# 話題構造に基づく放送とWebコンテンツの統合のための検索機構

馬 強[†] 田中克己[†,††]

ブロードバンド，デジタル放送およびインターネットの普及と発達にともなって，放送と通信の融合が着々と進んでいる．これらのメディア間の情報統合，相互補完によって，ユーザがより多様なコンテンツにアクセスすることが可能となる．本論文では，放送とWebコンテンツの動的統合のためのQuery-Free検索機構を提案する．Query-Free検索機構は，単なる類似検索ではなく，放送番組の内容を補完するWebページの検索が可能である．そのための，1）話題構造と話題構造によるコンテンツ結合モデルと，2）字幕データのようなテキストストリームのセグメンテーションと話題構造の抽出手法を提案する．さらに，実験とその結果について述べ，提案手法の検証を行う．

# Topic Structure Based Query-Free Web Retrieval Mechanism for Information Integration of Web and TV-program

Qiang Ma[†] and Katsumi Tanaka[†,††]

With the spreading of digital broadcasting and broadband internet connection services, the infrastructure for integration of TV-programs and the Internet is prepared. It is to say that it becomes possible to acquire interesting information from different media at the same time to improve the quality and detailedness of information. In this paper, at first, we propose a novel content-based join model for data streams (closed captions of videos or TV-programs) and web pages based on a notion called topic structure. Based on this model, we propose a query-free web retrieval mechanism by using topic structures of a video. One of the notable features of query-free retrieval mechanism is that the retrieved information is not just similar to the video, but also provides some additional information. We also show some experiment results in this paper.

## 1. はじめに

ブロードバンドの普及にともなって，高品質の映像や音声コンテンツをインターネットでも楽しめるようになってきている．また，デジタル放送では，本放送とともに，番組のメタデータなど関連情報が配信されることがある[1),2)]．このように，メディア間の情報統合，相互補完が着々と進み，ユーザはより多様なコンテンツにアクセスすることが可能となる[3)~10)]．テレビ放送とWebの融合を想定して，インターネットでテレビ番組のサイマルキャストを行うサービス[2)]やテレビとWebページが同時に見えるWebTV[11)]などのシステム・サービスがすでに存在する．つまり，放送とWebの統合のインフラストラクチャが整ってきている．

放送コンテンツは，高品質・高リアリティであるが，オンエア時間や不特定多数のユーザに情報を提供する必要があるなどの制約によって，情報の詳細や幅が限られている場合がある．一方，Webでは，品質はさまざまであるが，多種多様な情報が公開されている．このような性質の異なるメディアの情報を統合して，情報をより詳しく・より幅広く提供することが可能である．

本論文では，放送とWebコンテンツの動的統合のため，Query-Free検索機構を提案する．動的統合では，あらかじめインデックスを作成せず，関連情報をその場で検索し，オンラインで統合を行う．提案する機構は，ユーザのアクセスしているコンテンツ（Webまたは番組）に基づいて質問を自動的に生成する．ユーザが質問を意識する必要がないという意味で，この機構を，Query-Free検索機構と呼ぶ．Query-Free検索機構では，質問生成に用いられるコンテンツの範囲が不定であり，動的に決定される．なお，Query-Free検索機構によって検索されたWebページは，単に番組の内容と類似するのではなく，番組の内容をより詳し

† 独立行政法人情報通信研究機構
National Institute of Information and Communications Technology
†† 京都大学大学院情報学研究科
Graduate School of Imformatics, Kyoto University

くまたは別の視点から述べている．すなわち，番組の内容の補完を行うことが可能である．なお，本論文では，そのための，1）話題構造と話題構造によるコンテンツ結合モデルと，2）字幕データのようなテキストストリームのセグメンテーションと話題構造の抽出手法の提案を行う．

本論文では，1 つのイベントまたはアクティビティを話題（topic）と呼ぶ．番組や Web ページに含まれている話題を，3 章で定義される話題構造を用いて表現する．従来の研究[12)~16)] での話題（構造）は，コンテンツ（リソース）間の関係を解明するものであるが，我々は，1 つのコンテンツに述べられている話題を，語の役割に着目して構造化されたキーワード（群）で表現する．基本的に，話題構造は，番組や Web ページのタイトルを表す役割を持つキーワード（subject-term と呼ぶ）の集合と本体を表す役割を持つキーワード（content-term と呼ぶ）の集合のペアである．

我々は，話題構造を 1 つの連結成分からなる DAG（Directed Acyclic Graph）を用いて表現する．そして，2 つの話題構造の結合をグラフの和で表す．さらに，話題構造によるコンテンツ結合モデルを用いて放送と Web の統合の定式化を行う．このような話題構造に基づくコンテンツ結合モデルを利用して，我々は，放送と Web の動的統合のための，番組の内容を補完する Web ページの検索を行う Query-Free 検索機構を提案する．この検索機構では，単語の共起関係を考慮して，字幕データをセグメンテーションして，話題構造の抽出を行う．さらに，抽出された話題構造に対して，質問を自動的に生成し，Web から補完ページを検索してユーザに提示する．

以下，本論文の構成を示す．2 章では，関連研究について述べる．3 章では，話題構造とそれによるコンテンツ結合について述べる．4 章では，Query-Free 検索機構について述べる．評価実験については 5 章で述べる．6 章では，話題構造抽出手法と Query-Free 検索機構の議論を行う．7 章では，まとめと今後の課題について述べる．

## 2. 関 連 研 究

QBE（Query By Example）[17)] は，ユーザの与えられた例題に類似する情報を検索する手法である．例題に基づいて質問を生成する点では，我々の Query-Free 検索機構と同様である．しかしながら，Query-Free 検索機構では，話題構造とその結合モデルに基づいて，例題の単なる類似情報ではなく，補完情報（より詳細または別の観点の情報）の検索を行う点が異なる．

Henzinger らが Web から番組の類似ページを検索する手法を提案している[18)]．Henzinger らは，15 秒ごとに番組を分割して，字幕データから tf・idf ベースの手法を用いてキーワードを抽出し，番組の類似ページを検索する．Henzinger らの手法と比較して，我々の Query-Free 検索機構は，番組に類似するページだけではなく，番組の内容をより詳しく・より幅広く述べているページ，つまり内容補完のできるページを検索できる点が異なる．

トピックマップ[12)] 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，リソース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構造化されたキーワード群で表すものである．また，トピックマップでは，トピックは人手により定義されたものが多いが，本論文では，字幕データのセグメンテーションと話題構造抽出手法を提案している．TDT（Topic Detection and Tracking）[13)] では，ニュースのようなストリームデータからのトピック検出と追跡手法を研究開発している．TDT では，トピックはある重大なイベント・アクティビティおよびそれに関係するすべてのイベント・アクティビティを指し，トピックを構成するそれぞれのイベント（またはアクティビティ）をストーリ（story）と呼ぶ．ストーリが，我々の話題の概念と類似している．本論文では，1 つのイベントまたはアクティビティを話題と呼ぶ．その内容をキーワード集合のペアで表したものが話題構造である．さらに，我々は，話題の追跡や話題の相互関係を求めるのではなく，番組の補完ページを検索するために，字幕データのセグメンテーションと話題構造抽出手法の提案を行う．

結合（Join）は関係データベースにおける基本操作の 1 つである[19)]．Guha ら[20)] は構造やコンテンツの近似的なマッチングによる XML データソースの統合メカニズムを提案している．Bhowmick らは，異なる Web テーブルからの情報検索のため，Web ページのスキーマに基づく結合を提案している[21)]．彼らの結合演算は，WDM（Web Data Model）[22)] に基づくものである．本論文では，我々は話題構造という概念を用いた内容結合モデルを提案し，それに基づいて放送コンテンツを補完する Web ページの Query-Free 検索機構を提案している．

信号処理によるビデオのセグメンテーションを行う手法が多数提案されてきた[23)]．しかし，これらの手法は処理コスト（時間）がかかる場合がある．字幕デー

タの解析によるビデオセグメンテーションの手法も多数提案された[24),25)]．しかしながら，これらの手法は，字幕データの全体にアクセス可能であることや話題の切替えを表すマークの利用を前提としているため，受信中の字幕データのような絶えず追加・更新されるストリームデータには不十分である場合がある．本研究では，受信中の字幕データから話題構造を抽出して，その場で番組の補完情報を検索するため，オンラインの字幕データのセグメンテーション手法を提案している．実際，受信中の字幕データは，断片的な文章から構成されることが多い．従来手法では，このような字幕データに対して不十分である場合があると思われる．

データストリームに関する研究は，従来からさかんに行われてきた．最近，データベースコミュニティにおいて，データストリーム処理に新たな注目が集まっている[26),27)]．データストリームでは，データが絶えず追加されるので，データストリーム全体を通して処理を行うことが困難である．ゆえに，セグメンテーションはデータストリーム処理においては非常に重要な研究課題の 1 つである．スライディングウィンドウ関数[28)] は，時間幅やデータサイズを用いてデータストリームをいくつかのサブストリームに分割するには非常に有効である．これに対して，文献 29) では，ストリームにあるサブストリームの終了を表すマーク（punctuation）を導入する手法を紹介している．これらの研究に対して，本論文では，テキストストリームからの話題構造の抽出のため，語の共起関係を考慮した字幕データのセグメンテーション手法を提案し，それを用いた放送と Web の統合のための Query-Free 検索機構を提案する．

見出しに出現する語から本文に出現する語への関係を抽出して，情報の統合などに利用する試みは以前より行われてきた[30)〜32)]．これらの研究では，基本的に情報の断片を取扱いの単位として，同種メディアの情報整理を行う．見出しに出現する語と本文に出現する語の関係を考慮して，情報統合を行う点では，本論文と同様である．しかし，本論文では，それらの語の異なる役割を考慮した話題構造という概念を用いる点，および異種メディア（放送と Web）の情報統合・補完のための検索機構を提案している点が異なる．

## 3. 話題構造による内容結合

### 3.1 話題構造

話題構造が subject-term と content-term の集合のペアから構成される．subject-term は，Web ページやテキストストリームの話題において主題となる語である．本論文では，話題において，出現頻度の高い，かつ，その他のキーワードとの共起関係の強いキーワードを subject-term とする．一方，content-term は，同じページまたはテキストストリームに出現し，subject-term との共起関係の強いキーワードである．いい換えれば，subject-term はその話題のタイトルを表す役割があり，content-term は話題の本体を表す．Matsukura ら[33)] は話題構造の基本的な概念を次のように定義している．

$$topic = (S, C) \tag{1}$$

ただし，*topic* はある話題構造を表す．$S$ と $C$ は，それぞれ subject-term[☆]の集合と content-term の集合を表す．

Matsukura らのモデルを拡張して，我々は話題構造が階層的な構造を持つと考える．つまり，ある話題（構造）においては，別の話題（構造）を含むことが可能である．含まれているサブ話題は親話題の一部分を述べるという関係がある．我々の拡張話題構造は，次のように定義されている．

$$\begin{aligned} topic &:= '(' \ S, C \ ')' \\ S &:= '\{' \ (subject-term | topic)^{+} \ '\}' \\ C &:= '\{' \ (content-term | topic)^{+} \ '\}' \\ subject-term &:= keyword \\ content-term &:= keyword \end{aligned} \tag{2}$$

ただし，$S$ と $C$ はそれぞれ話題構造 *topic* の主題部と内容部であり，キーワード subject-term と content-term のほか，別の話題構造を含むことが可能である．また，定義のとおりに，subject-term と content-term は，キーワードである．さらに，あるキーワードは 1 つの話題構造においてたかだか 1 回しか現れないとする．ここでは，“+” は 1 回以上出現することを意味する．“|” は，“or” を意味する．

ここで注意すべきなのは，キーワード $k$ はある話題構造において，主体部 $S$ と内容部 $C$ に同時に属することができないが，どちらかに属することも可能である．たとえば，$(\{\text{環境}\}, \{(\{\text{ごみ}\}, \{\text{リサイクル}\})\})$ と $(\{(\{\text{環境}\}, \{\text{ごみ}\})\}, \{\text{リサイクル}\})$ は同じ話題構造を表している．キーワード「ごみ」が，前者では内容部のメンバであるが，後者では，主題部のメンバである．

### 3.2 話題グラフ

一般に，話題構造は 2 つ以上のノードを持つ，1 つの連結成分からなる DAG（Directed Acyclic Graph）を用いて表現できる．つまり，ある話題構造 $t$ は，キー

☆ Matsukura ら[33)] は，“thematic term” と呼んでいる．

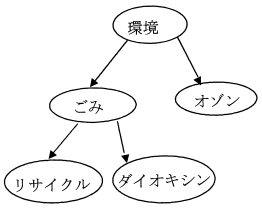


**図 1** 話題グラフの例
Fig. 1 Example of topic graph.

ワードを表す頂点の集合 $V$ ($|V| \geq 2$) と，キーワード間の subject-content 関係を表すエッジの集合 $E$ ($E \subseteq V \times V, |E| \geq 1$) を用いて表現できる．このような 1 つの連結成分からなる DAG を話題グラフと呼ぶ．

**定義 1（話題グラフ）** ある話題構造 $t$ の話題グラフ $G(t)$ は，次のように定義される：

$$G(t) = (V, E) \tag{3}$$

ただし，$V$ は頂点の集合であり，話題構造 $t$ に含まれるキーワードを表す．$E$ ($\subseteq V \times V$) はエッジの集合である．エッジ $e = (u, v)$ はキーワード $u$ と $v$ の間の *subject-content* 関係を表す．$u$ は，*subject-term* であり，$v$ は *content-term* である．ただし，$|V| \geq 2$, $E \neq \emptyset$ である．

**図 1** は，話題グラフの例を示している．この例では，$V$ = {環境, ごみ, オゾン, リサイクル, ダイオキシン}, $E$ = {(環境, ごみ), (環境, オゾン), (ごみ, リサイクル), (ごみ, ダイオキシン)}．この話題グラフは，次のような話題構造を表す．$t$ = ({ 環境 }, {({ ごみ }, { リサイクル, ダイオキシン }), オゾン }) である．

### 3.3 話題構造の結合

異なるメディアのコンテンツの統合を結合で表現することが可能である．たとえば，番組（データストリーム）と Web の関連コンテンツを統合することは，番組と Web の結合と見なすことができる．これを利用して，我々は，話題構造の結合を用いて情報統合の定式化を行う．

**定義 2（話題構造の結合）** 2 つの話題構造 $t$ と $t'$ の結合は，この 2 つの話題構造の話題グラフの和である．ただし，この 2 つの話題グラフの和は 1 つの連結成分からなる *DAG* である必要がある．つまり，2 つの話題構造の結合の結果は，話題構造である．

$$t \bowtie t' = \begin{cases} G(t) \cup G(t'), & G(t) \cup G(t') \text{ が 1 つの連結成分からなる } DAG \text{ である場合} \\ \phi, & \text{その他} \end{cases} \tag{4}$$

ただし，$G(t)$ と $G(t')$ は $t$ と $t'$ の話題グラフである．$\phi$ は空を表す．$t \bowtie \phi = \phi$ とする．

2 つの話題構造の結合が空でなければ，この 2 つの話題構造が結合可能であるという．結合の定義から，$t \bowtie t = t$ であることは明らかである．

明らかに，話題構造の結合は可換である．つまり，$t_1 \bowtie t_2 = t_2 \bowtie t_1$．**図 2** では，話題構造の結合の例を示している．例では，$(t_1 \bowtie t_2) \bowtie t_3 \neq t_1 \bowtie (t_2 \bowtie t_3)$．つまり，話題構造の結合は結合律を満たさない．

結合結果は 1 つの連結成分からなる DAG でなければ，空と見なす．これによって，結合結果も話題構造であることを保証する．したがって，結合結果である話題構造は，別の話題構造とのさらなる結合が可能である．1 つの連結成分という制約条件は，2 つの話題構造に共通要素のあることを保証する．また，DAG であることは，subject-term と content-term の区別を保つために必要である．たとえば，話題構造 ({a}, {b}) と ({b}, {a}) の結合を行う場合，DAG でないことを許すと，キーワード a とキーワード b の関係が矛盾となる．

### 3.4 放送と Web の統合の定式化

一般に，Web ページには，複数の話題がある．ゆえに，Web ページは話題構造の集合で表すことができる．たとえば，2 つの話題構造 $t_1$ と $t_2$ を含む Web ページ $p$ の話題構造を，$t_p = \{t_1, t_2\}$ のように表現することができる．したがって，Web ページのコレクションの話題（構造）は，話題（構造）の集合の和と見なすことができる．☆

$$T_P = t_{p_1} \cup t_{p_2} \cup \ldots \cup t_{p_n} \tag{5}$$

ただし，$T_P$ は，Web ページのコレクション $P$ の話題構造を表す．$t_{p_i}$ が $P$ に含まれるページ $p_i$ の話題構造の集合を表す．

番組の話題構造は，厳密に，話題構造の系列で表すべきであるが，本論文では，話題構造の間の順序関係を考慮せず，Web ページのコレクションと同様に，話題構造の集合で表す．

☆ 正確には，話題（構造）集合の集合である．本論文では，簡単のため，複数ページの話題（構造）を話題（構造）集合の和と見なす．つまり，話題（構造）の集合として取り扱う．

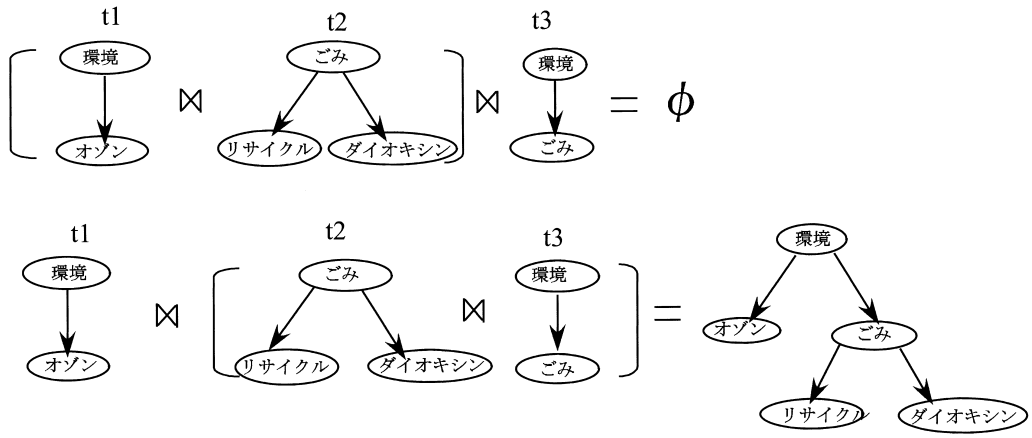


**図 2** 話題構造の結合の例
Fig. 2 Examples of join.

$$T_S = \{t_{s_1}, t_{s_2}, \ldots, t_{s_n}\} \quad (6)$$

ただし，$T_S$ は番組 $S$ $(= s_0 s_1 \ldots s_n)$ の話題構造を表す．$t_{s_i}$ は話題 $s_i$ の話題構造を表す．

したがって，放送と Web のコンテンツの統合は，2 つの話題構造の集合の結合として定式化できる．

**定義 3（話題構造集合の結合）** 2 つの話題構造集合 $T$ と $T'$ の結合は，次のように定義される．

$$T \bowtie T' = \{x \bowtie y \mid x \in T, y \in T'\} \quad (7)$$

Query-Free 検索機構では，このような結合を，番組の話題構造と結合可能な話題構造を持つ Web ページの検索と連動として実現している．

## 4. Query-Free 検索機構

本章では，Query-Free 検索機構について述べる．前述したように，Query-Free 検索機構では，番組の字幕データ（テキストストリーム）の話題構造を用いて，番組の内容を補完できる Web ページを検索するための質問を自動的に生成する．そのため，まず，字幕データのセグメンテーションと話題構造抽出を行う．そして，抽出された話題構造を用いて，コンテンツ結合モデルに基づいて質問を生成する．生成された質問を用いて，Google などの検索エンジンを通して必要な Web ページを獲得する．

### 4.1 テキストストリームのセグメンテーションと話題構造の抽出

ニュース放送の索引付けなどのためにニュース放送の字幕データを話題の単位に分割することは広く試みられている[24),25)]．本研究では，受信中の字幕データから話題構造を抽出して，その場で番組の補完情報を検索するため，オンラインの字幕データセグメンテーションの手法が必要である．実際，受信中の字幕データ（NHK ニュース 7 etc.）は，断片的な文章から構成されることが多い．従来手法は，このような字幕データモデルやオンライン利用を十分考慮していないため，受信中の字幕データのセグメンテーションには不十分である場合があると考え，独自の手法を用いた．

**(a) 共起関係**

本論文では，テキストストリームのセグメンテーションと話題構造抽出のため，以下の 2 種類の共起関係を定義している．1）無向共起度と，2）有向共起度である．

**定義 4（無向共起度）** ある話題コレクション☆において，語 $w_1$ と $w_2$ が同時に出現される話題（テキスト）が多いほど，この 2 つの語の共起関係が強いという．本論文では，語 $w_i$ と $w_j$ の無向共起度 $cooc(w_i, w_j)$ を次のように定義する．

$$cooc(w_i, w_j) = \frac{df(\{w_i, w_j\})}{df(\{w_i\}) + df(\{w_j\}) - df(\{w_i, w_j\})} \quad (8)$$

ただし，$df(\{w_i\})$ は，話題コレクションにおける，語 $w_i$ を含む話題（テキスト）の数である．$df(\{w_i, w_j\})$ は語 $w_i$ と $w_j$ を同時に含む話題（テキスト）の数である．

**定義 5（有向共起度）** ある話題のコレクションにおいて，語 $w_i$ と $w_j$ の有向共起度 $\overrightarrow{cooc}(w_i w_j)$ は，単語 $w_i$ が含まれる話題（テキスト）の中に単語 $w_j$ を

☆ 本論文では，一定期間内のすべての話題に対応するすべてのテキストの集合を話題コレクションと呼ぶ．

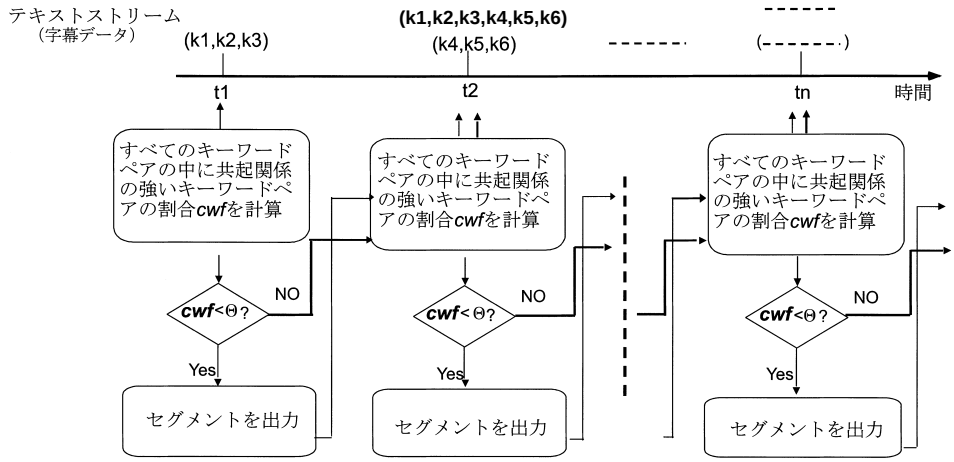


**図 3** テキストストリームのセグメンテーション
Fig. 3 Online segmentation of text streams.

含む話題（テキスト）の割合である．有向共起度が次のように計算される．

$$\overrightarrow{cooc}(w_i w_j) = \frac{df(\{w_i, w_j\})}{df(\{w_i\})} \tag{9}$$

ただし，$df(\{w_i, w_j\})$ は $w_i$ と $w_j$ を含む話題（テキスト）の数であり，$df(\{w_i\})$ は $w_i$ を含む話題（テキスト）の数である．

一般に，$cooc(w_i, w_j) = cooc(w_j, w_i)$ であるが，$\overrightarrow{cooc}(w_i w_j)$ と $\overrightarrow{cooc}(w_j w_i)$ は必ずしも等しいとは限らない．

**(b) テキストストリームのセグメンテーション**

テキストストリームには複数の話題があるので，話題構造を抽出するためには，まず，セグメンテーションが必要である．**図 3** では，テキストストリーム（字幕データ）のセグメンテーション手法の流れを示している．基本的に，受信された字幕データの中で共起関係の強いキーワードペアが多いほど，それらの字幕データは 1 つの話題について述べている可能性が高い．

セグメンテーションの手順を以下に示す．ここでは，$CT_i$ を時間 $t_i$ におけるキーワード集合とする．$ST$ と $ET$ は，それぞれ抽出されるセグメントの開始・終了時間である．

(1) $CT_0 = \emptyset$，$ST = 0$，$i = 1$ とする．
(2) 字幕データを受信する．データがなければ，終了する．
(3) 時点 $t_i$ $(i \geq 1)$ でデータを受信したら，キーワード集合 $K$ を受信された字幕データから抽出する．
(4) $CT_i = CT_{i-1} \cup K$ とする．
(5) $CT_i$ におけるすべてのキーワードペアの中に，共起関係の強いキーワードペアの割合 $cwf(t_i)$ を計算する．ここでは，共起関係の強いキーワードペアとは，無向共起度がある閾値 $\theta$ より大きい 2 つのキーワードのことである．$m$ は $CT_i$ におけるキーワードの数である．

$$cwf(t_i) = \sum_{j=1,k=j+1}^{j=m-1,k=m} cr(w_j, w_k) / \frac{m \cdot (m-1)}{2}$$

$$cr(w_j, w_k) = \begin{cases} 1, & cooc(w_j, w_k) \geq \theta \text{ のとき} \\ 0, & cooc(w_j, w_k) < \theta \text{ のとき} \end{cases}$$

(6) $cwf(t_i) \geq \Theta$ であれば，(9) へ．でなければ，次へ．ただし，$\Theta$ はあらかじめ定義された閾値である．
(7) $ET = t_i$ とする．開始時間と終了時間がそれぞれ $ST$ と $ET$ である字幕データのかたまりをセグメント $topic_i$ として出力する．$CT_i$ は $topic_i$ のキーワード集合として出力され，話題構造の抽出に利用される．
(8) $CT_i = \emptyset$，$ST = t_i$，$i = i + 1$ とする．
(9) 字幕データを受信する．これ以上のデータがなく，かつ，$CT_i = \emptyset$ であれば，終了する．これ以上のデータがないが，$CT_i \neq \emptyset$ であれば，(7) へ．その他の場合，(3) へ．

字幕データのセグメンテーション手法として，隣り合った字幕データ間の語の共起度が弱ければこの間で切るということも考えられる．ただし，この場合，次の字幕データの先読みが必要となる．また，本論文では，共起関係の弱いキーワードが混じりだしたところで，字幕データをまとめる操作を打ち切るという手順

でセグメンテーションを行っている．これに対して，共起関係の弱いキーワードが混じりだしたところの字幕データを次のセグメントの先頭データとすることも考えられる．いずれにせよ，セグメントに含まれるべきデータの欠落やノイズデータの混じりが発生する可能性が高いと思われる．共起関係の弱いキーワードが混じりだしたところの字幕データに対してさらなる解析は必要であると考えられる．今後の研究で，これについて検討する予定である．

上記のセグメンテーション手法は，テキストストリームのみではなく，Web ページにも適応可能である．たとえば，Web ページをパラグラフの系列と見なして，それぞれのパラグラフを 1 回の受信データとすれば，上記の手法の適応ができると考える．ただし，この場合，Web ページの構造を利用していないという短所がある．

**(c) 話題構造の抽出**

キーワードの subject-term である可能性を主題度という概念を用いて表す．語 $w_i$ の主題度は，1）話題（テキスト）におけるその他の語との有向共起度と，2）話題（テキスト）における出現頻度によって計算される．つまり，話題（テキスト）における出現頻度が高く，かつ，その他の語との有向共起度の強いキーワードは，主題度の高いキーワードであり，subject-term の可能性が高い．

語 $w_i$ の主題度 $sub(w_i)$ は，次のように計算される．

$$sub(w_i) = tf(w_i) + \sum_{j=1,j\neq i}^{n} \overrightarrow{cooc}(w_i w_j) \quad (10)$$

ただし，$tf(w_i)$ は話題（テキスト）における $w_i$ の出現頻度である．$\overrightarrow{cooc}(w_i w_j)$ は語 $w_i$ と $w_j$ の有向共起度である．$n$ は，話題（テキスト）に含まれているキーワードの数である．

話題（テキスト）に含まれている語の主題度をそれぞれ計算して，高い値を持つ $N$ 個の語は subject-term として選択される．

一方，content-term は，subject-term との無向共起度に基づいて求められる．すなわち，話題（テキスト）における，subject-term との無向共起度の強い語は，その話題の content-term である可能性が高い．語 $w_i$ は content-term である可能性を内容度 $con(w_i)$ とし，次のように計算される．

$$con(w_i) = \sum_{w_j \in S} cooc(w_i, w_j) \quad (11)$$

ただし，$S$ は抽出された subject-term の集合である．

内容度の高い $M$ 個の語を話題の content-term とする．**図 4** は，共起関係による subject-term と content-term の抽出例（$N = M = 1$）を示している（語の出現頻度は同じであるとする）．図では，ラベルはキーワード間の共起度を表す．

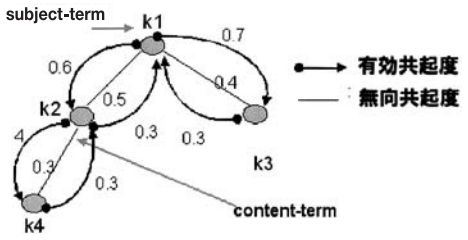


**図 4** subject-term と content-term の抽出例
Fig. 4 Example of extraction of subject and content terms.

前に述べたように，話題構造は階層的であるが，ここで抽出される話題構造の主題部と内容部が，別の話題構造を含まない．もちろん，上記の手法を再帰的に適応すれば，より複雑な話題構造の抽出が可能である．つまり，抽出された content-term を subject-term と見なして，さらに content-term を求めることが可能である．また，6.1 節で述べている簡約を用いた手法も考えられる．

### 4.2 質問生成

本節では，番組の補完 Web ページを検索するための質問生成について述べる．前に述べたように，Query-Free 検索機構では，テキストストリームの話題構造を用いて質問を生成し，番組の内容と単に類似するのではなく，より詳しくまたは別の視点から情報を述べているページの検索が可能である．ここでは，テキストストリームに含まれる話題構造は，主題部と内容部がその他の話題構造を含まないとする．検索される Web ページに含まれる話題構造も同様であるとする．

Oyama ら[34] は，HTML ソースの “title” と “body” タグが Web ページの話題構造抽出には有用であると報告している．彼らの研究成果を利用して，我々は，Web ページに含まれる話題構造は，subject-term が見出しに現れ，content-term が本文に現れると想定して検索質問を生成する．

Query-Free 検索機構では，次のような 4 種類の質問を生成する．1）CD（content-deepening）質問，2）SD（subject-deepening）質問，3）SB（subject-broadening）質問と，4）CB（content-broadening）質問である．

CD 質問と SD 質問は，次のような 2 つの話題構造の結合に基づいて定義されたものである．話題構造 $A$ の subject-term が別の話題構造 $B$ の content-term

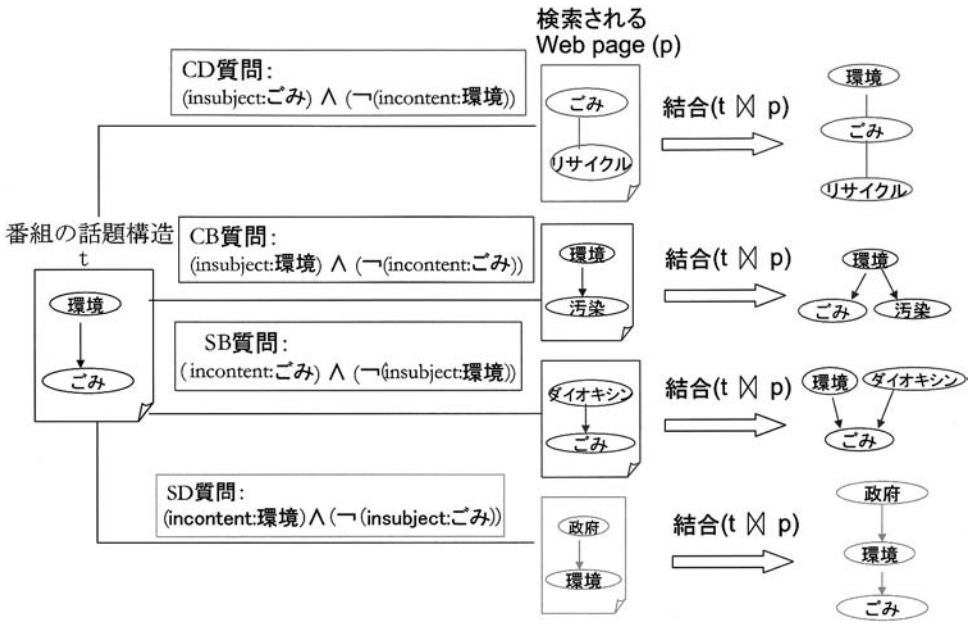


**図 5** 話題構造の結合に基づく質問の例
Fig. 5 Example of CD, SD, CB and SB queries.

に含まれる．つまり，$A$ は $B$ のある部分（content-term）について詳しく述べている．このような話題構造の結合が，元々の話題グラフの深さを増加する効果（deepening）があり，元の情報を詳細化することが可能である（**図 5**）．

SB 質問と CB 質問が，共通の subject-term または content-term を持つ 2 つの話題構造の結合に基づいて定義される．つまり，同じ内容（主題）であるが主題（内容）が別であるような 2 つの話題構造の結合である．このような話題構造の結合は，元々の情報の主題または内容の幅を広げる効果（broadening）がある（図 5）．つまり，結合によって別の視点の情報を提供することが可能である．

以下，テキストストリームのある話題構造 $t$ を $(\{s_1, s_2, \ldots, s_n\}, \{c_1, c_2, .., c_m\})$ とする．$m \geq 1$，$n \geq 1$ である．$s_i$ と $c_i$ は，それぞれ subject-term と content-term を表す．“insubject” と “incontent” に後置される検索文は，それぞれ Web ページの見出しと本文を検索対象とする．“∧” と “∨” はそれぞれ論理積と論理和を表す．“¬” は，論理否定を表す．たとえば，質問 $(insubject\colon k_1 \wedge k_2) \wedge (\neg(incontent\colon k_3 \wedge k_4))$ は，$k_1$ と $k_2$ が見出しに含まれ，$k_3$ と $k_4$ が本文に含まれないページを検索する．

- **CD（Content-Deepening）質問（$Q_{d_c}$）**:
$$Q_{d_c} = (insubject : c_1 \wedge c_2 \wedge \ldots \wedge c_m) \wedge (\neg(incontent : s_1 \vee s_2 \vee \ldots \vee s_n))$$
- **SD（Subject-Deepening）質問（$Q_{d_s}$）**:
$$Q_{d_s} = (incontent : s_1 \wedge s_2 \wedge \ldots \wedge s_n) \wedge (\neg(insubject : c_1 \vee c_2 \vee \ldots \vee c_m))$$
- **SB（Subject-Broadening）質問（$Q_{b_s}$）**:
$$Q_{b_s} = (incontent : c_1 \wedge c_2 \wedge \ldots \wedge c_m) \wedge (\neg(insubject : s_1 \wedge s_2 \wedge \ldots \wedge s_n))$$
- **CB(Content-Broadening Query)質問($Q_{b_c}$)**:
$$Q_{b_c} = (insubject : s_1 \wedge s_2 \wedge \ldots \wedge s_n) \wedge (\neg(incontent : c_1 \wedge c_2 \wedge \ldots \wedge c_n))$$

図 5 では，それぞれの質問の例およびそれに対応する結合を示している．

それぞれの検索式を用いて，与えられた話題構造（質問の生成に利用されたもの）と結合可能な話題構造（元の話題構造 $t$ 以外）を含む Web ページを検索する．Web ページに複数の話題構造がある場合，その Web ページに検索式で要求されていた話題構造を 1 つでも含めば，その Web ページが解となりうる．

本論文では，元の話題構造 $t$ を含むページを類似ページと見なす．このような類似ページが情報統合に有用であると考えられる．しかし，本論文では，単なる類似ではなく，元と似て非なる情報（たとえば，より詳細または別の視点の情報）を含むページを検索して，元のコンテンツの補足を行うことを目的としているため，類似ページ（$t$ と同じ話題構造を含む Web ページ）を検索するための質問を Query-Free 検索機

構の質問として定義しなかった☆．そのうえ，このような類似ページを排除している．さらに，結合結果が空になることをさけるため，それぞれの検索式には，否定条件部が定義されている．否定条件部は “¬” から開始する部分である．また，それぞれの検索式の前半部分にある，“¬” を含まない条件文を肯定条件部と呼ぶ．肯定条件部によって，2 つの話題構造に共通要素があることが保証される．否定条件部によって，$t$ との結合結果がサイクルとなるような話題構造を含むページ，および $t$ と同じ話題構造を含むページが排除される．CD と SD 質問は，元の話題構造との統合結果がサイクルとなることを抑制するために強めの否定条件になる．一方，CB と SB 質問は，統合結果がサイクルとなることがなく，元と同じ話題構造を含むページを排除できればよいので，弱めの否定条件にした．

このように，話題構造を利用して，構造化質問を記述することができる．これによって，主題と内容を区別して類似と非類似を考えることが可能となり，似て非なる（補完）情報の検索ができると思われる．また，Oyama ら[34] は，subject-term と content-term の区別の検索に及ぼす影響について詳しく報告しているが，本論文の実験結果からも同様なことが読み取れる．

## 5. 実　　験

本章では，4 つの実験とその考察について述べる．実験 *I* では，字幕データのセグメンテーションと話題構造抽出手法について評価を行う．実験 *II* では，Query-Free 検索機構について評価を行う．実験 *III* では，Query-Free 検索と類似検索の比較，および CD，SD，CB と SB 質問の比較を行う．実験 *IV* では，検索式の否定条件部による検索への影響について考察する．

我々は，2002 年 9 月から 2003 年 5 月までの 9 カ月間の NHK ニュース 7 の字幕データを利用して，共起度辞書を作成した．この共起度辞書を利用して，字幕データのセグメンテーションと話題構造の抽出に必要となる語の共起度を調べていた．共起度辞書には，語のペアおよび共起度（有向共起度と無向共起度）の値が登録されている．共起度辞書に登録されていない語のペアの無向共起度と有向共起度は 0 とする．1 日の字幕データの平均サイズは 26.53 KB であった．まず，1 カ月間（2002 年 9 月）の字幕データから手動でセグメンテーションを行い，初期の話題コレクションとそれに基づく初期の共起関係辞書を生成した．そして，それを用いて残りの 8 カ月間（2002 年 10 月から 2003 年 5 月まで）の字幕データのセグメンテーションを行い，話題コレクションと共起関係の辞書を更新した．日本語の形態素解析は茶筌[35] を利用して行われた．なお，ストップワードを省くため，940 ワード（そのうち 593 個は英単語である）のストップワードリストを作成した．いくつかの予備実験の結果によって，字幕データのセグメンテーションのための閾値 $\theta$ と $\Theta$ をそれぞれ 0.15 と 0.28 に設定した．すなわち，キーワードペアの無向共起度が $\theta$（$= 0.15$）以上であれば，共起関係の強いキーワードペアと見なす．また，共起関係の強いキーワードペアの割合は $\Theta$（$= 0.28$）より小さければ，セグメントを出力する．

実験では IBM の ThinkPad T30 を利用した．CPU は Pentium4-M 1.8 G であり，主メモリは 768 MB である．OS は Windows XP である．インターネット接続の実効スピードは 1.8 Mbps である．Microsoft Visual Studio .Net，GoogleAPI[36] と茶筌[35] を利用して開発を行った．

### 5.1 実験 *I*：字幕データのセグメンテーションと話題構造抽出の評価実験

85 日間（2003 年 5 月から 2003 年 7 月まで）の NHK ニュース 7 の字幕データを実験素材として利用した．これらのデータから，3,068 個のセグメントを出力した．そして，各々のセグメントの話題構造として，2 つの subject-term と 3 つの content-term を抽出した．

出力されたセグメントの文章の 8 割以上が 1 つの話題について述べていれば，我々は，そのセグメントのセグメンテーションが成功したという．逆に，成功しなかったものは失敗したという．同様に，抽出された話題構造に対して，セグメントに述べられている話題を表すために適する 4 つ以上の語が含まれていれば，その話題構造の抽出が成功したと判断する．このような判断基準で，3,068 個のセグメントのうち，2,460 件のセグメンテーションと 2,129 件の話題構造抽出が成功した．適合率はそれぞれ 0.802 と 0.694 である．ここでいう適合率は，処理（セグメンテーションまたは話題構造抽出）の総数の中での成功した処理の割合である．なお，セグメンテーションと話題構造抽出が同時に成功した件数は 1,893 件である．筆者による手動の分析によると，これらの字幕データから出力すべきセグメントが 3,506 件見い出された．それに対して，字幕データのセグメンテーションの再現率は 0.702 であった．再現率は，本来出力すべきセグメントの中での，セグメンテーションの成功したセグメントの割合

☆ 話題構造に基づく類似検索の定義式は，5 章に示されている．

である．

608 件のセグメンテーションの失敗があった．失敗のパターンは，次のように分類される．

- 新規イベントのような話題に対応する字幕データのセグメンテーションの失敗が 349 件あった．ここでいう新規イベントは，共起関係の辞書を生成するための話題コレクションに含まれる話題との類似の非常に低いものである．共起関係の強いキーワードペアをうまく検出できないので，セグメンテーションの失敗する可能性が高い．
- 2 つ以上の関連性の高い話題を述べている字幕データを 1 つのセグメントとして出力してしまった失敗が 201 件あった．セグメンテーションのパラメータと共起関係辞書による失敗であったと考える．
- セグメントが短かすぎるため，失敗と判断したのは 43 件あった．セグメントの最初と最終の字幕データの受信時間間隔が短い（10 秒以下），または，そのセグメントに 2 つ以下の文章しかなければ，セグメンテーションが失敗したと考えられる．利用された NHK の字幕データの中で，キャスタの喋る内容しか字幕データ化されない場合が多い．また，映像にテロップがある場合は字幕データを生成・配信しないことがある．このような字幕データの制約は，セグメンテーションの失敗原因となることがある．
- 字幕データの誤りによる 15 件の失敗があった．利用された字幕データには，脱字・誤字のある場合があった．さらに，それらの誤りを修正するためのメッセージ（字幕データ）が挿入される場合があった．このような字幕データの誤りや本来の番組内容と無関係な字幕データは，セグメンテーションに影響を及ぼす可能性がある．

一方，939 件の話題構造の抽出が失敗した．それぞれの失敗について，著者がその原因を推測した．そして，各々の失敗の最も可能性の高い原因で分類を行った．

- セグメンテーションの失敗による 503 件の誤抽出があった．
- 話題構造を抽出するための文章（セグメント）には，特定の話題について述べている文章のほかに，無関係のノイズ文章が含まれる場合がある．このようなノイズとなる文章の中のキーワードを subject-term または content-term として抽出してしまった場合があった．158 件のこのような失敗があった．共起関係によるものが多いと思われる．また，そのうちの 13 件は，字幕データの誤りによる誤抽出であった．
- ユーザが subject-term にふさわしいと思われる語は，セグメント（字幕データ）の中での出現頻度の低い可能性がある．これによる 113 件の失敗があった．たとえば，地名・組織名・人名などの固有名詞が，1 回しか現れない場合がある．このような単語が subject-term として抽出されなかった．しかし，その他の語と比べて，subject-term としてはより適切である場合があった．その対策としては，語の品詞を考慮する必要があると考える．
- 茶筌による人名・組織名のような単語の分割による 87 件の失敗があった．たとえば，“レッドソックス” を “レッド” と “ソックス” に分割してしまう場合があった．このような 1 つの単語として扱うべきものが 2 つの単語に分割され，そして，この 2 つの単語が同時に話題構造のキーワード（subject-term または content-term）として抽出される場合があった．これは適切でないと考える．
- 同義語または類義語を区別して取り扱っていたため，78 件の誤抽出があった．たとえば，“首相” と “総理” が同時に話題構造に現れる場合があった．このような場合，情報の冗長であり，適切ではないと考える．

実際，我々の手法が共起関係に依存するので，新しいイベントやアクティビティに対するセグメンテーションと話題構造の抽出は得意ではない．これは，上記の失敗例の解析結果からも明らかである．共起関係の辞書が，字幕データのセグメンテーションと話題構造抽出には非常に重要な役割を持つことが分かる．実験データに，共起関係の辞書の作成に利用された 28 日間（2003 年 5 月分）のデータが含まれた．これらのデータに対して，適合率（セグメンテーションと話題構造の抽出の両方が成功した場合）は 0.735 であった．一方，残りの 57 日間（2003 年 6 月，7 月分）のデータに対しては，適合率は 0.526 であった（**表 1**）．この結果からも，共起関係の重要性が分かる．

共起関係の辞書の工夫としては，次のようなことが考えられる．

- 学習による辞書更新をより頻繁に行う．
- Web 検索を利用して語の共起関係を求める．たとえば，語を検索して，検索されたページの総数を語を含む話題（テキスト）の数として，語の共起関係を求める．

また，新規話題（イベント）に対応する字幕データのセグメンテーションと話題構造抽出のため，改良策として，語の共起関係と $tf$（term frequency）値の

**表 1** 実験 $I$ の結果
Table 1 Results of evaluation $I$.

| | 適合率 | | | 再現率 |
|---|---|---|---|---|
| | 5 月分 | 6，7 月分 | 全体 | 全体 |
| 字幕データのセグメンテーション | - | - | 0.802 | 0.702 |
| 話題構造の抽出 | - | - | 0.694 | - |
| 字幕データのセグメンテーションと話題構造の抽出 | 0.735 | 0.526 | 0.617 | - |

**表 2** 実験 $II$ のための質問
Table 2 Query used in evaluation $II$.

| | 5-keys query | 3-keys query |
|---|---|---|
| SB 質問 | $intext:c_1\ intext:c_2\ intext:c_3\ -allintitle:s_1\ s_2$ | $intext:c_1\ intext:c_2\ -intitle:s_1$ |
| SD 質問 | $intext:s_1\ intext:s_2\ -intitle:c_1\ -intitle:c_2\ -intitle:c_3$ | $intext:s_1\ -intitle:c_1\ -intitle:c_2$ |
| CB 質問 | $intitle:s_1\ intitle:s_2\ -allintext:c_1\ c_2\ c_3$ | $intitle:s_1\ -allintext:c_1\ c_2$ |
| CD 質問 | $intitle:c_1\ intitle:c_2\ intitle:\ c_3\ -intext:s_1\ -intext:\ s_2$ | $intitle:c_1\ intitle:c_2\ -intext:s_1$ |

ほかに，さらに *idf*（inverse docuemnt frequey）値を利用することが考えられる．なお，この手法は，2 つの類似話題について述べている字幕データを 1 つのセグメントとして抽出してしまう失敗や，出現頻度の低い語の subject-term である可能性があるような問題を解決するにも有用であると考える．さらに，語の品詞を考慮することも改良案として考えられる．これらの改良案を，今後の研究で検証したいと思う．

### 5.2 実験 $II$：Query-Free 検索機構の評価実験

実験 $I$ で利用されたデータから，任意の 3 日間（2003 年 5 月 28 日，2003 年 6 月 20 日と 2003 年 7 月 20 日）のデータおよびそれらに対応するビデオを選択して，我々の Query-Free 検索機構の評価実験を行った．

この実験では，我々は，映像と同期された字幕データのセグメンテーションと話題構造抽出を行い，4.2 節で定義された 4 種類の質問を生成し，GoogleAPI [36] を通して Google に発行する．Google からの検索結果のトップページを番組の関連ページとした．我々は，まず，抽出された番組の話題構造の 2 つの subject-term と 3 つの content-term を利用して，質問を生成する（5-keys 質問と呼ぶ）．検索結果が 0 であれば，新たに 1 つの subject-term と 2 つの content-term を用いて質問を生成する（3-keys 質問と呼ぶ）．

4 種類の質問を実現するために，GoogleAPI の構造的な検索オプション，“*intitle*”，“*intext*”，“*allintitle*” と “*allintext*” を利用した．Google では，intitle と allintitle は，タイトル検索のオプションである．intext と allintext は，本文検索のオプションである．Google では，“intitle:” の直後のキーワードおよび “allintitle:” に後置されるキーワードを見出しに含む Web ページを検索する．同様に，“intext:” の直後のキーワードおよび “allintext:” に後置されるキーワードを本文に含む Web ページを検索する．なお，Google では，デフォルトは AND 検索である．“–” は，否定を意味する．

話題構造 $t = (\{s_1, s_2\}, \{c_1, c_2, c_3\})$ を用いて生成された，Google に発行する 4 種類の質問は**表 2** に示されている．

Google から 1 つ以上の検索結果が返されれば，その質問が有効質問であるという．上記のように，Google から返された検索結果のトップページを番組の関連ページとした．ユーザは，そのページを見て，正解であるかどうかを判断する．もしそのページが番組と関係し，しかも番組の内容を補足できれば，そのページを正解とする．実験では，2 人のユーザによる評価を行った．SB，SD，CB と CD 質問のそれぞれの検索ページが番組の内容を別の主題（視点）から述べているか，番組の主題を詳しく述べているか，番組の主題を別の内容（視点）から述べているかと番組の内容をより詳しく述べているかを基準とした．実験では，まず，被験者にそれぞれ判定を行ってもらった．検索されたページの主題・内容および元の番組との関連性に対する認識のずれによって，判定が分かれた場合があった（CD，SD，CB と SB 質問に対して，それぞれ 8，7，14，15 件があった）．被験者の協議によって判定を統一した．**図 6** に，判定結果の具体例（成功，失敗，微妙な判定となった検索結果の典型例）を示す．

**表 3** は実験の結果を示している．適合率とは検索された関連ページの中での正解ページの割合である．適合率の結果から，提案手法は，番組の内容を別の視点から述べているページとより詳細のページを検索することが可能であることが分かる．なお，表 3 の *SIM* 質問とその結果については，次節の実験 *III* で述べる．

**表 3** 実験結果（実験 *II* と実験 *III* (*a*)）
Table 3 Results of evaluation *II* and *III* (*a*).

| | Query-Free 質問 | | | | SIM 質問 |
|---|---|---|---|---|---|
| | SD 質問 | CD 質問 | SB 質問 | CB 質問 | |
| 話題構造 | 88 | 88 | 88 | 88 | 88 |
| 有効質問 | 86 | 66 | 88 | 86 | 74 |
| 有効 5-keys 質問 | 81 | 23 | 88 | 45 | 29 |
| 有効 3-keys 質問 | 5 | 43 | 0 | 41 | 45 |
| 正解数（主題をより詳細） | 55 | - | - | - | 15 |
| 正解数（内容をより詳細） | - | 43 | - | - | 12 |
| 正解数（同じ内容，別の主題） | - | - | 62 | - | 16 |
| 正解数（同じ主題，別の内容） | - | - | - | 45 | 17 |
| 適合率（主題をより詳細） | 0.625 | - | - | - | 0.170 |
| 適合率（内容をより詳細） | - | 0.489 | - | - | 0.136 |
| 適合率（同じ内容，別の主題） | - | - | 0.705 | - | 0.182 |
| 適合率（同じ主題，別の内容） | - | - | - | 0.511 | 0.193 |

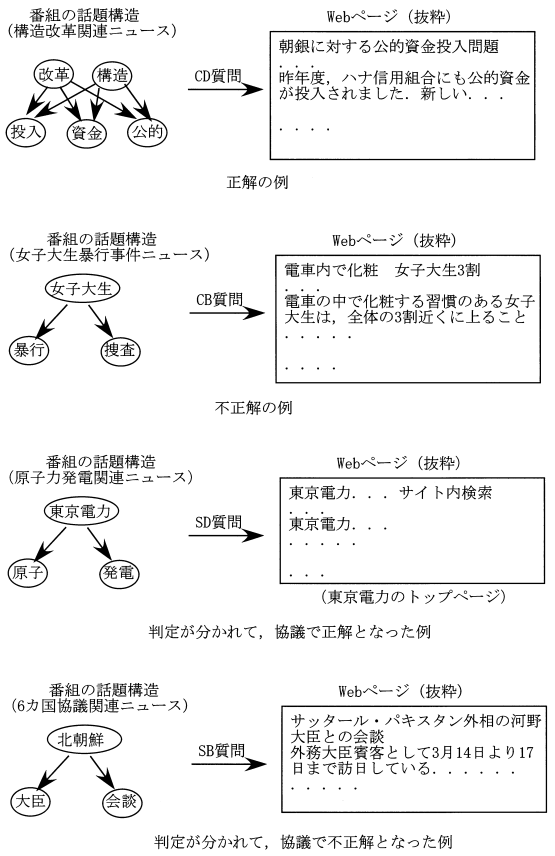


**図 6** 判定結果の具体例
Fig. 6 Examples of estimation results.

Query-Free 検索機構の適合率を改善するためには，以下のようなことが考えられる．

- 字幕データのセグメンテーションと話題構造抽出手法を改良する．抽出の成功した話題構造を用いた質問が良い結果が得られた．特に，話題構造に固有名詞が含まれた場合の検索は，より良い結果が得られた．したがって，セグメンテーションと話題構造抽出手法の改良によって，Query-Free 検索機構の適合率を改善できると考える．改良案としては，実験 *I* で述べたような，共起関係の辞書の作成方法の改良や語の idf 値の利用などが考えられる．また，話題構造の抽出を行う際，単語の品詞を考慮することも 1 つの案である．
- 検索結果の独自のランキング手法を開発する．特に，情報補完という観点からの Web ページのランキング手法の開発が必要かつ重要であると考える．たとえば，検索された Web ページの話題構造を抽出して，字幕データから抽出された番組の話題構造との比較に基づいて最終解を選択する手法が考えられる．

再現率（検索されるべきページにおける正しく検索されたページの割合）の計算が困難であるため，今回の評価は，適合率のみにとどまった．しかし，それぞれの検索式の肯定条件部は，“∧”（論理積）で定義されていることや，否定条件部を含んでいることは，検索結果の絞り込みに効果があると考えられるが，検索漏れの可能性も大きいと思われる．実験 *IV* では，否定条件部による検索結果への影響について考察を行う．また，上記の独自のランキング手法は，結果の絞り込みだけではなく，検索漏れを防ぐにも有効である可能性がある．これらについては，6 章で考察を行う．

今回の実験では，被験者の数が少なく，しかも，1 件目の検索結果のみを評価対象にしていた．これは，不十分であり，さらなる評価は必要であると思われる．

### 5.3 実験 *III*：検索方法の比較実験

本論文では，番組の内容と単に類似するのではなく，番組の内容をより詳細または別の視点から述べているページの Query-Free 検索機構を提案している．Query-Free 検索と従来の類似検索，および Query-

Free 検索機構の 4 種類の質問の違いを考察するため，実験 *III* を行った．

Google に対して，異なる質問を発行するので，検索結果が一致しないのは当然と考えられるが，Web ページに複数の話題が含まれる点や検索エンジンのアルゴリズムが不明確である点などを考慮すると，検索結果の重なる部分が多く存在することも考えられる．特に，今回の実験では，同じ話題構造に対してそれぞれの質問を生成しているため，これらの質問には関連性がある．そのため，意図どおりの異なる検索結果を得られない可能性は否定できない．実験 *III* の $(a)$ では，提案方式と類似検索の比較を行い，単なる類似検索ではないことを確認する．

実験 *II* の結果から，Query-Free 検索機構の 4 種類の質問はより詳細な情報または別の視点の情報を検索できることが分かった．実験 *III* の $(b)$ では，この 4 種類の質問の検索結果を比較して，異なる側面からの検索ができたかを考察する．

実験 *II* で使われた 88 個の話題構造に対して，それぞれの質問を生成してそれぞれ検索を行った☆．1 つの話題構造に対して，それぞれの検索質問は最大 10 ページを返すとした．異なる質問による検索結果の比較を行った．次の 3 つの項目で比較を行った．

- 解（検索されたページ）集合間の平均類似度：$A$ 質問と $B$ 質問の解集合間の平均類似度 $S(A,B)$ は，ベクトル空間モデルに基づいて次のように計算される．

$$S(A,B) = \frac{1}{N} \cdot \sum_{k=1}^{N} \frac{\sum_{i=1}^{m_k} \sum_{j=1}^{n_k} sim(p_i, p_j)}{m_k \cdot n_k}$$

$$sim(p_i, p_j) = \frac{V(p_i) \cdot V(p_j)}{|V(p_i)| \times |V(p_j)|} \qquad (12)$$

ただし，$N$ は，話題構造を用いて生成された質問の数である．今回の実験では，$N = 88$ である．$m_k$ $(\leq 10)$ と $n_k$ $(\leq 10)$ は，$k$ 番目の話題構造に基づいて生成された質問 $a_k$（A 質問）と $b_k$（B 質問）の解（検索されたページ）の数である．$p_i$ と $p_j$ は，質問 $a_k$ と $b_k$ の解である．$V(p_i)$ と $V(p_j)$ は，Web ページ $p_i$ と $p_j$ のキーワードベクトルである．

- 平均共通解（ページ）数：$A$ 質問と $B$ 質問を用いた検索の平均共通解数 $CR(A,B)$ は，次のように計算される．

$$CR(A,B) = \frac{1}{N} \cdot \sum_{k=1}^{N} |r_{a_k} \cap r_{b_k}| \qquad (13)$$

ただし，$r_{a_k}$ と $r_{b_k}$ は，それぞれ，$k$ 番目の話題構造に基づいて生成された質問 $a_k$（A 質問）と質問 $b_k$（B 質問）の解集合である．$|r_{a_k} \cap r_{b_k}|$ は，$r_{a_k}$ と $r_{b_k}$ の共通解の数を表す．

- 平均検索解（ページ）数：$A$ 質問の平均検索解数 $Num(A)$ は，次のように計算される．

$$Num(A) = \frac{1}{N} \cdot \sum_{k=1}^{N} m_k \qquad (14)$$

**(a) 類似検索と Query-Free 検索の比較実験**

本論文では，番組の話題構造と同じ話題構造を含むページの検索を類似検索とする．そのための質問を SIM 質問（$Q_{sim}$）と呼ぶ．話題構造 $t = (\{s_1, s_2, \ldots, s_n\}, \{c_1, c_2, \ldots, c_m\})$ に基づく SIM 質問が次のように定義される．

$$Q_{sim} = (\mathit{insubject} : s_1 \wedge s_2 \wedge \ldots \wedge s_n) \wedge (\mathit{incontent} : c_1 \wedge c_2 \wedge \ldots \wedge c_m)$$

Google の構造的な検索オプションを用いて SIM 質問の 5-keys と 3-keys 質問を生成した．5-keys 質問は，次のようになる．

$$\mathit{intitle}{:}s_1 \ \mathit{intitle}{:}s_2 \ \ \mathit{allintext}{:}c_1\ c_2\ c_3 \qquad (15)$$

3-keys 質問は，次のようになる．

$$\mathit{intitle}{:}s_1 \ \ \mathit{allintext}{:}c_1\ c_2 \qquad (16)$$

SIM 質問と Query-Free 検索機構の CD，CB，SB および SD 質問の検索結果の比較を行った．実験結果を**表 4** に示す．なお，SIM 質問の有効（検索されたページの数が 1 以上である）5-keys 質問と有効 3-keys 質問の数は，それぞれ 29 と 45 であった．

一般的に，Web ページの見出しに出現するキーワードは，本文にも出現する可能性が高い．一方，ページの本文に出現するキーワードの見出しに出現する可能性が低い．そのため，SD 質問と SIM 質問は，ともに，見出しに $s_1$，$s_2$ を含み，本文にキーワード $s_1$，$s_2$，$c_1$，$c_2$，$c_3$ を含むようなページを解とすることがある．一方，CD と SIM 質問の場合，共通解（ページ）となりうるのは，見出しにキーワード $s_1$，$s_2$，$c_1$，$c_2$，$c_3$ を含み，本文に $s_1$，$s_2$ を含まない（$c_1$，$c_2$，$c_3$ を含む）ページである．つまり，見出しに多くのキーワードを含み，しかもその中のキーワード（一部）が本文に含まないページである．このようなページは，めったにないと考える．つまり，CD と SIM 質問は，共通解（ページ）を持つ可能性が非常に低いと考える．

CB 質問と SB 質問の定義では，元と同じ話題構造

☆ 実験 *II* と同様，必要に応じて，5-keys と 3-keys 質問を生成する．

**表 4** Query-Free 検索と類似検索の検索結果の比較
Table 4 Comparison evaluation results of similar information retrieval and Query-Free information retrieval.

| | CD と SIM 質問 | SD と SIM 質問 | CB と SIM 質問 | SB と SIM 質問 |
|---|---|---|---|---|
| 解集合間の平均類似度 | 0.097 | 0.109 | 0.184 | 0.149 |
| 平均共通解数 | 0 | 0.125 | 1.568 | 0.318 |
| 平均検索解数 | CD 質問：5.11 件，CB 質問：7.19 件，SB 質問：9.90 件，SD 質問：7.71 件，SIM 質問：6.1 件 | | | |

**表 5** CD，SD，SB と CB 質問の検索結果の比較
Table 5 Comparison evaluation results of CD, SD, SB and CB queries.

| | CD と SD 質問 | CD と CB 質問 | CD と SB 質問 | SD と CB 質問 | SD と SB 質問 | CB と SB 質問 |
|---|---|---|---|---|---|---|
| 解集合間の平均類似度 | 0.096 | 0.098 | 0.149 | 0.105 | 0.134 | 0.132 |
| 平均共通解数 | 0 | 0 | 0.0568 | 0.193 | 0.171 | 0.091 |
| 平均検索解数 | CD 質問：5.11 件，CB 質問：7.19 件，SB 質問：9.90 件，SD 質問：7.71 件 | | | | | |

を含む Web ページを排除する．しかし，実験では，元の話題構造の部分を含む Web ページを排除できない場合があった．元の話題構造を $(\{a,b\},\{c,d,e\})$ とした場合，CB 質問の 5-keys 質問は，$(\{a,b\},\{c,d\})$ のような話題構造を含む Web ページを排除できない．同様に，SB 質問の 5-keys 質問は，$(\{a\},\{c,d,e\})$ のような話題構造を含む Web ページを排除できない．このようなページが，SIM 質問の 3-keys 質問で検索されることがある．したがって，実験では，SB 質問と SIM 質問，および CB 質問と SIM 質問が，それぞれ共通の検索結果を持つことがあった．理論上，両者の質問（CB と SB 質問）は，5-keys と 3-keys という質問の種別は同じである限り，SIM 質問との共通解は存在しない．しかし，実験では，5-keys と 3-keys 質問を区別して取り扱っておらず，しかも，Google の実装・仕様に依存しているため，比較的に多くの共通解が見られた．

実験結果から，SIM 質問と Query-Free 検索の質問（CD，CB，SB と SD 質問）の解集合間の平均類似度が低く，しかも，平均共通解数が少ないことが分かった．つまり，SIM 質問と Query-Free 検索の質問による検索結果の差が大きい．このことから，Query-Free 検索機構の質問は類似検索と異なり，単なる類似検索ではないと考えられる．

類似検索（SIM 質問）の検索結果を，実験 *II* と同様な方法で評価を行った．表 3 に実験結果を示す．計 47 件の補完ページがあったと判定した☆．また，46 件の類似ページがあったと判定した．その中，39 件は補完ページであった．これらの結果から，SIM 質問を用いた類似検索は，補完情報を獲得することが可能であるが，提案手法のような異なる側面（より詳細や別の観点など）を意識した補完情報の獲得は困難であることが分かる．

☆ 検索されたページの中に，複数の補完効果（より詳細，別の観点 etc.）のあるページがあった．

**(b) CD，CB，SD と SB 質問の比較実験**

Query-Free 検索機構の CD，CB，SD と SB 質問の比較を行った．**表 5** は，実験結果を示している．

実験結果から，これらの質問は，共通の検索解が少なく，しかも解集合間の類似度が低いということが分かった．単に Google のトップ検索結果を利用していることと適合率の結果（実験 *II*）を加えて考えると，本論文で提案している手法は，異なる側面から番組の補完ページを検索可能であると考える．しかし，異なる質問であるのに，共通の検索結果のある場合があった．また，検索式の肯定条件部の intitle と intext の違いによって，検索結果の数は差が大きかった．その原因としては，次のようなものがあると考える．

- Web ページに複数の話題が含まれていることがある．つまり，同じページに，番組の話題構造と結合可能な異なる話題構造が含まれる場合がある．したがって，同じページは，異なる質問の検索結果になりうる．
- 現状では，*title* タグを使わない Web ページが多数存在する．あるいは，*title* タグに，サイト名などの，ページの本来の見出しと違うものを書いている場合が多い．たとえば，すべてのページの *title* タグにサイトの名前が記述されているサイトがある．また，ニュースサイト（たとえば，asahi.com（http://www.asahi.com））では，ニュースのジャンルを *title* タグに記述する場合がある．Google のタイトル検索は，このような違いを区別できない場合が多い．
- 5-keys と 3-keys 質問の違いによって，共通の検索解を持つことがありうる．

**表 6** 否定条件部を含まない質問（実験 *IV*）
Table 6 Query without negative condition part.

| | 5-keys query | 3-keys query |
|---|---|---|
| $SB'$ 質問 | $intext:c_1$ $intext:c_2$ $intext:c_3$ | $intext:c_1$ $intext:c_2$ |
| $SD'$ 質問 | $intext:s_1$ $intext:s_2$ | $intext:s_1$ |
| $CB'$ 質問 | $intitle:s_1$ $intitle:s_2$ | $intitle:s_1$ |
| $CD'$ 質問 | $intitle:c_1$ $intitle:c_2$ $intitle:\ c_3$ | $intitle:c_1$ $intitle:c_2$ |

**表 7** 実験 *IV* の結果：結果否定条件部のある場合とない場合の検索結果比較
Table 7 Results of evalution *IV*.

| | | 解集合間の平均類似度 | 平均共通解数 | 検索結果の数（平均） | 正解数 | 適合率 |
|---|---|---|---|---|---|---|
| SB 質問 | 否定部のある場合 | 0.299 | 3.045 | 9.73 | 62 | 0.705 |
| | 否定部のない場合 | | | 9.89 | 52 | 0.591 |
| SD 質問 | 否定部のある場合 | 0.024 | 0.011 | 7.72 | 55 | 0.625 |
| | 否定部のない場合 | | | 10 | 43 | 0.489 |
| CB 質問 | 否定部のある場合 | 0.12 | 0.715 | 7.19 | 45 | 0.511 |
| | 否定部のない場合 | | | 8.61 | 38 | 0.432 |
| CD 質問 | 否定部のある場合 | 0.422 | 3.602 | 5.11 | 43 | 0.489 |
| | 否定部のない場合 | | | 5.30 | 42 | 0.477 |

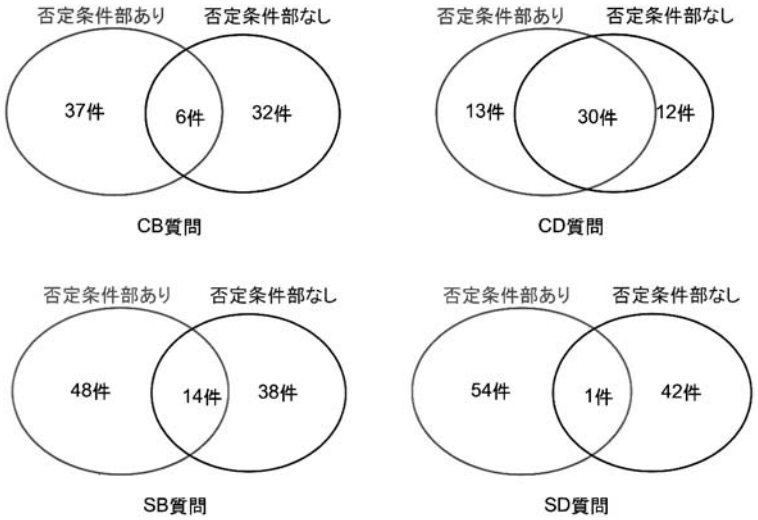


**図 7** 実験 *IV* の結果：否定条件部のある場合とない場合の正解ページの分布
Fig. 7 Results of evalution *IV*.

### 5.4 実験 *IV*：否定条件部の評価実験

Query-Free 検索機構の検索式の否定条件部による検索への影響を考察するため，実験 *III* と同様の実験環境で，実験 *IV* を行った．同じ 88 個の話題構造を利用して，否定条件部を含まない質問を生成して，検索を行った．否定条件部を含まない質問は**表 6** に示されている．Google から返された結果のトップページを番組の関連ページとした．実験 *II* と同様の基準で，そのページが正解であるかどうかを判断した．また，1 つの話題構造に対して，それぞれの検索質問は最大 10 ページを返すとした．実験 *III* の結果を利用して，否定条件部を含む場合と含まない場合の，解集合間の平均類似度，平均共通解数と平均検索解数を計算した．その結果を，**表 7** に示す．表 7 で示されている，否定条件部のある場合のそれぞれの質問の正解数と適合率は，実験 *II* の結果である．また，**図 7** に，否定条件部を含む質問と含まない質問による正解ページの分布を示す．図では，否定条件部を含む質問と含まない質問の正解ページの集合を，それぞれの楕円で示す．

実験結果から，否定条件部の有無によって，SB，SD，CB と CD 質問は検索結果が異なることが分かった．

Google を利用して Web 検索を行っているため，再現率の計算が困難である．そのため，検索漏れに関する評価が困難である．適合率の結果をみると，今回の提案式は，検索漏れの可能性があるが，検索結果の絞り込み効果があると分かった．SB 質問で，単純に content-term をキーに検索した場合より，subject-term に関する否定条件を加えた場合には，適合率が百分率で 11.4 ポイント向上した．SD，CB と CD 質問の場合，適合率がそれぞれ，13.6 ポイント，7.9 ポイントと 1.2 ポイント向上した．これらの結果から，提案した検索式は，番組の内容を補完できる Web ページの検索できるということが分かる．

本論文では，Web ページの話題構造の subject-term と content-term がそれぞれ見出しと本文に含まれると想定し，Google のタイトル検索と本文検索を用いて，元の話題構造と結合可能な話題構造（元の話題構造を除く）を近似的に検索している．しかし，Web ページの実際の話題構造がもっと複雑であり，これらの想定条件を満たさない場合がある．たとえば，検索式の否定条件部に含まれるキーワードは，必ずしもそのページの話題構造の subject-term（または content-term）であるとは限らない．そのため，統合情報となりうるページが検索されない可能性がある．実験では，そのようなページが多く見られた．その対策について，6 章で考案する．また，Query-Free 検索機構の各々の質問の肯定条件部は，検索結果を絞り込むため，“∧”（AND）で組み立てられている．そのため，結合可能な話題構造を含む Web ページの検索されない可能性がある．実際の応用の場に応じて，“∨”（OR）で組み立てることも考えられる．

## 6. 考　　察

### 6.1 話題構造集合の簡約

4.1 節で述べた手法で抽出される話題構造は，主題部と内容部がサブ話題構造を含まない単純なものであった．これは，オンラインで話題構造を抽出し，関連情報を検索するには適切であるが，話題構造の間の関連を考慮していない．特に，Web ページやテキストストリームには複数の話題構造が含まれるので，関連性のある（結合可能である）話題構造を含み，冗長である場合がある．たとえば，あるテキストストリームの話題構造は，$\{t_1 = (\{\text{環境}\}, \{\text{ごみ}, \text{オゾン}\}), t_2 = (\{\text{ごみ}\}, \{\text{リサイクル}, \text{ダイオキシン}\})\}$ であるとする．$t_1$ と $t_2$ が結合可能であり，関連性のある話題構造である．この場合，$t_1$ と $t_2$ を結合してより簡潔な話題構造を得ることができる．なお，この性質（話題構造の集合に結合可能な異なる話題構造を含む）は話題構造の集合の結合に，次のような影響を及ぼす．

- 価値のある結合を見逃す可能性がある．たとえば，**図 8** (a) に示されているように，$T$ に 2 つ結合可能な話題構造 $t_1$ と $t_2$ がある．$t_1 \ltimes t_3 = \phi$ なので，単に $T \ltimes S$ を行うと，我々は $t_1 \ltimes t_2 \ltimes t_3$ のような結合を見逃す可能性がある．
- 結合結果に結合可能な異なる話題構造がある．図 8 (a) の例では，$T^*$（$= T \ltimes S$）が 2 つの結合可能な話題構造を持つ．つまり，最大の結合結果が得られない．

これらを回避するため，我々は，話題構造の集合を

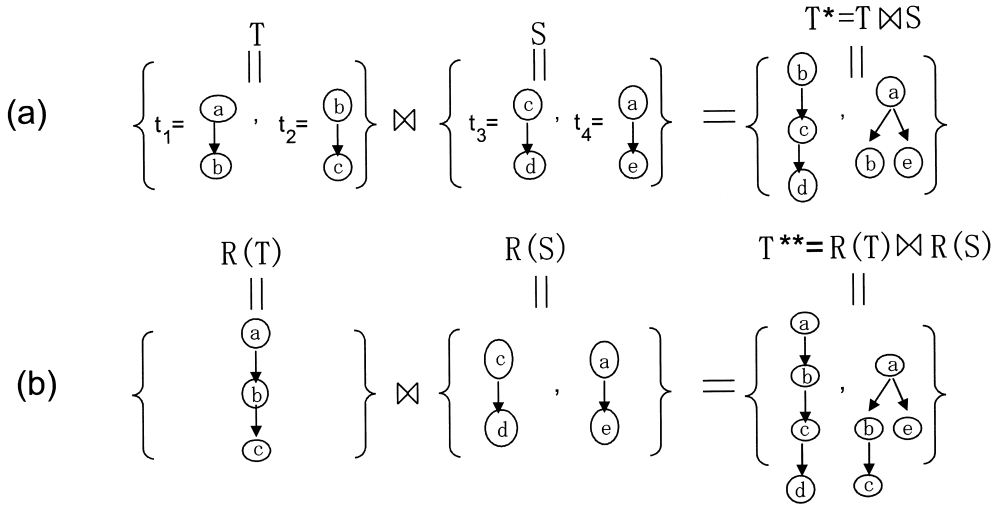


**図 8** 話題構造集合の簡約例
Fig. 8 Examples of reduction.

次のように簡約することを考えている．

**定義 6（話題構造の集合の簡約）** 話題構造の集合 $T = \{t_1, t_2, \ldots, t_n\}$ の簡約 $R(T)$ は次のように行われる．

$$R(T) = G(t_1) \cup G(t_2) \cup \ldots \cup G(t_n) \qquad (17)$$

ただし，$G(t_i)$ は $t_i$ の話題グラフである．$G(t_1) \cup G(t_2) \cup \ldots \cup G(t_n)$ は DAG である．

ある話題構造の集合 $X$ に対して，$R(X) = X$ であれば，$X$ が簡約済みであるという．図 8 では，いくつかの簡約の例を示している．ここで注意すべきなのは，例で示されているように，$T$ と $S$ を簡約しても，結合結果が簡約済みであるとは限らない．つまり，$R(R(T) \bowtie R(S)) \neq R(T) \bowtie R(S)$．したがって，価値のある結合を見逃さず，かつ，簡約済みの結合結果を得るためには，結合前後の両方で話題構造の集合の簡約化が必要である．

話題構造の簡約を利用して，複数の話題を持つ Web ページやテキストストリームの話題構造の集合を抽出することが可能である．たとえば，Web ページに対して，各段落が 1 つの話題について述べていると想定して，それぞれの話題構造を抽出する．そして，これらの話題構造から構成される話題構造集合を簡約して，Web ページの話題構造集合とすることが考えられる．

### 6.2 検索機構

Query-Free 検索機構の 4 種類の検索式は，結合モデルに基づいて組み立てられている．与えられる話題構造および検索される Web ページに含まれている話題構造は単純なものであると想定している．そのうえ，Web ページの話題構造を構成する，subject-term と content-term がそれぞれ見出しと本文に現れるとしている．実際，Web ページに含まれる話題構造がもっと複雑であり，subject-term と content-term は，必ずしも見出しと本文に出現するとは限らない．今回の検索式の前提条件を満たさないため，このようなページが検索されない可能性がある☆．また，我々は，Google の構造的な検索オプション（intitle，intext，allintitle，allintext）を利用してこれらの検索式を実現している．しかし，現状では，“title” タグの使われ方がさまざまであるため，検索結果の漏れや誤検索が多いと考える．なお，Google の検索サービスに依存しているので，検索失敗の原因の解析が困難である場合がある．

☆ 検索式の前提条件と検索式を満たす話題構造が別にあれば，検索可能である．

これらの問題点を解決するための 1 つの案として，Google のタイトル検索や本文検索に依存せず，とりあえず検索を行い，その検索結果のフィルタリングを行う手法が考えられる．たとえば，まず，元の話題構造にあるキーワード（全部または部分）を利用して検索を行う．あるいは，提案した 4 種類の質問を組み合わせて検索を行う．もちろん，5 章で述べた SIM 質問の利用も考えられる．そして，検索されたページの話題構造（集合）を抽出して，その中に元の話題構造と結合可能な話題構造があるか，どのぐらいあるかなどを調べて，それに基づく検索結果の絞り込みを行う．この手法は，Google のタイトル・本文検索に依存することが少なくなり，Web ページの “title” タグの使われ方の影響を緩和することが期待できると考える．また，検索式の否定条件部による検索漏れもある程度避けられると考える．さらに，こうすることによって，ページ全体だけではなく，ページの部分（結合可能な話題構造を含む部分）を情報統合の単位とすることも可能となる．これらについては，今後の研究で検討したいと思う．

## 7. おわりに

本論文では，放送と Web コンテンツの統合のための，話題構造に基づく結合モデルを提案した．さらに，テキストストリームのセグメンテーションと話題構造抽出手法を提案し，話題構造の結合モデルに基づいて，番組の内容を補完する Web ページを検索するための Query-Free 検索機構を提案した．また，評価実験を行った．実験結果から，我々の提案手法が放送コンテンツの補完情報を検索するためには有効であることが分かった．つまり，番組の内容と単に類似するのではなく，番組の内容をより詳しくまたは別の視点から述べているページを検索することが可能である．さらに，本論文では，放送と Web コンテンツの統合を，話題構造の結合モデルによって定式化した．

実験結果から，話題構造における subject-term と content-term の区別は有効であり，それを用いて従来の類似検索と異なる検索が可能であることが分かった．本論文で提案する補完情報の検索手法は，その 1 つの提案である．このほか，5 章で述べた類似検索のような，結合モデルに依存しない話題構造に基づく検索手法も考えられる．たとえば，主題に類似する情報，内容に類似する情報といった異なる側面からの情報検索が考えられる．

我々は，以前の研究で，放送と Web の動的統合システム WebTelop の提案と開発を行った[37]．WebTelop

は，Web から番組の関連情報を動的に取得し，番組と連動してユーザに提示するシステムである．本論文の提案手法および話題構造の結合の性質に基づいて，WebTelop の改良を行う予定である．

本論文では，受信中の字幕データのセグメンテーションと話題構造の抽出手法を提案しているが，その有効性を確認するため，従来手法との比較が必要であると思われる．今後，セグメンテーションと話題構造抽出手法の改良とともに，話題構造の結合の性質を解明し，Query-Free 検索機構とその応用システムの設計，解析および改良に用いる予定である．特に，結合モデルの性質，およびそれに基づいて定義された検索式の検索への影響をさらに解明する必要があると思う．

**謝辞** 著者の一部は，平成 16 年度文部科学省科学研究費特定領域研究（2）「Web の意味構造に基づく新しい Web 検索サービス方式に関する研究」および平成 14〜16 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」の助成をうけている．ここに記して謝意を表します．

## 参 考 文 献

1) MPEG-7: ISO/IEC JTC1/SC29/WG11 N4980.
2) 亀山 渉, 花村 剛：MPEG-7/MPEG-21/TV-Anytime デジタル放送教科書（上・下），IDG ジャパン (2002).
3) CANAL+TECHNOLOGIES (2003). http://www.canalplus-technologies.com
4) TV-Anytime Forum (2003). http://www.tv-anytime.org/
5) 馬 強, 角谷和俊, 田中克己：ストリームデータの統合・フィルタリング関数とその応用，電子情報通信学会信学技報 DE2002-6（2002-05），pp.29–34 (2002).
6) 馬 強, 角谷和俊, 田中克己：WebTelop：放送と Web コンテンツの動的統合システム，情報処理学会研究報告，Vol.2002, No.67, 2002-DBS-128-23, pp.169–176 (2002).
7) Ma, Q., Kondo, H., Sumiya, K. and Tanaka, K.: Virtual TV Channel: Filtering, Merging and Presenting Internet Broadcasting Channels, *Proc. ACM Digital Library Workshop On Organizing Web Space* (*WOWS*), pp.32–43 (1999).
8) Janevski, A. and Dimitrova, N.: Web Information Extraction for content augmentation, *Proc. IEEE International Conference on Multimedia and Expo 2002* (*ICME 2002*) (2002).
9) Tanaka, K., Nadamoto, A., Kusahara, M., Hattori, T., Kondo, H. and Sumiya, K.: Back to the TV: Information Visualization Interfaces based on TV-Program Metaphors, *Proc. IEEE International Conference on Multimedia and Expo 2000* (*ICME 2000*), Vol.3, pp.1229–1232 (2000).
10) 馬 強，田中克己：話題構造に基づくコンテンツ結合演算とその応用，情報処理学会研究報告，2003-DBS-131, pp.153–159 (2003).
11) MicrosoftTV (2003). http://www.microsoft.com
12) TopicMap (2003). http://www.topicmap.org
13) Wayne, C.L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, *Proc. Language Resources and Evaluation Conference* (*LREC*) *2000*, pp.1487–1494 (2000).
14) 今井 亨，リチャードシュワルツ，小林彰夫，安藤彰男：話題混合モデルによる放送ニュースからの話題抽出，電子情報通信学会論文誌，Vol.J81-D-II, No.9, pp.1955–1964 (1998).
15) 竹下 敦，井上孝史，田中一男：テキストの概要把握支援のための話題構造抽出，情報処理学会論文誌，Vol.37, No.11-007, p.1941 (1996).
16) 遠山義洋，西田豊明：話題構造の抽出と変形による対話録の自動要約，2000 年度人工知能学会全国大会（第 14 回）論文集，pp.157–160 (2000).
17) Zloof, M.: Query-By-Example: A Data Base Language, *IBM Systems Journal*, Vol.16, No.4, pp.324–343 (1977).
18) Henzinger, M., Chang, B.-W., Milch, B. and Brin, S.: Query-Free News Search, *Proc. 12th International World Wide Web Conference* (2003).
19) Mishra, P. and Eich, M.H.: Join Processing in Relational Databases, *ACM Computing Surveys*, Vol.24, No.1, pp.63–112 (1992).
20) Guha, S., Jagadish, H.V. and Koudas, N.: Aproximate XML Joins, *Proc. 2002 ACM SIGMOD International Conference on Management of Data*, pp.287–298 (2002).
21) Bhowmick, S.S., Ng, W.K., Lim, E.-P. and Madria, S.K.: Join Processing in Web Databases, *Proc. 9th International Conference on Database and Expert Systems* (*DEXA 98*), pp.647–657 (1998).
22) Ng, W.K., Lim, E.-P., Huang, C.-T., Bhowmick, S.S. and Qin, F.-Q.: Web Warehousing: An Algebra for Web Information, *Proc. Advances in Digital Libraries 1998*, pp.228–237 (1998).
23) Maybury, M.: *Intelligent Multimedia Information Retrieval*, AAAI Press and MIT Press

(1997).
24) Wactlar, H.D.: Informedia — Search and Summarization in the Video Medium, *Proc. Imagina 2000 Conference* (2000).
25) Mani, I., House, D., Maybury, M. and Green, M.: Towards content-based browsing of broadcast news video, *Intelligent multimedia information retrieval*, chapter 12, Maybury, M. (Ed.), pp.241–258 (1997).
26) Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J.: Models and Issues in Data Stream Systems, *Proc. 2002 ACM Symp. on Principles of Database Systems* (*PODS 2002*), pp.1–16 (2002).
27) Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Seidman, G., Stonebraker, M., Tatbul, N. and Zdonik, S.B.: Monitoring Streams — A New Class of Data Management Applications, *Proc. 28th International Conference on Very Large Dada Bases* (*VLDB 02*), pp.215–226 (2002).
28) Sullivan, M. and Heybey, A.: Tribeca: A system for managing large databases of network traffic, *Proc. 1996 USENIX Annual Technical Conference*, pp.13–24 (1998).
29) Tucker, P., Maier, D., Sheard, T. and Fegaras, L.: Punctuating Continuous Streams, Technical Report, Oregon Graduate Institute (2002).
30) 有田英一，岡　隆一：新聞記事テキストデーからの断片的知識の連鎖の抽出，電子情報通信学会技術研究報告，NLC93-66, pp.23–30 (1993).
31) 前田晴美，糀谷和人，西田豊明：連想構造を用いた情報整理システム，情報処理学会論文誌，Vol.38, No.3, pp.616–625 (1997).
32) 村上晴美，平田高志：WWW からの情報獲得・整理支援—思考・興味ブラウザ，情報処理学会研究報告 FI-142-23，pp.167–174 (2001).
33) Matsukura, T., Kondo, H., Hirata, Y. and Tanaka, K.: Discovery of Semantic Relationships among Web Pages Based on Web Topic Structures, *Proc. 9th IFIP 2.6 Working Conference on Database Semantics*, pp.184–199 (2001).
34) Oyama, S. and Tanaka, K.: Exploiting Document Structures for Comparing and Exploring Topics on the Web, *Proc. 12th International World Wide Web Conference* (*WWW2003*) (*poster tracks*) (2003).
35) ChaSen (2003). http://chasen.aist-nara.ac.jp/index.html.en
36) Google Web APIs (2003). http://www.google.com/apis/
37) Ma, Q. and Tanaka, K.: WebTelop: Dynamic TV-content Augmentation by Using Web Pages, *Proc. IEEE International Conference on Multimedia and Expo* (*ICME2003*), Vol.2, pp.173–176 (2003).

(平成 16 年 3 月 20 日受付)
(平成 16 年 7 月 14 日採録)

(担当編集委員　飯沢 篤志)

**馬　　強**（正会員）

1998 年広島県立大学経営学部経営情報学科卒業．2000 年神戸大学大学院自然科学研究科博士前期課程修了．2004 年京都大学大学院情報学研究科博士後期課程修了．同年独立行政法人情報通信研究機構（旧，独立行政法人通信総合研究所）入所，現在に至る．博士（情報学）．主に情報検索，情報統合とマルチメディア情報システムの研究に従事．IEEE，ACM 等各会員．

**田中　克己**（正会員）

1974 年京都大学工学部情報工学科卒業．1976 年京都大学大学院修士課程修了．1979 年神戸大学教養部助手．1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学科）．1995 年同大学大学院自然科学研究科情報メディア科学専攻専任教授．2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．2003 年から独立行政法人情報通信研究機構（旧，独立行政法人通信総合研究所）メディアインタラクショングループリーダー兼任．京都大学工学博士．主にデータベースとマルチメディア情報システムの研究に従事．人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．