社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
4C-i6

テキストストリームの文脈を考慮した補完情報検索

馬　強†　　田中　克己†,††

† 独立行政法人情報通信研究機構　メディアインタラクショングループ
〒 619-0289 京都府相楽郡精華町光台 3-5
†† 京都大学大学院情報学研究科
〒 606-8501 京都市左京区吉田本町
E-mail: †qiang@nict.go.jp, ††tanaka@dl.kuis.kyoto-u.ac.jp

あらまし　情報技術の進歩に伴い，アクセス可能な情報の種類や量が増大しつつあると同時に，ユーザの情報要求も多様化しつつある．従来の情報検索手法の多くは，ユーザの検索要求質問に最も合致・類似する情報をいかに効率よく検索するかに焦点を当ててきた．これに対して，我々は，類似情報の検索の性能向上のための方式ではなく，ある情報がユーザに呈示された場合に，この呈示された情報に対して，より詳細な情報や同一主題であるが異なる話題を含む情報を「補完情報」と呼び，このような補完情報の検索を行うための手法を提案してきた． 特に，情報の話題構造の抽出に基づき，テレビ番組の内容を補足する Web 情報の検索手法を開発してきた．本稿では，従来の我々の話題構造モデルの改良を行うと共に，番組の字幕情報のようなテキストストリームの「コンテキスト」を考慮した補完情報検索手法を提案する．テキストストリームのコンテキストとは，関連ある話題構造の系列であり，これを元にした，補完情報検索のための質問種別の自動選択，質問の自動修正，および，補完度計算の方法を提案する．実験結果から，提案する手法は，適合率の向上のほか，検索される補完情報の重複を防ぐ効果があることがわかる．
キーワード　情報補完，情報検索，話題構造，補完度，コンテキスト

Context-sensitive Complementary Information Retrieval

Qiang MA† and Katsumi TANAKA†,††

† Interactive Communication Media and Content Group
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo, Kyoto 606-8501,Japan
E-mail: †qiang@nict.go.jp, ††tanaka@dl.kuis.kyoto-u.ac.jp

Abstract　With the progressing of information technology, more and more information becomes available and the user's information needs are becoming more diverse. Most of the conventional information systems try to only provide information in which users are interested. In contrast, we are working on how to discover the information from both interest and necessary viewpoints. We have proposed some methods to search for complementary information of a TV program in term of providing information in detail or from different viewpoints to the users. In this paper, as one of the improvements, we propose a context-sensitive complementary information retrieval method. The "context" of text stream is considered as a series of topic structure. Based on such context, we propose the methods for search for complementary information of a TV-program, including query type selection, query modification and complementarity degree computation. The experiment results show that the context sensitive method could improve the precision ratio and avoid the overlap of complementary information.
Key words　Information Complementation, Information Retrieval, Topic Structure, Complementarity Degree, Context

1. はじめに

インターネットを代表とするIT技術の進歩に伴って，ユーザのアクセス可能な情報の量が莫大になってきた．ブロードバンドの普及に伴い，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，本放送と共に，番組のメタデータなどの関連情報が配信されることがある [1]．このように，放送と通信の融合が着実に進んでいる．すなわち，ユーザが，情報獲得のために利用されるデバイスを意識する必要が無くなる．放送と通信のような異なる性質を持つメディアのコンテンツを統合することによって，よりリッチな情報を提供することが可能である．

大量の情報から，ユーザの興味のある情報を獲得するには，情報検索とフィルタリング技術が有効である．従来の研究の多くは，ユーザの興味に即した情報の獲得手法に着目している．つまり，ユーザの興味を表す質問またはユーザプロファイルを手動・自動的に生成して，その質問またはユーザプロファイルにマッチした情報を提供するのである．しかしながら，ユーザの興味が時間の経過とともに変化することやユーザの検索に関するノウハウの不足などから，ユーザの興味に即した質問・ユーザプロファイルの生成が困難である場合がある．

明示的な質問記述を必要とせず，ユーザから与えられた例題に基づいて質問を自動生成して，その例題と類似するものを獲得（または排除）する手法として，QBE（Query By Example）[10] とその改良手法がよく知られている．また，Google ラボでは，ニュース番組の内容に類似する Web ページを自動的に検索する手法を提案している [11]．彼らは，時間幅で分割された字幕データのセグメントから，$tf \cdot idf$ ベースの手法を用いてキーワードを抽出して，番組の類似情報を検索する．

ユーザの興味に即した情報のみを提供することは，情報の“ 偏食 ”であり，知識のアンバランスとなりがちである．一方，テレビ番組などの放送コンテンツの場合，オンエア時間の制限や不特定多数のユーザに情報を提供する必要があるため，情報の詳細や幅が限られている場合がある．せっかく美味しい“ 食材 ”（番組コンテンツ）があるのに，満喫できない場合がある．このような情報の“ バランス ”および“ 美味しさ ”に着目して，我々は，情報補完について研究を行い，その一つの試みとして，ユーザの興味のある放送コンテンツをより詳しくまたは別の観点から述べている情報の検索およびフィルタリング手法を提案してきた [3]〜[6]．図 1 は，我々の放送コンテンツの補完情報の検索手法の処理フローを示している．まず，放送コンテンツの字幕データをセグメンテーションして，放送コンテンツの話題構造（系列）を抽出する．そして，それぞれのセグメントに対応する話題構造を用いて，補完情報を検索するための構造化質問を生成し，Web 検索を行う．さらに，検索されたページを，補完度という概念に基づいて再ランキングし，最終解を選択する．補完度は，元の情報を補完する程度を測る尺度であり，話題構造同士の比較に基づいて計算される．

我々の以前の研究では，番組の字幕データから抽出される話題構造はそれぞれ独立であると想定した．しかしながら，番組は，連続性の高いストリーム型コンテンツであるため，同じ話題構造を繰り返し抽出する可能性が高い．そのため，ユーザに同じページを提示する可能性があり，情報の重複となる場合がある．そこで，本稿では，放送番組の字幕データがテキストストリームであることに着目して，話題構造のコンテキストを考慮した補完情報検索手法を提案する．話題構造のコンテキストは，過去の関連する話題構造系列であり，話題構造の結合に基づいて計算される．このようなコンテキストを考慮することによって，番組の現在の内容をより正確に表現でき，同じ話題構造であっても異なる質問の生成が可能となる．つまり，検索される補完ページの重複を防ぐ効果が期待できる。コンテキストを考慮した補完情報検索は，主に以下のような３つの部分からなる．

- コンテキストによる質問種類の自動選択

コンテキストと現在の話題構造の比較を行い，生成される構造化質問の種類を決める．もし，今までのコンテンツ（コンテキスト）に，現在のコンテンツ（話題構造）を詳しく述べている部分を，より多く含んであれば，話題を広げるための情報を検索する質問を選ぶ．逆に，もし，今までのコンテンツは，いろんな話題を含んであれば，詳細情報を検索するための質問を生成する．

- コンテキストによる質問修正

一般に，字幕データから抽出される現在の話題構造を用いて質問を生成する．しかしながら，抽出された話題構造のストリームにリピートする話題構造がある場合，同じ質問の生成される可能性がある．このような重複を避けてより多くの情報を提供するため，我々は，必要に応じて，現在の話題構造またはコンテキストを利用して質問を生成する．

- コンテキストを考慮した補完度計算

基本的に，補完度は，検索されたページと現在の番組の話題構造の比較に基づいて計算される．コンテキストによって決められた質問は，より詳細な情報を求めるのであれば，話題構造の高さ（情報の詳細を表す）の差分に基づいて補完度を計算する．話題を広げるための質問であれば，話題構造の幅（話題のカーバーする範囲を表す）の差分で補完度を計算する．

以下，本稿の構成を示す．2. 節では，話題構造のグラフ表現について述べる．話題構造のコンテキストを考慮した補完質問の自動生成・修正，および補完度計算については，3. 節で述べる．4. 節では，実験と考察について述べる．5. 節では，まとめと今後の課題について述べる．

2. 話題構造

我々は，一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，話題構造を用いて表現する．従来の研究 [7]〜[9] 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，一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 とよぶ）の集合と本文を表す役割を

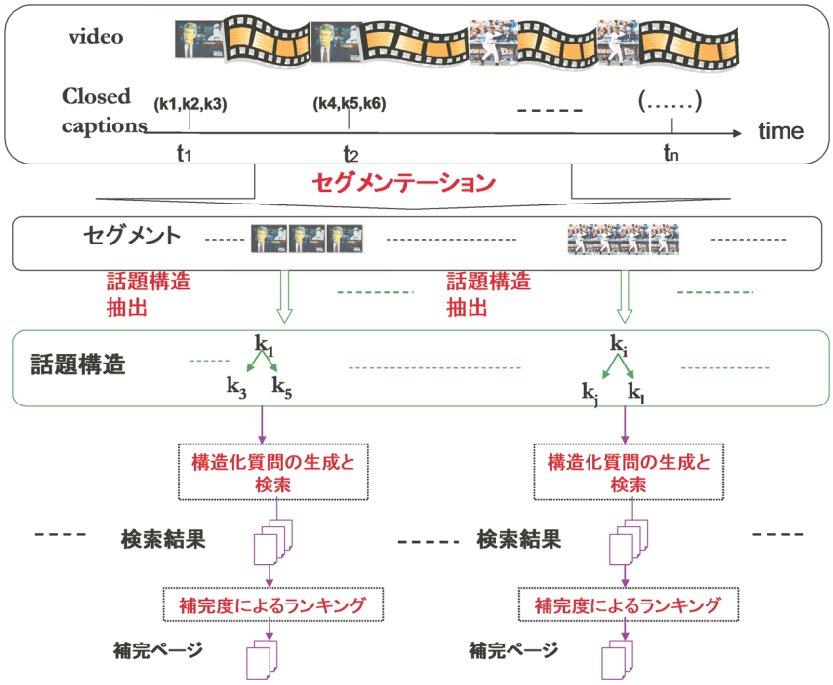


図 1　補完情報検索

持つキーワード（content-term とよぶ）の集合のペアである．

話題構造は，次のように定義されている．

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

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，定義の通りに，subject-term と content-term は，キーワードである．さらに，あるキーワードは一つの話題構造において高々1 回しか現れないとする．ここでは，“ + ”は一回以上出現することを意味する．“ | ”は，“ or ”を意味する．

文献 [5] で述べられているように，subject-term と content-term は，共起関係と tf(term frequency) によって抽出される．基本的に，ある文書の中に，その他の語との共起関係が強く，しかも，出現頻度の高い語が subject-term となる．その文書の中にある，subject-term との共起関係の強い語が content-term となる．なお，二つの語 w_i と w_j の共起関係 $cooc(w_i, w_j)$ が，次のように定義されている．

$$
cooc(w_i, w_j) = \frac{df(\{w_i, w_j\})}{df(\{w_i\}) + df(\{w_j\}) - df(\{w_i, w_j\})} \tag{2}
$$

ただし，$df(\{w_i\})$ は，あるコーパスにおける w_i を含む文書の数である．

2.1　話題グラフ

一般に，話題構造は二つ以上のノードを持つ，一つの連結成分からなる重み付き DAG（Directed Acyclic Graph）を用いて表現できる．このような一つの連結成分からなる重み付き DAG を話題グラフと呼ぶ．

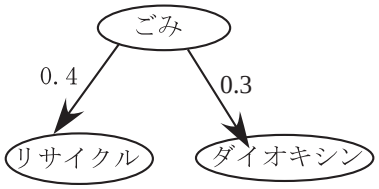


図 2　話題グラフの例

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．$E(\subseteqq V \times V)$ は重付き枝の集合である．枝の重みは，重み関数 $w : E \to R$ によって計算される．枝 $e = (u, v)$ 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．また，$|V| \geqq 2, E \neq \emptyset$ である．図 2 では，話題グラフの例を示している．

重み関数 w は，それぞれの応用場面に応じて定義できるものとする．本稿では，話題グラフの高さは，Web ページや番組などのコンテンツがその話題を詳しく述べている程度を表すと想定する．一方，話題グラフの幅は，Web ページや番組の内容のカバーする範囲を表すものである．この考えから，本稿では，話題グラフにおける二つのノード u と v の距離 $d(u, v)$ を，次のように定義する．

$$
d(u, v) = (1 - cooc(u, v)) \cdot \frac{min(tf(u), tf(v))}{max(tf(u), tf(v))} \tag{4}
$$

ただし，$tf(u)$ は，キーワード u の出現頻度を表す．$cooc(u, v)$

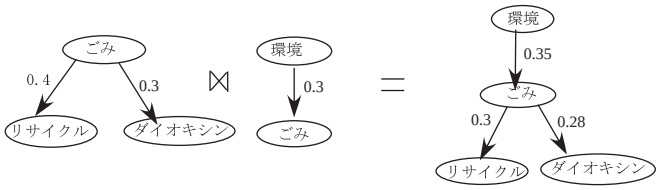


図 3 結合の例

は，u と v の共起関係を表し，予め用意した共起辞書を用いて調べられる．max と min は，それぞれ最大値と最小値を求める関数である．また，$e = (u, v) \in E$ の時，$d(u, v)$ を e の重みとする．

話題グラフの高さは，親を持たない節点（入次数は 0 であり，根節点と呼ぶ.）から，子供を持たない節点（出次数は 0 であり，葉節点と呼ぶ.）に到達までに通るパスの距離の最大値である．

一方，葉節点間と根節点間の距離の大きい方は，話題グラフの幅となる．葉節点（根節点）間の距離を求めるために，まず，葉節点（根節点）集合をノード集合とする完全グラフを構成する．この完全グラフの枝の重みを，話題グラフの重み関数（式 (4)）で計算する．そして，すべてのノードを結ぶ最短パスを求め，この最短パスの距離を葉節点（根節点）間の距離とする．話題グラフの根節点（葉節点）の集合を N とし，根（葉）節点間の最短距離 D を求める手順を以下に示す．

（1） 空のノード集合 M を生成し，D=0 とする．

（2） 任意の節点 $n \in N$ を選び，$M = M + \{n\}, N = N - \{n\}$ とする．

（3） N 内のノードで，n との距離が最短となるノード m を求める．この最短距離を $d(n, m)$ とし，$D = D + d(n, m), N = N - \{m\}, M = M + \{m\}$ とする．

（4） N が空であれば，(7) へ．

（5） N 内のノードで，M 内のノードとの距離が最短となるノード l を求める．この最短距離 を $d(x', l)$ とする．ただし，$x' \in M$.

（6） $D = D + d(x', l), M = M - \{x'\} + \{l\}, N = N - \{l\}$. (4) へ．

（7） D を最短距離として出力する．.

2.2 話題構造の結合

二つの話題構造 t と t' の結合は，この二つの話題構造の話題グラフの和である．ただし，この二つの話題グラフの和は一つの連結成分からなる DAG である必要がある．つまり，二つの話題構造の結合の結果は，話題構造である．

$$t \bowtie t' = \begin{cases} G(t) \cup G(t'), & G(t) \cup G(t') \text{ が一つの} \\ & \text{連結成分からなる} \\ & \text{DAG である場合} \\ \phi, & \text{その他} \end{cases} \quad (5)$$

ただし，$G(t)$ と $G(t')$ は t と t' の話題グラフである．ϕ は空を表す．$t \bowtie \phi = \phi$ とする．また，結合結果が空でなければ，結合された話題グラフの枝の重みは，重み付き関数によって再計算される．図 3 は，結合の例を示している．

結合結果は一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．一つの連結成分という制約条件は，二つの話題構造に共通要素のあることを保証する．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．例えば，話題構造 ({a,b},{(a,b)}) と ({a,b},{(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 b の関係が矛盾となる．

二つの話題構造の結合が空でなければ，この二つの話題構造が結合可能であるという．我々は，補完情報の検索を，与えられたコンテンツの話題構造と結合可能な話題構造を持つコンテンツの検索とする．

3. 文脈を考慮した補完情報検索

3.1 構造化質問

前述したように，我々は，補完情報の検索を，与えられた話題構造と結合可能な話題構造を持つ Web ページの検索とする．このため，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，ノード型質問と枝型質問を定義する．なお，本稿では，質問の生成に利用される話題構造は，二つの根節点と三つの葉節点しか含まないものとする．

これらの質問は，前半の肯定条件部と後半の否定条件部から構成される．肯定条件部は，与えられた話題構造（コンテンツ）の“ 同 ”を求め，否定条件部は,“ 異 ”を求める．つまり，元の情報に似て非なる情報の検索を行い，情報の補完を試みる．与えられた話題構造 t を $(\{s_1, s_2, c_1, c_2, c_3\}, \{(s_1, c_1), (s_1, c_2), (s_1, c_3), (s_2, c_1), (s_2, c_2), (s_2, c_3)\})$ とした場合，補完情報を検索するための構造化質問は，次のように定義される．ただし,“ 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．例えば，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．

- ノード型質問：ノード型質問の肯定条件部は，与えられた話題構造の subject-term または content-term を用いる．
  - CD(Content-Deepening) 質問 (Q_{cd}):

$$Q_{cd} = (insubject : c_1 \wedge c_2 \wedge c_3) \wedge (\neg(incontent : s_1 \vee s_2)) \quad (6)$$

  - SD (Subject-Deepening) 質問 (Q_{sd}):

$$Q_{sd} = (incontent : s_1 \wedge s_2) \wedge (\neg(insubject : c_1 \vee c_2 \vee c_3)) \quad (7)$$

  - SB(Subject-Broadening) 質問 (Q_{sb}):

$$Q_{sb} = (incontent : c_1 \wedge c_2 \wedge c_3) \wedge (\neg(insubject : s_1 \wedge s_2)) \quad (8)$$

  - CB(Content-Broadening Query) 質問 (Q_{cb}):

$$Q_{cb} = (insubject : s_1 \wedge s_2) \wedge (\neg(incontent : c_1 \wedge c_2 \wedge c_3)) \quad (9)$$

- 枝型質問：枝型質問の肯定条件部は，与えられた話題構造の部分グラフ（枝）を用いる．

$$Q_e = (insubject : s_1\ incontent : c_1) \wedge (\neg(insubject : s_2\ incontent : c_3)) \quad (10)$$

ただし，s_1, s_2, c_1, c_2, c_3 は，以下の制約条件を満たすとする．

$$\sum_{i=1}^{3}(cooc(s_1, c_i) - cooc(s_2, c_i)) \geqq 0$$

$$cooc(s_1, c_1) = max(cooc(s_1, c_1), cooc(s_1, c_2), cooc(s_1, c_3))$$

$$cooc(s_2, c_3) = max(cooc(s_2, c_2), cooc(s_2, c_3))$$

CD 質問と SD 質問は，次のような二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 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．SB 質問と CB 質問が，共通の subject-term または content-term を持つ二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）が別であるような二つの話題の結合である．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．

一方，枝型質問は，与えられた話題構造のある部分グラフを含むが，その他の部分グラフを含まない話題構造を有するページを検索するための質問である．つまり，与えられた話題構造に表現されているコンテンツと同じ（サブ）話題だけではなく，別の（サブ）話題についても述べているページを検索するための質問である．このようなページは，より幅広い情報とより詳細な情報を同時に含んでいる可能性がある．

構造化質問の肯定条件部によって，検索されたページが元のコンテンツとの関連を保つことが期待できる．一方，否定条件部は，元と同じものや結合結果が空となるものを排除する役割がある．

3.2 テキストストリームの文脈に基づく質問生成

文献 [5] では，我々は，テキストストリームから抽出された話題構造を，それぞれ独立したものと見なした．さらに，検索質問の種類をユーザによる指定とした．そこで，本稿では，テキストストリームの連続性（前後の関連）を考慮し，テキストストリームから抽出された話題構造ストリームの文脈に基づく質問の自動生成手法を提案する．これによって，テキストストリームから抽出された同じ話題構造に対して，できるだけ異なる質問を生成し，重複した補完情報の提供を避けることを目指す．

a） 話題構造ストリームにおける文脈

ある時点 i における，テキストストリームから抽出された話題構造のストリームを $T(= t_1 \cdot t_2 \cdot ... \cdot t_i)$ とする．話題構造 t_i のコンテキスト $C(t_i)$ は，過去の話題構造との結合結果が ϕ となる直前の結合結果である．

$$C(t_i) = t_i \bowtie t_{i-1} \bowtie ... \bowtie t_j \neq \phi$$
$$t_i \bowtie t_{i-1} \bowtie ... \bowtie t_j \bowtie t_{j-1} = \phi,\ j \geqq 1,\ t_0 = \phi$$

話題構造のコンテキストも話題構造であることが明らかである．図 4 は，話題構造のコンテキストのグラフ例を示している．なお，例題では，枝の重みを省略している．

b） 質問種類の自動選択

検索質問の種類は，t_i と t_i のコンテキスト $C(t_i)$ のグラフの幅と高さの比較に基づいて自動選択される．

- $(W(C(t_i)) - W(t_i)) - (H(C(t_i)) - H(t_i)) \geqq \theta$，つまり，幅の差が高さの差より大きければ，$t_i$ を用いて，話題グラフの高さを増加させる効果のある質問（CD と SD 質問）を生成する．ここでは，$H(x)$ と $W(x)$ は，それぞれ，話題構造（グラフ）x の高さと幅を表す．$\theta(> 0)$ は，予め定義された閾値である．
- $(W(C(t_i)) - W(t_i)) - (H(C(t_i)) - H(t_i)) \leqq -\theta$ であれば，t_i を用いて，話題グラフの幅を広げる効果のある質問（CB と SB 質問）を生成する．
- $-\theta < (H(C(t_i)) - H(t_i)) - (W(C(t_i)) - W(t_i)) < \theta$ であれば，t_i を用いて枝型質問を生成する．

c） 質問の修正

話題構造のストリームに現れる同じ話題構造に対して，上記の質問種類の選択手法のみでは，同じ質問の生成される可能性が残る．この場合，我々は，元の話題構造の代わりに，元の話題構造のコンテキストを用いて，質問を生成する．ただし，話題構造のコンテキストに多くのノード・枝が含まれている可能性が高いため，質問用の話題構造を抽出する必要がある．

話題構造のコンテキストのノード集合と枝集合をそれぞれ，V と E とする．すべてのノード $u \in V$ に対して，次のような subject-degree [5] を計算する．値の高いノードを二つ選んで，s_1 と s_2 とする．

$$sub(u) = \sum_{(u,v)\in E, v\in V-\{u\}} (cooc(u, v) * tf(u)) \quad (11)$$

そして，残りのノード $v \in V - \{s_1, s_2\}$ に対して，次のような content-degree [5] を計算し，値の高いノードを三つ選んで，c_1, c_2, c_3 とする．

$$con(v) = cooc(v, s_1) + cooc(v, s_2) \quad (12)$$

3.3 コンテキストを考慮した補完度計算

上記の構造化質問を用いて，与えられたコンテンツを補完するコンテンツの候補を検索することができる．これらの候補をランキングして，最も補完情報の多いページを選び出すため，我々，補完度という概念を用いる．基本的に，補完度は，検索されたページの話題グラフと元の話題グラフの比較に基づいて計算される．

話題グラフの高さと幅は，それぞれ，コンテンツの詳細と網羅の度合いを表すと考えられる．故に，話題グラフの幅と高さの差に基づいて，それぞれ，結合によるコンテンツのカーバーする範囲と詳細の増幅を計ることができる．幅の増幅が大きいほど，統合によって提供できる話題の幅が大きい．また，高さの増幅が大きければ，より多くの詳細情報を統合によって得られる．つまり，幅・高さの差が大きくほど，補完度が高い．

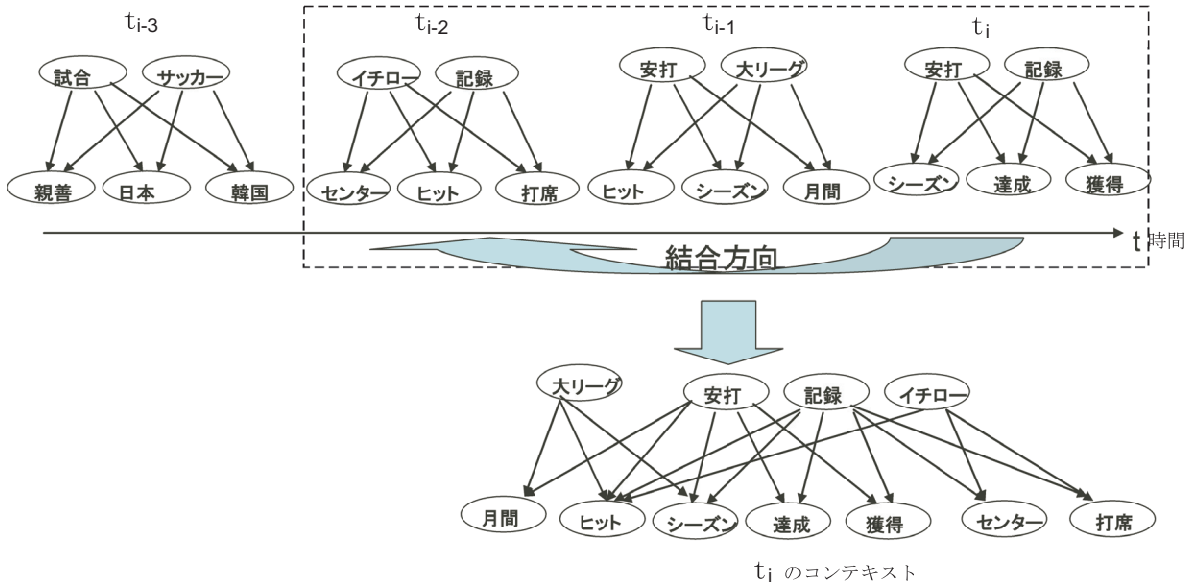


図 4　コンテキストグラフの例

話題構造 t' の話題構造 t に対する補完度 $comple(t,t')$ は，t のコンテキスト C_t および結合結果 $(t \bowtie t')$ と t の幅・高さの差に基づいて，次のように計算される．

- $(W(C_t) - W(t)) - (H(C_t) - H(t)) \geqq \theta$ の場合：

$$comple(t,t') = H(t \bowtie t') - H(t) \quad (13)$$

- $(W(C_t) - W(t)) - (H(C_t) - H(t)) \leqq -\theta$ の場合：

$$comple(t,t') = W(t \bowtie t') - W(t) \quad (14)$$

- $-\theta < (H(C_{t_i}) - H(t_i)) - (W(C_{t_i}) - W(t_i)) < \theta$ の場合：

$$comple(t,t') = H(t \bowtie t') * W(t \bowtie t') - H(t) * W(t) \quad (15)$$

一般に，Web ページなどのコンテンツには，複数の話題構造が存在すると考えられる [4]．これらのコンテンツ間の補完度は，話題構造集合のメンバーの補完度の総和である．二つの話題構造の集合 $S = \{s_1, ..., s_m\}$ と $T = \{t_1, ..., t_n\}$ が与えられた時，T の S に対する補完度 $com(S,T)$ は，次のように計算される．

$$com(S,T) = \sum_{i=1}^{i=m} \sum_{j=1}^{j=n} comple(s_i, t_j) \quad (16)$$

4. 実　　験

提案手法を評価するため，我々は，一日分の NHK ニュース 7 の字幕データを用いて，コンテキストを考慮する場合と考慮しない場合の補完情報検索手法の比較実験を行った．

実験では，2002 年 9 月から 2004 年 12 月までの字幕データを用いて，共起辞書を作成して利用した．字幕データのセグメンテーションと話題構造の抽出は，以前の提案手法 [5] を用いて行った．字幕データから 39 個の話題構造を抽出した．抽出された話題構造は，二つの subject-term と三つの content-term から構成される．Google の“ intitle ”,“ intext ”などの検索オプションを利用して，質問を実装した．

実験では，コンテキストを考慮する場合と考慮しない場合の補完情報検索手法の比較を行った．

- コンテキストを考慮しない場合の補完情報検索の実験：

39 個の話題構造をそれぞれ用いて，CD，SD，CB と SB 質問を生成して，検索を行った．検索エンジンから返された上位 10 ページにおいて，補完度の最も高いページをシステムの解とした．

- コンテキストを考慮する補完情報検索の実験：

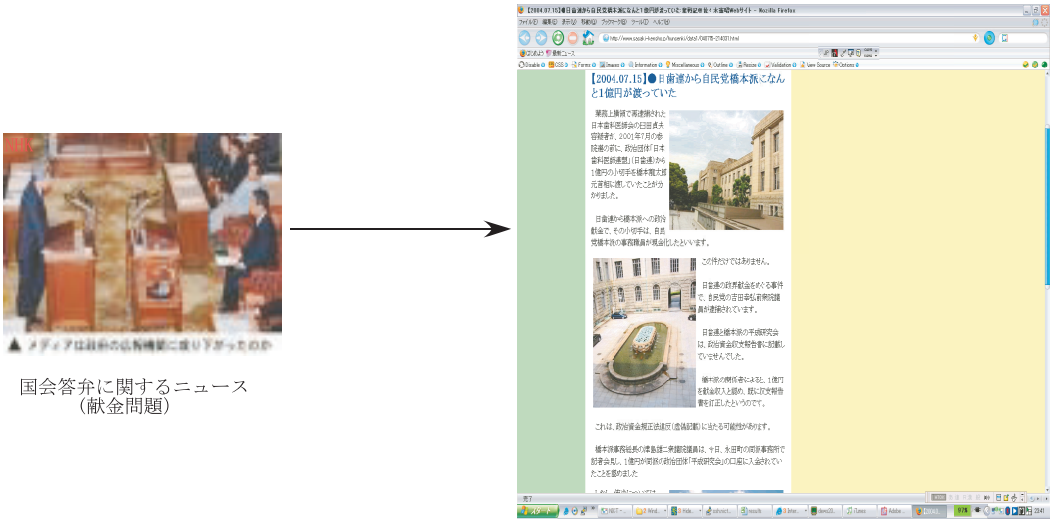
抽出された 39 個の話題構造をストリームと見なし，コンテキストに基づいて補完質問の種類を決定し，そして，必要に応じて質問修正を行って，検索を行った．同様に，検索結果の上位 10 ページの中で，質問種類に応じて計算された補完度の最も高いページを解とした．なお，実験では，以下のような 2 つ方法で質問の修正を行った．

– (A) 方式: 基本的に，字幕データから抽出された話題構造を利用して質問を生成する．過去の質問と同じ質問生成される可能性がある場合，コンテキストから質問グラフを抽出して質問を生成する．

– (B) 方式: 基本的に，コンテキストから抽出される質問グラフを用いて質問を生成する．同じ質問が過去にあった場合，字幕データから抽出された話題構造を利用して質問生成する．

実験では，システムから返された解に対して，番組の内容を補完できる内容があるかについて評価を行った．つまり，ページが番組の内容（主題）を別の主題（内容）から述べているか，番組の主題または内容を詳しく述べているかを基準として，正解ページの判断を行った．図 5 では，検索された補完ページの例を示している．左辺は，番組である．右辺は，詳細情報のある補完ページである．

表 1 では，それぞれの検索方式の適合率を示している．コンテキストを考慮した検索手法の適合率の向上が見られた．特に，コンテキストを考慮した (B) 方式では，従来の CD，SD，CB，



国会答弁に関するニュース
（献金問題）

図 5　補完情報の検索例

表 1　実験結果 (1)：適合率

実験方法		適合率	質問の数
コンテキストを考慮しない場合	CD 質問	0.452	CD 質問:39 件
	SD 質問	0.50	SD 質問：39 件
	CB 質問	0.474	CB 質問: 39 件
	SB 質問	0.528	SB 質問: 39 件
コンテキストを考慮する場合	(A) 方式	0.529	CD+SD 質問：17 件, CB+SB 質問:1 件, 枝型質問:21 件
	(B) 方式	0.583	CD+SD 質問：17 件, CB+SB 質問:1 件, 枝型質問:21 件

表 2　実験結果 (2)：生成される同じ質問の数

実験方法		生成された同じ質問の数
コンテキストを考慮しない場合	CD 質問	3
	SD 質問	3
	CB 質問	6
	SB 質問	6
コンテキストを考慮する場合	(A) 方式	1
	(B) 方式	2

SB 方式に比べて，適合率が百分率でそれぞれ，13.1 ポイント，8.3 ポイント，10.9 ポイントと 5.5 ポイント向上した．これは，コンテキストから質問グラフを抽出することによって，より適切なキーワードを用いて検索できたからと思われる．我々の字幕データのセグメンテーション手法 [5] によって得られたセグメントに，別の話題を述べる，ノイズとなる字幕データを含む可能性があるので，検索用の話題構造（キーワード）の抽出で失敗することがある．隣接する話題構造の結合によって生成されるコンテキストを用いることで，このようなノイズを省くことが可能であると思われる．

独立したニュース項目の話題構造や関連ニュースの最初項目の話題構造は，直前の話題構造と結合不可能であるため，これらの話題構造のコンテキストグラフは，自分自身である．よって，枝型質問を生成する可能性が高い．従って，コンテキストを考慮した検索方式では，質問の種類からみると，枝型質問が多かった．CB と SB 質問が少なかったのは，利用した番組では，国会答弁のようなイベントがいくつかの側面から報道されていたからであると思われる．つまり，国会の内容を幅広く報道されているので，より詳細の情報を検索すべきと判断したわけである．

表 2 で示されているように，コンテキストを考慮しない手法では，CD，SD，CB と SB 方式で生成された同じ質問(注1)の数は，それぞれ，3，3，6 と 6 である．一方，コンテキストを考慮する手法では，(A) と (B) 方式から生成された同じ質問の数が，それぞれ 1 と 2 である．従って，コンテキストを考慮した手法は，検索される補完ページの重複を防ぐ効果があると考えられる．しかしながら，コンテキストを考慮した，いずれの手法でも同じ質問を生成してしまう可能性が残っているので，検索結果のコンテキスト（履歴）を考慮するなど，さらなる対策が必要であると思われる．

5. ま と め

多種多様な情報から，ユーザの興味に即した情報の獲得が重

(注1)：本稿では，生成された質問の肯定条件部が同じであれば，それらの質問が同じであると見なす．

要であり，そのための研究開発が数多くあった．同時に，獲得する情報の質やバランスもますます重要になってきている．つまり，ユーザの現在の興味にマッチまたは類似した情報だけではなく，より幅広い，より詳しい情報をユーザに提供することが必要である．そのため，我々は，情報補完という観点からの情報検索および統合について研究を行っている．

本稿では，情報補完の一つの試みとして，字幕データのようなテキストストリームのコンテキストを考慮した補完情報検索手法を提案した．この手法を用いて，テキストストリームにある同じ話題構造に対して，異なる補完情報の検索ができる．また，ユーザに提示される補完情報の重複を避けられるといった効果も期待できる．その上，従来の手法より，適合率が向上したことも実験結果からわかった．今後，さらなる実証実験を行い，提案手法の改良を行う予定である．例えば，検索結果や前回の質問種類などを用いて，コンテキストの概念を拡張していくことが考えられる．また，提案手法に基づいて我々の開発している応用システム WebTelop [3] の改良を行う予定である．

文　献

[1] 亀山渋, 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).

[2] Qiang Ma, Akiyo Nadamoto, Katsumi Tanaka: Complementary Information Retrieval for Cross-Media News Contents. *Proc. of ACM MMDB 2004*, pp 45-54, 2004.11

[3] Qiang Ma and Katsumi Tanaka: WebTelop: Dynamic TV-content Augmentation by Using Web Pages, *Proc. of ICME2003 Vol.2*, pp. 173–176 (2003).

[4] Qiang Ma and Katsumi Tanaka: Topic-Structure Based Complementary Information Retrieval for Information Augmentation, LNCS3007, pp. 608-619 (2004).

[5] 馬強, 田中克己：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情報処理学会論文誌：データベース (TOD23), pp. 18-36 (2004).

[6] 馬強, 田中克己: 補完情報の検索に基づくコンテンツ統合, 情報処理学会研究報告, 2004-DBS-134, pp. 337-343 (2004).

[7] TopicMap: http://www.topicmap.org (2003).

[8] Wayne, C. L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, *Proc. of LREC2000*, pp. 1487–1494 (2000).

[9] 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).

[10] Zloof, M.: Query-By-Example: A Data Base Language, *IBM Systems Journal, Vol.16, No.4*, pp. 324–343 (1977).

[11] Henzinger, M., Chang, B.-W., Milch, B. and Brin, S.: Query-Free News Search, *Proceedings of The Twelfth International World Wide Web Conference* (2003).