D-009

ウェブページ分類用ジャンル体系の再利用適合性評価

Evaluation of Reuse-Suitability of Genre Systems for Classifying Web Pages

大森 晃†
Akira Ohmori

1. はじめに

近年，機械学習法を利用して，ウェブページの自動的ジャンル分類技術に関して研究が行われている[1][2][3][4][5]．ウェブページの自動的ジャンル分類技術（特にジャンル分類器）は，例えば，WWWユーザが情報収集を行う場合，指定した検索キーワードによる検索結果をジャンル分類器によって自動分類し，ユーザが所望するジャンルに限定して表示するというように，WWW検索を支援する技術として今後必要になってくると考える．

一般に，機械学習法を利用した自動的ジャンル分類技術の研究では，以下の手続きが実行されている．

(1) ウェブページのジャンル体系を設定する．
(2) WWW から収集したウェブページの集合（以下，ウェブページセットと呼ぶ）を対象にして，設定したジャンル体系に従って各々のウェブページを人間が分類し，ジャンル分類器の訓練用・テスト用データセットを作成する．
(3) 訓練用・テスト用データセットを用いて，各ウェブページに1つのジャンルを割り当てるジャンル分類器を訓練・テストし，その分類性能を評価する．

上記(3)は自動的ジャンル分類技術のこれまでの研究における中心的部分であり，よりよい分類性能を有するジャンル分類器を構築するために，どんな機械学習アルゴリズムが適当か，機械学習においてどんな素性（feature）や素性値が適当か，などが多く議論されてきた．しかしながら，自動的ジャンル分類技術の研究に着手する際にまず直面する問題は，どんなジャンル体系を設定するかである．どんなジャンル体系であれ，それを設定しなければ次の研究手続きには進めない．この意味で，上記(1)は自動的ジャンル分類技術の研究において重要な基礎的部分である．

独自のジャンル体系を新設する場合には，それなりに労力が必要になる．例えば大森[6]は，日本語ウェブページを主観的／非主観的に分類する分類器を構築する技術的研究のなかで，分類器の「ジャンル領域拡大化能力」を評価するために独自のジャンル体系を用いている．当該研究論文の付録部に詳述されているジャンル体系の設定法から，ジャンル体系の新設にはかなりの労力が必要であることが予想できる．また，独自のジャンル体系を利用する場合には，ジャンル分類器の構築技術について他研究との比較ができず，技術の優劣を判断できない．こうした状況は技術的研究において好ましいとは言えない．

一方，既設のジャンル体系が幾つかあり，そのうち再利用に適しているものがあれば，それを再利用する方が，ジャンル体系新設のための労力を削減できるので，得策である．その上，ジャンル分類器の構築技術について同一のジャンル体系のもとで優劣の比較が可能になるので，より優れた技術を追究していけるという技術研究上のメリットも生じる．

現状では，既設のジャンル体系はいくつか存在している[1][2][3][4][5][6][7]．しかしながら，それらが再利用に適しているか否かについて定性的にすら評価した研究が見当たらず，どのジャンル体系が再利用に適しているかは明確ではない．

そこで本論文では，いくつかの既設ジャンル体系を取り上げ，そのうちどれが再利用に適しているかを明らかにする．まず 2 節では，意味明瞭性，排他性，網羅性という性質について既設ジャンル体系を定性的に評価し，そのなかから再利用に最も適していると思えるジャンル体系を選択する．続いて 3 節では，選択したジャンル体系が再利用に適しているか否かを判断するために，意味明瞭性，排他性，網羅性とともに重要な評価指標となるジャンル決定率を導入する．そして 4 節では，当該ジャンル体系のジャンル決定率を実験的に求め，ジャンル決定率の許容性を評価する．5 節では当該ジャンル体系が再利用に適していると結論づけるとともに，ジャンル体系の再利用適合性評価の実施から得られた副産物に言及する．

なお，Finnら[8]は，ジャンルとは何か？について以下のように述べている．

(1) ジャンルは，文章の主題に関係するというよりも，むしろ一定の文章スタイルを反映する．
(2) ジャンルは，その文章が何について書かれているかというよりもむしろ，それがどんな種類の文章であるかを示す．
(3) ジャンルは主題とは直交する概念であり，同一主題について書かれた複数の文章が，異なるジャンルに所属し得る．同様に，同じジャンルに所属する複数の文章は，異なった主題を持ち得る．

このように，ウェブページのジャンルは，主題を意味するものではなく，主題とは直交し，文章の種類に関わる概念であると解釈できる．

2. 既設ジャンル体系の定性的評価

本節では，ウェブページ分類用の既設ジャンル体系[1][2][3][4][5][6][7]を，以下のような性質について，評価値「低」，「中」，「高」を用いて第 3 者視点にたって定性的に評価し，そのなかから再利用に最も適していると思えるジャンル体系を選択する．

(1) 意味明瞭性：ジャンルの意味が明瞭である．
(2) 排他性：ジャンルの意味が排他的である．
(3) 網羅性：実在するジャンルを網羅している．

自動的ジャンル分類技術の研究において，あるジャンル体系が再利用に適している（つまり再利用適合性を有する）と言えるためには，当該ジャンル体系が意味明瞭性と

†東京理科大学, Tokyo University of Science

排他性を有することは必須であると考える．その理由は，これらの性質を欠いたジャンル体系では，人手によるウェブページのジャンル分類が適切に行えないからである．人手によるジャンル分類が適切に行えないならば，ジャンル分類器の訓練用・テスト用に適切なデータセットを作成することが困難になる．

自動的ジャンル分類技術の研究において再利用に適するジャンル体系の性質として，網羅性は必ずしも必須であるとは言えないかもしれない．しかしながら，本論文では自動的ジャンル分類技術は WWW 検索支援にとって今後必要であるとの立場をとっており，網羅性は非常に重要な性質であると考える．なぜならば，ジャンル体系が多様なジャンルを網羅することにより，WWW 検索ユーザのジャンル分類に対する多様なニーズに応えられる可能性が高くなるからである．

表 1 に本論文で評価対象とする既設ジャンル体系の一覧を示す．第 1 列にはジャンル体系の設定者を示し，第 2 列にはジャンル体系を示してある．第 3 列から第 5 列には，以下で意味明瞭性，排他性，網羅性について定性的に評価した結果を示してある．なお，評価値は「低」，「中」，「高」を用いているが，評価するに価しない場合には記号「—」を用いている．

7 つのジャンルから成るDewdneyらのジャンル体系[1]と，7 つのジャンルから成るLeeらのジャンル体系[2]では，各ジャンルに説明も例示も与えられていない．11 のジャンルから成るDeweらのジャンル体系[7]においては，各ジャンルに説明は与えられていないが，いくつかのジャンルには例示が与えられている．これらのジャンル体系について，各ジャンルの名称から，当該ジャンルがどんな意味を持つかを推察することは可能であるかもしれない．しかしながら，推察はあくまでも推察の域を出ない．以上のことから，上記 3 つのジャンル体系の意味明瞭性は低いと評価するのが適当である．意味明瞭性が低いジャンル体系については，ジャンルの意味が確定していないことから，排他性や網羅性について適当な評価を行なうことは非常に困難である．また，推察されるジャンルの意味に基づいて排他性や網羅性を評価することに，大きな意味があるとも思われない．

3 つのジャンルから成る Kennedy らのジャンル体系[5]では，各ジャンルに簡単な説明が与えられている．Personal Home Page は営利を目的としない個人が発信するウェブページであり，Corporate Home Page は営利を目的とする企業が発信するウェブページであり，Organization Home Page は営利を目的としない団体が発信するウェブページである．ただし，各ジャンルの理解を助けるための例示が与えられていないことから，本ジャンル体系の意味明瞭性は，中程度と評価するのが適当である．本ジャンル体系は，発信主体の違いによってジャンルを設定しており，排他性は高いと評価してよい．しかしながら，網羅性については，実在するジャンルであると認め得るけれどもジャンル数が 3 つしかなく，低いと評価するのが適当である．

8 つのジャンルから成る Meyer zu Eissen らのジャンル体系[4]では，各ジャンルに簡単な説明が与えられており，いくつかのジャンルには例示も与えられている．したがって，本ジャンル体系の意味明瞭性は高いと考える．本ジャンル体系は，学生に対するアンケート調査結果を踏まえて設定されたもので，学生が頻繁に利用するジャンルによって構成されていると考えてよい．そのため，かなり重要と考えられる報道に関するジャンルが欠けている．網羅性については，このような問題を抱えていることと，設定されているジャンル数が 8 つであることから，中程度と評価するのが適当である．排他性の評価については，各ジャンルの説明や例示を吟味する必要がある．そのため，本ジャンル体系を以下に引用する．

(1) Help: All pages that provide assistance, e.g. Q&A or FAQ pages.
(2) Article: Documents with longer passages of text, such as research articles, reviews, technical reports, or book chapters.
(3) Discussion: All pages that provide forums, mailing lists or discussion boards.
(4) Shop: All kinds of pages whose main purpose is product information or sale.

表 1 既設ジャンル体系とその定性的評価
Table 1 Existing Genre Systems and Their Qualitative Evaluation

ジャンル体系の設定者	ジャンル体系	意味明瞭性	排他性	網羅性
Dewdneyら[1]	Advertisement, Bulletin Board, FAQ, Message Board, Radio News, Reuters Newswire, Television News	低	—	—
Leeら[2]	Reportage, Editorial, Technical Paper, Critical Review, Personal Homepage, Q&A, Product Specification	低	—	—
Deweら[7]	Informal-Private, Public-Commercial, Searchable Indices, Journalistic Materials, Reports, Other Running Text, FAQs, Link Collections, Other Listings and Tables, Asynchronous Multi-Party Correspondence, Error Messages	低	—	—
Kennedyら[5]	Personal Home Page, Corporate Home Page, Organization Home Page	中	高	低
Meyer zu Eissenら[4]	Help, Article, Discussion, Shop, Portrayal(non-priv), Portrayal(priv), Link Collection, Download	高	低	中
木村ら[3]	意見，説明・解説，紹介・案内，質問・回答，報知・報告，表現，その他	高	低	中
大森[6]	予想，標語，批評，解説，報道，独白・感想，情宣（情報宣伝），質問・相談・依頼（回答つき限定），記録（現代文限定），商品広告・宣伝，マニュアル，用語説明，案内・紹介，その他	高	高	高

(5) Portrayal (non-priv): Web appearances of companies, universities, and other public institutions. I.e., home or entry or portal pages, descriptions of organization and mission, annual reports, brochures, contact information, etc.
(6) Portrayal (priv): Private self-portrayals, i.e., typical private homepages with informal content.
(7) Link Collection: Documents which consist of link lists for the main part.
(8) Download: Pages on which freeware, shareware, demo versions of programs etc. can be downloaded.

本ジャンル体系に従うと，例えば，研究論文（research articles）や技術報告（technical reports）を記載するウェブページは，ジャンル Article に属するが，それらが個人によって発信されたウェブページに記載されている場合，ジャンル Portrayal(priv)にも属すると判定できる．また，ソフトウェアの製品情報（product information）を記載するウェブページはジャンル Shop に属するが，当該ソフトウェアがダウンロード可能な場合，ジャンル Download にも属すると判定できる．排他性については，このように問題を抱えており，低いと評価するのが適当である．

7 つのジャンルから成る木村らのジャンル体系[3]では，各ジャンルに簡単な説明と例示が与えられている．したがって，本ジャンル体系の意味明瞭性は高いと評価するのが適当である．網羅性については，実在するジャンルであると認め得るけれども設定されているジャンル数が 7 つであることから，中程度と評価するのが適当である．排他性の評価については，各ジャンルの説明や例示を吟味する必要がある．そのため，本ジャンル体系を以下に引用する．

(1) 意見：ブックレビュー，感想，主張など，書き手が自分の考えを述べているもの．
(2) 説明・解説：専門用語の説明，操作手順の説明など，読み手に深い理解をうながすもの．
(3) 紹介・案内：人物紹介，製品紹介，交通経路案内など，読み手に知覚してもらうことを目的としたもの．
(4) 質問・回答：疑問点・問題点の解決を求めているもの，およびその回答（FAQ などもここに含める）．
(5) 報知・報告：ニュース，レポートなど，読み手に何らかの事実を報せることを目的としたもの．
(6) 表現：日記，随筆，小説など，書き手が何かを表現しているもの．
(7) その他：上記のどのカテゴリの条件にも当てはまらないもの．

本ジャンル体系においては，「報知・報告」というジャンルは「読み手に何らかの事実を報せることを目的としたもの」と説明されているが，「紹介・案内」というジャンルも，その例示から，ある意味で人物，製品などに関する事実を報せるものであり，両者のジャンルの区別がはっきりしていない．さらに問題になるのは「表現」というジャンルである．これは「書き手が何かを表現しているもの」と説明されている．ウェブページは書き手が何かを表現しているメディアであるため，どのウェブページも「表現」というジャンルに属すると判定でき，「表現」というジャンルと他の全てのジャンルとの区別がはっきりしない．本ジャンル体系では，ジャンル「報知・報告」とジャンル「紹介・案内」との排他性に問題があると認められること，ジャンル「表現」が非常に包括的であることから，排他性は低いと評価するのが適当である．

14 のジャンルから成る大森のジャンル体系[6]では，各ジャンルに適度な説明が与えられているとともに，例示も与えられている．したがって，本ジャンル体系の意味明瞭性は高いと評価するのが適当である．本ジャンル体系は NTCIR-3 WEB[a]からのランダムサンプルをもとに設定されており，大森[6]によって詳述されている設定法から各ジャンルは実在するものとして認め得るものであり，ジャンル数が 14 であることから，網羅性は高いと評価するのが適当である．排他性の評価については，各ジャンルの説明や例示を吟味する必要がある．そのため，本ジャンル体系を以下に引用する[b]．

(1) 予想：ある物・事の今後の動きや結果についてあらかじめ想像したもの．例としては，株価予測や天気予報を記載したページがある．
(2) 標語：個人・団体のモットーやスローガン，商品のキャッチフレーズなど，人の注意をひくように工夫した簡潔な文言のみを書いたもの．例としては，「滅菌効果抜群！」というような商品のキャッチフレーズを記載したページがある．なお，簡潔な文言とともに，その説明が加えられている場合は，このジャンルには該当しない．
(3) 批評：ある物・事の善悪・優劣・美醜・良悪・是非などについて第三者として評価し論じたもの．例としては，業況批評を記載したページがある．
(4) 解説：ある物・事について，その内容（法文の条項，文学作品，ニュース記事，ある人の発言，株価動向など）を引用した上で，それを分析して，わかりやすいように客観的に説明したもの．例としては，ある法文を引用し解説したページがある．通常，引用は他の情報源からとられた文章の形式をとるが，場合によっては，グラフや表の形式をとることもある．引用内容が明示的でない場合は，このジャンルに該当しない．さらに，用語（単語，句）の意味を説明するページは，このジャンルに該当しない．
(5) 報道：社会の出来事などを報道機関が告げ知らせようとしたもの．例としては，ニュースの見出し（ヘッドライン），ニュース記事本文を記載したページがある．ニュースのような情報を記載はしているが，報道機関が明示的でない場合，このジャンルに該当しない．ここで報道機関は通常のニュース配信組織（テレビ局，新聞社，雑誌発行機関など）だけを意味していない．報道機関には，他のページに記載されたニュースを改めて報道する組織・機関（例えば，http://news.google.co.jp）も含まれる．ただし，個人は含まれない．
(6) 独白・感想：ある物・事について，個人あるいは複数の人がその人の立場で気ままに自分の考え・思いを語ったもの，あるいは個人的な体験記．例としては，経験談，私的な記録（日記など），私的なメッセージ，個人の布教文，自作の書き物（物語，雑談など），メール，を記載したページがある．

---

a) NTCIR-3 WEBについては以下のURLを参照されたい．http://research.nii.ac.jp/ntcir/permission/perm-ja.html#ntcir-3-web（参照 2009-05-16）

b) 大森のジャンル体系[6]では，各ジャンルの説明で用いられている「物」と「事」という用語は次のような意味を持つ．「物」とは形のある物体をはじめとして，存在を客観的に感知できる対象であり，「事」とは意識・思考の対象のうち，具象的・空間的でなく，抽象的に考えられるものである．

(7) 情宣（情報宣伝）：団体（圧力団体，宗教団体，町内会，労働組合など）が推し進めようとようとしている考え方・思想・経典についての情報を提供し，その有効性や有用性，危険性や有害性などを説明して理解・共鳴させようとしたもの．例としては，ウェブ版ニュースレター，ウェブかわら版，機関紙を記載したページがある．

(8) 質問・相談・依頼（回答つき限定）：ある物・事について，他人に意見を求め，返答を得たもの．例としては，FAQ，Q&A，不特定多数による一連の質問・回答を記載したページがある．質問だけ，あるいは，ある質問に対する回答だけを記載するページは，このジャンルに該当しない．

(9) 記録（現代文限定）：過去の事実や将来の計画について，発信者あるいは受け手が後々の証拠として使いうるように，書き残されたもの．ただし，現代文に限定する．例としては，研究報告，終了イベント，競技結果，史実，議事録，都市再開発計画を記載したページがある．たとえ事実や計画を記載していても，記載内容が単純に追加されるという更新を除いて，記載内容の一部あるいは全部が更新される可能性が高いページは，このジャンルに該当しない．

(10) 商品広告・宣伝：商品化されたものについて，人々に関心を持たせ，購買させることを目的として，最低限，その商品を特定する情報（商品名，品番など）と価格情報を知らせたもの．例としては，株式投信，講座，宿泊施設，駐車場の広告を記載したページがある．「オープンプライス」という語句は，価格情報と見なす．以下のようなページはこのジャンルに該当しない．

(a)誰かによって既に購買された商品の情報を記載するページ．

(b)製造中止とか販売終了とかの理由で，すでに購入できなくなっている商品の情報を記載するページ．

(11) マニュアル：ある物・事について，その使用法，調理法，摂取法，作成法，設定法，見方，進め方などの手順に関する技術的知識・情報のすべて，ないし一部を主として現在形で説明したもの，あるいはノウハウ．例としては，PC 関連，書類届出関連のマニュアル，ノウハウを記載したページがある．

(12) 用語説明：ある物・事について，それが何であるかの客観的説明に重点を置いたもの．例としては，PC 用語，歴史的人物，史跡，文化遺産を記述したページがある．

(13) 案内・紹介：情報の受け手にとって未知，既知を問わず，ある物・事へ導く情報，あるいは，物事の特徴的情報を羅列して知らせたもの．例としては，イベント案内，方法の紹介，新製品の紹介を行ったページがある．

(14) その他：上記のどのジャンルにも当てはまらないもの．例としては，英語ページ[c]，画像ページ，Not-Found ページがある．

本ジャンル体系では，一見して，ジャンル「批評」とジャンル「独白・感想」との排他性に問題があるように見受けられる．しかしながら，「批評」の説明では「第三者として評価し論じたもの」となっている一方，「独白・感想」の説明では「その人の立場で気ままに自分の考え・思いを語ったもの」となっている．ジャンル「批評」とジャンル「独白・感想」については，文章としてのスタイルの相違が明確になっており，排他性は確保されていると考えてよい．

ジャンル「商品広告・宣伝」とジャンル「案内・紹介」との排他性にも，一見して問題があるように見受けられる．例えば，ある商品に関する情報を，その商品名と価格を含めて記述したウェブページは，「商品広告・宣伝」というジャンルに属するのか，「案内・紹介」というジャンルに属するのか，判断に迷うかもしれない．しかしながら，「商品広告・宣伝」の説明では「人々に関心を持たせ，購買させることを目的として」となっている一方で，「案内・紹介」の説明では「ある物・事へ導く情報，あるいは，物事の特徴的情報を羅列して知らせたもの」となっている．したがって，当該ウェブページは，そこに「人々に関心を持たせ，購買させることを目的」とするような記述が含まれていれば「商品広告・宣伝」というジャンルに属し，そこに商品に関する情報が単に羅列されているのであれば「案内・紹介」というジャンルに属すると判断できる．このように，ジャンル「商品広告・宣伝」とジャンル「案内・紹介」についても，文章としてのスタイルの相違が明確になっており，排他性は確保されていると考えてよい．

本ジャンル体系における各ジャンルの説明から，排他性に問題がありそうなジャンルの組は，上記以外に見当たらない．したがって，本ジャンル体系では排他性は高いと考える．

以上，既設ジャンル体系の意味明瞭性，排他性，網羅性について定性的に評価した．評価結果は**表 1** の右側部分に示した通りである．**表 1** から分かるように，本論文で取り上げている既設ジャンル体系のうち，再利用に最も適していると思えるジャンル体系は大森のジャンル体系[6]であることが明らかになった．

3. ジャンル決定率の導入

あるウェブページ分類用ジャンル体系が意味明瞭性，排他性，網羅性について比較的優れているからといって，それが再利用に適していると判断するのは早計である．以下，ジャンル分類器の訓練・テスト用データセットの作成効率という面からその理由を述べる．

機械学習法を利用したウェブページの自動的ジャンル分類技術の研究においては，1 節で言及したように，あるウェブページセットを対象にして，あるジャンル体系に従って各々のウェブページを人間が分類し，ジャンル分類器の訓練用・テスト用データセットを作成する必要がある．自動的ジャンル分類技術の研究報告において必ずしも明記されているわけではないが，ウェブページを人手によってジャンル分類する際，分類は研究者たち自身，つまり複数の人間によって行われていると考えるのが自然である．

複数の人間がジャンル分類に関与する場合には個々人のジャンル分類結果が必ずしも一致するわけではないので，適当な方法によって，当該ウェブページセットにおける各ウェブページのジャンルを決定しなければならない．もちろん，ジャンル不明という決定もあり得る．したがって，当該ウェブページセットにおける全ウェブページ件数のうち，ジャンルを決定できるウェブページ件数の比率（以下，ジャンル決定率と呼ぶ）が，ジャンル分類器の訓

c) 本ジャンル体系は日本語ウェブページを対象としているので[6]，ジャンル「その他」の例として英語ページが示されていると考えてよい．

練・テスト用データセット[d]の作成効率という面から重要になってくる．

あるジャンル体系が意味明瞭性，排他性，網羅性について比較的優れていても，低いジャンル決定率をもたらすようなものを再利用に適していると判断することは適当ではないと考える．言い換えれば，ジャンル体系は，それが再利用に適していると判断できるためには，意味明瞭性，排他性，網羅性に加えて，許容し得る程度のジャンル決定率をもたらすことが重要な条件であると考える．

以上のことから，ジャンル体系の評価指標としてジャンル決定率を具体的に導入する必要がある．本論文では，実用性を考慮して，以下の 2 つを導入する．

多数決率＝（多数決ページの数/W）×100　(1)
過半数決率＝（過半数決ページの数/W）×100　(2)

ここで，W は人手による分類対象であるウェブページセットにおけるウェブページの件数である．多数決ページと過半数決ページの定義は以下の通りである．定義中，GS はあるジャンル体系を構成するジャンルの集合である．また，M はジャンル分類を行なう人間の数であり，2 以上である．

(a) 多数決ページ：あるウェブページをM人の人間がそれぞれ独立にジャンル分類した結果，あるジャンル$G \in GS$に分類した人間がN_G人とすると，$M = \Sigma_{\{G \in GS\}} N_G$である．この時，あるジャンル$G^* \in GS$が唯一つ存在して，他のすべてのジャンル$G \in GS$について$N_{G^*} > N_G (G^* \neq G)$となる場合，そのウェブページのジャンルを多数決によってジャンルG^*に決定する．そうでない場合には，ジャンルは未決定とする．多数決によってジャンルが決まるウェブページを多数決ページと呼ぶ．

(b) 過半数決ページ：あるウェブページをM人の人間がそれぞれ独立にジャンル分類した結果，あるジャンル$G \in GS$に分類した人間がN_G人とすると，$M = \Sigma_{\{G \in GS\}} N_G$である．このとき，$N_{G^*} > (M/2)$となるようなジャンル$G^* \in GS$が存在する場合，そのウェブページのジャンルを過半数決によってジャンルG^*に決定する．そうでない場合には，ジャンルは未決定とする．過半数決によってジャンルが決まるウェブページを過半数決ページと呼ぶ．

4. ジャンル決定率の許容性評価

2 節で明らかにしたように，本論文で取り上げているジャンル体系のうち，意味明瞭性，排他性，網羅性が高く，再利用に最も適していると思えるジャンル体系は大森のジャンル体系[6]である．3 節における議論から，当該ジャンル体系が再利用に適しているか否かを判断するためには，さらにジャンル決定率（多数決率と過半数決率）の許容性を評価する必要がある．本節では，当該ジャンル体系のジャンル決定率をジャンル分類実験によって求め，その許容性について評価する．なお以下では，大森のジャンル体系に言及する場合，引用文献番号の記載を省略する．

d) ジャンル分類器の訓練・テスト用データセットに含まれるウェブページはすべて，属するジャンルが決まっていなければならない．

4.1 ジャンル分類実験

大森のジャンル体系のジャンル決定率を求めることを実験目的として，どのようなジャンル分類実験を行なったかを以下に述べる．

(a) 実験回数

ウェブページのジャンル分類という判定作業では，判定の基準となるジャンル体系が与えられているとしても，ジャンルの説明に関する理解やウェブページにおける記載内容に関する理解にジャンル分類者（ウェブページをジャンル分類する人間）の主観が入り込み，しかもそれが変化するために，判定の一貫性を確保することが困難である．同じウェブページを，今回はあるジャンルに分類するかもしれないが，次回は別のジャンルに分類するかもしれない．そこで，平均的なジャンル決定率を知るために，ジャンル分類実験を 3 ヶ月に渡ってほぼ 1 ヶ月毎に 3 回実施した．具体的には，後述する(e)ジャンル説明と教示，(f)ジャンル分類を 3 回実施した．

(b) 被験者（ジャンル分類者）

ジャンル分類を行う被験者の人数は筆者を含めて 5 人であった．筆者以外の 4 人は，大学の夜間部学生であり，単位認定対象となっている教育科目の受講生であった．1 人は 20 歳代前半の女子学生であった．残り 3 人は男子学生で，そのうち 2 人は 20 歳代前半であり，1 人は 30 歳代前半であった．4 人の学生被験者は，単位を取得するためにジャンル分類という課題に真面目に取り組むことを了解した学生たちである．

自動的ジャンル分類技術の研究[1][2][3][4][5]に見られるように，通常，ジャンル分類器の訓練用・テスト用データセットの設定に関与するのは 2〜3 名の研究者であると考えられる．こうした状況に比べて，被験者 5 人という数は決して少ない数ではない．

(c) 予備的ウェブページセット

実験用ウェブページセットの作成に先立って，予備的ウェブページセットとして，大森[6]が利用した，ジャンル分類済みの 338 件からなるウェブページセットを用意した．338 件のウェブページは NTCIR-3 WEB からランダムに抽出されたものである．

予備的ウェブページセットにおいて，各ジャンルに分布するウェブページの件数，その件数が 338 件中に占める比率（%），および 99%信頼区間[e]を**表 2** の左側部分に示した．ウェブページ件数の少ないジャンルはあるが，ウェブページのジャンル分布に関しては，予備的ウェブページセットはNTCIR-3 WEBの縮図であると言ってよい．

ジャンル「その他」に分類されているウェブページが 87 件（約 25.7%）と多い．大森のジャンル体系においてジャンル「その他」以外のジャンルが不足しているのではないかという懸念があり，内訳を調べてみた．その結果，画像の痕跡のみからなるページが 21 件，意味のないリンク集が 18 件，Not-Found ページが 11 件，ウェブフォームが 11 件と，計 61 件はそれ自体では情報収集に役立たないものであった．英語ページが 20 件あったが，先に述べたようにこれらは本ジャンル体系の対象になっていないので，ジャンル「その他」に分類されるのは当然である．これら以

e) NTCIR-3 WEBを母集団とした場合の母比率の 99%信頼区間である．

表 2　予備的／実験用ウェブページセットのジャンル分布
Table 2 Genre Distribution in Preliminary/Experimental Web-Pages Sets

	予備的 (NTCIR-3 WEBからのランダムサンプル)			実験用	
ジャンル	件数	比率(%)	99%信頼区間	件数	比率(%)
予想	1	0.30	0.00 – 1.06	1	1.00
標語	1	0.30	0.00 – 1.06	1	1.00
批評	1	0.30	0.00 – 1.06	1	1.00
解説	2	0.59	0.00 – 1.67	2	2.00
報道	5	1.48	0.00 – 3.17	3	3.00
独白・感想	25	7.40	3.73 – 11.06	10	10.00
情宣	4	1.18	0.00 – 2.70	4	4.00
質問・相談・依頼	15	4.44	1.55 – 7.32	2	2.00
記録	37	10.95	6.57 – 15.32	9	9.00
商品広告・宣伝	31	9.17	5.13 – 13.22	5	5.00
マニュアル	18	5.33	2.18 – 8.47	7	7.00
用語説明	15	4.44	1.55 – 7.32	9	9.00
案内・紹介	96	28.40	22.08 – 34.72	22	22.00
その他	87	25.74	19.61 – 31.87	24	24.00
総計	338	–	–	100	–

外に，Q&A の質問だけあるいは回答だけが掲載されておりジャンル「質問・相談・依頼（回答つき限定）」に属する条件を満たさないページが 4 件，記載内容の主旨を理解しにくく分類困難なページが 2 件あった．

以上のことから，ジャンル「その他」に属するウェブページ件数が比較的多いのは，当該ジャンル体系におけるジャンルの不足に起因するものではないと言える．むしろ，重要性の低いジャンル「その他」に分類するのが適当であるようなウェブページが相当程度にあるということに起因していると言える．

(d)　**実験用ウェブページセット**

予備的ウェブページセットを基にして，**表 2** の右側部分に示すように 100 件のウェブページからなる実験用ウェブページセットを準備した．その際，予備的ウェブページセットにおける各ジャンルに対する件数比率の 99%信頼区間をある程度考慮しつつも，件数の少ないジャンルについては出来るだけ全てのウェブページを実験用として採用するように努めた．さらに，個々のウェブページの選別にあたっては，筆者からみてジャンルへの分類が容易なウェブページは極力避け，分類に迷いやすいであろうと思われるウェブページを取り上げるように努めた．これは，ウェブページの分類しやすさがジャンル決定率（多数決率と過半数決率）に有利に働くのを出来るだけ抑えるための工夫である．

結果として，ジャンル「解説」，「情宣」，「商品広告・宣伝」，「用語説明」，「案内・紹介」については，ウェブページ件数の構成比率が 99%信頼区間に収まらなかった．しかしながら，どのジャンルについても，ウェブページ件数の構成比率は 99%信頼区間の上限あるいは下限を大きく超えるものではない．したがって，実験用ウェブページセットは，NTCIR-3 WEB におけるウェブページのジャンル分布を相当程度に反映しており，ジャンル分布という点では NTCIR-3 WEB の近似的な縮図になっていると考えることができる．

実験用ウェブページセットを何件のウェブページによって構成するかについては，被験者にかける負担を考慮する必要がある．被験者に過度の負担をかけると，精神的な疲れからジャンル分類作業が投げやりになり分類結果の信頼性が低下すると予想できる．こうした事態に陥らないようにするために，筆者のジャンル分類経験とジャンル分類実験を 3 回行なうことを考慮して，実験 1 回あたりのウェブページ件数は 100 件が適当であると判断した．

(e)　**ジャンル説明と教示**

ミーティングを開き，大森のジャンル体系を記載した資料を学生被験者たちに配布したうえで，各ジャンルの説明文を読み上げた．各ジャンルの説明文については，人によって解釈が異なることは当然であるので各自で解釈するよう指示した．複数のジャンルに関連する情報を含むウェブページが存在することは分かっており，学生被験者たちがそのようなウェブページのジャンル分類に迷うであろうと予想できた．そこで彼らには，複数のジャンルに関連する情報が混在していて判断に迷った場合には，ウェブページの主要部に着目して，「その他」のジャンルも含めて，どちらかと言えばどのジャンルであるかを選択するよう指示した．これは，自動的ジャンル分類技術のこれまでの研究では 1 つのウェブページを 1 つのジャンルに分類しており，従来のやり方を踏襲したものである．さらに，他の被験者と相談することなくそれぞれの被験者が独立してジャンル分類作業を行うこととした．なお，筆者も被験者であるので，学生被験者たちへの教示は筆者への教示でもある．

(f)　**ジャンル分類**

前述したジャンル説明と教示を行なった日から 1 週間以内に，被験者は大森のジャンル体系に従って実験用ウェブページセットにおける各ウェブページをジャンルに分類した．

4.2 ジャンル決定率の分析

実験用ウェブページセットにおける 100 件のウェブページを 5 人の被験者が分類した結果に基づいて，多数決ページと過半数決ページを計数し，多数決率と過半数決率を求めた．3 回のジャンル分類実験から得られた多数決率と過半数決率を**表 3** に示す．同時に，**表 3** の右側部分には平均を示した．

表3 多数決率と過半数決率（%）
Table 3 Majority Rates and More-than-Half Rates (%)

	1回目	2回目	3回目	平均
多数決率	88	84	80	84
過半数決率	75	79	68	74

多数決率と過半数決率は，5 人の被験者が実験用ウェブページセットを分類した結果におけるジャンルの一致に基づいて計算されたものである．この場合，分類結果におけるジャンルの一致には偶然による一致もあり得ることから，得られた多数決率と過半数決率に偶然を超えた意味があるかどうかについて検討が必要である．カッパ統計量 K[9]は，分類結果におけるジャンルの一致について，偶然によらない一致度を表す統計量として用いることができる[f]．分類結果におけるジャンルの一致がすべて偶然によるものであれば，K=0 である．したがって，K の値が 0 より統計的有意に大きければ，得られた多数決率と過半数決率には偶然を超えた意味があると判断できる．

表 4 カッパ統計量と有意性検定
Table 4 Kappa Statistics and Significance Test

	1 回目	2 回目	3 回目
カッパ統計量 K	0.330	0.429	0.415
分散 Var(K)の推定値	0.000294	0.000158	0.000135
z 値	19.258	34.143	35.729
p 値 ＝Pr(z≧z 値)	p 値<0.001	p 値<0.001	p 値<0.001

表5 ジャンル「その他」の割合（%）
Table 5 Ratios of the Genre "Others" (%)

	1 回目	2 回目	3 回目	平均
多数決率	26.1	16.7	13.8	18.9
過半数決率	28.0	17.7	13.2	19.7

そこで，1 回目，2 回目，3 回目の分類結果について，カッパ統計量 K を検定統計量として，有意確率（p 値）を求めた．その結果を**表 4** に示す．表中の変数 z は以下の式により得られ，近似的に標準正規分布に従う．

$$z = K/\sqrt{Var(K)} \quad (3)$$

表中の z 値は，カッパ統計量 K に対応して，上式によって求めた値である．ただし，分散 Var(K)には表中の推定値を用いた．ここでの p 値は，変数 z が z 値以上の確率（つまり，Pr(z≧z 値)）である．z 値から分かるように，p 値はほとんど 0 に近い値である．したがって，分類結果におけるジャンルの一致は統計学的に偶然だけによるものではなく，**表 3** に示す多数決率と過半数決率には偶然を超えた意味があると判断できる．

表 5 に，ジャンルが決定されたウェブページのうち，ジャンル「その他」に決定されたものの割合（%）を多数決率と過半数決率に分けて示す．大森のジャンル体系において，ジャンル「その他」は，それ以外の 13 種類のジャンルのいずれにも属さないウェブページを分類するためのジャンルであり，その重要性は低い．もし「その他」に決定されたウェブページの割合が極端に大きければ，得られた多数決率と過半数決率に偶然を超えた意味があるとしても，それらの許容性を評価することには意味がないと思われる．しかしながら，**表 5** から，平均で見れば「その他」の割合は 20%未満である．実験用ウェブページセット（**表 2** 参照）ではジャンル「その他」に分類されたウェブページが 24%あることを考慮すると，これは極端に大きな割合ではない．

以上の議論から，ジャンル分類実験から得られた多数決率と過半数決率の許容性を評価することには意味があると言える．

4.3 許容性評価

多数決率，過半数決率というジャンル決定率が何%以上であれば許容できるかについては，客観的な基準はない．また，一般にジャンル決定率の許容性評価についての経験が積まれていないため，経験則に基づく基準も得がたい．ここでは，ジャンル決定率の許容性評価基準を提案し，その上で，ジャンル分類実験から得られた多数決率と過半数決率の許容性を評価する．

まず，ジャンル決定率 60%を境にして，それ以上であれば許容でき，それ未満であれば許容できないとする．60%という基準は，教育の分野で学力評価試験の合格基準としてしばしば用いられており，これを踏襲することは不自然なことではなかろう．ここで，許容できる場合について 4 段階の評価基準を設定する．人手による分類対象となるウェブページセットが極端に分類しやすいものだけから構成されていない限り，ジャンル決定率が 90%以上になることは極めて稀と考える．そのため，ジャンル決定率が 90%以上である場合，許容性は「極めて高い」と判断する．そして，この評価表現を基準にして，60%以上 90%未満のジャンル決定率については 10%刻みで，80%以上 90%未満である場合には「非常に高い」と判断し，70%以上 80%未満である場合には「高い」と判断し，60%以上 70%未満の場合には「やや高い」

f) カッパ統計量の適切な利用に関して，以下のURLに主要な注意点が掲載されているので参考にされたい．
http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm（参照 2009-05-16）

表6 ジャンル決定率の許容性評価基準
Table 6 Evaluation Criterion of Acceptability of Genre Decision Rate

ジャンル決定率(%)	60未満	60～70	70～80	80～90	90～
許容性	許容できない	許容できる			
	低い	やや高い	高い	非常に高い	極めて高い

と判断する．一方，許容できない場合については，細かく評価基準を設定する必要性はないので，一括して「低い」と判断する．

以上をまとめると**表 6** のようになる．**表 3** に示したように，ジャンル分類実験から得られた多数決率は，その平均が 84%であることから「非常に高い」許容性を有すると評価できる．また，過半数決率は，その平均が 74%であることから「高い」許容性を有すると評価できる．いずれにしても，大森のジャンル体系は高い許容性を有するジャンル決定率をもたらすということが明らかになった．

5. おわりに

本論文では，いくつかの既設ジャンル体系を取り上げ，それぞれが再利用に適しているか否かを評価した．意味明瞭性，排他性，網羅性という性質について既設ジャンル体系を定性的に評価した結果，再利用に最も適していると思えるのは大森のジャンル体系であることが明らかになった．そして，当該ジャンル体系のジャンル決定率を実験的に求め，その許容性を評価したところ，当該ジャンル体系は高い許容性を有するジャンル決定率をもたらすことが分かった．以上のことから，当該ジャンル体系は再利用に適していると結論づけることができる．

ウェブページ分類用ジャンル体系の再利用適合性を評価した研究は見当たらず，評価手続きはこれまで明らかではなかった．本論文では評価の実施を通じて，副産物として評価手続きを与えている．今後，新たなジャンル体系を含めて同様の研究を行う際に，本論文で与えた評価手続きは参考になると考える．

謝辞 NTCIR-3 WEBは国立情報学研究所の許諾を得て使用させて頂きました．この場を借りて深謝いたします．

参考文献

1) Dewdney, N., VanEss-Dykema, C. and MacMillan, R.: The Form is the Substance: Classification of Genres in Text, Proc. of the Workshop on Human Language Technology and Knowledge Management - Volume 2001, pp.1-8 (2001).
2) Lee, Y. and Myaeng, S.H.: Text Genre Classification with Genre-Revealing and Subject-Revealing Features, Proc. of the 25th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.145-150 (2002).
3) 木村託巳，山田寛康，島津明: WWW 探索支援のための記述意図によるテキスト分類，言語処理学会第 9 回年次大会発表論文集，pp.505～508 (2003).
4) Meyer zu Eissen, S. and Stein, B.: Genre Classification of Web Pages, Proc. of the 27th German Conf. on Artificial Intelligence (KI-2004), pp.256-269 (2004).
5) Kennedy, A. and Shepherd, M.: Automatic Identification of Home Pages on the Web, Proc. of the 38th Hawaii Int. Conf. on System Sciences (HICSS'05)-Track 4-Volume 04, p.99.3 (2005).
6) 大森晃: 日本語ウェブページを主観的か非主観的かに分類する分類器のジャンル領域拡大化能力の改善：実用的な分類器へ向けて，電子情報通信学会論文誌 D，Vol.J91-D, No.4，pp.978～992 (2008).
7) Dewe, J., Karlgren, J. and Bretan, I.: Assembling a Balanced Corpus from the Internet, Proc. of the 11th Nordic Conf. on Computational Linguistics, pp.28-29 (1998).
8) Finn, A. and Kushmerick, N.: Learning to Classify Documents According to Genre, Journal of the American Society for Information Science and Technology (JASIST), Special Issue on Computational Analysis of Style, Vol.57, No.11, pp.1506-1518 (2006).
9) Siegel, S. and Castellan, Jr., N.J.: Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, Second ed. (1988).