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

2004－DBS－134　(II)　(45)
2004／7／14

補完情報の検索に基づくコンテンツ統合

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伴って，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，ユーザの興味のあるコンテンツを補足する情報の検索手法を提案し，情報補完という観点からのコンテンツ統合を試みる．我々は，コンテンツの表現手法として，キーワードの役割に着目した話題構造という概念を用いる．本論文では，話題構造に基づく構造化質問を記述して，コンテンツの主題と内容を区別して取り扱う情報検索手法を提案する．この検索手法は，従来の類似検索と異なり，より詳細または別の観点の情報を検索可能である．

Content Integration Based on Complementary Information Retrieval

Qiang Ma ,[†]　and Katsumi Tanaka[††,†]

In this paper, we propose a new way of integrating cross-media content, such as television programs and web pages based on a notion called the ”topic structure”. Intuitively, a topic structure is made up of a pair of subject and content terms. Subject terms denote the dominant terms of a news item. A content term is a term having strong co-occurrence relationships with the subject terms. Based on the topic structure, we search cross-media content to find complementary items which can provide additional information to users interested in a particular topic. The complementary information searched for are not just similar to the item the user is interested in, but also provide information in more detail or from a different perspective.

1. はじめに

ブロードバンドの普及に伴って，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．映像コンテンツは，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．

情報統合に関する研究は，従来から数多く存在するが，本論文では，情報補完という観点から，コンテンツの統合について考案し，補完情報の検索手法を提案する．

本論文では，一つのイベントまたはアクティビティを話題（topic）と呼ぶ．コンテンツに述べられている話題を，3 節で定義される話題構造を用いて表現する．話題構造は，話題のタイトルと内容をそれぞれ表すキーワード subject-term と content-term のペアから構成される．我々は，話題構造を一つの連結成分からなる DAG（Directed Acyclic Graph) を用いて表現する．話題構造に基づくコンテンツの統合は，グラフの和で表すことが可能となる．

我々は，話題構造とその話題構造の結合に基づいて，ユーザの興味のある情報を補足できるコンテンツの検索手法を提案する．この手法では，まず，ユーザの興味のあるコンテンツの話題構造を抽出して，補完情報を検索するための構造化質問を生成する．そして，Web から検索されたページを，補完度という概念に基づいてランキングし，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，結合結果と元の話題構造との比較に基づいて計算される．

本論文で提案する補完情報の検索手法は，従来の類

† 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
Interactive Communication Media and Contents Group, National Institute of Information and Communications Technology
†† 京都大学大学院 情報学研究科 社会情報学専攻
Division of Social Informatics,
Graduate School of Informatics, Kyoto University

似検索と異なり，より詳細や別のアスペクトといった異なる観点からの情報検索が可能である．つまり，与えられたコンテンツの補完情報の検索ができる．

以下，本論文の構成を示す．2 節では，関連研究について述べる．3 節では，話題構造について説明する．4 節では，補完情報を検索するための構造化質問について考案する．5 節では，補完度という概念を紹介する．6 節では，予備実験の結果を示す．本論文のまとめと今後の研究課題については，7 節で述べる．

2. 関 連 研 究

QBE (Query By Example)[1] 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，我々の補完情報検索と同様である．しかしながら，我々の検索手法では，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 [2]．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て，番組の類似ページを検索する．Henzinger らの手法と比較して，我々の補完情報検索手法は，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．

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，情報の統合等に利用する試みは以前より行われてきた [3)~5)]．これらの研究では，基本的に情報の断片を取り扱いの単位として，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，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，およびクロスメディアの情報統合・補完のための検索手法を提案している点が異なる．

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．TDT（Topic Detection and Tracking）[7] 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す．TDT では，

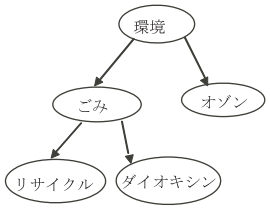

図 1 話題グラフの例

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，一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．

3. 話 題 構 造

3.1 話 題 構 造

話題構造は，subject-term と content-term の集合のペアから構成される．subject-term は，コンテンツに述べられている話題の主題となる語である．本論文では，ある話題について述べているコンテンツにおいて，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，同じコンテンツに出現し，subject-term との共起関係の強いキーワードである．言い換えれば，subject-term はその話題のタイトルを表す役割があり，content-term は話題の本体を表し，内容記述の役割がある．

以下，話題構造の定義を示す．

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

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．

3.2 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 DAG（Directed Acyclic Graph）を用いて表現できる．

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：

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

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．$|V| \geq 2, E \neq \emptyset$ である．

3.3 話題構造の結合

コンテンツの統合を結合で表現することが可能である．例えば，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 の結合とみなすことができる．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) 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．

$$t \bowtie t' = \begin{cases} G(t) \cup G(t'), & G(t) \cup G(t') \text{ が一つの連結成分からなる DAG である場合} \\ \phi, & \text{その他} \end{cases} \tag{3}$$

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．

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言う．結合の定義から，$t \bowtie t = t$ であることは明らかである．

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．したがって，結合結果である話題構造は，別の話題構造との更なる結合が可能である．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．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．例えば，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 話題構造の抽出

(a) 共起関係

本論文では，コンテンツの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．

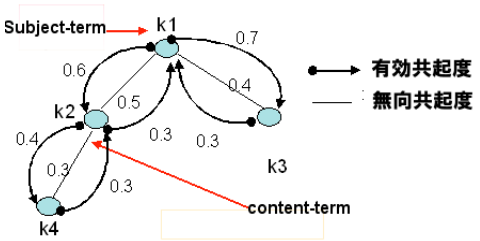

図 2 subject-term と content-term の抽出例

定義 3 (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 が同時に出現する話題（テキスト）が多いほど，この二つの語の共起関係が強いと言う．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．

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

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 4 (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．

$$\overrightarrow{cooc}(w_i w_j) = \frac{df(\{w_i, w_j\})}{df(\{w_i\})} \tag{5}$$

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) コンテンツの話題構造抽出

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が，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．

☆ 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．

$$sub(w_i) = tf(w_i) + \sum_{j=1, j\neq i}^{n} \overrightarrow{cooc}(w_i w_j) \quad (6)$$

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．

$$con(w_i) = \sum_{w_j \in S} cooc(w_i, w_j) \quad (7)$$

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 2 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．

主題度と内容度を計算することによって，話題グラフの高さは 1 であるような単純な話題構造を抽出できる．しかしながら，実際，コンテンツの話題構造はもっと複雑であると思われる．上記の手法を再帰的に適応すれば，高さが 2 以上の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．これに対して，本論文では，話題構造（集合）の簡約という操作に基づく手法を用いる．コンテンツをいくつかのユニット（話題抽出の最小単位と呼ぶ．Web ページの段落,etc.）に分けて，それぞれのユニットの話題構造を主題度と内容度を用いて抽出する．これらのユニットの話題構造をマージ（簡約）した結果を，コンテンツの話題構造（集合）とする．ユニットの分け方としては，従来からいろいろな手法が提案されているが，我々が提案している共起関係によるセグメンテーション手法を用いることも考えられる [8]．

話題構造の集合の簡約は，次のように定義されている．

定義 5 (話題構造の集合の簡約) 話題構造の集合 $T = \{t_1, t_2, ..., t_n\}$ の簡約 $R(T)$ は次のように行われる．

$$R(T) = G(t_1) \cup G(t_2) \cup ... \cup G(t_n) \quad (8)$$

ただし，$G(t_1) \cup G(t_2) \cup ... \cup G(t_n)$ 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言う．

4. 話題構造に基づく構造化質問

4.1 構造化質問

本節では，補完情報を含むコンテンツを検索するための質問生成について述べる．検索されるコンテンツは，与えられたコンテンツと単に類似するのではなく，より詳しいまたは別の視点からの情報を述べている．

与えられた話題構造を $t = (\{s_1, s_2, ..., s_m\}, \{c_2, c_2, ..., c_n\})$ であるとする．t を用いて，以下のような 4 種類の質問を定義する．それぞれの検索式では，“ insubject ”と“ incontent ”に後置される検索文は，それぞれコンテンツの話題構造の subject-term と content-term を検索対象とする．”∧” と ”∨” はそれぞれ論理積と論理和を表す．例えば，質問 $(insubject: k_1 \wedge k_2) \wedge (incontent: k_3 \wedge k_4)$ は，k_1 と k_2 が subject-term に含まれ，k_3 と k_4 が content-term に含まれるコンテンツを検索する．

- **CD(Content-Deepening) 質問 (Q_{cd})**: 話題構造 t に対して，我々は，CD 質問を用いて，次のような話題構造を含むコンテンツを検索する．t の content-term が subject-term に含まれる．
$$Q_{cd} = insubject : c_1 \wedge c_2 \wedge ... \wedge c_m \quad (9)$$
- **SD（Subject-Deepening) 質問 (Q_{sd})**: 話題構造 t に対して，次のような話題構造を含むコンテンツを検索する．t の subject-term が content-term に含まれる．
$$Q_{sd} = incontent : s_1 \wedge s_2 \wedge ... \wedge s_n \quad (10)$$
- **SB(Subject-Broadening) 質問 (Q_{sb})**: SB 質問を用いて，t の content-term が content-term に含まれるような話題構造を含むコンテンツを検索する．
$$Q_{sb} = incontent : c_1 \wedge c_2 \wedge ... \wedge c_m \quad (11)$$
- **CB(Content-Broadening Query) 質問 (Q_{cb})**: CB 質問を用いて，t の subject-term が subject-term に含まれるような話題構造を含むコンテンツを検索する．
$$Q_{cb} = insubject : s_1 \wedge s_2 \wedge ... \wedge s_n \quad (12)$$

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．

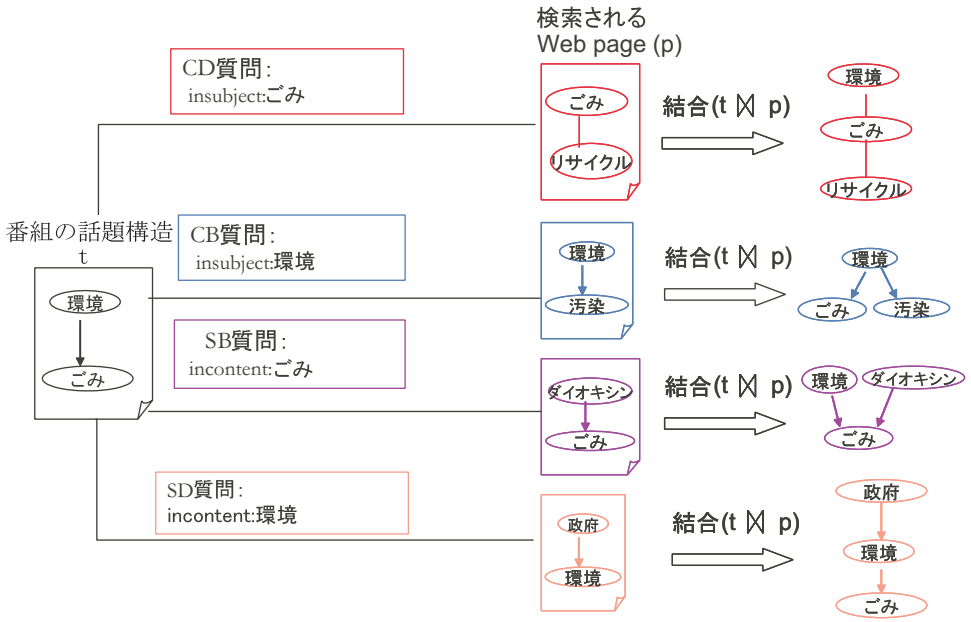

図 3 話題構造の結合に基づく質問の例

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をコンテンツを検索する．コンテンツに複数の話題構造がある場合，そのコンテンツに検索式で要求されていた話題構造を一つでも含めば，そのコンテンツが解となりうる．上記の検索式では，検索結果に，元の話題構造との結合が空となるような話題構造（CD と SD 質問）や元と同じ話題構造（CB と SB 質問）を含むコンテンツが含まれる可能性がある．元の話題構造との結合が空となる話題構造や元と同じ話題構造しか持たないコンテンツには，補完情報がないと考えられる．次節に述べられている補完度は，これらのコンテンツの排除ができる．

実際，CD と SD 質問の検索結果を，元のコンテンツと結合させると，話題グラフの高さが増加されることとなる．つまり，元の話題構造を詳細化 (deepening) する効果があると考えられる．また，CB と SB 質問の検索結果と元のコンテンツの結合結果は，話題グラフの幅を拡大（broadening）するので，より幅広く情報を提供できると考えられる．**図 3** 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．

4.2 構造化質問の実装について

小山ら [9] は，HTML ソースの“ title ”と“ body ”タグを利用した *intitle* と *intext* といった検索オプション [10] の有用性について報告している．これ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ことが可能であり，上記の構造化質問の簡易版を実装することが可能である．この手法は，検索エンジンへ依存していることや検索結果の精度が良くないことなどの欠点があるが，既存の検索エンジンを利用できる点が大きなメリットとなる．

表 1 構造化質問の実装

	query
SB 質問	$allintext: c_1\ c_2 \ldots c_n$
SD 質問	$allintext: s_1\ s_2 \ldots s_m$
CB 質問	$allintitle: s_1\ s_2 \ldots s_m$
CD 質問	$allintitle: c_1\ c_2 \ldots c_n$

Google の *intitle*，*intext* などの検索オプションを利用して，**表 1** のように構造化質問を実装することができる．

また，別の実装手法として，あらかじめ収集したデータを話題構造を用いて索引付けを行って，話題構造ベースの検索システムを開発するのが考えられる．しかし，この手法は，検索結果の精度が良いかもしれないが，開発のコストが掛かるという欠点がある．

5. 補 完 度

5.1 補 完 度

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，我々は，補完度という概念を用いる．

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．基本的に，補完度は，結合前後の話題構造の比較に基づいて計算される．

定義 6 (話題構造間の補完度) 話題構造 t' の与えられた話題構造 t に対する補完度 $comple(t,t')$ は，結合結果 $(G(t \bowtie t'))$ と $G(t)$ の幅・高さの差に基づいて，次のように計算される．つまり，幅・高さの差が大きくほど，補完度が高い．

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

ただし，$H(G(x))$ と $W(G(x))$ は，それぞれ話題グラフ $G(x)$ の高さと幅を表している．

一般に，一つのコンテンツには，複数の話題構造が存在すると考えられる．従って，コンテンツ間の補完度は，話題構造集合間の補完度である．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 と S 間の補完度 $com(S,T)$ は，次のように計算される．

$$com(S,T) = \sum_{i=1}^{n} \sum_{j=1}^{m} comple(t_i, s_j) \quad (14)$$

5.2 話題グラフの高さと幅

一般に，話題グラフの高さは，親を持たない節点（入次数は 0 である．以下，根節点と呼ぶ.）から子供を持たない節点（出次数は 0 である．以下，葉節点と呼ぶ.）に到達までに通る枝の数の最大値である．本論文では，単に枝の数を数えるだけではなく，隣接する二つの節点の tf 値も考慮する．枝の二つの節点のコンテンツでの出現頻度の比率を枝の重みとし，枝の長さとする．話題グラフの高さは，根節点から葉節点に到達までに通る枝の長さの最大値である．枝 $e = (u,v)$ が与えられた時，e の長さ $L(e)$ は，次のように計算される．

$$L(e) = tf(v)/tf(u) \quad (15)$$

ただし，$tf(u)$ と $tf(v)$ は，コンテンツでの u と v の出現頻度である．

一方，話題グラフの幅は，葉と根節点の数に基づいて計算できる．つまり，葉節点と根節点の数の最大値を N とした場合，幅は $N-1$ である．本論文では，高さの計算と同様に，葉（根）節点間の関連を考慮して計算を行う．二つの葉（根）節点 l_i と l_j の距離 $d(l_i,l_j)$ は，l_i と l_j の共起関係および出現頻度に基づいて，次のように計算される．つまり，出現頻度の差が小さく，しかも共起関係の弱い二つの語は，別の側面から情報を述べている可能性が高く，距離が大きいと考えられる．

$$d(l_i,l_j) = (1 - cooc(l_i,l_j)) \cdot \left(\frac{min(tf(l_i), tf(l_j))}{max(tf(l_i), tf(l_j))}\right) \quad (16)$$

ただし，$min(x,y)$ と $max(x,y)$ は，それぞれ x と y の最小値と最大値を取る関数である．

二つの節点間の距離を用いた節点集合 L の節点間の距離を計算する手順を，次に示す．

(1) L における任意の節点 l_i を始点 s とする．
(2) $L = L - \{s\}$; $L = \phi$ であれば，5 へ．
(3) L における，s との距離が最小である節点 l_j を選択する．
(4) $D = D + d(s,l_j)$，$s = l_j$ とし，2 へ．
(5) D を L の節点間の距離とする．

葉節点間と根節点間の距離をそれぞれ計算して，値の大きい方は，話題グラフの幅となる．例えば，葉節点の集合 $L = \{l_1, l_2, l_3\}$，$d(l_1,l_2) = 0.4$，$d(l_1,l_3) = 0.3$，$d(l_2,l_3) = 0.3$ とすれば，葉節点間の距離の値は 0.6 である．一方，根節点の集合 $R = \{r_1, r_2, r_3\}$，$d(r_1,r_2) = 0.3$，$d(r_1,r_3) = 0.3$，$d(r_2,r_3) = 0.2$ とすれば，根節点間の距離は 0.5 である．したがって，話題グラフの幅は，0.6 となる．

表 2 予備実験結果（適合率）

	補完度によるランキング無	補完度によるランキング有
SB 質問	0.591	0.856
SD 質問	0.489	0.753
CB 質問	0.432	0.825
CD 質問	0.483	0.857

6. 予 備 実 験

我々の構造化検索式による補完情報の検索手法を検証するため，番組の字幕データから抽出された 88 個の話題構造（2 つの subject-term と 3 つの content-term から構成される）を用いて，検証実験を行った．実験では，Google の”intitle”と”intext”などの検索オプションを利用して，構造化質問の実装を行った．我々の実験では，Google に返された 1 件目の検索結果をシステムの解として，二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．つまり，CD と SD 質問の検索ページに詳細な情報が含まれていれば，正

解ページとする．一方，CB と SB 質問の検索ページにヒントとなるような情報が書かれていれば正解ページとする．実験では，まず，二人にそれぞれ判定を行ってもらった．二人の判定結果に違いがあった場合，協議してもらって最終の判定結果を出してもらった．それぞれの検索式による検索結果の適合率は，**表 2** で示されている．

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，補完情報の検索には一定の効果があることが分かった．

また，話題グラフの高さと幅の計算を，語の出現頻度や共起関係を考慮せず，単に枝と葉節点の数で計算して，補完度について評価を行った．実験では，根節点から葉節点に到達までに通る枝の数の最大値を話題グラフの高さとした．根節点と葉節点の数の最大値から 1 を引いた値を話題グラフの幅とした．上記の 88 個の話題構造を用いて実験を行った．それぞれの質問の Google のトップ 10 の検索結果を対象に，補完度を計算して，補完度の値の最も高いページをシステムの解とした．上記と同様な方法でユーザに評価してもらった．SB，SD，CB と CD 質問の適合率は，それぞれ 0.807,0.693,0.613 と 0.688 であった（表 2）．単に構造化質問を用いた検索と比べて，適合率の改善が見られた．これは，補完度の概念は，補完情報の検索において有用であることを示していると考えられる．

7. ま と め

本論文では，話題構造というコンテンツの表現モデルと，それに基づく構造化質問式による補完情報の検索手法と補完度によるランキング手法を提案した．

話題構造は，コンテンツに述べられている話題を，タイトルを表すキーワードと内容を表すキーワードの集合のペアで表している．このように，主題と内容のキーワードを区別することによって，構造化質問が記述でき，主題類似や内容類似といった観点からの情報検索が可能となる．また，主題と類似するが，内容が異なるような似て非なる情報の検索も可能である．このような検索手法は，従来の類似検索の範疇を超えていると考えられる．

提案手法によって検索されるコンテンツは，元のコンテンツをより詳しくまたは別の観点から述べ，情報の補完を行える．また，本論文では，検索されたコンテンツの元のコンテンツへの補完の度合いを測るため，補完度という概念を提案した．実験結果から，これらの手法と尺度は，補完情報の検索には有効であることが分かる．

今後，本論文で提案されている手法の更なる検証を行う予定である．構造化質問の実装や補完度計算の改良も必要であると思われる．また，応用システムの開発も予定している．さらに，話題構造の抽出手法について考察を行う予定である．

参 考 文 献

1) Zloof, M.: Query-By-Example: A Data Base Language, *IBM Systems Journal, Vol.16, No.4*, pp. 324–343 (1977).
2) Henzinger, M., Chang, B.-W., Milch, B. and Brin, S.: Query-Free News Search, *Proceedings of The Twelfth International World Wide Web Conference* (2003).
3) 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, 信学技報 NLC93-66, pp. 23–30 (1993).
4) 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).
5) 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–, 情処研報 FI-142-23, pp. 167–174 (2001).
6) TopicMap: http://www.topicmap.org (2003).
7) Wayne, C. L.: Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, *Proceedings of the Language Resources and Evaluation Conference (LREC) 2000*, pp. 1487–1494 (2000).
8) Ma, Q. and Tanaka, K.: WebTelop: Dynamic TV-content Augmentation by Using Web Pages, *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2003) Vol.2*, pp. 173–176 (2003).
9) Oyama, S. and Tanaka, K.: Exploiting Document Structures for Comparing and Exploring Topics on the Web, *Proceedings of the 12th Internaional World Wide Web Conference (WWW2003) (poster tracks)* (2003).
10) Google: http://www.google.com (2003).