自然言語処理 120－12
(1997. 7. 25)

複数文章の融合

柴田 昇吾　上田 隆也　池田 裕治

キヤノン(株) 情報メディア研究所
〒211 神奈川県川崎市幸区鹿島田 890-12　Tel: (044)549-5111
E-mail: shibata@cis.canon.co.jp

あらまし:「情報洪水」の一因として，重複する内容を持った情報を複数の情報源から受信することが考えられる．しかし，複数の情報源から幅広く有用な情報を得ることも重要である．そこで，これらの情報を整理することで，情報量を適切に制御することができると考えた．具体的には，重複のある情報を集め，重複している部分を重要であるとして取り出すＡＮＤタイプと，重複部分と差分とを取り出してまとめるＯＲタイプと，さらには，自分の興味に近い文章に他の情報をまとめるＰＲＥＦＥＲタイプの三手法を考案した．最後に，複数の新聞を対象として，新製品に関する記事を融合する実験システムについて述べる．

キーワード: 複数文章，情報検索，要約，重複情報，情報抽出

A Fusion of Overlapped Documents

Shogo SHIBATA,　Takaya UEDA　and　Yuji IKEDA

Media Technology Laboratory, Canon Inc.
890-12, Kashimada, Saiwai-ku, Kawasaki-shi, Kanagawa, 211, Japan
Phone: +81-44-549-5111
E-mail: shibata@cis.canon.co.jp

Abstract: Overlapped documents from multiple information sources contribute to the "information overflow". It is also necessary to get valuable information from these sources widely. So we have found it is able to control volume by sorting out these information. To be more specific, we propose "AND-method" which gathers overlapped information because it must be important to be written multiply, and "OR-method" which gathers overlapped items and differences, and "PREFER-method" which gathers differences to user's preferred document.
Finally we describes the experimental system which fuses multiple news articles on new products.

key words: documents fusion, combining documents, multiple documents, information retrieval, information extraction

1 はじめに

インターネットをはじめとするオンライン情報の急増により，必要な情報を必要な時に入手できるようになってきた．これに伴って，オンライン情報をさらに活用したいという要求が高まっている．例えば，インターネット白書 [2] によれば，現在のところインターネットの利用は，製品・サービス情報の収集にとどまっているものの，今後利用したいサービスとして，生活情報の収集や情報配信サービスやオンラインショッピングなどがあげられている．

一方，情報発信側の動向を見ると，マスコミなどの情報提供会社やプロバイダなどのインターネットサービス会社の専従会社の利用から，有効な情報発信手段として意識した一般企業の利用への広がりを見せている．現在，インターネットで情報を発信している企業のうち，アフターケア，販売・予約などのサービスを行なっているのは 15 %にも満たないが，導入しようと考えている企業は 40 %を越えている．

このように，今後も情報の流通量が増えていくことが予想されるが，量が多過ぎて人間の処理能力を越える**情報洪水**の問題も生じつつある．そこで，情報量を削減し，我々人間に効率よく情報を理解させる情報加工技術が求められてきている．

2 背景

情報の発信が容易になり，情報源の数が増えていくと，同一の事柄について書かれた情報も増えてくる．例えば，新聞 (日経産業新聞 (表 1 では**新聞S** と表わす)) と WWW(INTERNET Watch(同 **WWW-I**), アスキーネット (同 **WWW-A**)) と電子メールによる配信サービス (PC Watch(同 **Mail-P**)) との間には，以下のような記事の重複が見られた．

表 1. 重複記事の比率

	新聞S	Mail-P	WWW-I	WWW-A	合計 (数)	記事 総数	重複 (%)
新聞S		52	38	92	146	2,654	5.5
Mail-P	52		14	53	91	216	42.2
WWW-I	38	14		20	54	152	35.5
WWW-A	92	53	20		127	397	32.0

(注) 1997年4月の1か月間のデータによる

日経産業新聞については，産業のあらゆる分野の記事が書かれているので 5.5 %と低くなっているが，コンピュータやインターネットの分野に限って言えば，アスキーネット同様，30 %程度の数値になる．

この比率が世の中の興味を反映しており不変であると仮定すれば，新聞やインターネットを用いた情報収集・調査において，30 %から 40 %の情報が重複していることが推定できる．従って，これらの重複している情報を整理することができれば，情報量を削減することができる．

新聞や雑誌で代表されるマスコミなど情報提供会社 (以下，情報源とする) としては，様々な手段で収集したデータを元に，特色のある視点・編集方法を用いて独自の記事情報にまとめて，それを商品としている (この商品価値である視点・編集方法・意見などの情報を以下，**論旨**と呼ぶ)．

一方，情報受信側は，時間，労力，資金力の有限資源の中で，検索目的，許容コストに合わせて，複数の情報源から以下の三種類の検索戦略を使用して，情報を受け取っている．

情報同定 (exploration search)
関心が高い事柄の場合には，原因や影響などについて，あらゆる情報源から知ろうとする．例えば，新聞記事の切り抜きなどを集める方法がこれに該当する．ここでは，情報源ごとの論旨の解釈にも時間と労力をかける．

情報鳥瞰 (bird's eye search)
時間が無かったり関心のない事柄であれば事実だけがわかればよい．新聞のリード部分だけを読んで以降は読み飛ばしたり，写真だけを見たりする方法がこれに該当する．ここでは，情報源ごとの論旨は不要である．

情報従覧 (guided search)
ある情報源の論旨に共感できれば，以降，情報源からの情報を信頼して提供を受ける．特定の新聞や書籍などを読んだりテレビを見るという方法がこれに該当する．

情報同定, 情報鳥瞰による情報検索には，一般に受信側での時間・労力が費やされるため，情報従覧による情報検索が従来の手法であったといえる．しかし，昨今の廉価な情報受信手段・経路の増加とともに，情報の質の格差も同時に生じていることから，情報従覧のみによる効率の良い情報収集の保証が得られなくなった．

以上のことから，情報の受け手であるユーザが，重複している情報をまとめて，いろいろな検索戦略に適した情報に加工する必要がある．

3 文章融合のコンセプト

情報は，テキストや映像や音などさまざまなメディアで構成されており，情報を統合する際にはメディ

アを問わない方式が理想である．例えば，情報鳥瞰であれば，テキストで説明するよりも映像の方が効果的な場合もあるし，情報同定であれば，映像では見過ごしがちな部分をテキストで補う方が好ましい場合もある．

本稿では，情報統合の第一歩として，テキスト情報(以下，**文章**と表わす) を扱う．文章は，あらゆるメディアの中でも記述性が高く，マルチメディア化が進んでも重要性は変わらない．従って，文章の統合を可能とすることで，これを手掛かりとして多様な情報の統合へと展開できると考えられる．このように，複数の文章を混ぜ合わせて，あたかも一つの文章に見せる技術を**文章融合**と呼ぶ．

文章融合は次の二つのフェーズから成る．

- **関連文章の検出 (フェーズ 1)**
  文章融合では同一の事柄について記述した**共通部分**を核として文章を融合する．フェーズ 1 では，共通部分を有する文章の候補として，何らかの関連がある関連文章を選択する．

- **融合文章の生成 (フェーズ 2)**
  検出した関連文章を一つの文章に融合するフェーズである．フェーズ 1 で検出した文章を，それぞれの共通部分と差分とに分け，整合性を取りながら文章へ再構成する．その際，利用目的に応じて情報量をコントロールすることが可能である．即ち，

  – すべての情報を漏らさずにまとめる

  – 情報を取捨選択してまとめる

  という方針の違いに応じて，共通部分と差分とを組み合わせたり共通部分だけを使用したりする．

  例えば，情報同定には前者の方針で融合文章を生成し，情報鳥瞰には後者の方針で生成する．

フェーズ 1 に関しては，記事の対応付けのために文章中の一致文字列を用いる方法 [3] などが知られている．また，**情報可視化**は，情報を平面や立体に配置することでそれぞれの情報の位置付けや情報間の関係などを把握しやすくする研究分野であるが，ここでも関連文章を検出するさまざまな手法が提案されている．

しかし，これらの研究は，関連文章の検出自体が目的であり，関連させた文章の内容を提示する方法については言及していない．従って，外面的な情報，即ち情報の分布や量などの調査を目的とした情報鳥瞰には有効な情報となり得るが，論旨を解釈する必要がある情報同定には有効ではない．

文章融合では，フェーズ 1 の結果をより有効に活用するために，フェーズ 2 を用意している．フェーズ 2 に関しては，文章にまとめるという手法以外の手段で関連情報を提示することも考えられる．

情報間リンクが付いたハイパーテキストをたどる方法は，必要に応じてリンクをたどって詳細情報へ飛んだり，概略情報へ戻ったりできるので情報鳥瞰と情報同定を切替えながら検索する場合に適している．しかし，リンクはあらかじめ設定されているので情報同定には不充分である場合が多い．また，深いリンクをたどると，自分がどこにいるかを見失いやすいという問題がある．

NetNews は，グループごとに関連する情報が集まるので，これに対象した研究でも関連情報を効率的に提示する方法が検討されている．例えば，NetNews を受け手側で加工する方法として，井佐原らによる「知的ニュースリーダ」[4] や佐藤らの「電子ニュースのダイジェスト自動生成」[5] がある．しかし，これらは，記事の要／不要を判断するのが主目的であり，フェーズ 1 の研究同様，必要だと判断した記事の内容を効率よく提示する方法については言及していない．

文章融合では，複数文章の提示手段として文章を用いる．この手法は，以下のような長所を持つ．

- **冗長度が少ない**
  差が少ない関連文章を保存するよりも共通部分を整理して保存した方が冗長性が少ない．また，再利用の際にひとつにまとまっている方が効率的である．

- **理解が容易**
  共通部分と差分とを分けることにより情報が断片的になる恐れがあるが，文章融合で一文章にまとめることにより前後の文脈が利用できるので理解しやすくなる．

- **閲覧が容易**
  文章間を移動することなく全体を見ることができる．また，ひとつの文章なので，例えば，音声による「読み上げ」で内容を把握することも可能である．

複数文章をひとつの文章にまとめる研究としては，要約を目的とした方法 [6]，[7] が知られている．文章融合は，情報鳥瞰には要約的な側面も持ち，情報同定には情報の詳細化の面を持つなど，目的に応じて適切な量の文章を作成できる．

4 実験システム

文章融合の第一歩として，我々が開発してきた**知的検索システム Fit** [1] に文章融合機能を埋め込んだ．Fit とは，新聞記事などのフロー情報をユーザ

の視点に応じてフォルダに整理し (収集フェーズ), フォルダ単位の検索機能や視点に応じた情報提示機能で情報の活用を支援する (活用フェーズ) システムである.

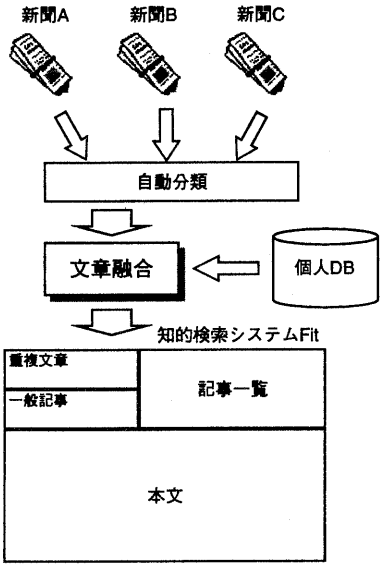

図1. 文章融合機能付き Fit の概要

このシステムでは, ベクトル空間モデルに基づいてフォルダや文章を表現し, それらの間の類似度を判定することで自動分類を行なっている. 従って, 文章を適切なフォルダに保存していれば分類精度が向上するが, 関連文章が異なったフォルダに保存されるといった不適切な保存を行なうと精度が低下する.

文章融合は, 収集フェーズで関連文章をまとめて一つの記事として提示する機能を実現した. この機能により, 関連文章が別々のフォルダに格納される危険性を回避することができた.

さらに, 複数の情報源から同じ内容に関する記事が発信されているということは重要度が高いと考えられるので, 関連文章を他の記事と区別して先に読むことができるようにした.

以下, 本実験システムの実装について述べる. なお, 対象とする情報としては, 新製品に関する記事を用いた.

4.1 関連文章候補の絞り込み

1日に300〜400記事がある新聞記事から共通部分を同定するには, 膨大な計算量を必要とする. そこで, 本システムでは, あらかじめ関連文章の候補記事を絞り込む.

絞り込みは, フロー情報である新着記事間で行なう. また, 関連文章でも情報源ごとに発信時間のずれがあるため, 新着記事と過去にユーザが保存したストック情報との間でも行なう.

絞り込みの手段として, 記事から「製品名」「メーカ名」「発売日」「特徴」などの情報を得る**情報抽出**を用いる. Fit では, 視点に応じた情報提示機能の一つとして, これらの情報を一覧表として提示する機能を実装している. 実装した情報抽出機能は, 記事の第1パラグラフの文を対象として,「は」「を」などの格情報や日付などの新聞特有の表現やパタンなどに着目して目的の情報を抽出する.

一般に情報抽出結果が同じであっても共通部分を持つとは限らないが, 新製品に関する記事を対象とするので, 記事間で情報抽出結果を比較し同一である記事を関連文章と仮定する.

4.2 重複文の同定

文章融合では関連文章から共通部分を同定する必要がある. 本実験システムでは, 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, または, 1対多の組み合わせを想定している.

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. 図2に, 想定している関連文章とそれぞれの文章中の出現頻度が少ない語の例を示す. 以後, この関連文章 (文章1と文章2) を用いて説明する.

次に, 文章1と文章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.

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

例えば, 複合語「プッシュ型」は, 文章1と文章2とに1回だけ用いられているので,「プッシュ型」という語を含む文が重複文である可能性が高い. 逆に, 文章1でN回, 文章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, 形態素の部分的な一致を許すことにより, 表記の揺れを吸収している.

A. キヤノンは十日、届いた電子…
B. 競合他社より約四割安い…
C. 利用者は、まずプッシュ型の…
D. その後各番号ボタンで…
E. 発声速度や音の高さは…
F. キヤノンは低価格を売り物に…

出現回数 … 1回
競合他社，プッシュ型，…
出現回数 … 2回
キヤノン，電子メール，…
：

図 2-a. 関連文章の例 (文章 1)

1. キヤノンは電子メールの読み…
2. 十日に発売する．
3. 利用者が外出先から…
4. すでに自社のプロバイダー…
5. 「きこめ～る」の名称…
6. 利用者は，まずプッシュ型の…
7. その後各番号ボタンで指示…
8. 読み上げの中断や繰り返しも…
9. 発声速度や音の高さは十一…
10. 音声は男性版の一種類．

出現回数 … 1回
キヤノン，プッシュ型，…
出現回数 … 2回
プロバイダー，電子メール，…
：

図 2-b. 関連文章の例 (文章 2)

表 2. 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

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 文章 (11 組の関連文章) 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 2 に形態素の一致による評価結果を示す．重複している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，文章 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．

4.3 融合文章の生成

本実験システムでは，2 節で述べた 3 つの検索戦略に対応する AND / OR / PREFER の 3 タイプの文章を作成する．以下でそれぞれの文章の作成方法を説明する．

4.3.1 AND タイプ

関連文章の中から最も短い文章を選び，重複文だけを残したものを雛形とする．図 2 の例では，文章 1 の方が短いので文章 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文章で用いられていないものを切り取る．本システムでは，構文情報を用いており，係り受けを考慮することで生成される文が不自然にならないようにする．例えば，以下の例文で,「届いた」が用いられていないとすると，これを切り取る.「ユーザに」は「届いた」に係っているので一緒に切り取ることになる．

文章 1 ユーザに届いた電子メールを読み上げる．
文章 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，関連文章の共通部分を取ることにより，主に論旨にあたる個別情報を排除できるので情報鳥瞰による情報検索に適している．

4.3.2 OR タイプ

AND タイプとは対照的に，関連文章の中から最も長い文章に，他の文章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，他の文章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 2 の例では，文章 2 の方が長いので，文章 2 に文章 1 の単独文である文 B を追加する．追加する位置は，文章 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，他の文章に用いられている情報を文に取り込む．例えば，同じ例文で，体言「電子メール」に着目して，文章 1 の「ユーザに届いた」という連体修飾節を連結する．

文章 2 電子メールの本文を読み上げる．
文章 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，関連文章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，複数の情報源の論旨を漏らさずに得ることができるので，情報同定による情報検索に適している．

4.3.3 PREFER タイプ

関連文章のベクトルと，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，最も距離が近い文章[1]を元に，他の文章の差分を句や節を単位として取り込む．図2の例では，文章1を選択した場合，重複文である文A，文C～文Fについて，ORタイプの融合と同様，文章2のみで用いられている修飾句や節を取り込む．単独文Bについては加工しない．

関連文章から自分の興味に近い文章を一つ選択することは情報従覧にあたる．PREFER タイプの文章融合は，他の文章の情報を付加することで，情報従覧の手軽さで情報同定に近い情報検索ができるようにしたものである．

4.4 視点によるタイプの自動切替

本実験システムでは，AND / OR / PREFER の3タイプの文章融合を視点別に設定できる機能を追加した．これにより，興味のある視点については情報同定のための詳しい情報，興味のない視点については情報鳥瞰のための概要的な情報が提供できるようになった．

4.5 実験システムの評価結果

本実験システムで日本経済新聞と日経産業新聞の22記事(11組の関連文章)を処理し，文を単位とした評価を行なった．人間が関連文章を読み較べて作成したものと文単位での有無を調べ，以下の再現率と適合率を得た．

なお，評価方法としては，この他に文の正しさ・読みやすさや，文を挿入する位置の正確さなどがあるが，これらについての評価は今後の課題とする．

表3. 文レベルの再現率，適合率

タイプ	再現率	適合率
AND タイプ	89%	96%
OR タイプ	98%	92%

[1]ユーザの必要とする文章は，ユーザが集めたフォルダのベクトルとの距離が近いと仮定した．

5 おわりに

本研究では，多数の文章から関連文章を検出し，なかでも共通部分を持つ関連文章から有用な情報を取り出す**文章融合**のコンセプトについて述べた．複数の文章を扱う技術は，今後の情報洪水を緩和するために不可欠な技術であり，**文章融合**のコンセプトは情報検索のさまざまな分野での応用が期待できる．

今後は，提案したコンセプトをもとに，実験システムの改良を行なっていく．具体的には，インターネットなどの情報が扱えるようなロバスト性の向上，関連文章の数が増大した場合の多数決の導入などを行なっていきたい．

参考文献

[1] 上田隆也，大谷紀子，伊藤史朗，柴田昇吾，池田裕治: 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，情報処理学会第53回全国大会,2T-8,1996.

[2] 日本インターネット協会:インターネット白書'97，インプレス,1997.

[3] 角田 達彦，大石 巧，渡辺 靖彦，長尾 真: キャプションと記事テキストの文字列照合による報道番組と新聞記事との対応づけの自動化，情報処理学会論文誌，Vol.38, No.6,pp 1149-1160, 1997.

[4] 井佐原 均，小作 浩美，内元 清貴:討論型ニュースグループを対象とする知的ニュースリーダの開発，情報処理学会，自然言語処理研究会，119-3，1997.

[5] 佐藤 理史，佐藤 円:ネットニュースグループ fj.wanted のダイジェスト自動生成，言語処理学会，自然言語処理，Vol.3 No.2，1996.

[6] K.McKeown, D.R.Radev: Generating Summaries of Multiple News Articles, SIGIR'95,pp 74-82,1995.

[7] 船坂 貴浩，山本 和英，増山 繁:冗長度削減による関連新聞記事の要約，情報処理学会，自然言語処理研究会，114-7，1996.