社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－131 (20)
2003／7／16

話題構造に基づくコンテンツ結合演算とその応用

馬 強† 田中克己†

本稿では，Webページやデータストリームの内容構造を表す話題構造モデルを提案し，それに基づくストリームデータとWebのコンテンツ結合演算を提案する．話題構造は，話題のタイトルを表すキーワードと本体を表すキーワードの集合のペアで定義され，DAG（Directed Acyclic Graph）を用いて表現可能である．複数の話題構造の結合（⋈）が，新たな話題構造を生成し，DAGの和で求められる．我々は，このような話題構造ベースの結合を用いて，複数情報ソースのコンテンツ統合問題をモデル化し，その解析を図る．そのうち，有向木に限定した，話題構造およびそれの結合は，コンテンツ結合演算をより詳細な情報を提供できる垂直結合とより幅広い情報を提供する水平結合に分類でき，情報統合における補完・増殖の機能を有することを示す．さらに，話題構造による結合を用いて，我々が提案しているWebと放送の動的統合システムWebTelopを説明する．

Topic-structure-based Join Operation and Its Applications

QIANG MA † and KATSUMI TANAKA †

This paper proposes the notion of a topic structure for a given web page or given data stream, which intuitively is a pair of subject and content terms. The subject term denotes the most dominant term of the web page or data stream while the content term means a term whose co-occurrence ratio with a subject term is very high. We represent such topic structure as a DAG (Directed Acyclic Graph) and define several types of topic-structure-based join operations of data streams and web pages and discuss how to use these joins in information integration. Based on the proposed joins of topic structures, we explored an application system *WebTelop*, which integrates a TV program with its metadata and related web pages and discuss issues in implementing this system.

1. はじめに

多種多様な情報ソースの統合は，より詳しい・より幅広い情報がほしいといったユーザの多様な情報要求に応えるためにはますます重要となる．デジタルテレビジョンは，放送とコンピュータの技術を融合した新しいメディアであり，新しいテレビの視聴方式をもたらした[1),2)]．同時に，ブロードバンドなど高速・常時接続環境の普及に伴って，オンラインで映像などリッチコンテンツをインターネットを通して楽しむことが可能となりつつある．つまり，放送とWebの融合のインフラストラクチャは整えってきている．放送番組は，品質がよく，直感性のあるコンテンツであるが，放送時間などの制約で，詳細かつ広範囲的に情報を伝えない場合がある．一方，Webでは，多種多様な情報が公開されているが，その品質はさまざまであり，わかりやすさや直感性では放送コンテンツに劣る場合が多い．このような異なる性質をもったメディアを統合して，相互補完をさせながらユーザに情報呈示を行うのは，多彩な情報を提供でき，ユーザのより多様な情報要求に応えることが可能である．

本稿では，話題構造モデルに基づくコンテンツの結合（⋈）演算を提案し，それを用いて情報統合のモデル化を試みる．話題構造は，Webページや映像シーンの内容を表すものであり，タイトルをあらわすタイトル要素と本体を表す内容要素のペアから構成される．タイトル要素は，キーワード（*subject-term*）の集合である．一方，内容要素は，キーワード（*content-term*）と別の話題から構成される集合である．話題構造がタイトル要素を入口とする連結のDAG（話題グラフ）で表現される．話題構造の結合が，話題グラフの和で表現され，新たな話題構造を生成する．

一つのWebページが複数の話題を含み，話題構造の集合である．一方，放送コンテンツなどのストリーム型データは，話題構造のストリームとして表現することが可能である．従って，Webと映像の情報統合は，話題構造の集合とストリームの結合によって表現でき

† 京都大学大学院情報学研究科社会情報学専攻
Graduate School of Imformatics, Kyoto University

る．本稿では，これらの結合のいくつかの性質について述べる．

話題グラフの特殊ケースとして，有向話題木について考察を行う．我々は，有向話題木による結合を水平結合と垂直結合に分類し，情報補完・増殖における話題結合の役割の解析を試みる．そのうち，水平結合（*horizontal-join*）は，共通の subject-term を有する二つの話題の結合を表し，情報のカバーする範囲（幅）を広げることができる．垂直結合（*vertical-join*）は，共通の content-term を有する二つの話題の結合を表し，情報の詳細（深さ）を増やすことが可能である．これをベースに，結合情報の鮮度を追求し，結合情報の重複を避けるために，動的に結合のタイプを変えて繰り返しのある話題に対して異なる内容結合を行う時間依存型話題結合機構を提案する．さらに，水平結合と垂直結合を用いて，我々が提案している放送と Web コンテンツの動的統合システム *WebTelop*[3] について解析を行う．WebTelop は，有向木に限定した話題構造およびそれによる結合をベースにしたシステムであり，リアルタイムに抽出される放送番組の話題構造を用いて，Web からその映像コンテンツを補完・増殖できるページを検索してきて，連動してユーザに呈示する．我々は，このシステムを用いて，情報統合における話題結合の役割について述べる．

以下，2 節では，関連研究について述べる．3 節では，話題構造モデルとそのグラフ表現，および話題構造の抽出手法について述べる．4 節では，話題構造ベースの結合演算について述べる．5 節では，有向木に限定した話題結合を，応用システム WebTelop を用いて説明し，話題結合の情報統合における役割の一部を説明する．6 節では，まとめと今後の課題について述べる．

2. 関連研究

松倉らはウィンドウ関数を用いた話題構造抽出手法を提案している[4]．彼らのアプローチは複数のサブトピックを有する Web ページの話題抽出には有効であるが，コスト（処理時間）は非常に高い．小山らは，検索エンジンの構造的な検索オプションを利用した話題構造の抽出手法を提案している[5]．しかしながら，これらの提案手法では，“title”タグの利用，つまり Web ページの書き方に大きく依存している．

トピックマップ[6] は情報リソースを管理，検索と閲覧のための新しい ISO 基準である．名前，リソース，関係はトピックマップの３つの基本概念であり，トピック間の関係を明確にすることが目的である．これに対して，本論文で提案する話題構造は，コンテンツの内容を表すものである．また，トピックマップでは，トピックは人手による定義されたものが多いが，本論文では，話題構造の自動抽出手法を提案している．

Shasha らは，キーワードグラフやキーワード木による情報検索のアプリケーションとアルゴリズムを紹介している[7]．紹介された手法では，XML データに対して，木やグラフ同士のマッチングまたは近似マッチングによる情報検索を行うものが多い．一方，我々は，話題構造という概念を用いて，情報の内容構造を表す．また，それをベースに，情報統合のための結合演算を提案している．

Bhowmick らは，異なる Web テーブルからの情報検索のため，Web ページのスキーマベースの結合を提案している[8]．彼らの結合演算は，WDM（Web Data Model）[9] をベースとする．本稿では，話題構造モデルをベースに，情報補完・増殖のためのインスタンスベースの結合を提案している点が異なる．

3. 話題構造モデル

3.1 話題構造モデル

松倉ら[4] は，話題構造の基本概念を提案した．話題構造は，話題のタイトルを表すタイトル要素と本体の内容を表す内容要素のペアである☆．内容要素がタイトル要素を記述するという関係がある．ある話題 topic は次のように表現される．

$$
\begin{aligned}
topic &:= (S, C) \\
S &:= (subject-term)^{+} \\
C &:= (content-term|topic)^{+} \\
subject-term &:= keyword \\
content-term &:= keyword
\end{aligned} \tag{1}
$$

ただし，S と C は $\emptyset$ でない，かつ，互いに素である．ここでは，互いに素とは，共通キーワードのないことを意味する．言い換えれば，あるキーワードは，話題構造のなかに高々一回出現する．

Subject-term は，ある話題のタイトルを表す役割を有するキーワードであり，その話題での中心的なキーワードである．一方，content-term は，その話題にある，subject-term との共起関係の強いキーワードであり，subject-term を記述する役割がある．

3.2 話題構造のグラフ表現

一般に，話題構造を二つ以上のノードを持つ弱連結の DAG（Directed Acyclic Graph）を用いて表現できる．つまり，ある話題 t は，キーワードを表す頂点の

☆ 松倉らのモデル[4] では，タイトル要素と内容要素は，それぞれキーワードの集合である．

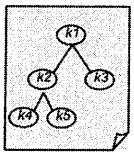


図 1　話題グラフの例

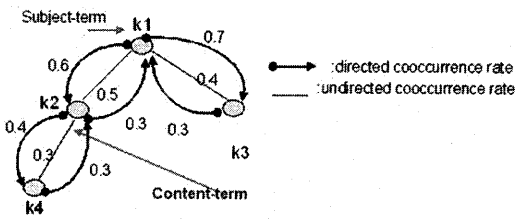


図 2　話題構造（subject-term と content-term）の抽出例

集合 $V(|\,V\,|\geq 2)$ と，キーワード間の subject-content 関係を表すアークの集合 A $(A\subseteq V\times V, |\,A\,|\geq 1)$ を用いて表現される．このような弱連結 DAG を話題グラフと呼ぶ．

$$G(t)=DAG(V,A) \tag{2}$$

図 1 は話題グラフの例を示している．例で示されている話題グラフでは，$V=\{k_1,k_2,k_3,k_4,k_5\}, A=\{(k_1,k_2),(k_1,k_3),(k_2,k_4),(k_2,k_5)\}$ である．

ある話題グラフを与えた時，その話題グラフの入口（source）は，その話題の subject-term となる．図 1 では，k_1 が subject-term となる．その話題構造は，タイトル要素が $\{k_1\}$ であり，内容要素が $\{(k_2,\{k_4,k_5\}),k_3\}$ である．

実際には，アーク $e=(u,v)\in A$ は，一つの話題と見なし，始点 u と終点 v が，それぞれ subject-term と content-term であると考えることができる．図 1 では，(k_1,k_2) を一つの話題と見なした場合，k_1 は subject-term であり，k_2 は content-term である．一方，(k_2,k_4) では，k_2 が subject-term であり，k_4 が content-term である．つまり，あるキーワード（例では，k_2）が subject-term と content-term の両方であることがありえる．

3.3　話題構造の抽出

本稿では，以下のような共起関係を定義している．

- 単語 w_i,w_j の有向共起関係：

$$\overrightarrow{cooc}(w_iw_j)=df(\{w_i,w_j\})/df(\{w_i\}) \tag{3}$$

ただし，$df(\{w_i,w_j\})$ は単語 w_i と w_j を同時に含む話題の数であり，$df(\{w_i\})$ は w_i のみを含む話題の数である．

- 単語 w_i,w_j の無向共起関係 $cooc(w_i,w_j)$：

$$cooc(w_i,w_j)=\frac{df(\{w_i,w_j\})}{df(\{w_i\})+df(\{w_j\})-df(\{w_i,w_j\})} \tag{4}$$

ただし，$df(\{w_i,w_j\})$ は単語 w_i と w_j を同時に含む話題の数であり，$df(\{w_i\})$ は w_i のみを含む話題の数である．

話題の subject-term を語の共起関係と出現頻度に基づいて抽出することが可能である．

- 出現頻度：話題において出現頻度の高いキーワードは，その話題の subject-term である可能性が高い．
- 共起関係：話題におけるその他のキーワードとの有向共起関係の強いキーワードは，その話題において中心的な役割があり，subject-term である可能性が高い．

図 2 では，共起関係による subject-term と content-term の抽出例を示している．図では，ラベルがキーワード間の共起度を表す．

一方，content-term は，subject-term との無向共起関係に基づいて求められる．すなわち，話題における subject-term と無向共起関係の強いキーワードは，そのトピックの content-term である可能性が高い．語の共起関係は，あらかじめ用意されているトピックのコーパスから作成されていた共起関係の辞書を用いて調べる．話題構造の抽出手法の詳細が，文献 [3] で述べられている．

4.　話題構造に基づくコンテンツ結合

4.1　話題構造の結合

定義 1 (話題結合（$\bowtie$))　二つの話題（t と t'）の結合 $\bowtie$ は，二つの話題グラフの和を意味する．

$$t\bowtie t'=\begin{cases} G(t)\cup G(t'), & G(t)\cup G(t')\text{ は弱連結の }DAG\text{ である} \\ \emptyset, & \text{その他} \end{cases} \tag{5}$$

定義から，$t_1\bowtie t_1=t_1$ であることが明らかであり，特に，このような結合を自己結合（self-join）と言う．話題構造 t_1 と t_2 の結合が $\emptyset$ でない場合は，t_1 と t_2 は結合可能であると言う．

明らかに，$t_1\bowtie t_2=t_2\bowtie t_1$ が成り立つ．つまり，話題結合は可換である．

話題構造 $t_1=(\{a,b\},\{(a,b)\})$，$t_2=(\{c,d\},\{(c,d)\})$，$t_3=(\{b,c\},\{(b,c)\})$ に対して，$t_1\bowtie t_2=\emptyset$ であるので，$(t_1\bowtie t_2)\bowtie t_3\neq t_1\bowtie(t_2\bowtie t_3)$ である．つまり，任意の話題構造の結合が結合律を満たさない．ただし，任意の結合可能な話題に対しては，結合律が成り立つ．つまり，話題構造 t_1,t_2,t_3 に対して，

$t_1 \bowtie t_2 \neq \emptyset, t_2 \bowtie t_3 \neq \emptyset, t_1 \bowtie t_3 \neq \emptyset$ であれば，$(t_1 \bowtie t_2) \bowtie t_3 = t_1 \bowtie (t_2 \bowtie t_3)$ が成立つ．

4.2 話題結合によるコンテンツ統合

一般に，Web ページには，複数の話題が含まれている．故に，Web ページ T_p は話題構造の集合で表すことができる：$T_p = \{t_1, t_2, ..., t_n\}$．ただし，$t_i$ $(1 \leq i \leq n)$ は Web ページに含まれている話題 i の構造を表す．一方，映像などストリームデータは，話題のストリーム T_s と見なして，次のように表される：$T_s = s_1 s_2 ... s_m$．ただし，s_j $(1 \leq j \leq m)$ はストリームの j 番目の話題構造を表す．

内容ベースのデータストリームと Web ページの情報統合は，話題ストリームと話題集合の結合によって表現できる．

(a) 話題集合の結合

二つの話題集合 T_p と T_q が与えられたとする．T_p と T_q の結合は，次のように定義される．

$$T_p \bowtie T_q = \{x \bowtie y | x \in T_p, y \in T_q\} \quad (6)$$

二つの話題集合の結合は，それに対応する二つの Web ページの統合であると見なすことができる．この時，$x \in T_p, y \in T_q, x \bowtie y \neq \emptyset$ を満たす x, y が存在すれば，T_p と T_q が統合可能であると言う．統合可能な二つの Web ページは，共通のキーワード（x, y の共通ノード）またはサブトピック（x, y の共通アーク）について述べている部分がある．つまり，ある Web ページを用いて別のページの内容を補完・増殖することが可能である．

(b) 話題ストリームの結合

二つの話題ストリーム $T_s = s_1 ... s_n$ と $T'_s = s'_1 ... s'_m$ の結合は，次のように定義される．

$$\begin{aligned} T_s \bowtie T'_s &= (s_1 \bowtie T'_s)(s_2 \bowtie T'_s)...(s_n \bowtie T'_s) \\ &= (((s_1 \bowtie s'_1)(s_1 \bowtie s'_2)...(s_1 \bowtie s'_m)) \\ &\quad ((s_2 \bowtie s'_1)(s_2 \bowtie s'_2)...(s_2 \bowtie s'_m)) \\ &\quad ... \\ &\quad ((s_n \bowtie s'_1)...(s_n \bowtie s'_m))) \end{aligned} \quad (7)$$

話題ストリームの結合は，放送コンテンツなどデータストリームの統合・連動を表すことができる．ある $s_i \in T_s$ に対して，$s_i \bowtie s'_j \neq \emptyset$ を満たす s'_j が T'_s にあれば，s_i が T' と統合可能であると言う．特に，任意の $s_i \in T_s$ に対して，s_i と T' が統合可能であれば，データストリーム T_s と T'_s が統合可能であると言う．

(c) 話題のストリームと集合の結合

Web と映像の相互補完・増殖は話題集合と話題ストリームの結合で実現できる．

話題ストリーム $T = (s_1 s_2 ... s_n)$ と，話題集合 $T' = \{s'_1, s'_2, ..., s'_m\}$ が与えられたとする．$T \bowtie T'$ は次のように定義される．

$$\begin{aligned} T \bowtie T' &= (s_1 \bowtie T')(s_2 \bowtie T')...(s_n \bowtie T') \\ &= (\{s_1 \bowtie s'_1, s_1 \bowtie s'_2, ..., s_1 \bowtie s'_m\} \\ &\quad \{s_2 \bowtie s'_1, s_2 \bowtie s'_2, ..., s_2 \bowtie s'_m\} \\ &\quad ... \\ &\quad \{s_n \bowtie s'_1, s_n \bowtie s'_2, ..., s_n \bowtie s'_m\}) \end{aligned} \quad (8)$$

一方，$T' \bowtie T$ は次のように計算される．

$$\begin{aligned} T' \bowtie T &= \{s'_1 \bowtie T, s'_2 \bowtie T, ..., s'_m \bowtie T\} \\ &= \{((s'_1 \bowtie s_1)(s'_1 \bowtie s_2)...(s'_1 \bowtie s_n)), \\ &\quad ((s'_2 \bowtie s_1)(s'_2 \bowtie s_2)...(s'_2 \bowtie s_n)), \\ &\quad ... \\ &\quad ((s'_m \bowtie s_1)...(s'_m \bowtie s_n))\} \end{aligned} \quad (9)$$

前者の結合結果は，話題集合のストリームである．映像の各シーン（データ項目）に対して，Web ページを用いて増殖・補完を行うことを意味する．一方，後者の結合結果は，話題ストリームの集合である．これは，ある Web ページの個々の話題に対して，結合可能なシーン（ストリームのデータ項目）を用いて補完・増殖を行うことを意味する．

4.3 話題集合の簡約

Web ページ（データストリーム）にある話題構造は，そのページ（データストリーム）の論理構造（段落，映像のシーンなど）に対応している．つまり，その話題構造集合（ストリーム）は，結合可能な異なる話題を含む可能性がある．例えば，ある Web ページの話題構造集合が $\{t_1, t_2, t_3\}$ であり，$t_1 = (\{a, b\}, \{(a, b)\})$，$t_2 = (\{c, d\}, \{(c, d)\})$，$t_3 = (\{b, e\}, \{(b, e)\})$ であるとする．t_1，t_2，t_3 がそれぞれ，ページの第 1 段落，第 2 段落と第 3 段落に対応する．この場合，t_1 と t_3 は，ページの構造上では別々であるが，話題構造による結合が可能である．つまり，そのページの話題集合が結合可能な話題構造を含む．このような話題集合は，簡約可能であると言う．

話題集合 T の簡約（$\times$）は，つぎのように定義される．

$$T \times T = G(t_1) \cup G(t_2) \cup ... \cup G(t_n) \quad (10)$$

ただし，$G(t_1) \cup G(t_2) \cup ... \cup G(t_n)$ が DAG であり，結合可能な異なる 2 つの要素を含まない．t_i $(1 \leq i \leq n)$ は，T に含まれる話題構造を表す．

話題集合と話題のストリームの結合では，結合結果に話題集合が含まれる場合があるので，アプリケーションによって簡約する必要のある場合がある．

話題構造による結合が結合律を満たさないため，一般に，話題集合 T と T' に対して，

$$(T \times T) \bowtie (T' \times T') \neq (T \bowtie T') \times (T \bowtie T')$$

である．つまり，分配律が成立たない．

例えば，$T = \{(\{a,b\},\{(a,b)\}),(\{b,c\},\{(b,c)\})\}$ と $T' = \{(\{c,d\},\{(c,d)\}),(\{d,e\},\{(d,e)\})\}$ の時，

$$T \bowtie T' = \{(\{b,c,d\},\{(b,c),(c,d)\})\}$$

である．一方，$T \times T = \{(\{a,b,c\},\{(a,b),(b,c)\})\}$ と $T' \times T' = \{(\{c,d,e\},\{(c,d),(d,e)\})\}$ であるため，

$$(T \times T) \bowtie (T' \times T') = \{(\{a,b,c,d,e\},\{(a,b),(b,c),(c,d),(d,e)\})\}$$

である．

目的によって，結合前に話題集合の簡約を行うアプリケーションと，結合結果を簡約するアプリケーションがそれぞれ存在する．例えば，Web ページ同士の結合は，結合前にそれぞれのページの話題集合を簡約することが考えられる．一方，話題構造のストリームの場合，すべての話題構造にアクセス不可能なので，結合結果を簡約する方法が考えられる．これらのシステムは結果の異なる場合が多く，等価性がない．

実際には，任意の話題構造集合（その集合は DAG でもある）T に対して，T が $T \times T$ を包摂するため，$T \bowtie T'$ が $(T \times T) \bowtie (T' \times T')$ を包摂する．従って，一般的には，情報の増殖・補完という意味では，結合前に話題集合の簡約がより有効であると考える．今後，二者の関係の更なる解析を行う予定である．

5. 有向話題木の結合と応用システム WebTelop

本節では，話題グラフの特例として有向木を用いて，話題結合の情報統合における役割を説明し，その応用システムとして，放送と Web の動的統合システム WebTelop について述べる．図 3 は WebTelop のコンセプトを示している．デジタル放送で，本放送と同時に，データ放送では，番組に関するメタデータを配信しているとする．このようなメタデータストリームを利用して，リアルタイムに放送コンテンツの話題を抽出して，話題構造に基づいて番組の補完情報として Web ページを検索する．検索結果を元々の番組コンテンツと連動させて，放送コンテンツの増殖と補完を行いながらユーザに呈示を行う．また，検索された Web ページの重要な部分をガイドする仮想キャラクターを導入している．従来の検索手法では，コンテンツ間の類似に基づいて関連情報を獲得する場合が多いので，情報が冗長となる場合がある．これに対して，WebTelop が，話題構造の結合をベースに，より詳しい情報を提供する D（deepening）モードと，より幅広い情報をユーザに提供する B（broadening）モードを有する．

WebTelop は，話題ストリームと話題集合の結合の応用システムである．字幕データストリームから抽出された番組の話題構造を $T_s = s_1 s_2 ... s_n, n \geq 1$ とする．ただし，s_i はシーン i☆の話題構造を表す．一方，Web ページの話題構造 T_p は $\{tp_1, tp_2, .., tp_m\}$ とする．tp_j は，ページ p_j の話題構造を表す．1 ページに複数の話題あるので，tp_j は話題の集合である．

WebTelop は，話題構造による結合を用いて次のように表現される．

$$T_s \bowtie T_p \tag{11}$$

実際には，WebTelop では，話題結合が，結合可能な Web ページの検索と連動呈示として実装されている．

5.1 有向話題木の結合

WebTelop では，話題構造を，有向木で表現できるものに限定している．その有向木のルートが話題構造の subject-term に対応する．このような話題グラフを有向話題木と呼ぶ．それに対応する話題構造 topic が次のように限定される．なお，特別の説明のない限り，本節では，話題構造をグラフ表現ではなく，subject-term と内容要素のペアで表現する．

$$\begin{aligned} topic &:= (s, C) \\ s &:= subject - term \\ C &:= (content - term | topic)^+ \\ subject - term &:= keyword \\ content - term &:= keyword \end{aligned} \tag{12}$$

このように限定した話題構造の結合は，下記で定義される水平結合と垂直結合に分類することができる．二つの有向話題木が結合可能（結合結果は $\emptyset$ でない）の必要条件は，いずれかの有向話題木のルート要素は共通ノードであるのである．そのため，共通ノードの位置によって，有向話題木の結合を次の 2 種類の基本結合に分類できる：水平結合（horizontal-join, $\bowtie^h$）と垂直結合（vertical-join, $\bowtie^v$）．WebTelop のより詳細の情報を提供する D モードは，垂直結合の一実装である．一方，B モードは，水平結合と対応する．

以下，これらの結合演算について述べる．説明のない限り，二つの結合可能な話題構造 t_1, t_2 が与えられたとし，$t_1 = (s_1, C_1), t_2 = (s_2, C_2)$ とする．図 4(a) と図 4(b) では，それぞれ垂直結合と水平結合の例を示している．

（a）水平結合 ($\bowtie^h$)

$s_1 = s_2 = s$，つまり，共通ノードは t_1 と t_2 のルートである場合の結合（$\bowtie$）を水平結合と呼び，次のように計算できる．

☆ WebTelop では，1 シーンに 1 つの話題しかないとする．文献 3) では，受信中の字幕データからシーンとシーンの話題構造抽出手法の詳細について述べている．

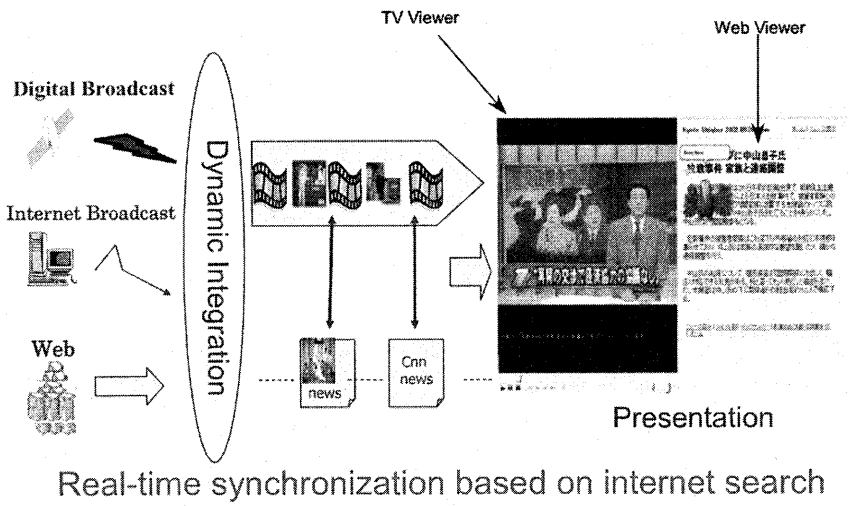
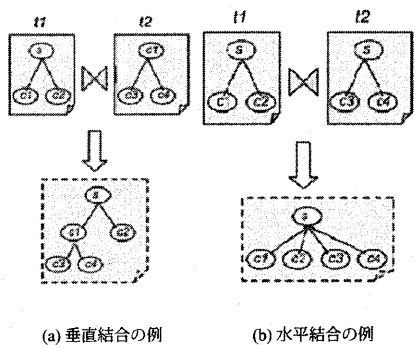


図 3　WebTelop の概念図



(a) 垂直結合の例　　(b) 水平結合の例

図 4

$$t_1 \bowtie^h t_2 \quad = (s, C_1) \bowtie^h (s, C_2)$$
$$= (s, C_1 \cup C_2) \tag{13}$$

(b) 垂直結合 ($\bowtie^v$)

$s_1 \in C_2$ または $s_2 \in C_1$ の場合の結合 ($\bowtie$) を垂直結合と呼ぶ.

$s_2 \in C_1$ の場合,

$$t_1 \bowtie^v t_2 = (s_1, C_1) \bowtie^v (s_2, C_2)$$
$$= (s_1, \{(s_2, C_2)\} \cup C_1) \tag{14}$$

である.

一方, $s_1 \in C_2$ の場合,

$$t_1 \bowtie^v t_2 = (s_1, C_1) \bowtie^v (s_2, C_2)$$
$$= (s_2, \{(s_1, C_1)\} \cup C_2) \tag{15}$$

となる.

特に, 前者を左垂直結合 ($\bowtie^v_l$), 後者を右垂直結合 ($\bowtie^v_r$) と呼ぶ. t_2 (t_1) は t_1 (t_2) のルート以外のノードをルートとする, t_1 (t_2) の部分木の場合, $t_1 \bowtie^v_l t_2 = t_1$ ($t_1 \bowtie^v_r t_2 = t_2$) となる.

水平結合は, 話題木の幅を広げる効果があると考えられる. つまり, 水平結合された話題は, ある subject-term をより広範囲から記述を行う. 水平結合は, 話題の視点・内容のカバーする範囲を広げて, オリジナル情報の増殖・補完を行うことができる. 一方, 垂直結合は, 話題木の深さを増加し, 話題をより詳しく述べる効果がある. つまり, 垂直結合は, より詳細の情報を提示し, オリジナル情報の補完・増殖を行える.

5.2 時間依存型結合機構

一般的に, 我々は, より詳細な情報またはより幅広い情報がほしいかに応じて, 水平結合 (WebTelop の B モード) または垂直結合 (WebTelop の D モード) を選択する. 結合候補の集合が決まると, ある特定の話題に対して, 結合の結果は常に一定であり, 結合された情報の重複する場合がある. 同じ話題構造が複数の Web ページに含まれる場合は, その一例である. 特に, 野球の試合映像などのデータストリームの場合, 同一話題の繰り返す場合が多いと考えれるので, 同一タイプ (水平または垂直) の結合を繰り返すと, 統合した情報の鮮度を損なう場合がある.

このような話題結合の重複を避けるための時間依存型結合機構を提案する. 基本的な考えは, 同一話題構造の繰り返しに対して異なるタイプの結合 (水平結合か垂直結合か) を選択するのである.

話題tの，今までの繰り返し回数（今回を含む）をrとする．また，tに対して，水平結合と垂直結合を実行した回数をそれぞれhとvとする．結合された話題木とtの高さの差をdとし，幅（葉ノードの数）の差をwとする．tと結合可能な話題の集合をTとする．$t_i \in T$に対して，つぎのような垂直結合価値$Vvalue(t, t_i)$（垂直結合に対する）と水平結合価値$Hvalue(t, t_i)$（水平結合に対する）を計算する．

$$Vvalue(t, t_i) = (1 - v/r) \times v_d \times d \qquad (16)$$

$$Hvalue(t, t_i) = (1 - h/r) \times v_w \times w \qquad (17)$$

ただし，v_dとv_tはパラメタである．v_dは話題木の高さの変化の単位価値であり，v_wは話題木の幅の変化の単位価値である．情報の詳細を重要としたい場合は，v_dを高く設定して，v_wを低くする．逆に，情報の幅を重要とする場合は，v_wを高く，v_dを低く設定する．

そして，tの，Tにあるすべての話題との，垂直結合価値の和と水平結合価値の和をそれぞれ計算して，垂直結合価値の和が水平結合価値の和より大きければ，tに対して垂直結合を行なう．そうでなければ，水平結合を実行する．

6. おわりに

本稿では，話題構造モデルとそれに基づく結合演算を提案した．話題構造は，Webページや映像シーンの内容を表すものであり，タイトルの役割を果たすタイトル要素と本体の内容を示す内容要素のペアから構成され，話題グラフと呼ばれる弱連結のDAGで表現される．二つの話題構造の結合（$\bowtie$）は，二つの話題グラフの和であり，新たな話題グラフ（話題構造）を生成する．$\bowtie$が結合律を満たさないため，話題グラフの和で簡単に表現される3つ以上の話題構造の結合を，うまく表現できない場合がある．これについて，今後の研究で究明して行く予定である．

一つのWebページが話題構造の集合である．これに対して，映像などのデータストリームが話題構造のストリームで表される．本稿では，話題構造の集合とストリームの結合を用いて，Webや放送コンテンツなどの複数情報ソースの内容統合を表現し，それらのいくつかの性質と特徴を明らかにした．今後，これらの演算の性質の更なる解明を行う予定である．

また，話題グラフの一種である有向話題木について考察を行った．有向話題木の結合を垂直結合と水平結合に分類した上で，情報の鮮度を求め，重複を避けるための時間依存型話題結合機構の提案を行った．垂直結合と水平結合を用いて，より詳しい・より幅広い情報の提供が可能となる．つまり，これらの演算を用いて情報の補完・増殖が可能となる．なお，これらの演算に基づいて我々の放送とWebの統合システムWebTelopの解析を行った．

有向話題木に限定していたが，話題構造の結合による情報補完・増殖の役割の解析の一部ができたと思う．今後，話題結合の分類や結合可能条件などの解明を行いながら，情報統合における内容統合の機能を解析していく予定がある．なお，話題構造による結合の関数的な性質などについて検討していく予定がある．

謝　辞

本研究の一部は，文部科学省科学研究費基盤研究(A)(2)「モバイル環境におけるコンテンツのマルチモーダル検索・呈示と放送コンテンツ生成」，科学研究費補助金（特別研究員奨励費）およびび平成14・15年度基盤技術研究促進事業（民間基盤技術研究支援制度）「クロスメディアコンテンツ基盤技術の研究開発」による．ここに記して謝意を表します．

参考文献

1) Digital Video Broadcasting Project: http://www.dvb.org (2003).
2) WebTV: http://www.webtv.com (2003).
3) Ma, Q. and Tanaka, K.: WebTelop: Dynamic TV-content Augmentation by Using Web Pages, *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2003) (to appear)* (2003).
4) Matsukura, T., Kondo, H., Hirata, Y. and Tanaka, K.: Discovery of Semantic Relationships among Web Pages Based on Web Topic Structures, *Proceedings of 9th IFIP 2.6 Working Conference on Database Semantics*, pp. 184–199 (2001).
5) Oyama, S. and Tanaka, K.: Exploiting Document Structures for Comparing and Exploring Topics on the Web, *Proceeding of the 12th Internaional World Wide Web Conference (WWW2003) (to apper, poster tracks)* (2003).
6) TopicMap.org: http://www.topicmap.org (2003).
7) Shasha, D., Wang, J. T. and Giugno, R.: Algorithmics and Applications of Tree and Graph Searching, *Proceedings of the 2002 ACM Symp. on Principles of Database Systems (PODS 2002)*, pp. 39–52 (2002).
8) Bhowmick, S. S., Ng, W. K., Lim, E.-P. and Madria, S. K.: Join Processing in Web Databases, *Database and Expert Systems Applications*, pp.647–657 (1998).
9) Ng, W. K., Lim, E.-P., Huang, C.-T., Bhowmick, S. S. and Qin, F.-Q.: Web Warehousing: An Algebra for Web Information, *Advances in Digital Libraries*, pp. 228–237 (1998).