情 報 学 基 礎 61−21
自 然 言 語 処 理 142−21
（2001. 3. 6）

目録データベースとWebコンテンツの統合的利用方式

杉田 茂樹
北海道大学附属図書館

江口 浩二
国立情報学研究所

Web上に分散する図書情報とくに書評コンテンツの探索において，目録データベースを用い，より有効な検索を行う方法について述べる．提案する手法は，利用者が特定した図書に関する書誌情報を目録検索システムに問合せ，それを手がかりにWeb上で提供される書評を検索し，その結果に対して自動編集を施した上で，利用者に提示するものである．試作システムによる評価を行い，基礎的な実験に基づいて本方式の有効性を確認した．

Integrated Utilization Method of Library Catalog Database and Web Contents

Shigeki SUGITA
Hokkaido University Library

Koji EGUCHI
National Institute of Informatics

We propose an effective method for searching information on books, especially book reviews, which is scattered on the Web, by making use of library catalog database. Our proposed method retrieves the bibliography record of a user specified book using a bibliographic information retrieval system. Using them, it retrieves book reviews on the web, which are then automatically edited and finally presented to the user. We implemented the prototype system and evaluated the effectiveness through preliminary experiments.

1 はじめに

ユーザが特定した図書に関連する，Web上の非定型な書評に着目し，それを対象とした情報検索インタフェースを考える．Web上には，出版社や流通業者など出版サイドから，また，個人や読書サークルなどの読者サイドから，大量の書評（以下，オンライン書評）が公開されており，それらの書式は統一されていない．現状，こうしたオンライン書評に効率的にアクセスするためにはWebサーチエンジンを用いる方法が考えられるが，さまざまな関心を持った不特定多数のユーザの検索要求に対応すべく設計されたWebサーチエンジンからは，書評を探索するユーザにとって必ずしも最適な検索結果集合が得られるとは限らない．

難点のひとつは，ユーザにとってクエリの選定が困難である点である．たとえば書名による検索では，とくに書名が一般的な語彙である場合など，当該図書についてのオンライン書評以外の大量の文書がヒットしてしまう．また，逆に特定性が高いと考えられるISBN(国際図書標準番号）による検索では，ISBNが記述されていない文書が検索結果から洩れてしまう．

もう一つの難点として挙げられることに，仮に，適切な検索結果集合が得られたとしても，ユーザはそこで発見したオンライン書評を順次アクセスしなければならず，それに加え，オンライン書評が非定型であることもあいまって，ユーザの負荷を増長させていることがある．

本研究は，上記の難点を解消することを目指し，書籍に関する二次情報を集積した目録検索システムを援用することを提案する．目録検索システムには，代表的なものに国立情報学研究所の提供する目録所在情報サービス [1](以下，NACSIS-CAT) や，図書館の提供する OPAC (オンライン蔵書検索システム，例えば [2]) などがあり，図書の書誌情報・所在情報の提供を主眼としている．提案するシステムでは，検索処理中，こうした目録検索システムに自動的にアクセスし，そこで得られる書誌情報等を用いてユーザのクエリの補完，検索結果のフィルタリング，ランキングを行うことにより第 1 の難点を，さらに，自動編集によりオンライン書評の一覧を整理された状態で提示することにより第 2 の難点を克服し，ユーザの要求に応えようとするものである．さらに目録検索システムから得られる所在情報により，ユーザは図書館等における関心対象の図書の所蔵状況を併せて入手することができる．

以下，2 章で提案手法の概要について述べ，3 章では提案手法の実装・評価について述べ，最後に 4 章でまとめを行う．

2 提案手法の概要

特定された図書の固有情報[1]を入力とした以下の方式を提案する (図 1)．

1. 入力をもとに目録検索システムより当該図書の書誌，所在情報を取得する．
2. 取得した書誌情報 (書名・著者名[2]) でユーザのクエリを補完し，オンライン書評のインデックスに対して問合せを行う．
3. 問合せ結果に含まれるオンライン書評コンテンツを順次走査し，検索結果の自動編集を施す．
4. オンライン書評の編集結果に加えて，所蔵データベースの問合せ結果である書誌，所在情報[3]をも併せてユーザに提示する．

[1]ISBN を用いる．
[2]書評の書き手により書名・著者名の表記には揺れがあるので，柔軟な文字列照合の技術 ([3] など) の援用が望ましい．
[3]図書館 OPAC を用いた館内所在情報や，あるいは，NACSIS-CAT を用いた全国大学図書館の所在情報などが考えられる．

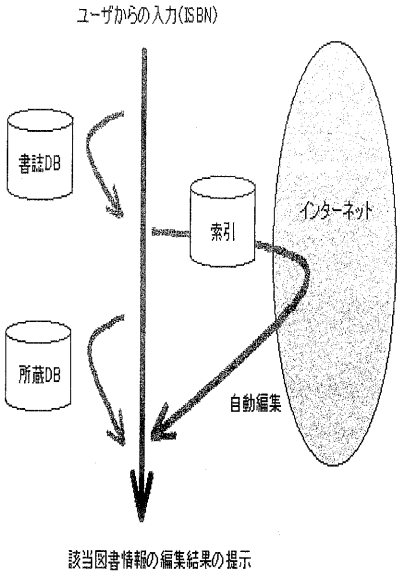

図 1: 提案手法の概要

2.1 前提となる要素技術

提案する手法では，あらかじめ，オンライン書評のインデックスを作成しておくことが前提となる．ある主題に特化した資源収集を行う手法については Chakrabarti らによる研究 [4] などがなされているところであるが，本研究では，係る技術を要素技術として用いるなどにより，オンライン書評の URL や特徴語などの情報を含むインデックスが既知である環境を想定し，以下検討を行う．

2.2 オンライン書評について

検索結果の自動編集のための予備調査を行った結果を以下に記す．

調査は，ある書評関連 WWW サイトリンク集からリンクされた 87 サイトを対象に，書評を内容とする HTML 文書を 1 サイトにつき 1 ページとりあげて，その特性を調べたものである．これらの文書は主として個々人により作成，運用され，その多くは現代文学を対象としていた．

	サイト数 (87 サイト中)	
複数	71 サイト	(81%)
単一	16 サイト	(19%)

表 1: 一つの HTML 文書中に含まれる書評の数

オンライン書評では，一つのHTML文書内で複数の図書を扱っていることが多かった(表1)．なお，以下，一つのHTML文書に含まれる，単一の図書に対するまとまりのある書評の部分を「書評単位」と呼ぶ．

一つのHTML文書が複数の書評単位で構成される場合，システムが自動的にこれを書評単位ごとに分割し，該当図書に関する書評単位を切り出して提示するといった，自動編集が必要となる．書評単位の区切りを機械的に判定しうるタグ情報の有無を見ると(表2)，ほぼ3分の2が各書評単位の区切りにNAME属性を持つアンカータグを置いていることがわかる．さらに，こうした場合では，自サイト内に書評にたどりつくための目次ページを持ち，

目次 <A HREF="URL#XXX">書名</A>
↓
書評単位 <A NAME="XXX">書名</A>

というリンク構造を有していることが多かった(48サイト中43サイト)．

	サイト数(87サイト中)	
<A NAME>	48サイト	(67%)
<TR>	8サイト	(11%)
<TABLE>	5サイト	(7%)
その他・判別不能	10サイト	(15%)

表2: 複数書評単位の区切り

書評単位に含まれる書誌的事項について調査を行なった．その結果を表3に示す．

	サイト数(87サイト中)	
書名	87サイト	(100%)
著者名	69サイト	(79%)
ISBN	7サイト	(9%)
出版元	68サイト	(68%)

表3: 書評単位に含まれる書誌的事項

	サイト数(87サイト中)	
書名	50サイト	(57%)
『書名』	24サイト	(27%)
「書名」	13サイト	(16%)

表4: 書名の提示方法

すべての書評単位は，対象図書の書名を，その要素として含んでいた．ただしその記述方法は多様であり(表4)，特徴的な傾向は見い出せなかった．また，「」や『』を使わない場合も，書名文字列は，見出しタグやフォント指定タグにより，視認性を高める工夫がなされていることが多かった．しかし，その方法はサイトによってさまざまであり，とくに傾向は見出せない．以上のことから，書評単位に対し，どの部分が書名をあらわす文字列であるかを機械的に判別するのは容易でないと考えられる．

また，出版元を記したサイトの多くは「○○文庫」「△△ノベルス」のように，出版社そのものの名称でなく，文庫・叢書名を記していた．書誌的事項の抽出にはこれをふまえ，文庫・叢書名と出版社名とを同等に扱うなどの配慮が必要と考えられる．

2.3 検索結果の自動編集

インデックスに含まれるオンライン書評が，ユーザの要求に対する充分な適合性を備えていない場合も想定し，検索結果の編集にあたっては，問合せの結果得られた文書群に対し，次の四段階の処理が必要と考える．

1. 書評単位を切り出す．これについて2.3.1で述べる．
2. 検索結果提示時の視認性の向上のため，書評単位から書誌的事項を抽出する．これについて2.3.2で述べる．
3. 不要と考えられる文書や文書構成要素を破棄するようフィルタリングを行う．これについて2.3.3で述べる．
4. 検索結果のランキングを行う．これについて2.3.4で述べる．

2.3.1 書評単位の切り出し

複数の書評単位を含む文書の場合，当該図書に関する書評単位の位置を推定し抽出する必要がある．

オンラインコンテンツ，とくにWebコンテンツの内容を判定し，機械的な編集を施す処理については，いくつかの研究がなされている[5][6]．これらは，ユーザの要求に備え，前処理として，対象コンテンツを走査し，要素ごとに整理されたテーブルを構築する手法であると思われる．

一方，本研究では，2.2に述べたように，書評中の重要な要素である書名に相当する文字列の出現を判定することが困難であることから，ユーザの要求を契機に，実時間で対象コンテンツを走査することとする．次の手順で書評単位の抽出を試みる．

1. <A NAME="XXX">タグが存在すれば，これをデリミタとして文書を分割し，書名を含む部分を，目的の書評単位であると推定し抽出する．

2. <A NAME="XXX">タグが存在しなければ，書名の出現位置の前後に存在するHTMLタグを前後数個取り出し，次に同じ順序で同じHTMLタグの組合せが出現する位置までを，目的の書評単位であると推定し抽出する．

2.3.2 書誌的事項の抽出

書評単位から著者名，出版社，ISBNの抽出を行う．

著者名の抽出には，目録検索システムから得られる著者名との文字列のマッチングにより行う．

出版社名の抽出には，NACSIS-CAT総合目録データベース上の関連項目より抽出・加工したデータに基づく辞書を用いる．辞書に採録したのは，書誌データベースの出版者フィールドおよび文庫・叢書名フィールドである．

ISBNの抽出は，「0-123456-78-9」および「0123456789」の2種類の表記がなされる場合があることから，ハイフンを中間に含むか含まない連続した10個の数字からなるパターンにより行う[4]．

2.3.3 不要文書の破棄

抽出された書評単位に対し，著者名により絞込みを行う．これは，別著者による同一書名の図書についての書評単位を検索結果から除外するため，また，単に一般語彙として出現した書名と同一の文字列のに起因する影響を軽減するためである．

また，書名が含まれていても，あきらかに書評ではなく，重要度は低いと判断できる文書として，2.2でも触れたサイト内の書評の目次にあたる機能を果たしている文書がある．その多くは，箇条書の形で列挙された多数の書名のみからなっており，ほとんど文章を持たない．こうした文書を検索結果から除外するため，句点または句点に相当する機能を備えた記号 (。!?) を備えた文の数が，ある閾値に満たない文書は破棄することにした．

2.3.4 ランキング

ユーザの要求に対する適合度が比較的低いと考えられる文書も存在する．たとえばWeb作者の日記に相当する文章中に，読書の記録が記されているが，書評ではないようなケースである．こうした文書よりも，より書評らしい内容の文書を優先的に提示することが望ましい．

ランキングを行うときに，情報検索研究の分野においては，頻度や確率の情報を用いることが多い [7]．本研究では書評という限定された対象領域を扱うにあたり，対象文書の書評らしさを尺度としたランキングを試みる．具体的には，国立国語研究所による『分類語彙表』[8, 9] を用い，書評に関連した語彙の出現数で検索結果のソートを行うこととした．採用した語彙は同表の以下のカテゴリに含まれる505の語彙である．

- 1.3134 批評・弁解
- 1.3136 説明
- 1.3150 読み書き・読み
- 1.3160 文献・図書
- 2.3151 読み
- 1.3832 出版・放送・興行
- 2.3832 出版・放送

なお，後二者からは出版に係る部分のみを採用した．

3 実装と評価

3.1 実装

前章の検討に基づき試作システムの実装を行った．

試作システムは書誌・所在情報を取得するための目録検索システムとして，書誌情報についてはNACSIS-CAT，所在情報については北海道大学のOPACを用いた．また，オンライン書評のインデックスとしては，ある書評関連WWWサイトリンク集に収められた266サイト内の11,632のHTML文書の索引を作成し，これを用いた．試作システムへの入力方法は，携帯端末等からの接続を想定した手動入力，ならびに，既存目録システムの検索結果からの自動リンクの二種類を試作した．

図2に試作システムの検索結果出力画面例を示す．上部が書誌情報，下部左欄がオンライン書評の一覧，下部右欄が所在情報等である．

3.2 評価

評価には，オンライン書評のインデックス中に多くの文書が含まれる，知名度の高い，現代ミステリ作品を用いた．具体的なタイトルの選定には，ミステリ作品の秀作を紹介する「このミステリがすごい」(宝島社より毎年刊行) を用い，同書で最近5年間に第1位を獲得した次の五作品を選出した．

[4] 出版社名ならびにISBNについても，著者名の抽出と同様に，目録検索システムから得られる書誌情報を用いて，文字列のマッチングにより抽出を行うことも考えられるが，本研究では上記の方法をとる．これは，同一作品が複数の出版社から刊行されているケースなどで，どちらに対する書評であっても適合文書として許容するためである．

	書名のみによる検索				試作システムによる検索			
	総ヒット数	A	B	C	総ヒット数	A'	B'	C'
ホワイトアウト	87	30	36	21	28	21	6	1
不夜城	30	10	9	11	15	7	2	6
OUT	127	6	15	106	10	4	6	0
レディ・ジョーカー	19	5	13	1	1	1	0	0
永遠の仔	88	22	42	24	29	19	8	2
合計	351	73	115	163	83	52	22	9

表 5: 書名のみによる検索結果と試作システムによる検索結果

図 2: 試作システムの検索結果出力

- ホワイトアウト/真保裕一 (1996)
- 不夜城/馳星周 (1997)
- OUT/桐野夏生 (1998)
- レディ・ジョーカー/高村薫 (1999)
- 永遠の仔/天童荒太 (2000)

以下，文書を以下の 3 カテゴリに分けて記す．

A: 当該図書に関するオンライン書評

B: 当該図書に言及してはいるがオンライン書評ではない文書

C: 当該図書に言及してない文書

ここで，各文書のカテゴリ分けは，

- 当該図書に関する記述が存在するか．
- 当該図書に関する感想・書評を記した 1 行以上の日本語文章が存在するか．

を基準に著者の視認により行った．

ブーリアン型検索システムを用いて書名のみにより検索した場合[5]のヒット数と試作システムによるヒット数，および，それぞれにおける A，B，C の数 (内数) を表 5 に示す．

試作システムによる検索の場合において，以下の評価指標を定義する．

$$\text{擬似再現率} = A'/A \quad (1)$$

$$\text{精度} = A'/(A'+B'+C') \quad (2)$$

同様に，比較のため，書名のみによる検索の場合において，次の評価指標を定義する．

$$\text{擬似再現率} = A/A \quad (3)$$

$$\text{精度} = A/(A+B+C) \quad (4)$$

書名のみによる検索ならびに試作システムにおける擬似再現率，精度を表 6 に示す．

	擬似再現率	精度
書名のみによる検索	100%	20.8%
試作システム	71.2%	62.7%

表 6: 擬似再現率，精度の比較

試作システムにおける擬似再現率は 71.2%とであった．この主な要因は，著者名の未記載，誤記などであった．

試作システムにおける精度は 62.7%となった．この要因としては，B ではサイト内目次ページに日本語文章が含まれるケース，また，C では，評価対象の図書のうち

[5]実際には，情報検索システム Namazu[10] を利用した．なお，試作システムへの入力である書名を示す文字列に対して形態素解析を行なうことで名詞および未定義語を抽出し，それらの And 条件での検索を行った．形態素解析には茶筅 [11] を用いた．

いくらかは映画やテレビドラマなど，別メディアでも公開されていることから，映画評やテレビドラマ評などがヒットしているケースが多く見られた．また，書名のみによる検索における精度と比較すると，とくに『OUT』のような，書名が一般的な語彙であるケースでは，とくに有効であると見ることができる．

試作システムの検索結果において，総ヒット数をランキングの上位と下位に等分割して，正解文書の出現位置を調べた結果を次に示す．

検索結果上位に占める正解文書の割合: 68.3%

検索結果下位に占める正解文書の割合: 54.8%

書評関連語彙の有無に基づくランキングによる若干の効果が見られた．

4 おわりに

以上，目録検索システムを援用したオンライン書評検索について検討を行った．

提案手法による試作システムでは書名のみによる検索に比し，約3倍の精度で目的の文書を抽出することができた．しかしながら，試作システムでは3割弱の抽出洩れが発生した．今後，提案手法の詳細な評価，および，本稿での提案以外のランキング手法等に関する検討が必要と考える．

検索処理中に利用した目録検索システムであるが，関心に合致した図書をいかに発見するかについて，従来の検索方式に留まらない高度な検索機能の研究 ([12, 13]) がなされてきているところである．また，その検索結果についても所在情報の提示だけでなく，近年では，当該図書の目次情報や原文，つまり一次情報そのものの提供の試みも現れてきている．

本研究の応用として，上のように高機能化しつつある目録検索システムに連動させることにより，その検索結果に，関連する書評情報を組み合わせることなどが考えられる．

謝辞 本研究は，平成12年度国立情報学研究所セミナーにおける研究成果である．

参考文献

[1] 国立情報学研究所. 目録所在情報サービス NACSIS-CAT. ⟨http://www.nii.ac.jp/CAT-ILL/welcome.html⟩.

[2] 北海道大学附属図書館. Online Catalog. ⟨http://www.lib.hokudai.ac.jp/opac/⟩.

[3] 山本英子, 武田善行, 梅村恭司. 情報検索性能と表記の揺れへの寛容性を持つ類似度. 情報処理学会研究報告, No. 2000-FI-71, pp. 9–15, 2000.

[4] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks: The International Journal of Computer & Telecommunications Networking*, Vol. 31, No. 11-16, pp. 1623–1640, 1999.

[5] 大槻洋輔, 佐藤理史. ワールドワイドウェブを知識源とした地域情報の自動編集. 情報処理学会研究報告, No. 2000-ICS-119, pp. 165–172, 2000.

[6] 山本あゆみ, 佐藤理史. ワールドワイドウェブからの人物情報の自動収集. 情報処理学会研究報告, No. 2000-ICS-119, pp. 173–180, 2000.

[7] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[8] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[9] 国立国語研究所. 分類語彙表 [フロッピー版](国立国語研究所言語処理データ集5). 秀英出版, 1996.

[10] 全文検索エンジン Namazu. ⟨http://www.namazu.org/⟩.

[11] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』version 1.0 使用説明書, 1997. ⟨http://chasen.aist-nara.ac.jp/⟩.

[12] 木伸征, 黒橋禎夫. 自然言語入力と目次との柔軟な照合による図書検索システム. 情報処理学会論文誌, Vol. 41, No. 4, pp. 1162–1170, 2000.

[13] 茂出木理子, 杉田いづみ, 前田朗. 東京大学ブックコンテンツ・データベースサービスの紹介. 薬学図書館, Vol. 44, No. 4, pp. 342–344, 1999. ⟨http://contents.lib.u-tokyo.ac.jp/contents/top.html⟩.