知的検索システム Fit での重複記事の統合*

5N－7

柴田昇吾　大谷紀子　伊藤史朗　上田隆也　池田裕治
キヤノン(株) 情報メディア研究所

1　はじめに

情報の発信が容易になり情報源の数が増えると，情報の量の増加だけでなく**情報の重複**が問題となってくる．現在，コンピュータや通信の分野では，電子メールによるニュースサービスやWWWにおいて，産業系の新聞と同様の情報が発信されており，これらの間の重複率は30%～40%(記事数比) にもなる．今後，インターネット上での push 型の情報発信やデータ放送が普及することになると，ますますニュース的な情報が増え，重複率はさらに上昇していくことが予想される．

この問題を解決するために，我々は，重複している記事を統合することが有効だと考えた．そこで，フロー情報を収集する際に，内容の重複がある**重複記事**を選別し，これらを統合する機能を実現し，知的検索システム Fit [1] [2] [3] に追加した．本稿では，重複記事の統合機能とその効果について報告する．

2　関連記事の検出

1日に数百ある記事全体から内容の重複を検出するとなると，膨大な計算量が必要となる．そこで，本システムでは，あらかじめ重複記事の候補となる関連記事を絞り込むことにした．絞り込みの対象は，フロー情報である新着記事間で行なう．また，情報源ごとに発信時間のずれがあるため，ユーザがフォルダに保存したストック情報も対象とする．

Fit では，視点に応じた情報提示機能の一つとして，新製品記事を対象として**情報抽出**を行ない，「製品名」「メーカ名」「発売日」「特徴」などを一覧する機能を実装している．ここで抽出した情報が他の記事の情報と一致する記事を関連記事とした．また，新製品記事以外は括弧で括られた情報で代用している．

関連記事を対象として次の重複文の同定を行ない，実際に重複があるものを重複記事とする．

*A Fusion of Overlapped Documents in Fit, an Intelligent Retrieval System
SHIBATA Shogo, OTANI Noriko, ITOH Fumiaki, UEDA Takaya and IKEDA Yuji (Media Technology Laboratory, Canon Inc.)

3　重複文の同定

図1に示すような記事を統合するためには，関連記事から共通部分を同定する必要がある．本システムでは，文を単位として共通部分を同定し，これを**重複文**と呼ぶ．重複文は，同一の内容について記述してある文であり，1対1，1対多の組み合わせを想定した．

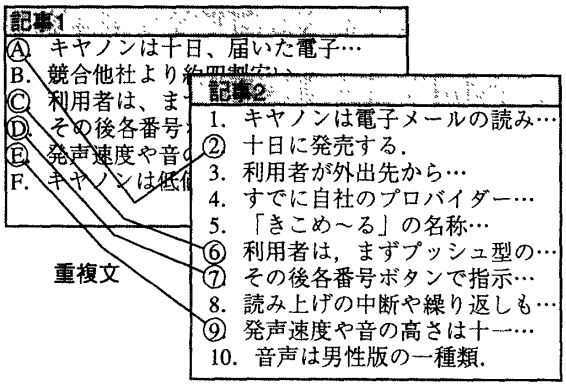

図1．関連記事の例

重複文の同定に先立ち，文章の全文に形態素解析を行ない，各形態素の出現頻度を調べる．その際，汎用的に用いられる頻度の高い形態素は除外しておく．記事1と記事2の各文の組み合わせについて，一致する形態素を洗い出し，出現頻度に応じて以下のような評価を行なう．

$$\frac{100}{(\text{記事1中の出現回数}) \times (\text{記事2中の出現回数})}$$

例えば，記事1と記事2とでそれぞれ1回だけ用いられる語があれば，この語を含む文が重複文である可能性が高い．逆に，記事1でN回，記事2でM回というように多く用いられていれば，N×Mの組み合わせが考えられるので，その分重複の可能性は低くなる．なお，複合語については，複合語全体と構成要素との両方で評価を行なう．

評価結果を形態素の数で正規化 (理想的な最大値は100) し，閾値と比較しこれを越えたものを重複文とし，達しなかったものを単独文とする．22記事 (11組の関連記事) で評価実験を行なったところ，閾値を5.0とすることで84組の重複文が検出でき，再現率96%，適合率96%となった．

表1に形態素の一致による評価結果を示す．重複文の組み合わせを灰色に着色してある．文Aについては，文2と文3とに重複しているが，これは文Aの内容が，記事2では二文に分けて述べられていることを表わしている．

表1．文対応テーブルの例

	A	B	C	D	E	F
1	1.27	0	0	0.47	0	0.8
2	28.57	0	0	0	0	0
3	12.38	0	0	0.15	0	0
4	0.28	1.15	0	0	0	0
5	0	0	0	0	0	15.00
6	0	0	32.29	0	0	0
7	0.12	0	0	24.91	0	0
8	0	0	0	0	0	0
9	0	0	0	0	35.29	0
10	0	0	0	0	0	0

4 記事の統合

本実験システムでは，情報の利用目的に応じて，AND / OR / PREFER の3タイプの記事を作成する．

4.1 ANDタイプ

関連記事の中から最も短い記事を選び，重複文だけを残したものを雛形とする．図1の例では，記事1の方が短いので記事1から単独文Bを除く．そして，重複文については，形態素レベルで他の記事で用いられていない句を切り取る．その際，構文情報を用いることで生成される文が不自然にならないようにした．例えば，以下の例文で，「届いた」が記事2で用いられていないので，「届いた」に係っている「ユーザに」も一緒に切り取る．

記事1 ユーザに届いた電子メールを読み上げる．
記事2 電子メールの本文を読み上げる．
結果 電子メールを読み上げる．

この方法で作成された文章は，複数の記事に記述された情報から成るので，情報の全体像をざっと掴みたい場合に有効である．

4.2 ORタイプ

関連記事の中で最も長い記事に，他の記事の単独文を取り込んで雛形を作成する．文を取り込む際には，他の記事で直前に重複している文の直後に挿入し，連接が不自然にならないようにする．図1の例では，記事2の方が長いので，記事2に記事1の単独文である文Bを追加する．追加する位置は，記事1で文Bの直前である文Aが文2，文3に対応するので，文3の直後とする．ただし，文Aと文Bとの間で段落が切れている場合には，文Cと対応する文6の直前に追加する．

重複文に対しては，ANDタイプの場合とは逆に，他の記事のみに用いられている句を文に取り込む．例えば，同じ例文で，体言「電子メール」に着目して，記事1の「ユーザに届いた」という連体修飾節を連結する．

記事2 電子メールの本文を読み上げる．
記事1 ユーザに届いた電子メールを読み上げる．

結果 ユーザに届いた電子メールの本文を読み上げる．

この方法で作成した文章は，関連記事の共通部分に加えて全情報の差分を取り込んでいる．従って，複数の情報源からの情報を漏らさずに読みたい場合に有効である．

4.3 PREFERタイプ

関連記事のベクトルと，ユーザが作成したフォルダのベクトルの類似度を判定し，最も距離が近い記事[1]を選択する．この記事を元に，ORタイプと同様，他の記事との差分を取り込む．

5 結果

日本経済新聞と日経産業新聞とを複数情報源として1週間(5日間)分の記事を対象として評価を行なった．関連記事の検出には括弧に括られた文字列を用い，総記事数1,683記事の全記事を対象とした．重複は，46組(99記事)を検出し適合率は74%となった．重複検出で失敗した要因として，偶然に文字列が一致し，関連のない記事を選んだ場合が多かったが，製品の発表記事とその解説といった共通部分が僅少である場合もあった．

正解した重複記事の組を対象としてANDタイプで作成された237文を評価したところ，86%が内容的に正しい文であった．誤りの内訳は，述語が欠落して文の体裁をなさない非文が3%，情報を削り過ぎや解析の失敗による不自然な文が11%であった．

参考文献

[1] 大谷他:知的検索システム Fit での類似度判定の改良，第55回情処全大，5N-6，1997.
[2] 上田他:フロー情報収集・活用のための知的検索システム Fit(1) コンセプト，第53回情処全大，2T-8，1996.
[3] 柴田他:フロー情報を対象にした情報検索システム (3)–文章圧縮–，第50回情処全大，4F-8，1995.
[4] 柴田他:複数文章の融合，情報処理学会，自然言語処理研究会 120-13，1997.

[1] ユーザの必要とする記事は，ユーザが作成したフォルダのベクトルとの距離が最も近い記事と仮定した．