目録データベースを用いた WWW における図書情報自動編集システム

3V−3

江口 浩二† 杉田 茂樹‡
† 国立情報学研究所 ‡ 北海道大学

1 はじめに

WWW 上に分散する図書情報，特に書評コンテンツを対象として，目録データベースの援用による自動編集方式を提案する．提案する方式は，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 WWW 上で提供される書評を検索し，検索結果に対して自動編集を施した上で利用者に対して適切に提示するものである．試作システムを構築し，それを用いた基礎的な実験を実施した結果，概ね良好な有効性を確認した．

2 提案システムの概要

提案するシステムでは，WWW 検索エンジンを利用するか，或は，予めオンライン書評のインデックスを作成しておくことを前提とし，ユーザにより特定された図書に関する固有情報（ISBN など）を入力とする（図 1）．

(1) 図書の固有情報の入力をもとに目録検索システムより当該図書の書誌情報を取得する．

(2) 取得した書名・著者名等の書誌情報でクエリを構成し，予め作成された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，多様な観点に基づいて自動編集を施す．

(4) オンライン書評の編集結果に加えて，目録データベースの問合せ結果である書誌，所在情報をも併せてユーザに提示する．

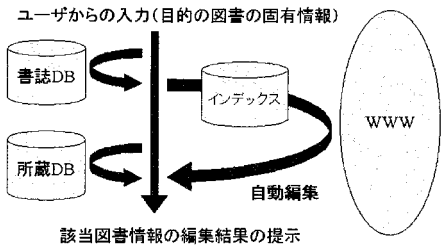

図 1: 提案システムの概要

上記 (3) の検索結果の自動編集について説明する．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，以下に述べる四段階の処理が必要と考える．

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．次の要領で書評単位の切り出しを試みる．

- ⟨A NAME="X"⟩ 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．
- ⟨A NAME="X"⟩ 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 タグを数個取り出し，次に対応する順序で同種の 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．

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．まず，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 [1] 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」の二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み得る連続した 10 個の数字からなるパターンにより行う．

フィルタリング　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，明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，サイト内の書評の目次にあたる機能を果たす文書がある．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，一定量の文を含まない書評単位は破棄する．本論文では，文の数が，ある閾値に満たない文書を破棄することとした．

ソーティング　情報検索分野における適合度順のランキングとは異なり，本論文では，より書評らしい内容の文書を優先的に提示することを目指す．書評を含まない読書記録等よりも，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と考えた．本論文では，対象文書の「書評らしさ」を尺度としたランキング（以下，ソーティング）を実現するため，国立国語研究所による『分類語彙表』[2] を用いた．書評に関連した語彙の出現数に基づいて書評単位のソーティングを行うこととした．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

Automatic Editing System of Book Information on the WWW using Library Catalog Databases

Koji EGUCHI† and Shigeki SUGITA‡, † National Institute of Informatics, ‡ Hokkaido University

表 1: 従来型システムの検索結果と提案システムによる検索結果

	従来型システム						提案システム						
書名	総ヒット件数	A	B	C	D	精度	総ヒット件数	A'	B'	C'	D'	精度	疑似再現率
『ホワイトアウト』	87	30	36	21	0	0.345	28	21	6	1	0	0.750	0.700
『不夜城』	30	10	9	11	0	0.333	15	7	2	6	0	0.467	0.700
『OUT』	127	6	15	106	0	0.047	10	4	6	0	0	0.400	0.667
『レディ・ジョーカー』	19	5	13	1	0	0.263	1	1	0	0	0	1.000	0.200
『永遠の仔』	88	22	42	24	0	0.250	29	19	8	2	0	0.655	0.864
合計件数	351	73	115	163	0	–	83	52	22	9	0	–	–
平均値	–	–	–	–	–	0.248	–	–	–	–	–	0.654	0.626

／ 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

実装 前章に述べた提案システムの実装を行った．提案システムにおいて，書誌・所在情報を取得するために目録検索システムを用いるが，書誌情報については NACSIS-CAT[1]，所在情報については北海道大学の OPAC[3] 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 文書を対象としてインデックスを作成し，これを用いた．なお，提案システムにおける検索処理には，書名を手がかりとしたクエリ語を構成した上で，対象文書におけるクエリ語の出現の有無とその論理演算に基づくブーリアン型検索システムを用いて行なった[1]．

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ていた，比較的知名度の高い，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，同書で最近五年間に第一位を獲得した次の五作品を選出した．『ホワイトアウト』，真保裕一 著 (1996) ／『不夜城』，馳星周 著 (1997) ／『OUT』，桐野夏生 著 (1998) ／『レディ・ジョーカー』，高村薫 著 (1999) ／『永遠の仔』，天童荒太 著 (2000)．

以下，文書を次の三つのカテゴリに分けて記す．A: 当該図書に関連するオンライン書評／ B: 当該図書に言及して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おらずオンライン書評でもない文書／ D: 当該図書に関連しないオンライン書評．ここで，各文書のカテゴリ分けは，(i) 当該図書に関する記述が存在するかどうか，並びに，(ii) 当該図書に関する感想・書評を記した一行以上の日本語文章が存在するかどうか，を基準に著者の視認により行った．

提案システムを用いた場合と，通常の検索機能のみを用いた従来型システムの場合において，総ヒット件数および前述のカテゴリ A，B，C，D それぞれの数（内数）を表 1に示す．また，評価指標としては，提案システムを用いた場合において次に定義される**擬似再現率**および**精度**を用いた．精度 $= A'/(A'+B'+C'+D')$，擬似再現率 $= A'/A$．同様に，比較のため，従来型システムの検索結果に対して次の評価指標を用いた．精度 $= A/(A+B+C+D)$，擬似再現率 $= A/A$．提案システム，従来型システムそれぞれの検索結果に対する精度と，提案システムに対する擬似再現率を表 1に示す．

表 1に示される通り，従来型システムにおける精度の平均値が 24.8%であるのに対して，提案システムにおいては 65.4%であった．最終的な出力結果に非正解文書（非書評文書あるいは非関連文書）が混入する要因としては，B ではサイト内目次ページに一定量以上の文章が含まれるケースがあったこと，また，C では，評価対象の図書のうちいくらかは映画やテレビドラマなどの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ことが挙げられる．また，従来型システムにおける精度と比較すると，『OUT』のような，書名が一般的な語彙であるケースでは，特に提案手法が有効であると見ることができる．

また，提案システムにおける擬似再現率の平均値は，従来型システムを基準としたとき，62.6%であった．正解文書（書評文書）の洩れの主な要因は，著者名の未記載，誤記などであることを確認した．

次に，提案システムの検索結果において，総ヒット件数をソーティング結果の上位部と下位部に等分割して，それぞれにおける正解文書の割合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%．このように書評関連語彙の有無に基づく「書評らしさ」のソーティングによって若干の効果が見られた．

謝辞 本研究の一部は，平成 12 年度国立情報学研究所セミナー，および，日本学術振興会科学研究費補助金（奨励 12780322）による．

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/⟩.

[2]国立国語研究所: 分類語彙表, 秀英出版 (1964).

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

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

[5]Matsumoto, Y., Kitauchi, A., Yamashita, T. and Hirano, Y.: Japanese Morphological Analysis System ChaSen version 2.0 Manual, Technical Report NAIST-IS-TR99009, NAIST (1999).

[1]情報検索システム Namazu[4] を用いた．なお，提案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 [5] を用いた．