レファレンス事例の分析による論文検索に効果的な要素の調査

難波 英嗣

広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1

E-mail: nanba@hiroshima-cu.ac.jp

あらまし 本研究では，第一回 NTCIR ワークショップ情報検索タスクの検索課題およびレファレンス協同事例や Yahoo! 知恵袋の回答欄に学術論文データベースへのリンクがあるものを対象に，論文調査目的のレファレンス事例を分析し，求める論文を特定するのにどのような情報が効果的かという観点で，検索状況および検索手法を体系的に調査する．

キーワード 情報検索, 情報要求, 学術論文，レファレンス事例

1. はじめに

学術論文検索では，情報要求を満たす論文を網羅的に探す必要がある．しかし，初学者にとって，それはなかなか容易ではない．そこで，著者らは，初学者が論文を検索する際，適切な検索語を推薦することで，検索作業を支援するシステムの開発に取り組んでいる．

このような検索支援システムを開発するには，「どのような情報要求に対し，検索者がどのようなプロセスで検索しているのか」を調査する必要がある．そこで，本研究では，論文調査目的のレファレンス事例を大量に収集し，情報要求およびプロセスを体系的に分析する．

本論文の構成は以下のとおりである．次節では，論文検索および情報要求の分類に関する関連研究について述べる．3 節では，論文調査目的のレファレンス事例の収集および分析結果について報告し，4 節で本稿をまとめる．

2. 関連研究

2.1 学術論文を対象とした情報検索

学術論文を対象とした情報検索の研究は古くから行われており，Cranfield，Medlars，CACM，NPL，LISA など，数多くのテストコレクションが作られてきた．より最近の研究プロジェクトとして，TREC で行われた Chemical IR トラックがある[1]．このトラックの中で実施された技術サーベイタスクでは，化学のある分野の動向を知るために必要な論文と特許を検索することを目的としている．このタスクでは，例えばある化合物の画像ファイルおよび構造ファイルが検索システムの入力として与えられ，その化合物に関する特許と論文を検索することが求められる．

上述のプロジェクトは英語を対象にしているが，日本語論文を対象にしたものに，第一回および第二回 NTCIR ワークショップで行われた情報検索タスクがある[2, 3]．図 1 は，実際に NTCIR-1 情報検索タスクで使われた検索課題の例であり，本研究では，NTCIR-1 情報検索タスクのデータを分析に用いる．

```
<TOPIC q=0005>
<TITLE>
特徴次元リダクション
</TITLE>
<DESCRIPTION>
クラスタリングにおける特徴次元リダクション
</DESCRIPTION>
<NARRATIVE>
オブジェクトのクラスタリングを行なうとき、オブジェクトを特徴ベクトルで表現することが望まれる。アプリケーションによっては、オブジェクトの次元は数千、数万となることがある。このような場合、事前に次元を落とすことが必要になる。正解文書は、特徴次元リダクションの方法について、理論面から、または実験によって、提案、比較などを行なっているもの。画像処理などの実験の操作の一部として特徴次元リダクションを用いているだけでは要求を満たさない。
</NARRATIVE>
<CONCEPT>
特徴選択，主成分分析，グラフ理論，情報の粒度，幾何クラスタリング
</CONCEPT>
<FIELD>
1.電子・情報・制御
</FIELD>
</TOPIC>
```

図 1 NTCIR-1 検索課題の例 ([2]より抜粋)

2.2 情報要求の分類

渡邊ら[4]は，QA サイトの質問を，以下の 5 つのタ

イプに分け，機械学習に基づく手法で，自動的に分類する手法を提案している．

- **事実**：事象の定義，真実，客観的な理由や手法を問う質問
- **根拠**：客観的な根拠，理由を問う質問
- **経験**：回答者の経験や体験がなければ回答できない質問
- **提案**：問題の解決方法を問う質問や情報提供を依頼する
- **意見**：推測，嗜好など，主観的に回答をしてよい質問

上述のものとは定義は若干異なるものの，林ら[5]も，渡邉らと同様に，質問を「事実」,「根拠」,「経験」,「提案」,「意見」の 5 つのタイプに分類している．本研究でも，QA コンテンツを分類するという点ではこれらの研究と共通するが，論文調査目的のレファレンス事例に対象を限定し，学術に特化したカテゴリを扱う点が異なる．

3. レファレンス事例の収集と分析

3.1 レファレンス事例の収集

以下の 3 種類のデータベースを論文調査目的のレファレンス事例として収集し，分析に用いる．

- 第 1 回 NTCIR ワークショップ情報検索タスクの検索課題
- Yahoo!知恵袋データベース
  知恵袋の各エントリで，回答欄に NII 学術情報ナビゲータ(CiNii)へのリンクがあるものは論文調査目的のレファレンス事例と考えて，質問と回答の対を収集する．
- レファレンス協同事例データベース(レファ協)[1]
  知恵袋と同様に，回答欄に NII 学術情報ナビゲータ(CiNii)へのリンクがあるものは論文調査目的のレファレンス事例と考えて，質問と回答の対を収集する．

上記の 3 種類のデータベースから収集したレファレンス事例数を表 1 に示す．

表 1　収集した論文調査目的のレファレンス事例数

データベース	事例数
NTCIR-1	83
Yahoo!知恵袋	442
レファ協	3032
計	3557

[1] http://crd.ndl.go.jp/reference/

また，実際に知恵袋とレファ協から収集したレファレンス事例を図 2 および図 3 に示す．知恵袋の各エントリは，エントリ投稿時に投稿者自身が知恵袋で設定されているカテゴリを付与することになっている．知恵袋カテゴリは，3 階層から構成されており，最下層で約 300 カテゴリ存在する．図 2 の事例は，「教養と学問、サイエンス>生物、動物、植物>動物」というカテゴリに分類されている．一方，レファ協の各エントリには，日本十進分類法(NDC)の分類コードが付与されている．NDC は，3 階層から構成されており，最下層で約 950 カテゴリ存在する．このうち，図 3 の事例には 2 つのコード「食品．料理」と「衛生学．公衆衛生．予防医学」が付与されている．レファ協レファレンス事例には，さらに，回答を導き出す過程(回答プロセス)についても記載されている場合がある．3.2 節および 3.3 節では，質問と回答だけでなく各事例に付与されたカテゴリや回答プロセスを用いたレファレンス事例の分析結果について報告する．

質問：皇居の生物相について皇居にはたくさんの生物が生息していると知ったのですがどのくらいの生物が生息しているのですか
回答：1996～2000 年度の生物相調査で 3,638 種の動物と 1,366 種の植物が見いだされたそうです。動物のうちの殆どは昆虫をはじめとした節足動物です。この報告は「国立科学博物館専報」に掲載されていまして、一部パソコン上で無料で公開されているのでご覧になってはいかがでしょう。 http://ci.nii.ac.jp/naid/110004313456 http://ci.nii.ac.jp/naid/110004313489 http://ci.nii.ac.jp/naid/110004313511
カテゴリ：教養と学問、サイエンス>生物、動物、植物>動物

図 2　Yahoo!知恵袋のレファレンス事例
(http://detail.chiebukuro.yahoo.co.jp/qa/question_detail/q12115380622)

質問：サフランの代わりにクチナシの乾燥果実を使ってもち米を蒸してみたところ、緑がかった黄色に染まってしまった。黄色をきれいに出す方法はないか。
回答：山口加代子,木咲弘「クチナシ果実で黄色に染めたおこわの緑変について」(調理科学 22-3)、同「クチナシ果実で黄色に染めたおこわの緑変について―ゲニポサイドの青変反応」(調理科学 26-3) をご覧いただいた。
回答プロセス：料理書コーナーや栄養学の書架を確認したが、関連した記述のある図書を見つけられなかった。そこで CiNNi(ママ) Artlicles 日本の論文をさがす（http://ci.nii.ac.jp/）にて「クチナシ」をキーワードに検索した中から 2 論文を紹介した。
NDC：食品．料理　(596　9 版)，衛生学．公衆衛生．予防医学　(498　9 版)

図 3　レファレンス協同事例のレファレンス事例 (http://crd.ndl.go.jp/reference/modules/d3ndlcrdentry/index.php?page=ref_view&id=1000163931)

3.2 分析対象とするレファレンス事例の選定

3.1 節の表 1 で示した 3 種類のレファレンス事例について，カテゴリごとの分布を調査した．それぞれの調査結果を表 2～表 4 に示す.いずれの表においても，レファレンス事例の件数が多い順に上位最大 10 カテゴリを示している．

表 2　NTCIR-1 のカテゴリごとのレファレンス事例数

カテゴリ名	件数
1. 電子・情報・制御	68
8. 人文・社会	21
7. 医学・歯学	10
6. 工学	10
4. 生物学・農学	7
3. 建築・土木・造園	4

表 3　Yahoo!知恵袋のカテゴリごとのレファレンス事例数

カテゴリ名	件数
教養と学問、サイエンス>言葉、語学>英語	58
教養と学問、サイエンス>宿題	35
健康、美容とファッション>健康、病気、病院>病気、症状、ヘルスケア	34
子育てと学校>大学、短大、大学院>大学	23
教養と学問、サイエンス>一般教養	23
教養と学問、サイエンス>生物、動物、植物>植物	20
教養と学問、サイエンス>歴史>日本史	20
教養と学問、サイエンス>数学、サイエンス>化学	19
教養と学問、サイエンス>言葉、語学>日本語	18
教養と学問、サイエンス>数学、サイエンス>工学	17

表 4　レファ協のカテゴリごとのレファレンス事例数

カテゴリ名	件数
日本史　(210　9 版)	136
日本文学　(910　9 版)	56
中国　(222　9 版)	45
詩歌　(911　9 版)	44
個人伝記　(289　9 版)	41
小説．物語　(913　9 版)	40
経営管理　(336　9 版)	39
社会福祉　(369　9 版)	26
朝鮮　(221　9 版)	26
詩歌．韻文．詩文　(921　9 版)	26

まず，NTCIR-1 の事例について述べる．表 2 において，今回は科学技術系の分野を想定しているため，「8. 人文・社会」カテゴリの事例は分析対象から外した．なお，NTCIR の各検索課題には 2 つ以上のカテゴリが付与されているものもある．表 2 に示す数字は，

次に,知恵袋の事例について述べる.表 3 において，一番件数の多い「教養と学問、サイエンス>言葉、語学>英語」カテゴリは，ある日本語の専門用語を英語でどのように表現するのかについて質問したものである．このカテゴリの事例は，回答中で CiNii の論文が引用されているものの，必ずしも論文を引用する必要性はなく，今回の分析に馴染まないものが多いため，分析対象から外した．また，NTCIR-1 と同様，社会科学系などの分野の事例も分析対象から外した．

最後に，レファ協の事例の傾向について述べる．表 4 からわかるとおり，レファ協の事例は社会科学系の事例が多い．上述のとおり，本研究では科学技術系の分野を想定しているため，該当するカテゴリ(4. 自然科学，5. 技術，6. 産業)のレファレンス事例を分析対象とした．ただ，これらの事例の中には，以下の例のように,目的の文献はすでに書誌情報が判明しており，その入手方法について質問するという場合が数多く存在した．

『京都大学人文科学研究所創立二十五周年記念論文集』の倉田淳之助「「説郛」版本諸説と私見」を探している。

このような質問はレファ協事例にのみ現れ，その多くは「[文献情報]を探して」または「[文献情報]をさがして」というパターンで記述されていた．そこで,「を(探|さが)して」という表現を質問文中に含む事例は分析対象から外した．最終的に，表 5 に示す全〇〇件のレファレンス事例を対象に分析を行うことにした.次節で，分析結果について報告する．

表 5　分析対象となるレファレンス事例数

データベース	事例数
NTCIR-1	62
Yahoo!知恵袋	268
レファ協	466
計	796

3.3 レファレンス事例の分析

レファレンス事例の分析は，2 種類の分析：(1)質問の分析および(2)回答プロセスの分析を行った．以下に，その結果を報告する．

質問の分析

NTCIR-1，2 の検索課題，Yahoo!知恵袋およびレファ協の質問文を対象にした．レファレンス事例を，その内容に応じて以下の 4 種類に分類した．

(1) コト(手順，手法，仕組み)に関する質問

手順，手法，仕組みに関する質問で，以下はその例である．

- 衣服の虫喰いについて、予防できる方法が知りたい。（染色・加工など）また、虫の好む色、におい、布の種類なども知りたい。(レファ協：衣服．裁縫(593))
- ピルビン酸を直接、呈色させ検出・定量する方法はありませんか？(知恵袋：教養と学問、サイエンス>サイエンス>化学)
- TCP を無線通信制御に適用するための改良方法についての論文はないか。(NTCIR-1：1. 電子・情報・制御)

(2) モノに関する質問

モノに関する質問で，そのモノが持つ意味，性質，効果，あるいはそのモノに関する具体的な事例などが回答として望まれる．以下は，その例である．

- 頭蓋骨早期癒合症という病気や治療について知りたい。(レファ協：医学(490))
- ブリルアン散乱光について知りたいのですが？ブリルアン散乱光とは何なのか？また、発生原因を教えていただきたいのですが？(知恵袋:教養と学問、サイエンス>天気、天文、宇宙)
- 動画像圧縮を行なう知能化イメージセンサに関する研究が知りたい。(NTCIR-1：1. 電子・情報・制御，6. 工学)

(3) モノの属性値に関する質問

モノに関する質問であるが，質問文中にモノとその属性名が記載されており，属性値が回答として望まれる．以下の例において，下線がモノ，破線が属性名を示す．

- 玄米の吸水率が載っている本はないか。(レファ協：食用作物(616))
- 材料力学についての質問です．鋼材が火災などにより加熱された時鋼材の強度は低下しますが，その後鋼材の加熱冷却後の強度はどのようになるのか教えていただけませんか．また、参考文献などがあれば教えていただけませんか．よろしくお願いします．(知恵袋：教養と学問、サイエンス>芸術、文学、哲学>建築)
- モバイル環境におけるグループウェアの問題点とはなにか。(NTCIR-1：topic 0036)

(4) その他の質問

(1)から(3)のいずれにも該当しない質問で，以下はその例である．

- くらげの水分を利用して化粧品として開発することが可能かどうか知りたい。(レファ協：漁撈．漁業各論　(664))
- FET 回路の実験で、周波数と、入力波形と出力波形の位相のずれは何か関係があるのでしょうか？（比例関係とか反比例の関係とか）(知恵袋：教養と学問、サイエンス>サイエンス>工学)
- 設計・製造・保全などの人工物のライフサイクルの異なる部門間での情報共有および知識共有について報告した論文・記事が欲しい。(NTCIR-1：1. 電子・情報・制御，6. 工学)

回答プロセスの分析

質問の分析における(1)～(4)を考慮し，レファ協の回答プロセスを分析した．

(a) より適切なクエリへの修正

以下は「(2)モノに関する質問」の例で，クエリ「頭蓋骨早期癒合症」を，より一般的な表現である「頭蓋骨縫合早期癒合症」や「頭蓋縫合早期癒合症」に修正し，検索している．

質問：頭蓋骨早期癒合症という病気や治療について知りたい。
回答プロセス：1.医学辞典で病名を確認する。『南山堂医学大辞典』p1318～1319 に「頭蓋骨縫合早期癒合症(ずがいこつほうごうそうきゆごうしょう)」の記述あり。2.Google で「頭蓋骨縫合早期癒合症」をキーワードに検索すると多数ヒット。(略) 4.GeNii で「頭蓋縫合早期癒合症」で検索すると、論文３３件、WebcatPlus で資料３件がヒット。

(b) 同義語，上位語拡張

以下は，「(3) モノの属性値に関する質問」の例で，「玄米」を，その上位語である「コメ」に拡張して検索している．上位語を使う代わりに検索対象文書に付与されているカテゴリで結果を絞ったり，特定ドメインのデータベースで検索できなければ，Google で検索したりする，といった事例も見つかった．なお，属性語の「吸水率」は，ここでは「吸水」や「浸漬」に拡張されているが，これについては，後の(d)で述べる．

質問：玄米の吸水率が載っている本はないか。
回答プロセス：Webcatplus で「玄米　吸水」「コメ　吸水」「玄米　浸漬」「コメ　浸漬」をキーワードに検索→ヒットしない。2．自館 OPAC で「炊飯」「米飯」「コメ」をキーワードに検索し、載っていそうな本を見る。

(c) 書籍を対象にした検索

以下は，「(1) コト(手順，手法，仕組み)に関する質問」の例であり，特に一般性の高い手順について調べる時には，書籍を対象にするケースがあった．

質問 1：家庭で出た生ごみをダンボールに入れて堆肥にする方法を知りたい
回答プロセス：当館蔵書検索 書名＝“生ごみ”で検索

質問 2：外来昆虫のアルゼンチンアリの駆除方法についてまとめられた文献はあるか。
回答プロセス：自館 OPAC を使って、Kw：“外来&事典”で検索する。

(d) 属性語の拡張・省略

以下は，「(3) モノの属性値に関する質問」の例である．このケースでは，属性語である「デメリット」をクエリから除外して検索している．上述の(b)の例では，「吸水率」を「吸水」や「浸漬」などに言い換えて検索している．なお，以下の例において，「肝臓移植」は「肝移植」で検索されているが，これは上述の(a)に該当する．

質問：肝臓移植のドナーのデメリットについて書かれた本はあるか。
回答プロセス：49（医学）の所蔵資料を「肝移植」で検索し、取寄せて内容を確認。

4. おわりに

本研究では，NTCIR 情報検索タスク，Yahoo!知恵袋，レファレンス協同事例データベースから収集した論文調査目的のレファレンス事例を 3557 件収集し，このうちの 796 件を用いて，質問および回答プロセスを分類した．今後は，この分析で得られた知見，すなわち，質問の種類に応じて検索方法やクエリ拡張の方法が異なることを考慮した検索システムを実装し，検索精度が向上できることを実証する．

謝辞

本研究の一部は科学研究費補助金(基盤研究(A))(研究課題番号:15H017214)および広島市立大学受託研究費(科学技術振興機構)の支援を受けて行われた．Yahoo!知恵袋データは，ヤフー株式会社から提供いただいた．論文データは，国立情報学研究所および科学技術振興機構から提供いただいた．論文検索テストコレクションは，国立情報学研究所主催の第 1 回，第 2 回 NTCIR ワークショップのものを利用させていただいた．ここに記して謹んで感謝の意を表する．

参　考　文　献

[1] M. Lupu, F. Piroi, J. Huang, J. Zhu, and J. Tait, “Overview of the TREC Chemical IR Track”, Proceedings of the 18th Text Retrieval Conference, 2009.

[2] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka, “Overview of IR Tasks”, Proceedings of the 1st NTCIR Workshop Meeting, 1999.

[3] N. Kando, K. Kuriyama, and M. Yoshioka, “Overview of Japanese and English Information Retrieval Tasks”, Proceedings of the 2nd NTCIR Workshop Meeting, 2001.

[4] 渡邊 直人，島田 諭，関 洋平，神門 典子，佐藤 哲司，“QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討”，DEIM Forum 2011，2011.

[5] 林 秀治，山本 和英，“質問意図による QA サイト質問文の自動分類”，電子情報通信学会技術研究報告，思考と言語，113(82)，pp.51-56，2013.