しかしながら、ホットな話題という観点では、これら2つの話題は関連はあるものの別の話題として捉えるべきである。KEIFISでは個々の単語の出現頻度よりも、2つの単語が同じ情報単位の中で出現する共起関係に着目する。前述の例では、「○×党」と「大勝利」、「□△党」と「不正発覚」とをそれぞれ組み合わせ、別々の話題として示すようにしている。

2.3 話題による検索と続報の監視

KEIFISでは、話題をキーとした検索を行える。例えば、ユーザは表示された話題の中から興味のある話題をリモコン操作で選択するだけで、検索キーワードの入力といった複雑な操作をすることなく、情報検索や収集を行える。例えば、図1の表示に対してユーザが上下ボタンで話題を選択すれば、厳密に一致する単語が含まれていなくても、類似単語を考慮しながら関連する情報単位を検索、表示できる。また、選択された話題に関連する新たな情報(続報と呼ぶ)の収集をボタンなどで指定するだけで、続報の到着を監視し、到着すれば効果音と画面によりユーザに通知する。

3. システムの構成

図2にKEIFISのシステム構成の概要を示す。KEIFISは、(1)話題構造生成部、(2)ホットキーワード抽出部、(3)情報収集部の3部分から構成されている。なお、KEIFISはUNIXワークステーション上のC言語で実装され、コア部分のソースコードは約2,000行である。以下、各部分について詳細に述べる。

図1 ホットキーワードの提示例

次に、対象データ中に含まれる類似単語を、類似単語抽出ルールを用いて抽出する。KEIFISは、異なる情報ソースを処理の対象とするため、単語表現の多様性に対応する必要がある。例えば、実際のデータでは「インタフェイス」、「インターフェース」、「インターフェイス」など、意味的に同じであっても、情報ソース毎に異なる表現(類似単語と呼ぶ)が用いられている。KEIFISではファジィ推論ルールを用いることにより、

NII-Electronic Library Service

を表している。これは、情報ソースが同じであれば表現の統一が図られており、わずかな表現の違いで2つの単語が別の事柄を表している可能性が高いことを示している。これに対して R 24 は、一致する文字の割合と、共起単語の部分は R 23 と同じ条件であるが、情報ソースが異なるために (S=false)、類似度 SI が大きいことを表している。これは、情報ソースが異なれば同じ事柄でも表現がまちまちなことが相対的に多いことを示している。

上記の類似単語抽出ルールでは、ルール調整の簡単さを考慮してメンバーシップ関数は固定にし、後件部の実数値のみを図4に示すような教師データを用いて調整する[2]。一旦調整された後は、例えば、国際ニュースからスポーツニュースへとといった情報ソースの変更を行わない限り、特に調整することなく KEIFIS を運用できる。

調整は、予め人手によって大まかな値を後件部に設定しておき、全ての教師データに対するトータルの誤差平均が極小になるよう再急降下法で調整する。この調整の方法は、[1]で著者らが用いたものと同様の方法であり、本論文では詳細な説明は行なわない。また、

単語A	単語B	被験者の付けた類似度	単語Aの情報単位	単語Bの情報単位
(伊達公子	伊達選手	0.9	情報ソース1の3番目の記事	情報ソース2の2番目の記事)
(伊達公子	伊達杏子	0	情報ソース1の3番目の記事	情報ソース3の2番目の記事)
(湾岸戦争	湾岸危機	0.5	情報ソース4の1番目の記事	情報ソース5の5番目の記事)
⋮	⋮		⋮	

図4 教師データの一例

図5 ファジィ共起頻度

この調整を行なうのは、KEIFIS を運用する分野によって類似単語の特徴が異なるためである。例えば、教師データの分野として、国際ニュースを選択した場合、「米朝高官協議」と「米朝高官会談」のように表現が比較的多様な訳語を多く含むため、表現の違いに対する判定が緩やかになるよう調整される。

・話題構造データ生成処理

話題構造データを生成するために、2つの単語 a, b が同時に同一情報単位内に出現する回数をファジィ共起頻度 D(a, b) として求める。

$$D(a, b) = d(a, b) + \alpha \sum_{x \neq a} \{ SI(a, x) \cdot d(x, b) \}$$

ただし、 x : a との類似単語

$d(a, b)$: a, b の共起頻度

$SI(a, x)$: a と x の類似度

2つの単語 a, b の共起頻度 $d(a, b)$ を、a, b の両方を含む情報単位数として定義する。例えば、図5の情報単位 a で「サッカー日本代表」は2回出現していますが、情報単位 a では「サッカー日本代表」と「アトランタオリンピック」の共起頻度は1回しかカウントされません。また、ファジィ共起頻度 $D(a, b)$ は、2つの単語 a, b の共起頻度 $d(a, b)$ に、a と類似した単語 x との共起頻度 $d(x, b)$ を前述の類似単語処理で求めた類似度 $SI(a, x)$ に従って加算し得られる。なお、類似した単語の、ファジィ共起頻度 $D(a, b)$ への影響を、類似度 $SI(a, x)$ の学習と独立して調節できるよう、加算の際に定数 α (ここでは経験値 0.5) を乗じている。この結果、1回、2回とカウントされる従来の共起頻度に比べて、1.2回、3.8回のように、単語間の類似度の曖昧性が反映される。これにより、同じ意味の単語が異なる表現で出現した場合にでも、類似単語辞書なしで類似単語を考慮したファジィ共起頻度を求められる(図5)。

次に、ファジィ共起頻度が0より大きい単語同士をリンクで繋ぎ話題構造データを生成する。話題構造データは単語をノードとするネットワークであり、リンクの属性としてファジィ共起頻度を持つ。図6に話題構造データの例を示す。例えば、「サッカー日本代表」と「アトランタオリンピック」と間のファジィ共起頻

図6 話題構造データの例

度は5.3である。

更に、2.2で述べたように、複数の話題に共通して含まれる単語が存在する場合、1つの話題構造データに複数の話題が対応する可能性がある(図7)。図7(a)は、「大統領」が2つの話題に共通して含まれるために、話題構造データが話題ごとに生成されない例を示している。

KEIFISは、少数のノードを複製することで分割されるような話題構造データが存在するかを判定する。例えば図7(b)の話題構造データは、1つのノード「大統領」を複製することにより分割される。次に、分割された話題構造データそれぞれに対して、ホットキーワードを抽出する。ただし、複製により分割可能かを判定する際に複製するノード数の最大は、もとの話題構造データに含まれる単語数に対する単調増加関数としている。これは、単語数が少ない話題構造データは、ノード間の結合が疎であるので、必要以上に話題構造データが分割される危険性が高いからである。今回は、この関数値として、話題構造データに含まれる単語数に実データによる経験値0.1を乗算した値を用いている。

3.2 ホットキーワード抽出部

ホットキーワード抽出部は、以下の処理を行なうことでホットキーワードを抽出する。

- (Step 1) 各話題構造データそれぞれに対し、最も共起頻度の大きいリンクを得る。
- (Step 2) Step 1で得られた全てのリンクを共起頻度順にソートし、上位を取り出す。
- (Step 3) Step 2で取り出されたリンクの両端の単語を2つのホットキーワードとして出力する。
- (Step 4) Step 3で取り出されたホットキーワードの両方に対して共起頻度の大きい単語を3つ目のホットキーワードとして出力する。

例えば図5の話題構造データでは、サッカー日本代表とアトランタオリンピックとの共起頻度が5.3と最も大きいので、この2つのキーワードが図6の話題構造データに対するホットキーワードとして選択される。Step 4まで実行することにより得た3つのホットキーワードは、最もホットな話題に対応するホットキーワードの組として図1のように話題の1つ目として表示する。次に、これらのホットキーワードの組を含む話題構造データを除く話題構造データに対して再びStep 1～Step 4を実行する。これにより、次にホットな話題に対するホットキーワードが3個得られる。図1のようにホットな話題を3個得るにはStep 1～Step 4を3回繰り返し実行する。

(a) 分割前

(b) 「大統領」を複製することによる分割後

図7 複数話題に対応する話題構造データの例

3.3 情報収集部

KEIFISは、話題を検索キーとする対象データの検索を行ない、関連する情報単位の集合(関連情報と呼ぶ)を得る。話題の指定はホットキーワードの組の選択により行なう。検索は、選択した組に含まれるホットキーワードの組HKに対する情報単位Iの関連度Mを用いて行われる。関連度Mの式を以下に示す。

$$M = \left[\sum_{t \in HK} \max_{k \in K(I)} \{SI(k, t)\} \right] / N(HK)$$

ただし、t: HK中の1単語

K(I): 情報単位Iに含まれる単語集合

N(HK): HKに含まれる単語数

関連度Mは、ホットキーワードkが情報単位Iに含まれない場合でも、類似度SI(k,t)の高い単語tが含まれれば、その類似度に従って値が大きくなる。KEIFISは、情報単位を関連度Mの降順にソートし、規定値以上の関連度Mをもつ情報単位をソート順に関連情報として出力する。このため、ユーザは検索結果を先頭から読み進めることで、関連度合いの大きい順に情報を得ることができる。

また、KEIFISは、指定した話題に関する新たな情報(続報と呼ぶ)の到着の監視を実現している。既に検索済みの関連情報が続報として重複して現れないようにするため、過去の検索結果の関連情報をLRU(Least Recently Use)アルゴリズム[10]を用い管理するバッファに格納し、新たに得られた関連情報と、バッファに格納済みの各情報単位とについて、同一か否かを判定している(図8)。同一か否かの判定は、情報単位に含まれる漢字、片仮名、算用数字などを1文字ずつ順に比較して行なっている。

図8 続報の抽出

図9 ホットキーワードの評価法

4. 適用実験

本章では実データに対する本手法の適用実験について述べる。

・話題からのホットキーワード抽出

話題に対して適切なホットキーワードが抽出されたか評価する実験を行なう（図9）。

あらかじめ10人の被験者が対象データを見て、最もホットと判断した3つの話題を抽出しておく。次に、各話題からホットキーワードとしてふさわしいと思われる単語3個ずつを抽出し、最も多くの人が抽出した上位3個の話題とそのホットキーワードを理想的な出力と定義する。

次に、出力されるホットキーワードと、理想的な出力との一致度を、ホットキーワードの一致する個数で行なう。ただし、比較する2単語の一方が他方の真部分文字列となっている場合は、一致と見なす。なお、最高得点は9個のホットキーワード全てが一致する場合の9点である。

実験では、実際のテレビ文字多重放送のニュース番組を対象データとした。対象データから3章で示した字種に基づく単語切り出しを行なった場合、3週間分のデータから75%以上が固有名詞になるという結果を得た。従って、同義語辞書など辞書的情報を用いる手法は現実的でないために、辞書的情報を用いない手法について、適切なホットキーワードが抽出されるかを比較する。以下に、比較した手法を示す。

- (1) 本手法。
- (2) ファジィルールなしの手法：本手法でファジィ共起頻度 $D(a,b)$ の代わりに共起頻度 $d(a,b)$ を用いた手法。
- (3) 出現頻度による手法：話題ごとに出現頻度の高い上位3単語を抽出する手法。

表1 ホットキーワード抽出評価

	ニュース	スポーツ
本手法(ファジィルールあり)	6.3	6.8
ファジィルールなしの手法	4.8	4.7
出現頻度による手法	3.7	3.6

ただし、手法(3)は情報単位の話題への分類は人手により行うものとする。

KEIFISで用いるホットキーワード抽出アルゴリズムの特徴として、ファジィルールを用いて類似単語処理を行なっている点と、共起頻度に着目している点が挙げられる。従って、比較する手法として、ファジィルールによる類似単語処理を行わない手法(2)と、共起頻度ではなく単純な単語の頻度を扱う手法(3)とを採りあげた。

実験では、5放送局の文字多重放送のデータを集めて1セットとし、20セットのデータを用い、各セットごとの点数の平均を求めた。実験の結果、表1に示すように本手法の方が単語の出現頻度による手法に比べ、より適切なホットキーワードを抽出できることが明らかになった。また、ファジィ推論ルールの導入の効果についても有用性が認められた。なお、手法(3)の得点が低いのは、出力に同義語を含むことが比較的多いからと考えられる。

5. おわりに

本論文では、著者らの開発した、日本語テキストからホット情報を抽出するシステム、KEIFISについて述べた。KEIFISは話題を特徴付けるホットキーワードを自動抽出し、ホットキーワードの組でホット情報を表現する。そして、ホットキーワードの組を用いる

ことで話題単位の高度な検索ができる。

また、KEIFIS は、ファジィ推論ルールを用いたことで類似語辞書など辞書的情報を必要とせず、より広い分野を適用範囲として扱える。しかし、適用分野によってはチューニングを含めた変更を要する可能性がある。例えば、どの話題にも共通して含まれる出現頻度の高い単語が存在する場合には、この単語の影響を除去する必要がある。しかしながら、本論文で実験対象とした文字多重放送のニュース番組では、抽出された単語の多くが話題に関連する固有名詞となったため、實際上、問題とならなかった。

最後に、今回は日本語テキスト情報を処理の対象としたが、今後は、英語などの他言語への適用を検討する。英語では単語はスペースによって明確に分離されているが、複数の単語の組み合わせによって、1つの事柄を表す。このため、複数単語の組み合わせをフレーズとして取り扱えるようにルールの拡張を検討している。

参 考 文 献

- [1] 今中 武, 片岡 充照, 若見 昇: ファジィを用いたトピック情報抽出手法, 第13回ファジィシステムシンポジウム講演論文集, 早稲田大学, pp.479-482 (1996).
- [2] 今中 武, 若見 昇: 入出力データからの推論ルール自動合成における記号属性の処理, 日本ファジィ学会誌, Vol.5, No.6, pp.1408-1423 (1993).
- [3] 今中 武, 片岡 充照, 松浦 聡, 若見 昇: ファジィを用いた類似キーワード処理手法の提案, 計測自動制御学会第38回自動制御連合講演会, pp.395-396 (1995).
- [4] 梶 博行, 新田 義彦: 文献検索システムのファジィモデルとその実現, 電子情報通信学会論文誌, Vol.D-43, pp.297-304 (1979).
- [5] 神尾 達也: 新聞記事データベースにおけるキーワード自動抽出, 情報管理, Vol.32, No.4, pp.283-293 (1989).
- [6] 内山 恵三, 中村 正規: 重要キーワードとその活用方法, 電子情報通信学会データ工学研究会資料, Vol.DE91-23, pp.83-93 (1991).
- [7] 木本 晴夫: 日本語新聞記事からのキーワード自動抽出と重要度評価, 電子情報通信学会論文誌, D-1, Vol.J74-D-1, pp.556-566 (1991).
- [8] Schank.R.C.: Conceptual Information Processing, North-Holland (1975).
- [9] 丹羽 寿男, 荻嶋 一弘, 木 泰治: キーワード抽出と最適文節選択による文字認識後処理, 電子情報通信学会論文誌, D-II, Vol.J76-D-II, No.5, pp.940-948 (1993).
- [10] Knuth,D.E.: The Art of Computer Programming, Vol.1: Fundamental Algorithms, Addison-Wesley (1973).

(1996年10月22日 受 付)

(1997年 6月17日 再受付)

[問い合わせ先]

〒567 大阪府茨木市松下町1番1号
松下電器産業(株) AVC社 AVC商品開発研究所
コンシューマ機器開発グループ
第1テレビ開発チーム
片岡 充照
TEL : 0726-24-7736
FAX : 0726-23-7034
E-mail : kataoka@tvrl.mei.co.jp

著 者 紹 介

片岡 充照 (かたおか みつてる)

松下電器産業(株) AVC社 AVC商品開発研究所

1989年 大阪大学・基礎工学部・情報工学科卒業。1991年 同大学大学院博士前期課程修了。同年 松下電器産業(株)入社。中央研究所を経て、現在、同社AVC商品開発研究所、及び、(株)次世代情報放送システム研究所に勤務。知的インタフェース、次世代放送型サービスの研究に従事。情報処理学会会員。

今中 武 (いманаか たけし)

松下電器産業(株) 中央研究所

1985年 大阪大学・基礎工学部・情報工学科卒業。1991年 同大学大学院博士後期課程修了。同年松下電器産業(株)入社。現在、中央研究所に勤務。工学博士。高次推論、データベース検索、知的インタフェースの研究に従事。電子情報通信学会、情報処理学会、人工知能学会各会員。

水谷 研治 (みずたに けんじ)

松下電器産業(株) BIU

1968年生まれ。1990年 京都大学・工学部・情報工学科卒業、1992年 同大学院修士課程修了。同年 松下電器産業(株)入社。現在、BIUに勤務。自然言語処理、対話処理、音声認識の研究に従事。情報処理学会、人工知能学会各会員。

若見 昇 (わかみ のぼる)

松下電器産業(株) 海外R&D推進センター

1971年 大阪大学・基礎工学部・制御工学科卒業。同年、松下電器産業(株)入社。中央研究所を経て、現在海外R&D推進センター所長。VTR、光ディスクドライブ、ファジィ理論とその応用に従事。工学博士。日本ファジィ学会、システム制御情報学会、現代経営学会などの会員。

A System of Automatic Keywords Extraction and Related Data Collection for Japanese Text

by

Mitsuteru KATAOKA, Takeshi IMANAKA, Kenji MIZUTANI and Noboru WAKAMI

Abstract :

Rapid progress of information services has been increasing the amount of text information which users receive on Internet, teletext broadcasting and so on. A system extracting important part automatically is useful for users to get their interest information easily in such a case. Some methods in the field of natural language processing can extract important keywords or phrases from text information by using word dictionaries. However, these methods can hardly deal with unexpected words such as new proper nouns.

On the other hand, the methods based on word occurrence are useful for dealing with unexpected words and they have high applicability for many kinds of text information because they use no dictionary. However, the results are insufficient as a summary in some cases.

In this paper, we describe our newly developed system KEIFIS (Keyword Extracting and Information Filtering System). KEIFIS extracts important keywords which represent major topics from a large amount of Japanese text information and collects related data with no dictionary.

KEIFIS has the following features :

- (1) It automatically extracts important keywords and combines some of them to give the users major topics.
- (2) It retrieves and collects information according to the topic specified by the users, and informs them of its arrival in real time.

KEIFIS employs fuzzy information processing in computing the similarity between words. The similarity is utilized in counting co-occurrence of the related words to extract the major topics in KEIFIS. The similarity is also used to define relationship between the specified topic and newly provided information.

We applied KEIFIS to news programs on the teletext broadcasting in Japanese, and made sure that KEIFIS was capable of extracting the important keywords which represent current major topics appropriately.

Keywords : Topic Extraction, Information Filtering, Keyword Extraction, Fuzzy Information Processing

Contact Address : **Mitsuteru KATAOKA**

*AVC products development Laboratory, Matsushita Electric Industrial Co., Ltd.
Matsushita, Ibaraki, Osaka 567, Japan.*

TEL : 0726-24-7736

FAX : 0726-23-7034

E-mail : kataoka@tvrl.mei.co.jp