インターネットオークション古文書取引情報抽出への機械学習手法の適用

寺澤正直
筑波大学大学院図書館情報メディア研究科

文化行政機関では史料の散逸防止の目的から市場の史料取引状況の把握を行うことがある．近年，インターネットオークションという新たな市場内で史料取引が確認される．しかし日々行われるオークション取引から古文書（記録資料）の取引のみ抽出することは困難である．本研究の目的は機械学習手法を適用して取引情報の自動抽出を試みることである．はじめに一定期間における古文書取引情報を自動分類のテストコレクションとし，決定木による分類ルールを作成する．次に異なる期間の古文書取引情報への分類ルールの適用実験を行う．古文書取引情報収集，分類ルールの作成，分類ルールの適応実験結果をもとに問題点と可能性について検討する．

Application of the machine learning technique of extracting the dealings information on historical records from an Internet auction

Masanao TERASAWA
Doctor Program, Gradeate School of Library, Information and Media Sutudies,

A cultural administration grasps the historical-records trading conditions of a market from the purpose of loss prevention of them. In recent years, their dealings are watched in Internet auction. However, it is difficult to extract only them from a daily auction. The purpose of this paper is to try automatic extraction of dealings information with the application of the machine learning technique. First, a classification rule is created from the test collections which are them of a fixed period. Next, the classification rule applies to them of a different period. As mentioned above, the problem of this trial is considered.

1．はじめに

図書館，文書館，博物館などの文化行政は，史料の散逸防止を目的として史料収集を行うことがある[1]．文化行政による史料収集方法は寄贈，寄託，購入のいずれかである．寄贈，寄託の方法による史料収集は，個人や団体を収集対象とするが，購入による史料収集は，個人，団体以外に，古書店，故紙業者などの収集対象が存在する．図 1 は国文学研究史料館の収集史料数と収集史料群数に分け，史料収集方法別に表記したものである．同施設は戦後より日本国内全般を対象として史料収集を行ってきた文部科学省直轄の機関である．

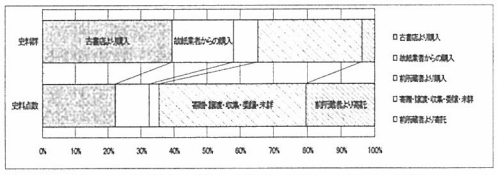

図 1　収集史料の群と点数の割合比較

図 1 より，購入による収集史料群数約 60%弱，史料点数約 30%強あり，購入は史料収集において無視できる手段ではない．

史料保存の観点からは，購入は一種の非常手段であり，史料取引を成立させない環境作りや，市民一人一人の保存意識を高めることが求められていた．しかし，表 1 は都道府県レベルで再度行われた史料所在調査による一定年数あたりの史料散逸の割合を示している．表 1 より史料は明らかに消失，もしくは散逸している事実がわかる．少なくとも現在まで行われてきた文化行政の史料保存の対応では不十分である．

時期	調査主体と調査範囲	結果(散逸割合/年)
1994	大分県立先哲史料館	23% / 20 年[2],[3]
2002	埼玉県立文書館	13% / 約 30 年[2],[3]
2007	和歌山県立文書館	17% / 約 30 年[4],[5]

第 1 表　散逸状況確認目的の史料所在調査

近年インターネットオークション（以下，ネットオークション）上で史料の取引が見られるようになった．しかし，その正確な取引件数について明らかではない．今後，文化行政が史料取引によって現地保存を継続できない史料の対策を検討するためにも，ネットオークションの史料取引件数の時系列的変化や出品者の傾向をより正確に把握する必要がある．

しかし，文化行政の職員が日常業務の中で，膨大な取引の中から史料取引を判別することは困難である．さらに，古文書には「こもんじょ」「こぶんしょ」の 2 種類の意味が内在することや，地域史料，民間所在史料など同義であるが異なる名称が複数存在することが問題をより複雑にさせている．

そこで，本研究では，ネットオークションの全取引から古文書を扱う取引を自動的に判別するための分類ルールを検討することを目的とした．

2．ネットオークションの検索システム

本研究の調査対象は，Yahoo!オークション[6]とした．それは，Yahoo!オークション，ビダーズ，楽天オークション，モバオクを対象に検索キー「古文書」で検索したところ，Yahoo!オークション以外に史料を扱う取引情報がヒットしなかったからである．

Yahoo!オークションの常時出品数は約 1400 万である．出品カテゴリ数はトップの直下に指定されたメインカテゴリで 24，サブカテゴリで約 400 であり，小項目においてさらに詳細な分類わけがなされる [7]．

Yahoo!オークション上の取引情報の検索には，次の 3 つの方法がある．（ｉ）出品タイトル，（ii）出品カテゴリ，（iii）Yahoo! JAPAN ID，を検索対象とした方法である．（ｉ）出品タイトル検索は，Yahoo!ではキーワード検索と呼ばれる．（ii）出品カテゴリは，Yahoo!が指定した出品カテゴリで，出品物はそのいずれかに含まれる．（iii）Yahoo! JAPAN ID は出品者に必ず付与される ID で，出品物を出品者の Yahoo! JAPAN ID で検索できる．また上記 3 つの方法の組み合わせによって，取引を探すこともできる．

出品カテゴリに「古文書」を主題とする項目はない（2008 年 11 月現在）．したがって，キーワード検索で「古文書」という語が含まれている取引を検索するしかない．しかし，出品タイトルは出品者が自由に付与できるため，実際の古文書取引であっても，必ずしも出品タイトルに「古文書」という語が付与されるとは限らない．一方，古文書取引ではないのに「古文書」と付与した取引も存在する可能性がある．以上のことから，どのような語が適切であるのか検討する必要がある．

そこで全取引から古文書取引を自動的に判別するために，テキスト自動分類システムの手順を参考にして，自動的に意図する取引を集めることはできないかと考えた．

3．テキスト自動分類システム

先行研究より，テキスト自動分類システムの対象とするテキストは，図書のタイトル[8]やウェブサイト[9]，など，様々である．

自動分類システムは，分類済みのテキスト集合を用いて分類ルールの作成を行う学習フェーズと，この作成した分類ルールを用いて分類対象テキストの分類先を決定する分類フェーズがある．学習フェーズと分類フェーズの関係，については図 2 に図示した[8]．

本研究では学習フェーズにあたる分類ルールの作成までを行う．具体的には一定期間内に行われた取引の中で，古文書を取り扱った取引かどうかの判断を終えた取引情報の集合より，特徴素を抽出し分類ルールを作成する．

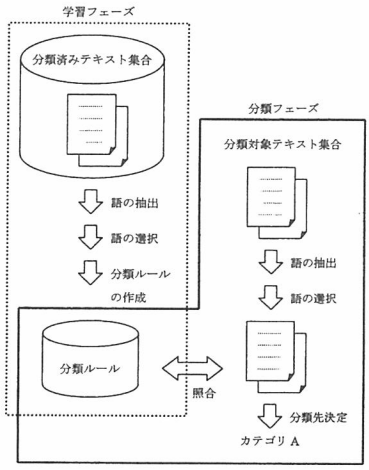

図 2　自動分類システムのしくみ

4．テストコレクションの作成

分類ルールを構築するためには，分類済みのテキスト集合が必要である．この集合をテストコレクションという．本研究のテストコレクションは，出品タイトルを検索キー「古文書」でヒットした取引の中で，一定期間の実際に古文書を取り扱かった取引であるかどうかの判断をした集合である．このテストコレクションの設定には，ネットオークションの検索機能の限界と，事前に行った「実務として文書館職員がネットオークション内の史料取引情報を取得するために，どのような検索方法で行うか」の聞き取り調査を参考にした．本研究で用いた取引情報の期間は 2008 年 7～9 月中に完了した取引である．テストコレクションの収集は以下の手順で行った．

（1）検索キー「古文書」でヒットした取引を収集する．
（2）出品時に画像のない取引は，古文書取引であるか判断を行えないので除外する．
（3）（2）の集合を，アーカイブズ学の記録資料に関する専門知識を持つ判定者が，（a）古文書（明治期初期以前に，相手に要件を伝える目的作成された記録文書），（b）和本，写本（出版物の手書き複写物）類，（c）明治後期以降作成の古紙類，（d）その他，に分類する．表 1 はその分類を行った結果である．

分類	件数	割合
(a) 古文書	769	25.52%
(b) 和本,写本類	1636	54.30%
(c) 明治後期以降の古紙類	357	11.85%
(d) その他（新書，箱等）	251	8.33%
総計	3013	100.00%

表 1　検索キー「古文書」取引集合の分類分布

5. 出品タイトルテキストのトークン化

SVM などの機械学習手法で，出現語を属性として用いて自動判定を行う場合，日本語は膠着語であるため，テキストデータをトークン（文字列や単語）に分割する必要がある．出品タイトルも同様の理由からトークンに分割する必要がある．本研究のトークン化には Yahoo! JAPAN 日本語形態素解析 Web サービス[11]を用いた．

表 2 はテストコレクション全集合における出品タイトルテキスト内の頻出語句である．

順位	出現語	品詞	出現数	頻度
1	和本	名詞	1079	2.06%
2	古書	名詞	594	1.14%
3	年	接尾辞	501	0.96%
4	郡	名詞	363	0.69%
5	江戸	名詞	350	0.67%
6	冊	名詞	331	0.63%
7	明治	名詞	306	0.59%
8	江戸時代	名詞	279	0.53%
9	文書	名詞	245	0.47%
10	蔵出	名詞	240	0.46%
出現語総数			52305	100%

表 2　出品タイトルに含まれる頻出語句

頻出語句の抽出に当たり，テストコレクションの全取引に「古文書」が内在しているので，本研究では除外した．また助詞と特殊文字（括弧，記号）は語句に意味を持たず，分類ルールを作成できないと判断し，該当語句を除外した．

6. テストコレクションの特性

テストコレクションの（a）～（d）の判断された分類別に，（ｉ）出品タイトルをトークン化した語彙集合，（ii）出品カテゴリ，（iii）出品者の Yahoo! ID，を出現頻度順に集計し，それぞれの特徴をまとめる．

表 1 より，テストコレクションでもっとも多くを占めるのは(b) 和本,写本類で 54.3%にあたる．つづいて（a）古文書が 25.52%であり，このことからもテストコレクションと同様の手順で古文書を取り扱う取引情報の取得を試みても効率が悪いことがわかる．

（a）古文書

順位	出現語	品詞	出現数	頻度
1	**文書**	名詞	238	1.78%
2	**蔵出**	名詞	236	1.77%
3	**秩父**	名詞	236	1.77%
4	**江戸時代**	名詞	223	1.67%
5	**家**	接尾辞	199	1.49%
出現語総数			13357	100%

表 3　（a）出品タイトルに含まれる頻出語句

出品カテゴリ	件数	頻度
アンティーク,コレクション/印刷物/その他	466	60.8%
ホビー,カルチャー/美術品/書/その他	133	17.3%
ホビー,カルチャー/美術品/書/掛軸	40	5.2%
本,雑誌/文学,小説/古典文学/日本古典	37	4.8%
ホビー,カルチャー/美術品/版画/木版画	26	3.4%
上記以外のカテゴリ(16 カテゴリ)	65	8.5%

表 4　（a）頻出出品カテゴリ

Yahoo! JAPAN ID	件数	頻度
goroncho55	236	30.8%
fwf28429	52	6.8%
abh_940	30	3.9%
abcd_mart2000	27	3.5%
aoaiauay	26	3.4%
上記以外出品者(115 人)	396	51.6%

表 5　（a）頻出出品者の Yahoo! ID

出現語「秩父」は Yahoo！ID「goroncho55」の取引情報に必ず付与した語句である．また（a）の特徴として出現語「家」がある．この語句は古文書の出所を示す情報として付与している場合が多い．

（b）和本

順位	出現語	品詞	出現数	頻度
1	和本	名詞	862	10.3%
2	古書	名詞	464	5.5%
3	冊	接尾辞	269	3.2%
4	江戸	名詞	245	2.9%
5	郡	名詞	220	2.6%
出現語総数			8383	100%

表 6　（b）出品タイトルに含まれる頻出語句

出品カテゴリ	件数	頻度
アンティーク,コレクション/印刷物/その他	493	30.2%
本,雑誌/文学,小説/古典文学/日本古典	428	26.2%
ホビー,カルチャー/美術品/版画/木版画	258	15.8%
ホビー,カルチャー/美術品/書/その他	76	4.7%
本,雑誌/文学,小説/時代,歴史/日本の歴史	69	4.2%
上記以外のカテゴリ(16 カテゴリ)	310	19.0%

表 7　（b）頻出出品カテゴリ

Yahoo! JAPAN ID	件数	頻度
aoaiauay	346	21.2%
kadono777jp	250	15.3%
iruman41777	210	12.9%
mekunmon	69	4.2%
koukakurou99	56	3.4%
上記以外出品者(131 人)	703	43.0%

表 8　（b）頻出出品者の Yahoo! ID

出現語「郡」は Yahoo！ID「kadono777jp」の取引情報に必ず付与した語句である．しかし，旧所在地をしめす「郡」の場合もあるので，一概に特性と言いにくい．

(c) 明治古紙

順位	出現語	品詞	出現数	頻度
1	明治	接頭辞	148	6.3%
2	年	接尾辞	126	5.4%
3	資料	名詞	75	3.2%
4	郡	名詞	61	2.6%
5	商業	名詞	56	2.4%
出現語総数			2343	100%

表 9　(c) 出品タイトルに含まれる頻出語句

出品カテゴリ	件数	頻度
アンティーク,コレクション/印刷物/その他	237	66.6%
ホビー,カルチャー/美術品/版画/木版画	63	17.7%
ホビー,カルチャー/美術品/書/その他	26	7.3%
アンティーク,コレクション/雑貨/その他	8	2.2%
ホビー,カルチャー/美術品/書/掛軸	4	1.1%
上記以外のカテゴリ(11 カテゴリ)	18	5.1%

表 10　(c) 頻出出品カテゴリ

Yahoo! JAPAN ID	件数	頻度
kadono777jp	59	16.6%
t_s_u_auction	54	15.2%
fwf28429	22	6.2%
homeplan36	16	4.5%
cupidsan2000	14	3.9%
上記以外出品者(131 人)	191	53.7%

表 11　(c) 頻出出品者の Yahoo! ID

(d) その他

順位	出現語	品詞	出現数	頻度
1	アンティーク	名詞	49	1.2%
2	フランス	名詞	47	1.2%
3	本	名詞	46	1.1%
4	年	接尾辞	37	0.9%
5	手紙	名詞	31	0.8%
合計			8383	100%

表 12　(d) 出品タイトルに含まれる頻出語句

出品カテゴリ	件数	頻度
住まい,インテリア/キッチン,食器/収納/缶	43	17.2%
本,雑誌/人文,社会/歴史/日本史	41	16.4%
アンティーク,コレクション/印刷物/その他	40	16.0%
本,雑誌/文学,小説/古典文学/日本古典	9	3.6%
本,雑誌/人文,社会/文化,民俗	9	3.6%
上記以外のカテゴリ(46 カテゴリ)	108	43.2%

表 13　(d) 頻出出品カテゴリ

(d) その他で頻繁に見られる取引上には，古文書崩し字読解を取り扱った新書や，以前古文書を収納していた木箱などが多く見られた．そのため出品カテゴリが多岐にわたる．

Yahoo! JAPAN ID	件数	頻度
gyhcr239	26	10.4%
tanbaji80	17	6.8%
kanro30	16	6.4%
pepegannzo	15	6.0%
escalade50jp	9	3.6%
上記以外出品者(85 人)	105	42.0%

表 14　(d) 頻出出品者の Yahoo! ID

7. 実験環境

本研究では分類ルールの作成を試みる．分類ルールの作成に用いた属性は（ｉ）出品タイトルをトークン化した語彙集合の上位 20 語と（ii）出品カテゴリの出現回数上位 5 カテゴリ，（iii）出品者の Yahoo! ID の出現回数の多い上位 5 件の出品者 ID を用いた．（ｉ）を用いた理由は，出品タイトルの中に出現する語を手がかりとして，取引情報内容を判定できるのではないかと考えたからである．特にテストコレクションの特性より，使用目的の異なる語が確認できる．例えば史料が何であるかを表す語（出現語「文書」，「和本」ほか），史料の状態を表す語（出現語「年」，「冊」ほか），史料の出所を表す語（出現語「郡」，「家」ほか），特定の出品者が付与する語（出現語「蔵出」，「秩父」ほか）など様々である．これらの詳細についてはまだテストコレクションの収集と分析が不可欠であるが，分類ルールの属性として有用であると考える．（ii），（iii）もテストコレクションの特性より，特徴的な出現をするため属性に加えた．テストコレクションには（ｉ）～（iii）以外にも，落札日（時間），落札価格，落札者数などの属性も抽出できたか，取引内容に深く係わる属性か判断できなかったため，本研究では除外した．

カテゴリ	属性
出現キーワード (上位 20 位)	「和本」「古書」「年」「郡」「江戸」「冊」「明治」「江戸時代」「文書」「蔵出」「浮世絵」「秩父」「家」「太郎」「久馬」「写本」「牧原」「資料」「江戸期」「杜」
カテゴリ（上位 5 位）	･･･/印刷物/その他 ･･･/古典文学/日本古典 ･･･/美術品/版画/木版画 ･･･/美術品/書/その他 ･･･/美術品/書/掛軸
出品者 ID(上位 5 位)	「 aoaiauay 」「 kadono777jp 」「 goroncho55 」「 iruman41777 」「fwf28429」

表 15　実験で使用した属性

分類ルールの作成には Weka[10]に実装される C4.5（モジュール名は J48 ）の決定木アルゴリズムを用いた．Weka では出現語は高次元の属性群であり取り扱うことができない．そこで Weka に取引情報を読み込ませるため，（ii），（iii）に含まれる属性は個別の属性として全て列挙し，出現語の属性と同

様にその有無で判定を試みた．実験に使用する属性を表 15 に示した．

8. Weka による決定木の結果

本研究のテストコレクションは 2008 年 7～9 月中に検索キー「古文書」ヒットし，古文書取引の分類判断を終えた取引集合である．図 3 はテストコレクションを Weka の決定木アルゴリズムによって分類ルールを作成した様子である．

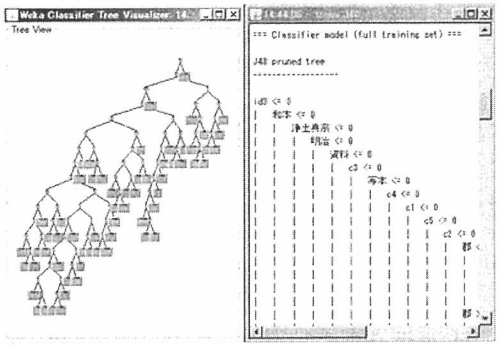

図 3　Weka による決定木作成の様子（4 分類判断）

本研究で作成された決定木は Weka に実装される交差検証法において，図 4 より 74.75%の値が得られた．

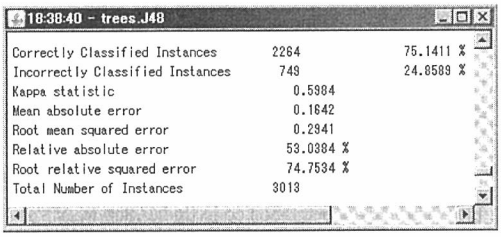

図 4　交差検証法による評価（4 分類判断）

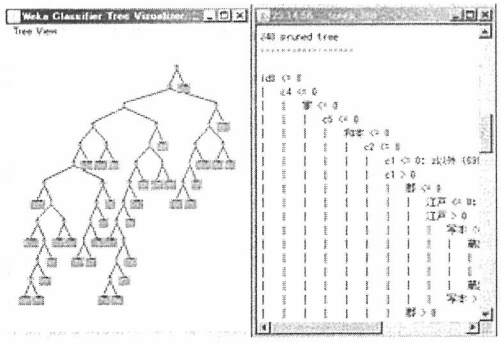

図 5　Weka による決定木作成の様子（2 分類判断）

また，（a）から（d）までの分類判断の内，（a）古文書以外の（b）から（d）までを「（z）古文書以外の取引」にまとめ，（a）と（z）のみの分類判断で Weka の決定木アルゴリズムによって分類ルールを作成すると図 3 のような決定木が得られた．そして（a）古文書と（z）古文書以外で作成した決定木は Weka に実装される交差検証法において，図 4 より 85.7%の値が得られた．図 3 と図 5 を比較すると決定木の構造が可視化しやすくなり，交差検証法で導かれる数値も 10%ほど上昇していることからも明らかである．

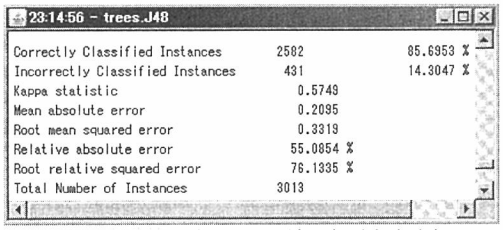

図 6　交差検証法による評価（2 分類判断）

図 5 において，決定木の上位から概観すると以下の事柄がわかる．はじめに最上位の分岐基準はユーザーID の一つです．このことは古文書のみ出品している出品者の存在を示している．次に最も古文書以外の取引集合を特定する条件を確認すると，カテゴリ「･･･/**美術品/書/その他**」，カテゴリ「･･･/**美術品/書/掛軸**」，**出品タイトルの出現語「家」を属性に持たず，且つ「和本」を含む取引集合の 994 件中 949 件が古文書を取り扱わない取引である．**

9. 考察と今後の課題

本研究は，ネットオークションの全取引から古文書を扱う取引を自動的に判別するための分類ルールの検討を，Weka に実装される C4.5 の決定木アルゴリズムを用いて作成する試みを行った．この分類ルールを作成するために使用した（ｉ）出品タイトルをトークン化した語彙集合の上位 20 語，（ii）出品カテゴリの出現回数上位 5 カテゴリ，（iii）出品者の Yahoo! ID の出現回数の多い上位 5 件の出品者 ID，の属性は，図 6 の交差検証法の数値から判断しても適当であったと考える．

今後，交差検証法の数値を高めるための方法として，属性の数を増やすことで，より詳細な分類ルールの作成が行えると考える．現段階で属性数を増やす方法として，（ｉ）～（iii）の頻出属性の数値を増やすことが考えられる．そのためには，異なる期間の取引を対象とした分類ルールと，その適用実験の蓄積が必要である．

また Weka 以外の分類器に本研究で収集できたデータを適用することで，決定木以外の方法からの自動判定の可能性を模索する必要がある．

10. おわりに

本研究では副産物的な成果がいくつかあり，それらについて言及して終わりとする．

1 つ目はネットオークション内で古文書を扱う取引の規模の一端が把握できたことである．本研究のテストコレクションより，古文書を扱う取引数は 2008 年 7 月～9 月までの 3 ヶ月間で少なくとも 769 件の取引が成立している．これは 1 ヶ月あたり 30 日で計算すると，1 日あたり 8.5 件の古文書を扱う取引が成立していることになる．テストコレクションは出品タイトル内に「古文書」を含む取引のみであるので，
実際の取引数はより多いことが予想される．

2 つ目はネットオークションで古文書を扱う取引特有の属性に関する追跡調査が必要である．これまで多くの文化行政機関ではそれぞれの史料収集方針を設定し，現在もそれぞれの機関において収集活動を行っている．その活動においては旧所在地名や旧支配者などの歴史学に関する知識や史料の形状，書状の型式など古文書学に関する知識が求められる．しかし本研究の（ｉ）出品タイトルをトークン化した語彙集合を確認すると，歴史学や古文書学では導きだせないキーワードの存在がわかる．現在確認できる語として「蔵出」「珍品」など落札者の落札意識を刺激するキーワードの存在や，テストコレクションの特性で記載した特定の出品者が付与する特定のキーワードなどが存在する．今後定期的に取引情報を観測することで，以上のようなネットオークション特有の語彙を収集することができると考える．以上の 2 点が副産物的な成果である．

本研究では利用者として想定している文化行政機関の職員からの分類ルールの作成，並びに本研究の着眼点に関する評価を行えていないことが心残りである．しかし，本研究では出現語「古文書」を含む取引情報のみを対象としているので，出現語「古文書」を含まない古文書を取り扱う取引情報の集合が欠けている．

今後はテストコレクションより抽出できた属性から，出現語「古文書」を除外した状態で分類ルールを作成し，全取引情報に適用することで，出現語「古文書」を含まない古文書を取り扱う取引情報の集合の収集を試みる予定である．そして収集できうる全ての古文書を取り扱う取引情報を分析することで，文化行政機関の職員の評価と日々取引される史料の対策を検討するために必要な基礎情報の一つを提供できると考える．

11. 謝辞

記録資料の流通，分類判定に関してご助言をいただいた大友一雄氏（国文学研究資料館）,綿抜豊昭氏（筑波大学），また自動分類，データマイニングに関してご助言をいただいた谷口祥一氏（筑波大学），緑川信之氏（筑波大学）に感謝いたします．

参考文献

[1] 国文学研究資料館史料館編: 史料館収蔵史料総覧．名著出版，1996.

[2] 新井浩文．文書館における民間所在資料(古文書)の取り扱いをめぐって (小特集 公文書館専門職員養成課程修了研究論文)，埼玉県立文書館文書館紀要，2002，no.15，p. 39-54.

[3] 平井義人．阪神・淡路大震災の教訓は生かされたのか：文化財保護法を柱にした「地域史料」調査の実践．地方史研究，2005，vol.55，no.2，p.26-36.

[4] 藤隆宏．民間所在資料保存状況調査の中間報告．和歌山県立文書館紀要，2003，no. 8，p. 103-117.

[5] 藤隆宏．民間所在資料保存状況調査結果報告．和歌山県立文書館紀要，2007，vol.12，p.190-167.

[6] Yahoo! オークション. http://auctions.yahoo.co.jp/, (参照 2008-09-01).

[7] 文部科学省. インターネット・オークションについて． http://www.mext.go.jp/b_menu/shingi/bunka/gijiroku/013/07072002/006.htm，(参照 2008-09-01).

[8] 石田 栄美ほか: 目次と帯を用いた図書の自動分類(情報検索・分類,テーマ：「デジタルアーカイブの活用(応用)」および一般). 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2006, No.33, p.85-92, 2006.

[9] 安形 輝ほか: WWW ページの自動分類：NDC の分類体系と Yahoo のカテゴリを使った分類. 情報処理学会研究報告. データベース・システム研究会報告．Vol.99，No.39，p.113-120，1999.

[10] Waikato 大学にて開発されたデータマイニング・ツール. http://www.cs.waikato.ac.jp.nz/ml/, (参照 2008-09-01).

[11] http://developer.yahoo.co.jp/jlp/MAService/V1/parse.html, (参照 2008-09-01).