集中型横断検索システムのための自動書誌同定処理の検討

上田　洋[*1] (d06tb001@ex.media.osaka-cu.ac.jp)
村上　晴美[*2]　辰巳　昭治[*1]
[*1] 大阪市立大学 大学院工学研究科 電子情報系専攻
[*2] 大阪市立大学 大学院創造都市研究科 都市情報学専攻

概要: 本研究では、集中型横断検索システムのための固有識別情報を用いない自動書誌同定処理を検討する。蔵書タイトル、責任表示などの表記に曖昧性を持つ情報を主に用いて特徴ベクトルの作成を行い、機械学習の一種であるサポートベクターマシン（SVM）によって書誌同定判定を行う。性能評価を行った結果、固有識別情報を用いず、曖昧性を持つ情報のみを用いる手法としては結果が良かった。

1. はじめに

近年、インターネットの普及に伴い、Web 上で情報検索を行うことが一般的となり、図書館でも Web 上で検索が可能な OPAC（以下、Web OPAC）が設置されるようになった。現在、Web OPAC を設置する図書館は、2009 年 3 月現在、公共図書館に 1085 館 [1]、大学図書館に 611 館 [2]ある。Web OPAC を設置する図書館が増えるにつれて、1 図書館だけではなく、複数の図書館の蔵書を横断的に検索することへのニーズが増加している。そのため、複数の図書館の蔵書を横断的に検索できる蔵書検索システム（以下、横断検索システム）が増加している。

現在公開されている横断検索システムは、大きく分けて 2 種類存在する。1 つは、検索要求毎に複数の Web OPAC の検索を同時に行い、取得した検索結果を集約するシステム（分散型横断検索システム）、もう 1 つは、あらかじめ複数の図書館から書誌情報を集め、1 つのデータベースシステムに格納し、検索を提供するシステム（集中型横断検索システム[1]）である。これら横断検索システムの問題の一つに「書誌割れ」がある。書誌割れとは、同じ書誌情報が重複して検索される現象を指す。分散型横断検索システムは、リアルタイム処理により書誌同定を行わなければならないため、書誌同定を行わないシステムが多い。一方、集中型横断検索システムでは、書誌同定のリアルタイム処理が必要とされないため、書誌同定処理を行うシステムが多い。集中型横断検索システムにおける書誌割れの解決方法としては、人手により書誌同定を行う方法と、自動的に書誌同定を行う方法がある。前者は、国立情報学研究所の NACSIS-CAT[2]が採用し、後者は、国立国会図書館の総合目録ネットワークシステム（以下、「ゆにかねっと」）[3]が採用している。

NACSIS-CAT では、以下のように書誌同定を行う。図書館員が書誌情報を登録する際に、登録する書誌情報を検索し、既に登録されていれば登録済み書誌情報に所蔵情報のみを登録する。人手による登録であれば、高い精度で重複した書誌情報の処理が行えるが、人的コストは非常に大きい。

「ゆにかねっと」では、ISBN や MARC 番号など蔵書に固有に割り当てられた情報（以下、固有識別情報）を主に用いた自動書誌同定処理を行っている。自動処理のため人的コストは少なく済むが、ISBN や MARC 番号などは全ての書誌情報に付与されておらず、記載ミスがある可能性も考えられる。これらの要因により、人手で処理を行う場合よりも、書誌割れが多く発生する。書誌割れの発生は、蔵書検索時に利用者に無駄な情報を閲覧させることになり、利用者の負担となる。また、データベースの容量増加や検索速度の低下など、検索システム内で悪影響が出る可能性がある。

本研究では、集中型横断検索システムのための自動書誌同定処理を検討する。既存システムは固有識別情報を用いて自動書誌同定処理を行っているが有効性が限定的であるため、本研究では固有識別情報以外の情報を用いる自動書誌同定処理を検討する。

2. 自動書誌同定処理

本研究では、集中型横断検索システムにおける自動書誌同定処理を検討する。本研究の特徴は、固有識別情報を用いず、表記に曖昧性を持つ情報を用いて書誌同定を行う点にある。自動書誌同定処理には、機械学習の一種であるサポートベクターマシン（以下、SVM）を用いる。

2.1. 処理概要

自動書誌同定処理の概要を図 1 に示す。

---

[1] 一般的には、「総合目録システム」や「総合目録データベース」と呼ばれるが、本研究では「分散型」との対比として「集中型横断検索システム」と呼ぶ。

[2] http://www.nii.ac.jp/CAT-ILL/

[3] http://unicanet.ndl.go.jp/

まず、事前処理として SVM に同じ書誌情報、異なる書誌情報の特徴ベクトルを学習させる。学習により得た教師データを元に、未判定の 2 つの書誌情報を SVM により判定を行う。

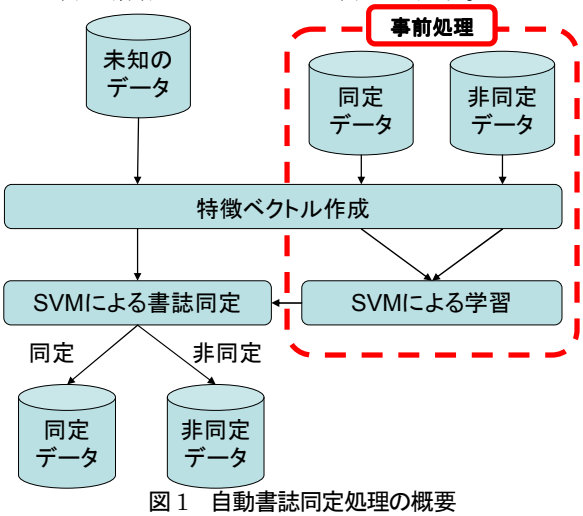

図 1　自動書誌同定処理の概要

2.2.　教師データの作成

2.2.1.　書誌情報の取得

未判定の書誌情報が同じものかを判定するために使用する教師データを事前に作成する。

教師データ作成に用いる書誌情報として、「ゆにかねっと」に蓄積された書誌情報を使用した。「ゆにかねっと」とは、国内の公共図書館における図書館資料資源の共有化、書誌サービスの標準化と効率的利用を図ると共に、公共図書館の県域を超える全国的な図書館相互貸借等を支援することを主な目的[3]として設置された蔵書検索システムである。「ゆにかねっと」には、平成 21 年 3 月末現在、64 の参加図書館から収集した和図書の書誌データ約 3872 万件が蓄積されている[4]。

本研究では、Web 上で公開されている一般利用者用検索画面[4]から、検索語を入力し書誌情報を取得した。使用した検索語は、Wikiepdia の「ベストセラー」[5]より、「シリーズ作品」、「漫画」の項目に記述された 10 タイトルである。使用検索語と取得した書誌情報を表 1 に示す。10 の検索語から、1777 の書誌情報を得た。この 1777 の書誌情報から教師データを作成した。

まず、ISBN を用いた書誌同定を行った。同じ検索語で得られた書誌情報毎に、ISBN が同じかどうかを判定した。ISBN が同じであれば、同じ書誌情報とし、ISBN が異なる場合は、異なる書誌情報とした。書誌同定の結果、同じ書誌情報のペアが 1028、異なる書誌情報のペアが 24028、計 25056 の書誌情報のペアが得られた。この 25056 の書誌情報のペアを用いて、SVM により学習させて教師データを作成した。

2.2.2.　正規化

取得した書誌情報は、様々な公共図書館で作成されたものであり、表記に揺れが存在する。表記の揺れがあるまま特徴ベクトルの作成、学習を行うと、同義の特徴であるにもかかわらず異なる特徴として認識される可能性が高い。できるだけ同一表記になるように文字列を正規化する。

まず、文字の半角、全角、大文字、小文字を統一する。次に、蔵書タイトルに含まれる巻号の情報を正規化する。「ゆにかねっと」から取得した書誌情報を調査したところ、例えば、

- ハリー・ポッターと謎のプリンス . **上巻**
- ハリー・ポッターと謎のプリンス . **上**

- 「ハリー・ポッター」新サイドブック . **ｖ．１**
- 「ハリー・ポッター」新サイドブック . **ＶＯＬ．１** パロディ・ノベル&ファンクイズ

- ドラゴンボール . **０１**
- ドラゴンボール . **１**

のように、特に巻号の記述に表記の揺れがあるものが多いことがわかった。また、「「ハリー・ポッター」新サイドブック . ＶＯＬ．１ パロディ・ノベル&ファンクイズ」のように、巻号の後に、「パロディ・ノベル&ファンクイズ」のようなサブタイトルが付記されているものがあった。これらは、後述の特徴ベクトル作成や SVM による学習に悪影響を与えるため、蔵書タイトル、巻号、サブタイトルを切り出し、それぞれ統一した表記に正規化する。

表 1　使用検索語と取得した書誌情報の件数

シリーズ作品	漫画
「アンパンマン」(200件)	「SLAM DUNK」(71件)
「スレイヤーズ」(133件)	「ドラゴンボール」(200件)
「ハリーポッター」(196件)	「美味しんぼ」(200件)
「鬼平犯科帳」(200件)	「名探偵コナン」(200件)
「三毛猫ホームズ」(174件)	
「竜馬がゆく」(200件)	

表 2　使用レコード

「タイトル」	「出版者」
「タイトルよみ」	「出版年」
「責任表示」	「形態」
「出版地」	「定価」

2.2.3.　特徴ベクトルの作成

正規化された書誌情報から SVM に用いる特徴ベクトルを作成する。本研究では、主に各書誌情報から得られるレコード同士の一致パターンを用いる。書誌情報からの特徴ベクトルの作成には、表 2 の 8 レコードを用いる。8 レコードは、取得されたほとんどの書誌情報に情報が記述されていた。

---

4 「ゆにかねっと」には、一般利用者用と参加図書館用の 2 種類の検索画面が存在する。一般利用者用から得られる書誌情報の上限は 200 である。

5 http://ja.wikipedia.org/wiki/%E3%83%99%E3%82%B9%E3%83%88%E3%82%BB%E3%83%A9%E3%83%BC

これらレコードのうち、「出版年」には「２００２．１２」のように出版年と月、「形態」には「２１４ｐ；２１ｃｍ」のようにページ数と資料の大きさが併せて記述されている。そのため、「出版年」については「２００２」と「１２」のように出版年と月を、「形態」については、「２１４ｐ」と「２１ｃｍ」のようにページ数と資料の大きさを分離する処理を行う。

表３　SVM に用いる特徴

1. 書籍1と2両方にタイトルがあるかどうか
2. 書籍1のタイトルの文字数
3. 書籍2のタイトルの文字数
4. 書籍1、2のタイトルの一致率
5. 書籍1と2両方に巻号があるかどうか※1
6. 書籍1の巻号の文字数※1
7. 書籍2の巻号の文字数※1
8. 書籍1、2の巻号の一致するか※1
9. 書籍1と2両方にタイトルよみがあるかどうか
10. 書籍1のタイトルよみの文字数
11. 書籍2のタイトルよみの文字数
12. 書籍1、2のタイトルよみの一致率
13. 書籍1と2両方に巻号があるかどうか※2
14. 書籍1の巻号の文字数※2
15. 書籍2の巻号の文字数※2
16. 書籍1、2の巻号の一致するか※2
17. 書籍1と2両方に出版年があるかどうか
18. 書籍1に出版年があるかどうか
19. 書籍2に出版年があるかどうか
20. 書籍1、2の出版年の差※3
21. 書籍1と2両方に出版月があるかどうか
22. 書籍1に出版月があるかどうか
23. 書籍2に出版月があるかどうか
24. 書籍1、2の出版月の一致するか
25. 書籍1と2両方に出版者があるかどうか
26. 書籍1の出版者の文字数
27. 書籍2の出版者の文字数
28. 書籍1、2の出版者の一致率
29. 書籍1と2両方に出版地があるかどうか
30. 書籍1の出版地の文字数
31. 書籍2の出版地の文字数
32. 書籍1、2の出版地の一致率
33. 書籍1と2両方にページ情報があるかどうか
34. 書籍1にページ情報があるかどうか
35. 書籍2にページ情報があるかどうか
36. 書籍1、2のページ情報の一致するか
37. 書籍1と2両方に資料のサイズ情報があるかどうか
38. 書籍1に資料のサイズ情報があるかどうか
39. 書籍2に資料のサイズ情報があるかどうか
40. 書籍1、2の資料のサイズ情報の一致するか
41. 書籍1と2両方に価格があるかどうか
42. 書籍1に価格があるかどうか
43. 書籍2に価格があるかどうか
44. 書籍1、2の価格の差※4
45. 書籍1の著者の数
46. 書籍2の著者の数
47. 書籍1、2の一致する著者の数

※1「タイトル」レコードから切り出された巻号
※2「タイトルよみ」レコードから切り出された巻号
※3 例えば 書籍1が2008、書籍2が1998だった場合、「10」(=2008−1998)となる
※4 例えば 書籍1が2000円、書籍2が1000円だった場合、「1000」(=2000−1000)となる

これまでの処理により得た書誌情報を用いて、レコード毎に特徴を取得する。取得する特徴を表 3 に示す。これらの特徴を、25056 の書誌情報のペアについて作成し、SVM により学習させる。なお、表 3 の 4、12、28、29 の一致率は、文字列の一致する割合である。文字列の一致を割り出す方法として、形態素解析により得られた形態素の一致数を計算する方法と、1 文字ずつ決められた文字数毎に切り出し（N-Gram）、切り出した文字列の一致数を計算する方法がある。本研究では、後者の N-Gram を用いた方法を採用した。なお、本研究では、N=2（2 文字毎に切り出す）とした。

表４　性能評価に使用した検索語と取得した書誌情報の件数

「ONE PIECE」(200件)	「日本沈没」(73件)
「ロードス島戦記」(84件)	「いないいないばあ」(189件)
「こちら葛飾区亀有公園前派出所」(200件)	「青春の門」(200件)
「ゴルゴ13」(200件)	「ドラえもん」(200件)
「銀河英雄伝説」(200件)	「ぐりとぐら」(107件)

3. 性能評価

本研究の自動書誌同定処理の性能評価を行った。

3.1. 方法

「ゆにかねっと」の一般利用者用検索画面から書誌情報を取得した。書誌情報の取得に用いた検索語は、Wikipedia「ベストセラー」から選択した 10 のタイトルである。使用検索語と取得件数を表 4 に示す。取得した 1653 の書誌情報を性能評価に用いた。

得られた書誌情報について、2.2 節の処理に基づき検索語毎に書誌情報のペアを作り、正規化して特徴ベクトルを作成した。作成した特徴ベクトルを 2.2 節にて作成した教師データを用いて SVM により書誌同定判定を行う。

評価尺度については、精度、適合率、再現率と、適合率と再現率の総合的な尺度である F 値を用いた。各尺度の定義は以下の通りである。

$$\text{精度} = \frac{\text{SVMにより正しく判定された書誌情報のペアの数}}{\text{入力した書誌情報のペアの数}}$$

$$\text{適合率} = \frac{\text{SVMにより同定と判定された書誌情報の中で実際に同じ書誌情報のペアの数}}{\text{SVMにより同定と判定された書誌情報のペアの数}}$$

$$\text{再現率} = \frac{\text{SVMにより同定と判定された書誌情報のペアの数}}{\text{性能評価に用いた全ての書誌情報のうち同じ書誌情報のペアの数}}$$

$$\text{F値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$$

3.2. 結果

精度は 98.2%、適合率は 81.7%、再現率は 54.1%、F 値は 56.9%であった。検索語別の集計では、7 つの検索語（タイトル）で適合率 90%以上であった。再現率も 7 つの検索語で 50%以上であった（表 5 参照）。

適合率については、多数の検索語で 90%以上の評価が得られたこと、再現率も概ね 50%を超える評価が得られたことなどから、ISBN、MARC 番号などの固有識別情報を用いず、曖昧性を持つ情報のみを用いる手法としては結果が良かった。

集計の結果、検索語により評価に大きな差が生まれていることがわかった。例えば、「こちら葛飾区亀有公園前派出所」では、再現率は100.0%であったが、適合率は 7.9%と 1 割にも満たなかった。また、「ぐりとぐら」では、適合率は100.0%であるが、再現率は16.7%と他の検索語と比べるとかなり低かった。今後、これらの検索語別の書誌同定結果の分析を行い、評価結果の低下の原因を探りたい。

4. 議論

本研究では、固有識別情報を用いない自動書誌同定処理を検討した。蔵書タイトルや責任表示など、表記に曖昧性がある情報を用いて、書誌同定自動処理を行った。性能評価の結果、適合率 81.7%、再現率 54.1%と、曖昧性を持つ情報を用いた手法としては結果が良かった。本研究では、ISBN や MARC 番号など、一致すれば高い確率で同じ書誌情報と見分けられる情報（固有識別情報）をあえて用いなかった。固有識別情報を用いる処理と本研究の自動書誌同定処理を併せて使用することにより、固有識別情報のみを用いる場合より、書誌割れの発生を抑えられると考える。

本研究で検討した自動処理は、処理の高速化によりリアルタイム処理が可能となれば、分散型横断検索システムでも利用可能であると考える。

5. 関連研究

相澤らの調査[5]では、本研究のような異なるデータベースから取得したレコードの自動同定の研究の多くで、Fellegi-Sunter [6]モデルを参照しているとされる。今後、このモデルを用いた手法との比較実験を行いたい。

6. おわりに

本研究では、集中型横断検索システムのための固有識別情報を用いない自動書誌同定処理を検討した。蔵書タイトル、責任表示などの表記に曖昧性を持つ情報を主に用いて特徴ベクトルの作成を行い、機械学習の一種であるサポートベクターマシン（SVM）によって書誌同定判定を行った。性能評価の結果、適合率が 81.7%、再現率 54.1%と、固有識別情報を用いず、曖昧性を持つ情報のみを用いる手法としては結果が良かった。

今後は、(1) 書誌同定性能の向上、 (2) 処理の高速化、(3) 性能評価の結果から有益に機能する特徴の調査、など行う予定である。

謝辞

本研究を行うにあたり、国立国会図書館関西館図書館協力課の皆様には多大なご協力を頂きました。深く感謝いたします。また、本研究について有益な助言を頂いた奈良学園登美ヶ丘ライブラリーの村上幸二氏には感謝いたします。

参考文献

[1] 公共図書館 Web サイトのサービス, http://www.jla.or.jp/link/public2.html (2009.04.20 参照)

[2] 大学図書館 OPAC の動向, http://www.slis.keio.ac.jp/~ueda/libwww/libwwwstat.html (2009.04.20 参照)

[3] 国立国会図書館総合目録ネットワーク事業実施要綱, http://www.ndl.go.jp/jp/library/pdf/yoko.pdf (2009.05.20 参照)

[4] 総合目録ネットワーク｜国立国会図書館-National Diet Library,http://www.ndl.go.jp/jp/library/backlist_network.html (2009.05.20 参照)

[5] 相澤彰子，高須淳宏，大山敬三， 安達淳：異種データベース間でのレコード照合に関する研究動向，NII Journal No.8, pp.43–51, 2004.

[6] Ivan P. Fellegi and Alan B. Sunter: A Theory for Record Linkage, Journal of American Statistical Association, Vol. 64, No. 328, pp. 1183-1210 , 1969.

表 5 性能評価の結果

	ONE PIECE	ロードス島戦記	こちら葛飾区 亀有公園前派出所	ゴルゴ13
精度	99.6% (11655/11700)	98.3% (3709/3772)	92.6% (9013/9735)	99.8% (958/960)
適合率	97.9% (46/47)	71.6% (48/67)	7.9% (62/784)	100.0% (4/4)
再現率	51.1% (46/90)	52.2% (48/92)	100.0% (62/62)	66.7% (4/6)
F値	67.2%	60.4%	14.7%	80.0%

	銀河英雄伝説	日本沈没	ドラえもん	青春の門
精度	99.1% (15803/15939)	99.4% (346/348)	99.7% (9700/9728)	98.5% (1014/1029)
適合率	91.5% (86/94)	100.0% (2/2)	98.4% (61/62)	100.0% (15/15)
再現率	40.2% (86/214)	50.0% (2/4)	69.3% (61/88)	50.0% (15/30)
F値	55.8%	66.7%	81.3%	66.7%

	いないいないばあ	ぐりとぐら	平均
精度	98.4% (5884/5980)	96.1% (1248/1298)	98.2%
適合率	50.0% (43/86)	100.0% (10/10)	81.7%
再現率	44.8% (43/96)	16.7% (10/60)	54.1%
F値	47.3%	28.6%	56.9%