レファレンスデータを用いた情報探索過程段階化手法の提案

嫁兼　弘修†　　北原沙緒理††　　波多野賢治†

† 同志社大学文化情報学部　〒 610–0394 京都府京田辺市多々羅都谷 1-3
†† 同志社大学大学院文化情報学研究科　〒 610–0394 京都府京田辺市多々羅都谷 1-3
E-mail: †{yomegane,kitahara}@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

あらまし　近年，検索エンジンの性能向上は，求める情報を取得する情報探索行為の簡素化に貢献しているが，ユーザによる情報源特定は検索エンジンの提示順序に依存する．その結果，ユーザが検索結果上位の情報源だけを閲覧するようになる状況を生み，情報源の特定という情報探索行為の中でも重要な過程を排除することとなり，最終的には求める情報の正確な取得には至らない．その対策として，検索エンジンの検索ログ等を用い利用者が求める情報を取得する過程をユーザに提示し，その過程を真似て情報源の特定を支援する仕組みが必要である．本稿では，Web 検索エンジンよりも小規模な図書館での蔵書検索において，情報探索過程提示の際の問題の一つである，探索過程の段階化に焦点を当て，その精度向上に貢献する提案を行う．

キーワード　プロセスの視覚化，レファレンスデータ

Hironao YOMEGANE†, Saori KITAHARA††, and Kenji HATANO†

† Faculty of Culture and Information Science, Doshisha University
1–3 Tatara Miyakodani, Kyotanabe, Kyoto 610–0394, Japan
†† Graduate School of Culture and Information Science, Doshisha University
1–3 Tatara Miyakodani, Kyotanabe, Kyoto 610–0394, Japan
E-mail: †{yomegane,kitahara}@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

1. はじめに

近年，World Wide Web (Web) へのアクセス環境や検索エンジンの性能の向上に伴い，Web 上の膨大なデータを利用し求める情報を取得するという形での情報探索を行う機会が増加している．Web ページの増加と多様化により，現在の Web 上の Web ページ群は様々なジャンルのデータを有している．そして Web を通じてあらゆるジャンルのデータが扱われるようになった現状は，検索エンジンで発見できるデータの網羅性を向上させている．また，あらゆるデータを網羅する Web 上のデータを扱うことが出来るようになった検索エンジンは，Web 上の膨大なデータの中からユーザが求める情報を高い精度で発見，提示している．

Web のデータの網羅性の向上と高度に発達した検索エンジンの登場は，情報を取得するために何冊もの図書にあたる以外の選択肢がなかった頃に比べ，情報探索行為の簡素化を大幅に促進させた [1].

また，ユーザの検索エンジン利用について，アイレップ他 2 社による共同調査によると，ユーザによる検索結果の閲覧傾向として主として Web ページの上段から中段へ集中することが報告されている．(注1)

前述のユーザの情報取得が検索エンジンの提示順位に依存するという点と，共同調査報告にあるユーザが検索結果の上段から中段までに注目するという状況が重なった場合を仮定すると，ユーザは情報源の対象として検索結果上位のみしか閲覧しないという状況が発生する．つまり,「信頼できる情報源の特定」という情報探索行為の中でも非常に重要な過程 [2] を排除した結果，ユーザが検索結果の上位に登場する Web ページの記述しかみない状況が発生する．その結果，求める情報の正確な取得に至ることが出来ないという問題が起こり得る．

求める情報の正確な取得に至ることができないという状況への対応策として，過去に行われた情報探索の過程をユーザに提示し，その過程を模倣させることで情報源の特定を補助することが考えられる．過去にユーザ自身と類似する情報探索を行った検索エンジン利用者の，情報源を特定し求める情報を取得するまでの過程をサンプルとし，その過程を模倣させることで情報源の特定を補助する，という形式で情報探索を支援すること

(注1)：アイレップ プレスリリース 2009 年 12 月 24 日: http://www.irep.co.jp/press/pdf/20091224.pdf，2013 年 1 月 12 日閲覧

が必要となる．

情報探索行為における前例を模倣するという形式での情報探索支援の利点には，他者の行為を模倣させることにより未知の情報探索パターンの学習が可能となる点が挙げられる．更に，以降情報探索を行う際に，かつて学習した情報探索パターンを利用できる事が挙げられる．そこで，情報探索支援を業務として行っている図書館のサーチャ(注2)に着目し，最終的には情報探索過程における前例を活用する．情報探索過程における前例を活用することにより，ユーザにとって未知の情報探索行為に対してその指針を与えることを我々の研究目標の一つとしている．

本稿では，情報探索行為の中でも情報探索の際に参照した資料と，資料を基にサーチャがとった情報取得行動に着目し，行動パターンの分析を行う．このような分析を行うことにより，サーチャの情報探索における資料の選択と，選択した資料からの情報取得の繰り返しがどのような規則に基づき行われているのかを解明する．そして情報取得の繰り返し規則を解明することによりいくつかのパターンに分類することが可能となる．そして情報探索におけるサーチャの行動規則がパターン化できれば，パターン化したサーチャの行動規則を用いて，ユーザの情報探索行動に類似する情報探索前例を提示できる．結果として提示した前例を基に探索方法の模倣を行わせることで探索支援を行うシステムの構築に用いることが可能となる．

この「前例となる情報探索過程の模倣」という形の情報探索支援のひとつとして，本稿では Web を対象とした検索エンジンよりも小規模な図書館での蔵書検索における情報探索過程提示に焦点をあてる．そして図書館での情報探索過程の前例を利用する際必要となる，情報探索過程の段階化を試みる．

蔵書検索に焦点を当てた理由は，図書館業務の一つに図書館利用者からの質問への回答内容が記録として残っているからである．これを 2.1 で述べるレファレンスデータというが，これを分析，その内容を時系列に並べる (情報探索過程の段階化) ことでサーチャの行動を抽出可能となる．

2. 基本的事項

本稿ではレファレンス質問への回答に要した蔵書検索の調査過程を情報探索過程として捉え，その情報探索過程の詳細を抽出し分析を行うため，レファレンス質問に対する調査プロセスについてのデータを調査，つまり情報探索の過程に基づき段階化する手法について提案する．その提案を行う上で非常に重要となるレファレンスサービス，サーチャについてのデータ (レファレンスデータ) ，および情報探索についての説明を行う．

2.1 レファレンスデータ

レファレンスデータとは公立図書館や大学図書館において質問回答サービス (レファレンスサービス) として蔵書問合せや調べ方マニュアルの作成などを請負うサーチャが，その業務内容を記述したものである．1. 節のように，本稿では図書館における蔵書を見つけ出す際の探索過程に着目するため，レファレンスデータから情報探索過程の抽出を行う必要がある．

そこで本稿において利用するレファレンスデータを公開しているデータベースであるレファレンス協同データベースについての説明を行う．またレファレンスデータと同様に情報探索の過程についてのデータであるアクセスログデータの，データの性質におけるレファレンスデータとの差異，そしてレファレンスデータを 2 次利用する際の問題点についても説明する．

2.1.1 レファレンス協同データベース

レファレンスデータは，国立国会図書館が全国の図書館と共同で運営しているレファレンス協同データベース(注3)によって公開，提供されているため，このレファレンスデータを情報探索過程の段階化を行う際に利用する．

レファレンス協同データベースで公開されているレファレンスデータは，11 種類の項目を持っている．

- 管理番号
- 質問
- 回答
- キーワード
- 参考資料
- 回答プロセス
- 照会先
- 事前調査項目
- 寄与者
- 備考
- 提供資料館名

このデータ項目の中で，レファレンスデータとしてデータベースに登録するために最低限必要なデータは「管理番号」,「質問」,「回答」である．さらにレファレンス質問に対する回答のための調査について記述されている「回答プロセス」も不可欠なデータ項目であるため，これら四項目を使用する．

2.1.2 レファレンスデータとアクセスログデータ

2.1 で説明したように，レファレンスデータはファレンス業務を請負うサーチャによって記録された業務内容である．よってレファレンスデータは，Web 検索のコンテキストで言えば，一種のアクセスログデータと言うことができる．つまりアクセスログデータを使用しても，情報探索過程の抽出自体は適切な手段を講じれば可能である．

しかし，アクセスログデータから「何について調べたのか」や「どのようなページ遷移をしたのか」といった事実は取得可能であっても，次のような情報は取得するのが非常に難しい．

- 情報探索者が閲覧した Web ページのデータから取得した情報
- 閲覧した Web ページのデータから取得した情報に基づきとった行動
- 情報取得後の，情報探索目的に対しての達成条件の判断

もちろん，大量のアクセスログデータを収集し，ページ遷移の様子からユーザが求める情報の推定を統計的に行うことは可能ではある．しかし実際にそのアクセスログデータが取得された場面でユーザがどのような情報を取得，解釈し，次の行動に移ったかは，アクセスログデータが単にユーザがとった行動を記録したものに過ぎない．よってユーザのデータ解釈や取得した情報についての思考といった内的な活動の部分は，ユーザに直接質問することでしか分からない．

レファレンスデータにおいても，情報探索の情報源となった

(注2)：蔵書問合せなどの質問に対し調査を行う人．2.2 にて説明．

(注3)：http://crd.ndl.go.jp/jp/public/，2013 年 1 月 12 日閲覧

資料群のデータであるという点はアクセスログデータと同様である．しかし，レファレンスデータはサーチャによるレファレンス業務の記録であるために，どのような経緯で情報源となる資料を参照し，またその資料からサーチャが取得した情報をもとにどのような解釈，判断を行い次の資料を参照したか，あるいは調査を終了したかが記録されている．よって情報探索行動の内容に関する詳細さという点において，ユーザの内的な活動記録の有無という意味で，レファレンスデータと通常のアクセスログデータは異なる．

2.2 サーチャ

サーチャとは，図書館などでレファレンスサービス利用者による依頼に基づき，レファレンスサービス利用者が要求する情報自体や，要求する情報が含まれる情報源を探索することを業務とする人のことである．

一般的に情報探索が行われる場合，要求される情報に関連する情報源が選択される．情報源の選択について，通常は資料のタイトルや目次，あらすじ，著者，ジャンル，あるいは内容に対して斜め読みが行われた結果,「関連する情報が記述されているかもしれないと判断する」という過程を伴う．よって情報源選択が行われる際には,「なぜその情報源が選択されたのか？」という，情報源選択の根拠が付随するといえる．その「情報源を選択する」という部分に対して，試行錯誤が行われつづける結果，我々が常日頃行う情報探索行動が存在する．

情報源選択が行われる際に発生する，情報探索のプロであるサーチャと，情報探索を行う一般ユーザとの差異は，情報源の選択についての知識や経験の違いであるといえる．よって，レファレンスサービス利用者の情報探索行動中の「情報源選択の基準となった要因」に関して，そのサンプルを提示するという形での支援が可能となれば，情報探索のプロが行うものに近い情報源選択を行うことができる．

2.3 情報探索

本稿では情報探索という用語を,「求める情報を，参考資料の選択と資料からの情報取得を繰り返し試行錯誤しながら得てゆく行為」として用いている．つまり，情報探索という行為の実態は，情報源となり得る資料の選択と，選択した資料からの情報取得という二つのステップが繰り返されることで形成されていると考えることができる．よって，情報探索行為よりこの資料選択，情報取得ステップを抽出することができれば，情報探索行為の過程を，参照した資料の順に複数の段階に分割することができる．

そこで資料選択と情報取得の部分を，情報探索過程における一つの行動単位として，以降「情報検索」という言葉で表す．その結果，情報探索は情報検索の繰り返しであるということができ，情報探索の構造は図 1 のようになる．

図 1 のように情報探索が開始されると，まず資料選択が行われる．資料選択段階においては，資料となり得る図書，Web ページなど多種多様な情報源の中から，情報を取得する対象となる資料が情報探索者によって選択される．資料が選択されると，資料からの情報取得の段階へと移行する．情報取得段階においては，選択した資料から，目的とする情報，もしくは目的

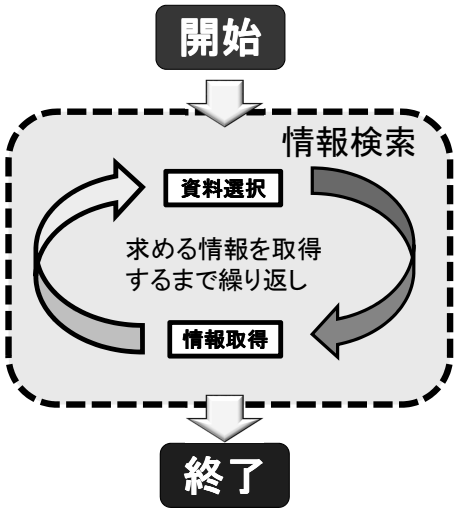

図 1 情報探索

とする情報に関連する情報の取得が情報探索者によって試みられる．そして，情報取得が試みられた結果，目的とする情報が取得されたのであれば探索が終了され，目的とする情報に関連する情報が得られた場合や，何の情報も得られなかった場合には再度資料を選択する段階に移行する．この行程が繰り返され，目的とする情報が取得されるか，情報の取得が断念されると，情報探索が終了する．なお，この資料選択と情報取得をひとつのまとまりとして見たものが情報検索となっている．

この情報探索過程に関するデータを取得する際，一般的にはアクセスログデータが用いられることが多い．アクセスログデータを用いる場合，計算機上のアプリケーションを間に挟むことになるが，データの生成者はそのアプリケーションを利用した一般ユーザの情報探索過程が用いられることになる．しかし，一般ユーザの情報探索過程は，明確な理由なしに資料を選択している場合が多く，どのような資料が信頼に足るかを判断できない場合が多い．そのため，これらの判断が適切に行えるユーザによってアクセスログデータが生成されている必要がある．

そこで，アクセスログと同じく情報探索過程を含んでいるデータであるレファレンスデータに着目する．レファレンスデータは 2.2 において説明した，情報探索のプロであるサーチャによって生成されている．そのため，アクセスログデータと比較してデータ生成者の知識量が多いレファレンスデータは信頼に足るということができる．そのため本稿ではレファレンスデータをアクセスログデータの代替となる情報探索過程データとして用いることにしている．

このレファレンスデータを用いることにより，探索過程において参照された資料とその前後に参照された資料の繋がりに規則性を見出し，資料の特徴に応じて特定のパターンに落とし込むことができれば，情報探索過程に合わせて提示する資料の扱う範囲や専門性を変えることが可能となる．

本稿において実現を目指している，レファレンスデータを用いた情報探索過程の提示システムは，図 1 の情報検索にあた

る部分をシステムのユーザに提示することで，探索方法のサンプルを提示するという形の情報探索支援を行う．アクセスログデータを用いた情報探索過程の記録とは異なり，一定以上の信頼性が担保される情報探索を行うサーチャによる資料へのアプローチ方法を，ユーザ自身の情報探索のサンプルにさせ，探索方法の理解，学習を促すことによる探索支援を目的とする．このとき各情報検索のサンプル提示を行うに当たり必要となる，レファレンスデータを資料選択関連文と情報取得関連文に分割するシステムについて，4. にて提案を行う．

3. 関 連 研 究

本節では，情報探索行動，テキストセグメンテーション及び文書中における話題境界同定に関する関連研究について述べる．

3.1 情報探索行動に関する研究

國本は人間の情報探索行動について，普段接触することのない医学・医療情報を対象とした情報探索を行う中で，情報探索行動の開始メカニズムを説明するモデルを構築した [3].

（1） イベント発生：情報探索を行う原因の発見
（2） 基本課題設定：探索目的の具体化
（3） 探索課題設定：探索対象の具体化
（4） 情報源選択：データを取得する資料の選択
（5） 情報取得開始

國本によると，情報探索行動とは，まず基本課題として情報探索を開始するそもそもの切っ掛け，動機が与えられるイベント発生の段階を経て，情報探索の対象を決定する段階に移行し，次に探索課題として対象のどのような情報を調べるのかを決定，情報源となる資料を選択し，ようやく情報の取得が開始されるものと定義している．

また，人間が対処しなければならない何かしらの状況に置かれた際に，対処行動を促すメカニズムである「文脈の活性化」を重要視しており，要求する情報が明確なものとする段階においても，その後求める情報をどの情報源から得るか選択する段階においても，関連情報の取得や情報源の発見という形での「文脈の活性化」を経て次の情報探索行動の段階へと移行するとしている．逆に，要求する情報が明確なものとならず，課題の設定が明確になされなければ後の情報源選択へと行動が繋がらないという状況が生み出されるため「文脈の活性化」は起こらない．同様に，情報源の選択についても，情報の受け手側が受け取った情報についての理解が困難であると認識する場合にも，情報の受け手側の情報源への関心が低下するため「文脈の活性化」が起こらない．結局，このように文脈の活性化が起こらないときには，次の段階である情報探索の開始へと繋がらない．

3.2 テキストセグメンテーションに関する研究

レファレンスデータの「回答」,「回答プロセス」中に格納されている情報探索過程データの特徴として，データの記述方式が統一されてない点がある．たとえば自然文や書誌情報リストといったさまざま記述方式がレファレンスデータの分析を困難なものとしている．

そこでレファレンスデータの分析を行うための文書の段落分割に関する研究として，中野らによる出現する語と語の近接性に基づいた意味段落の境界判定手法 [4] や，平尾らによる文書中での単語の共起頻度を考慮した出現単語の語彙的結束性と重要度に基づくセグメント法 [5] を用いる手法がある．これらの手法において説明されていた語の近接性，語彙結束性を利用することで参考資料の転換点で話題の切り替えが起こり，特徴的に語彙的結束性が低くなるのではないかと考えられたため情報探索過程の段階化に利用できるのではないかと考えられる．

3.2.1 文書中における話題境界同定に関する研究

松村らは議事録等の議論データから，話題を単位 (セグメント) ごとに切り出し，セグメント間の関連を調べることにより，議論データを構造化し可視化した [6]．彼らの研究では複数文からなるセグメントごとに含まれる単語に対し，tf–idf を基にした値を算出することにより，各セグメントの特徴ベクトルを導出する．具体的には文書中に登場する文を $s_1, s_2, \ldots, s_n$ とし，連続する文の集合をウィンドウ $W(s_i, s_{i+1})$ として表す．そのウィンドウ中で出現する形態素 w について，文中での w の出現頻度を $tf(w, W(s_i, s_{i+1}))$ とした上で，ウィンドウ $W(s_i, s_{i+1})$ に含まれる語それぞれについて，その重みを式 (1) により求める．

$$\log \frac{tf(w, W(s_1, s_n))}{tf(w, W(s_1, s_n)) - tf(w, W(t_i, t_{i+1}))} + 1 \qquad (1)$$

文中での w の出現頻度 $tf(w, W(s_i, s_{i+1}))$ と式 (1) の積を計算することで，w の重みを算出し，隣接する文が同じ話題について取り扱っているか否かを文 i と $i+1$ の特徴を表現するベクトル V_i と V_{i+1} のコサイン類似度 $sim(V_i, V_{i+1})$ を用いて判断する．

$$sim(V_i, V_{i+1}) = \frac{V_i \cdot V_{i+1}}{|V_i| \cdot |V_{i+1}|} \qquad (2)$$

設定した閾値を上回るか否かでテキストに起こされた議論データが各文間においてどのような繋がりを持つかを導出し，議論構造の可視化に用いている．

しかし，この手法は出現する形態素の出現頻度に基づくために，一定以上の長さを持つ文書データでなければ上手く情報探索過程の段階化を行う事ができない．本稿で扱うレファレンスデータは，データ記述者によっては 2，3 文でデータの記述を終了していることもある．そのため，レファレンスデータから資料選択関連情報と情報取得関連情報を取得するのにより適したデータ分割手法を模索する必要がある．

3.2.2 語彙的結束性と単語重要度

テキストセグメンテーションの代表的な手法として，TextTiling 法 [7] という手法が存在する．この手法では文書の意味的に関連の深い部分には，同一の語が繰り返し出現するという性質を利用しており，文書中のある基準点に対してその左右に同数の単語を包含した単語リストを設け，左右の単語リストの類似度を求める．そして基準点を一定間隔でずらしながら類似度の変化に着目し，グラフにおける類似度の極小点を話題の境界と推定する手法をとっている．

しかし，このセグメント手法は文書が一定以上の長さを持つということが前提となっている．比較的短い文書に対して語彙

的結束性に基づき文書のセグメントを行う場合，隣接する単語リストによる類似度が顕著に低くなるため，正確な類似度の計算が行えない．そのためセグメンテーションは十分な精度を得られるとは限らない．

そこで平尾らは，比較的短い文書に対しても適用可能なように文書内における単語重要度の変化に基づくセグメンテーション法を提案した [8]．彼らの手法では，文書集合 DB 中のある文 s に出現する単語 t の重要度 $W(t)$ は，単語 t の出現文書頻度 $df(t)$ と文書集合 DB の大きさ $|DB|$ を用いて以下のように定義される．

$$W(t) = \log \frac{|DB|}{df(t)} \tag{3}$$

したがって，文書集合 DB に含まれる任意の 1 文 $I(s)$ の重要度は，文 s に含まれる単語 t の重要度 $W(t)$ の総和と，s を含む文書に出現する総単語数 K を用いてることで以下のように定義される．

$$I(s) = \sum \frac{W(t)}{K} \tag{4}$$

このように算出された各文の重要度 $I(s)$ の値に着目し，話題境界となる文は重要度が低いという仮定を基に文の重要度 $I(s)$ の変化に着目し，$I(s)$ が極小値となる文で話題が切れると考えセグメンテーションを行っている．

4. 提 案 手 法

レファレンス協同データベースではデータ項目は指定されているが，各項目においてどのように記述を行うかはサーチャの判断に委ねられている．したがって，レファレンス協同データベース中のレファレンスデータはサーチャにごとに異なった記述パターンに基づき構成される．そこで，本節ではまずレファレンスデータを扱う上での問題点と，問題点を考慮した記述内容の区分についてふれた上で，提案するレファレンスデータの段階化手法について述べる．

4.1 レファレンスデータを扱う上での問題点

2.1.2 で述べたように，レファレンスデータには既存のアクセスログデータでは記録されていなかった情報探索過程に関するデータが含まれている．加えて，レファレンスデータは情報探索のプロであるサーチャによるレファレンス業務に基づき記述されている．したがって，2.3 で述べたように，レファレンスデータは探索過程の模倣を行うことを目的としている本研究の対象データとして，非常に信頼できるものである．

また，レファレンス協同データベースを運営している国立国会図書館からサーチャによるレファレンスデータ作成を支援するための標準フォーマットやデータ作成マニュアルはあらかじめ用意されている (注4)．しかし，このマニュアルでは各データ項目における記述形式は厳密に統一されていないため，前述の通り各項目における記述形式はサーチャの判断に委ねられる．そのため現在取得可能なレファレンスデータは，各サーチャが考えた独自の記述形式により，それぞれの項目ごとに自然文で書かれている．一方で，アクセスログデータでは各項目に対する記述形式は，計算機上のアプリケーションにより明確に定められている．このようにレファレンスデータでは記述形式が統一されていないため，アクセスログデータと比較してレファレンスデータからのデータ抽出は非常に困難である．したがって，レファレンスデータから資料情報やサーチャの情報探索行動などの要素を分析するためには，抽出対象である情報探索過程データが格納されているデータ項目ごとに自然文を解析していく必要がある．

4.2 段階化対象項目

2.1.1 のレファレンスデータ項目の説明において，レファレンスデータを作成する上で最低限必要な項目は「管理番号」,「質問」,「回答」の三つであると述べた．また，データ項目には情報探索過程に関するデータが記述されることが想定される「回答プロセス」という項目も存在する．したがって，本稿では「回答プロセス」にも着目する．しかし，レファレンス協同データベース中のレファレンスデータでは,「回答プロセス」は必ずしもデータを記入する必要がない項目である．各項目における記述形式はサーチャの判断に委ねられているため，この項目に何も記述せず，代わりに「回答」内に情報探索過程を含む調査結果データを一括で記入するサーチャも存在する．

よって「回答プロセス」にデータが記入されているか否かによって，まず情報探索過程を段階化するレファレンスデータ中のデータ項目 (段階化対象項目) を以下のように設定する．

- 項目「回答プロセス」にデータあり
  段階化対象項目：「回答プロセス」
- 項目「回答プロセス」にデータなし
  段階化対象項目：「回答」

この条件により段階化対象項目として使用する項目を「回答プロセス」「回答」のうちどちらか片方に設定する．ここで，各段階化対象項目中に存在するデータを俯瞰すると，各サーチャによる各項目に対する記述形式はいくつかのパターンに分類できることがわかった．よって，ここでは各段階化対象項目の記述内容を紹介し，その後各段階化対象項目の記述形式を定義する．

ここで「回答プロセス」「回答」の情報探索過程記述内容について説明する．このデータ項目において記述されるデータはサーチャが行った情報探索の過程であり，探索にあたって利用した情報源と取得した情報が記述されている．このデータ項目における記述形式は以下の 2 パターンに分割できる．

- 自然文回答型
- 資料一覧提示型

図 2 の自然文回答型は,「回答プロセス」「回答」中の文書が自然文による説明形式で記述されている．自然文回答型は，主として二重カギ括弧 (『』) の中に資料名が記述され，次の文で資料から取得した情報の説明が行われる．

図 3 の資料一覧提示型は,「回答プロセス」「回答」中の資料情報が，数字や中黒点 を用いたリストの形式で記述されている．資料一覧提示型では先に記述された資料情報と次に記述さ

(注4)：データ作成 お役立ちツール: http://crd.ndl.go.jp/jp/library/tools.html，2013 年 3 月 22 日閲覧

自然文回答型

参考文献とされている資料①『○○』を確認すると，～である。”□□”の項では～となっている。
また関連する資料として『△△』が見つかった。ここでは～という記述が見つかった。

図 2　記述形式による分類:自然文回答型

資料一覧提示型

参考文献を確認
・資料①『○○』著:～
・資料②『△△』著:～
資料①では～とある。また資料②では～とある。

図 3　記述形式による分類:資料一覧提示型

れる資料情報の分割を主として改行により行っている．資料から取得した情報の記述は，図 3 のように資料□，資料□という形や，自然文回答型のように二重カギ括弧 (『』) の中に資料名を記述するという形で資料紹介が行われたのちに行われる．

4.3　段階化手法

本節では，4.2 で挙げたレファレンスデータの記述形式ごとに適すると考えられる段階化手法について説明するために，まず，情報探索過程の段階化手順における，資料選択関連文，情報取得関連文の抽出と，抽出した各文を情報検索という単位にまとめる部分について詳述する．

レファレンスデータの記述形式はサーチャによって異なるが，データ内に出現する単語や資料名の前後に付随して現れる様々な種類の記号から，レファレンスデータの性質や表示の規則性を見いだすことができる．そのため，単語や記号の機能に基づく単語リストを作成することで，資料選択関連文の抽出を行う．

例を挙げると，括弧 (「」) が持つ性質として，左括弧と右括弧の間に文字をはさみこむことで強調の性質を表し，複数の行に渡って一列に並んでいる中黒点 (・) の後ろに単語が来る場合，中点にはリストの先頭となる性質が生まれる．特にレファレンスデータの場合，リストとして提示されるのは資料に関する情報を表すため，該当する文を資料選択関連文として扱うことが可能となる．

また，レファレンスデータは，基本的に「資料選択を行った後に，その資料から情報取得を行った結果を記述」という規則に基づき記述されている．例外として，資料選択関連文の一文中に資料が複数登場することもあるが，その場合その資料についての情報取得関連文でも，複数の資料から取得した情報を纏めて記述する場合や資料ごとに分けて情報取得文を記述するなどバラバラである．しかし，一文中に資料が複数登場する場合，情報取得関連文において扱っている内容も類似していることが多く，同一のテーマを扱っている情報取得関連文であると考えることが可能である．

さらに，レファレンスデータは情報探索過程について記述されたデータであるため，資料名を含む文が登場してから，次に資料名を含む文が登場する場合には，話題が完全に次の資料の内容にシフトすることが多い．したがって，資料名が登場する文，つまり資料選択関連文が登場してから，次の資料選択関連文が登場するまでの範囲を同一のテーマについて記述しているひとつの情報検索として考えることができる．その結果，次の資料選択関連文の登場を直前の資料からの話題の転換点として考え，情報検索自体の区分点として扱う．

次に，どのようなプロセスを踏んで情報探索過程の段階化を行うのかについて詳述すると,「回答」,「回答プロセス」のデータを文単位に分割し，それらをデータ記述パターンごとに設定した条件に基づき，各文の文頭に資料選択関連文，情報取得関連文のタグを付与することで各文を分類する．その後，文書中の資料選択関連文から次の資料選択関連文が登場するまでを情報検索の一行程として扱い，情報探索過程を情報検索ごとに段階化する．

段階化の具体的な手順は以下の通りである．

（1）　該当データ項目の各文文末に文末タグ $[\mathrm{STC}_1], [\mathrm{STC}_2], \ldots, [\mathrm{STC}_n]$ を付与

（2）　文書中に HTML 改行タグ
 が連続して挿入された直後に開始される文の文頭に改行後文タグ $[\mathrm{NL}_1], [\mathrm{NL}_2], \ldots, [\mathrm{NL}_n]$ を付与

（3）　データ記述形式ごとに設定した条件に基づき，資料名が記述されている文の文頭に資料選択関連文タグ $[\mathrm{SRC}_1], [\mathrm{SRC}_2], \ldots, [\mathrm{SRC}_n]$ を付与

（4）　資料選択関連文タグの付与されていない文の文頭に情報取得関連文タグ $[\mathrm{GI}_1], [\mathrm{GI}_2], \ldots, [\mathrm{GI}_n]$ を付与

この四ステップを経て「回答プロセス」や「回答」の文書に四種類の順序タグを追加し，これらの順序タグを基にレファレンスデータの分析を行う．この順序タグ付与の基準について，自然文回答型のデータでは，文書中の文は主に次のパターンに分かれる．

- タイトルや URL など情報源に関する情報を含む文
- 情報源に関する情報を含まない文

書誌情報や外部サイトの URL といった情報源に関する情報を含む文においては，主にカギ括弧や二重カギ括弧の中，中黒点の直後などに書誌情報を記述している．これら書誌情報や外部サイトの URL といった情報源に関する情報が記述されている文から，次に同様の情報が記述されている文の文頭までを情報探索における一過程として順序タグを付与する．

ここで自然文回答型の情報探索過程データへ各タグを付与する手法の具体例として，図 2 で紹介した自然文回答型のデータへのタグ付け結果を図 4 に示す．

（1）　各文の文末に文末タグ $[\mathrm{STC}_1][\mathrm{STC}_4]$ を付与

（2）　改行後の文の文頭に $[\mathrm{NL}_1]$ を付与

（3）　資料名記述用の二重鍵括弧『』を検出

（4）　二重鍵括弧を持つ文を資料選択関連文として検出

（5）　資料選択関連文の文頭に $[\mathrm{SRC}_1], [\mathrm{SRC}_2]$ を付与

自然文回答型

[SRC$_1$] 参考文献とされている資料①『○○』を確認すると，～である。[STC$_1$] [GI$_1$] "□□"の項では～となっている。[STC$_2$]
[NL$_1$] [SRC$_2$] また関連する資料として『△△』が見つかった。[STC$_3$] [GI$_2$] ここでは～という記述が見つかった。[STC$_4$]

図 4　タグ付与済み自然文回答型データ

（6） 資料選択関連文以外の文を情報取得関連文として検出
（7） 情報取得関連文の文頭に [GI$_1$], [GI$_2$] を付与

一方，資料一覧提示型へのタグ付与を行う場合，文書中の文は主に次のパターンに分けられる．

- 改行後の中点，ハイフンなどの直後の資料情報記述文
- 上記以外の文

資料一覧提示型の場合，資料選択に関する情報と取得した情報についての文が明確に分割されている．そのため改行タグや中点，ハイフンといった，資料から次の資料への話題転換が明確に行われている文と考えられるものに対して資料選択関連文タグを付与し，そうでない文に情報取得関連文タグを付与する．

資料一覧提示型の情報探索過程データへ各タグを付与する手法の具体例として，図 3 で紹介した資料一覧提示型のデータへのタグ付け結果を図 5 に示す．

資料一覧提示型

[GI$_1$] 参考文献を確認 [STC$_1$]
[NL$_1$] [SRC$_1$] ・資料①『○○』著：～ [STC$_2$]
[NL$_2$] [SRC$_2$] ・資料②『△△』著：～ [STC$_3$]
[NL$_3$] [GI$_2$] 資料①では～とある。[STC$_4$]
[GI$_3$] また資料②では～とある。[STC$_5$]

図 5　タグ付与済み資料一覧提示型データ

（1） 各文の文末に文末タグ [STC$_1$] ～ [STC$_5$] を付与
（2） 改行後の文の文頭に [NL$_1$] ～ [NL$_3$] を付与
（3） 資料名記述用の改行タグ隣接中点・を検出
（4） 改行タグ隣接中点保有文を資料選択関連文として検出
（5） 資料選択関連文の文頭に [SRC$_1$], [SRC$_2$] を付与
（6） 資料選択関連文以外の文を情報取得関連文として検出
（7） 情報取得関連文の文頭に [GI$_1$] ～ [GI$_3$] を付与

このようにタグを付与してレファレンスデータの項目「回答」「回答プロセス」より情報探索過程の各過程の分割ポイントを指定する．

資料選択関連文タグ [SRC] から文末タグ [STC] の間の文は資料情報関連文として分類され，情報取得関連文タグ [GI] から文末タグ [STC] の間の文は情報取得関連文として分類される．このようにして分類された各文に対して，資料選択関連文 [SRC$_n$] から [SRC$_{n+1}$] までの間の文をひとつの情報検索として認識，各タグのナンバリングに基づき段階化を行う．

5. 評 価 実 験

本稿では情報探索過程の段階化の精度を測るために，情報探索過程の段階化が行えているか否かを，実際にユーザの評価を基に測定した．

本稿で行う評価実験における比較手法として，3.2.1 で述べた文書中における話題境界同定に関する研究 [6] にて提案されている話題境界同定手法を用いた．ただし，レファレンスデータは比較手法で従来用いられていた議論データと比較すると括弧や中点といった記号が頻繁に登場することがある．しかし，文の内容を表すのは名詞や動詞であり，記号はデータ中の語の重みづけを行う際に内容とは全く関係ないところで重みづけされる可能性がある．したがって，記号は重みづけの対象から除外する必要がある．よって，比較手法においては話題境界の同定に用いる形態素から記号，助詞，助動詞そして接続詞を除外した．

また，評価実験に用いるデータはレファレンス協同データベースからランダムに取得したレファレンスデータである．そして，このレファレンスデータに対して，実際にユーザが情報探索過程の段階化を行った文の切れ目 (段階化位置) を基に提案手法及び比較手法における段階化精度の比較を行った．

本稿では段階化精度の評価について，情報検索分野の評価基準として用いられる精度，再現率 [9] を基に作成した正確性と網羅性という基準を設定する．この基準はレファレンスデータごとに計算される値である．そのため，各レファレンスデータに対し正確性と網羅性の計算を行い，その調和平均を算出し，各手法毎に算出された調和平均の平均をもとに精度の比較を行う．

正確性は，人間による段階化位置と各手法による段階化位置が，どの程度正確に合致しているかを測る指標である．一方，網羅性は，人間により設定された段階化位置を，各手法がどの程度網羅できているかを測る指標である．ここで，レファレンスデータ d_i における，人間による文書分割数を $n(A_i)$，各手法による文書分割数を $n(S_i)$，人間による段階化位置と各手法による段階化位置との合致位置数を $n(A_i \cap S_i)$ とする．よって，正確性 P_i 及び網羅性 R_i は以下の式 (5) 及び式 (6) のようになる．

$$P_i = \frac{n(A_i \cap S_i)}{n(S_i)} \tag{5}$$

$$R_i = \frac{n(A_i \cap S_i)}{n(A_i)} \tag{6}$$

したがって，レファレンスデータ d_i における正確性と網羅性の調和平均 H_i は，以下の式 (7) のようになる．

$$H_i = \frac{2}{\frac{1}{P_i} + \frac{1}{R_i}} \tag{7}$$

正確性及び網羅性は 1 を超えないため，調和平均が 1 に近ければ近いほど，ユーザの情報探索過程段階化のアプローチに類似した段階化を行うことができるといえる．したがって，調和平均の平均が大きい手法がユーザの情報探索過程段階化のアプローチに類似した段階化手法である．そこで，提案手法と比

較手法による調和平均の平均を両側 t 検定 [10] にかけることで，提案手法が比較手法よりも人間の情報探索過程段階化のアプローチに類似するものであるかどうかを評価するため，手法間の正確性および網羅性による調和平均の平均が，偶然誤差の範囲に存在する場合を帰無仮説とする検定を行った．

この帰無仮説が棄却される場合，比較手法よりも提案手法による調和平均の平均の値が，基準として設定されている，人間により行われた段階化に近いといえる．

表 1　提案手法及び比較手法に基づくレファレンスデータの段階化

	提案手法	比較手法
調和平均の平均	**0.902**	0.627

表 5. の結果より，帰無仮説を採択できる確率 p が $p < 0.05$ となり帰無仮説は有意水準 5% において棄却され，比較手法よりも提案手法が有意となる．よって，比較手法よりも提案手法が人間が行う情報探索過程の段階化への類似という観点において優れているといえる．

このような結果が得られた理由としては，資料提示における一定の形式がサーチャ間で共有されているためであると考えられる．レファレンス協同データベースは，情報探索支援を目的として運営されているため，特に資料情報の提示が重要である．そのため，データ作成にあたって，どのような資料からどのような情報を取得したかについて明確に記述する必要がある．その手段の一つがレファレンスデータ記述における一定の記述形式の共有である．今回その記述形式をルールとして表現できたため，人間による段階化アプローチに近い段階化を行う事が可能となったといえる．

6. おわりに

本稿では，図書館での蔵書検索において，情報探索過程提示の際の問題の一つである情報探索過程の段階化を行い，その結果レファレンスデータの段階化においてルールベースでの段階化が有効であることが分かった．

今後は，今回表現したルールを用いて，レファレンスデータにおけるサーチャの情報探索行動データ抽出に取り組む．また評価実験に用いるデータ数を増やし，さらなる精度向上を図る．これからの課題として，利用者が求める情報を取得する過程をユーザに提示し，その過程を真似て情報源の特定を支援する仕組みを作成するために，段階化を行った情報探索過程を用いたユーザへの情報探索支援を行う方法を検討する．

謝辞　本研究の一部は，独立行政法人日本学術振興会 科学研究費補助金 若手研究 (B) (課題番号：22700248) によるものである．ここに記して謝意を表す．

文　献

[1] 寺井仁, 種市淳子, 逸村裕. 情報要求と情報源利用に関するプランニングが情報探索行動に与える影響. 名古屋大学附属図書館研究年報, No. 6, pp. 39–45, 2007.

[2] 種市淳子, 逸村裕. Web の探索行動と情報評価過程の分析. 名古屋大学附属図書館研究年報, Vol. 3, pp. 1–13, 2005.

[3] 國本千裕. 情報探索行動の開始メカニズム–医学・医療情報の探索実例を通じて. *Library and information science*, No. 64, pp. 55–79, 2010.

[4] 中野滋徳, 足立顕, 牧野武則. 語の近接性に基づいた意味段落境界の判定手法 (解析). 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2005, No. 22, pp. 23–30, 2005.

[5] 平尾努, 北内啓, 木谷強. 語彙的結束性と単語重要度に基づくテキストセグメンテーション. 情報処理学会論文誌. データベース, Vol. 41, No. 3, pp. 24–36, 2000.

[6] 松村真宏, 加藤優, 大澤幸生, 石塚満. 議論構造の可視化による論点の発見と理解. 知能と情報 (日本知能情報ファジィ学会誌), Vol. 15, No. 5, pp. 554–564, 2003.

[7] Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, Vol. 23, No. 1, pp. 33–64, March 1997.

[8] 平尾努, 北内啓, 木谷強. 語彙的結束性と単語重要度に基づくテキストセグメンテーション. 情報処理学会論文誌. データベース, Vol. 41, No. 3, pp. 24–36, may 2000.

[9] R.A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., second edition, 2011.

[10] 宿久洋, 村上享, 原恭彦. 確率と統計の基礎. Minerva 数学講義, 第 2 巻. ミネルヴァ書房, 2011.