多言語論文データベースを用いたサーベイ論文検出
- サーベイ論文自動作成の実現に向けて -

難波 英嗣† 奥村 学†
† 東京工業大学 精密工学研究所
〒 226-8503 横浜市緑区長津田 4259
E-mail: {nanba,oku}@pi.titech.ac.jp

あらまし　われわれは，サーベイ論文の自動作成を目指して研究を行っており，その第一歩として，本研究では，論文データベースからのサーベイ論文の自動検出を取り扱う．サーベイ論文は，他の論文と比べてその分野の多くの重要論文を参照するという特徴がある．この特徴を用いてサーベイ論文を検出するには，まずある分野における重要論文を特定し，次にそれらを多く参照している論文を探せば良い．このような処理を行うため，本研究では HITS アルゴリズムに着目する．学術論文において，オーソリティはある分野の重要論文に，ハブはサーベイ論文に相当すると考えられるため，論文データベースに HITS アルゴリズムを適用し，ハブ値の高い論文を選択すれば，それがサーベイ論文の検出になっていると考えられる．しかし，HITS アルゴリズムは，文書間の参照・被参照関係にのみ着目し，個々の文書の内容は考慮していないため，たまたま多くの関連論文を参照する論文もサーベイ論文として検出されてしまう可能性がある．そこで本研究では，論文の内容を考慮することで，HITS アルゴリズムによるサーベイ論文検出の精度向上を試みた．提案手法の有効性を調べるため，実験を行った．実験の結果，HITS アルゴリズムはサーベイ論文検出に有効であり，また，提案手法は HITS アルゴリズムを上回る検出精度が得られることがわかった．
キーワード: サーベイ論文検出, 多言語論文データベース，HITS アルゴリズム

Automatic Detection of Survey Articles from a Multilingual Database - Towards Automatic Generation of Survey Articles -

Hidetsugu NANBA† Manabu OKUMURA†
† Precision and Intelligence Laboratory, Tokyo Institute of Technology
4259 Nagatsuta, Yokohama, 226-8503, Japan
E-mail: {nanba,oku}@pi.titech.ac.jp

Abstract　In this paper, we propose a method to detect survey articles from a multilingual database, as a first step towards automatic generation of a survey article. Generally, a survey article refers many important papers in the domain. Using this feature of survey articles, it is possible to detect them from a database, if important papers could be detected. We pay attention to HITS, which is an algorithm to retrieve web pages using the notions of 'authority' and 'hub'. Here, it is considered that important papers and survey articles correspond to authority and hub, respectively. It is therefore possible to detect survey articles, by applying HITS to academic databases and by selecting papers whose hub scores are high. However, as HITS does not take account of the contents of each paper, authority papers that refer many important papers might be detected as survey articles. We therefore improve HITS by taking account of the contents of each paper. In order to investigate the effectiveness of our method, we conducted an examination. As a result, we found that our method could improve HITS for detection of survey articles.
key words: detection of survey article, multilingual database, HITS

1 はじめに

サーベイ論文とは，特定の研究分野に関連した情報が整理・統合された文書である．特定分野のサーベイ論文を読めば，その分野の研究動向を効率的に把握することができる．しかし，サーベイ論文作成にかかる作業負荷の高さから，論文全体に対して占めるサーベイ論文の割合は極端に少ない．だが，今後の学術情報量の増加を考えれば，サーベイ論文の需要は益々高まっていくものと思われる．そこで，われわれはサーベイ論文の自動作成を目指して研究を行っている．

本研究におけるサーベイ論文作成の方法は，まず，論文データベースから人間の作成したサーベイ論文を検出し，次にサーベイ論文に含まれていない新しい情報を追加することで，最新の情報を含んだサーベイ論文を作成する．複数の論文をまとめるには，まとめるための観点が必要となるが，その際，サーベイ論文の著者の観点を利用するのがこの方法の特徴である．

このような枠組でサーベイ論文の自動作成を行うための第一歩として，本論文では，論文データベースからのサーベイ論文の自動検出を取り扱う．

本研究では，サーベイ論文検出の際，論文間の参照・被参照関係に着目する．サーベイ論文は，他の論文と比べてその分野の多くの重要論文を参照しているという特徴がある．この特徴を用いてサーベイ論文を検出するには，まずある分野における重要論文を特定し，次にそれらを多く参照している論文を探せば良い．このような処理を行うため，本研究では HITS アルゴリズム [1] に着目する．HITS アルゴリズムとは，WWW 文書を対象にした文書検索アルゴリズムで，リンク集に相当するハブと呼ばれるページと多くのハブから参照されるオーソリティと呼ばれる 2 種類のページを考慮している．学術論文において，オーソリティはある分野の重要論文に，ハブはサーベイ論文に相当すると考えられるため，論文データベースに HITS アルゴリズムを適用し，ハブ値の高い論文を選択すれば，それがサーベイ論文の検出になっていると考えられる．

しかし，HITS アルゴリズムは，文書間の参照・被参照関係にのみ着目し，個々の文書の内容は考慮していないため，たまたま多くの関連論文を参照する論文もサーベイ論文として検出されてしまう可能性がある．そこで本研究では，論文の内容を考慮することで，HITS アルゴリズムによるサーベイ論文検出の精度向上を試みる．

また，ある分野の網羅的なサーベイ論文を作成するには，特定の言語で書かれた論文だけではなく，多くの言語で書かれた論文を対象にする必要がある．本研究では，WWW 上に存在する日本語と英語で記述された論文データを収集して多言語論文データベースを構築し，このデータベースからサーベイ論文の検出を試みる．

本論文では，まず関連研究について述べ，次にサーベイ論文の検出方法を提案する．また，多言語論文データベースを用いて，サーベイ論文検出の実験を行ったので結果を報告する．

2 関連研究

Chakrabarti ら [1] は，WWW 文書を対象に，リンク集に相当するハブと呼ばれるページと多くのハブから参照されるオーソリティと呼ばれる 2 種類のページを考慮した文書検索アルゴリズム (HITS) を提案している．その基本的な考え方は，「多くのハブから参照されるオーソリティは重要である」「多くの重要なオーソリティを参照するハブは重要である」という 2 つの仮定に基づいており，以下に示す 2 つの式を用いて各 WWW 文書のハブ値とオーソリティ値を計算する．

$$x_p = \sum_{q\ such\ that\ q \to p} y_q$$

$$y_p = \sum_{q\ such\ that\ p \to q} x_q$$

ここで，$p \to q$ は，p が q にリンクしていることを示しており，あるページ p について，オーソリティ値 x_p は，p にリンクするすべてのページ q のハブ値 y_q の総和で表される．一方，あるページ p のハブ値 y_p は，p がリンクするすべてのページ q のオーソリティ値 x_q の総和で表される．これらの計算を再帰的に行うことで，オーソリティ値とハブ値を得る．

学術論文において，ハブがサーベイ論文に，オーソリティがその他の一般論文に相当すると考えられるため，HITS アルゴリズムを論文データベースに適用し，ハブ値の高い論文を選択すれば，サーベイ論文の検出が可能になると考えられる．

一方，データベースが様々な分野 (コミュニティ) から構成されている場合，HITS アルゴリズムは，データベース中で最も大きな分野に含まれるページに関しては適切なオーソリティ値やハブ値を算出出来るが，その他の小さな分野に含まれるページは，適切な値が算出出来ないという問題点が指摘されている [2]．この問題点を改良するために，

Cohnらは HITS アルゴリズムに主成分分析を取り入れている．一般に n ページから構成されるデータベースにおいて，ページ間の参照関係は n 次元の行列で表現できる．オーソリティ値とハブ値の算出は，この n 次元行列とその転置行列の積の固有値を求めることと等価であることがわかっている．Cohn らは，この行列に主成分分析を適用し，成分毎にオーソリティ値とハブ値を算出する手法を提案している．この手法を学術論文データベースと WWW 文書に適用した結果，各主成分とデータベース中の分野にある程度の関連が認められ，また，分野毎のオーソリティページ (論文) がより適切に検出できることを示している．

3 サーベイ論文検出の方法

本研究では，論文データベースに HITS アルゴリズムを適用し，ハブ値の高いものをサーベイ論文として検出する．その際，サーベイ論文の持つ特徴を考慮することで，HITS アルゴリズムを改良する．

3.1 節では，まずサーベイ論文検出に用いる 4 種類の素性 (特徴) と，その抽出方法について説明する．次に抽出された素性を用いたサーベイ論文の検出方法について述べる．

3.1 サーベイ論文検出に用いる素性

- 論文表題 (**TITLE WORD**)

  論文表題に「サーベイ」「レビュー」「動向」「Survey」「Review」「Trend」「state-of-the-art」を含む論文の多くは，サーベイ論文であると考えられる．そこで，これらの語句を論文表題に含む論文はオーソリティ値を 1/10 (w_{auth}) に，ハブ値を 10 倍 (w_{hub}) にする．

- 参照タイプ (参照の理由)(**CITATION TYPE**)

  サーベイ論文は，既存の研究成果を用いて新しい理論を提案するというスタイルの論文ではないので，このような理由での参照は一般に少ないと考えられる．そこで，論文中の総参照数のうちで論説根拠の理由での参照が占める割合 $r\,(0 \leq r \leq 1)$ を調べ，ハブ値を r 倍 (w_{hub})，オーソリティ値を $1/r$ 倍 (w_{auth}) にする．ただし，このような理由での参照がその論文中でまったく出現しない場合にはオーソリティ値が無限大になってしまうため，オーソリティ値は最大でも 10 倍までとした．

  なお，英語論文の参照の理由の解析には難波ら [4] の方法を用いた．難波らは，論文中で参照の出現する文脈を，表層的な手がかり語を用いて解析し，以下に示す 3 種類の参照の理由 (以後，参照タイプ) を自動的に決定する方法を開発し，評価用データで 83%の分類精度を得ている．今回は，同様の方法で日本語論文の参照タイプを解析するルールを作成した．

  - **type C (問題点指摘型)**
    他の論文の理論や方法等の問題点を指摘するための参照．
  - **type B (論説根拠型)**
    既存の研究成果を用いて，新しい理論を提案したり，システムを構築する場合の参照．
  - **type O (その他型)**
    type B にも C にも当てはまらない参照．

- 参照の偏り (**DEVIATION**)

  サーベイ論文では，論文全体にわたって関連論文が参照される傾向にあるが，他の論文は，論文の前半 (例えば関連研究や提案方法の説明個所等での参照) で多くの論文を参照する傾向にあると言える．そこで，サーベイ論文の検出に，このような論文中での参照の出現傾向の違いを考慮する．まず論文中で参照の出現する文と次に出現する文との間のスパンの長さ (文数) $Span_i$ と総スパン数 n を調べ，次に以下の式を用いて参照の偏りの度合を計算する．

$$D = \sqrt{\frac{\sum_{i=1}^{n}(\overline{Span} - Span_i)^2}{n}} \times \frac{1}{text\_len}$$

  上式において，$\overline{Span}$ は，論文中の全スパンの平均長を表している．参照の偏り (D) は，スパンの標準偏差を論文の総文数 $text\_len$ で正規化することで得られる．

  D は，スパンの長さのばらつきが大きくなるにつれて大きくなり，逆に，スパンの長さがほぼ均等の場合は値がゼロに近くなる．各論文のオーソリティ値を D 倍 (w_{auth}) に，ハブ値を $1/D$ 倍 (w_{hub}) にする．

- 論文のサイズ (**SIZE**)
  サーベイ論文は，他の論文と比べ，一般的には長めであると考えられる．そこで，論文の平均長 $\overline{L}$ と個々の論文の長さ L_i を比較し，オーソリティ値を $\overline{L}/L_i$ 倍 (w_{auth})，ハブ値を $L_i/\overline{L}$ 倍 (w_{hub}) にする．

3.2 HITS アルゴリズムの改良

3.1 節で説明した 4 種類の素性を用いて HITS アルゴリズムを改良する．各論文のハブとオーソリティ値を，以下の式を用いて計算する．式中の w_{auth_j} と w_{hub_j} は，4 種類の素性のオーソリティとハブへの重みを表している．以下の式を用いてハブ値とオーソリティ値を再帰的に計算し，最終的にハブ値の高いものをサーベイ論文として検出する．

$$x_p = \prod_{j=1}^{4} w_{auth_j} \times \sum_{q\ such\ that\ q \to p} y_q$$

$$y_p = \prod_{j=1}^{4} w_{hub_j} \times \sum_{q\ such\ that\ p \to q} x_q$$

なお，実際の HITS アルゴリズムの計算では，各論文のオーソリティ値とハブ値の計算は，1 サイクル終わる度に全論文のオーソリティ値およびハブ値の 2 乗和の平方根の大きさで割って，オーソリティ値とハブ値が正規化される．本研究でも同様に 1 サイクル毎に正規化を行う．

4 実験

提案方法の有効性を調べるため，実験を行った．本節では，まず実験に用いる多言語論文データベースについて述べる．次に実験および評価方法とその結果について報告する．

4.1 多言語論文データベースの構築

近年，数多くの電子化された論文が WWW 上から入手可能になってきている．本研究では，実験に用いる多言語論文データベースを以下の手順で構築した．

(1) 論文データの収集:
WWW 検索エンジン[1] [2] を用いて，キーワードの 6 種類の組み合せ ((“業績” or “研究” or “publications”) and (“postscript” or “pdf”)) で検索する．検索結果の URL を 2 階層までたどり，Postscript と PDF ファイルを収集する．

(2) テキストファイルへの変換:
Postscript は prescript[3] を，PDF は pdftotext[4] をそれぞれ用いて，テキストファイルへの変換を行う．なお，prescript の日本語パッチは国立情報学研究所の片山紀生氏より提供していただいた．

(3) 論文構造の解析:
「参考文献」「References」等の文字列に着目して，個々の論文ファイルが参照している論文を抽出する．次に論文中の参照位置を，1)，(1)，[1] といった参照パターンに基づいて特定する．また，テキストファイルの先頭 5 行以内から論文の書誌情報 (論文表題や著者名など) を抽出する．

(4) 論文間の参照・被参照関係の解析:
ステップ (2) で抽出された情報から，あるテキストファイルの書誌情報と，それが参照する論文の書誌情報がそれぞれ分かる．ここで，ある論文の参考文献に含まれている論文が別の論文の参考文献にも含まれている可能性がある．そこで，ステップ (2) で抽出された個々の論文ファイルの書誌情報や参考文献の書誌情報について，表層的なパターンマッチングで重複する書誌情報を同定する．具体的には，個々の書誌情報から 6-gram を 2 文字づつスライドさせながら抽出し，これらが一対の書誌情報間である閾値回以上一致すれば，それらの書誌情報対は同一であると判断する．このような同定処理が，結果としてステップ (1) で収集された論文データ全体の参照関係の解析になっている．

(5) 参照情報の抽出:
3.1 節で説明した手がかり語に基づくルールを用い，参照個所の抽出および参照タイプの決定を行う．

以上の手順により，フルテキスト論文を約 20,000 件 (日本語:2100 件, 英語:17900 件)，書誌情報を約 296,000 件含んだ多言語論文データベースが構築

[1] http://www.google.com
[2] http://www.goo.ne.jp
[3] http://www.nzdl.org/html/prescript.html
[4] http://www.foolabs.com/xpdf/

された．論文の分野は，計算機科学や物理学が中心であり，他に化学・天文学・材料科学・電気工学等の分野の論文も含んでいる．なお，論文データベースに関する詳細な情報は [5] を参照されたい．

4.2 実験方法

3 節で説明した方法を，4.1 節で述べた多言語論文データベースに適用し，サーベイ論文検出実験を行った．論文毎にハブ値とオーソリティ値を計算し，ハブ値の高いものから順に検出した．

実験では，3.1 節で述べた素性をすべて用いた提案方法の他に，各素性の有効性を調べるため素性を単独で HITS アルゴリズムに適用した 4 種類の方法，素性を全く考慮しない純粋な HITS アルゴリズムを用いた場合の計 6 通りでサーベイ論文の検出を行った．

4.3 評価方法

正解セットの作成

検出システムを評価するには，実験に用いる 20,000 論文の中から人手でサーベイ論文を探しておく必要がある．しかし，20,000 論文すべてに目を通して評価用データを作成するのは非常に困難である．そこで，本研究では評価用データの作成にプーリング法 [3] を利用する．プーリング法とは，検索課題毎に，複数の異なる検索結果の上位一定数の文書をプールし，それを人間の正解判定者が課題毎に正解か不正解かを判定して，正解文書のリストを作成する方法であり，大規模テストコレクション構築における正解文書候補の効率的な収集方法の一つとして知られている．本研究では，上で述べた 6 種類の方法で収集された検出結果をプーリングに利用した．ここで，検索結果毎に上位何論文をプールすべきかが問題になるが，プール全体の論文数が現実的に判定可能であろうと想定された 1000 論文程度になるように，各検出結果からプールする論文数を決定した．結果として，各検出結果から 700 論文づつプールすることになった．また，論文表題に「サーベイ」「レビュー」「動向」「Survey」「Review」「Trend」「state-of-the-art」を含む論文にはサーベイ論文が多く含まれていると考え，これらの論文もプーリングに加えた．最終的に，1125 論文をプーリングし，サーベイであるかどうかの判定作業を人手で行った．その結果，サーベイ論文と判定された論文が 112 論文あった．以後の実験では，これらの 112 論文を正解データとして用いる．

評価尺度

評価は，以下に示す式を用いて，システム毎に評価値を算出する．また，これらの値の算出には trec_eval[5] というツールを用いる．

$$\text{再現率 } (Recall) = \frac{\left(\begin{array}{l}\text{検索システムにより検索され}\\ \text{た論文の中で正解の論文数}\end{array}\right)}{\text{正解の論文総数}}$$

$$\text{精度 } (Precision) = \frac{\left(\begin{array}{l}\text{検索システムにより検索された}\\ \text{論文の中で正解の論文数}\end{array}\right)}{\left(\begin{array}{l}\text{検索システムにより検索された}\\ \text{論文総数}\end{array}\right)}$$

4.4 結果

提案方法による検出結果を図 1 に示す．図において，すべての素性を考慮した提案方法は ‘ALL’ で，素性を考慮しない純粋な HITS アルゴリズムによる結果は ‘HITS’ で表してある．また，ベースラインとして，ランダムに抽出した場合の結果を ‘RANDOM’ で示す．

図より，‘HITS’ は ‘RANDOM’ を上回っていることから，HITS アルゴリズムがサーベイ論文検出に有効であることが分かる．また，‘HITS’ 以外の検出方法のうち，単一の素性を用いた 4 種類の HITS の中では，‘WORD’ と ‘SIZE’ は ‘HITS’ よりも全体的に高い検出精度が得られている．また，すべての素性を組み合わせた ‘ALL’ は，Recall 値が 0〜0.2 の区間で若干 ‘WORD’ より小さいものの，概ね一番良い検出精度が得られていると判断できる．しかし，すべての手法において Precision 値が著しく低い．

4.5 考察

すべての手法において Precision 値が低くなっている原因を調査するため，6 システムの出力した結果を調べた．その結果，全体的に核物理学の論文が上位にランクされる傾向にあることが分かった．特に，核物理学の分野では相対的にそれほど

[5] ftp://ftp.cs.cornell.edu/pub/smart

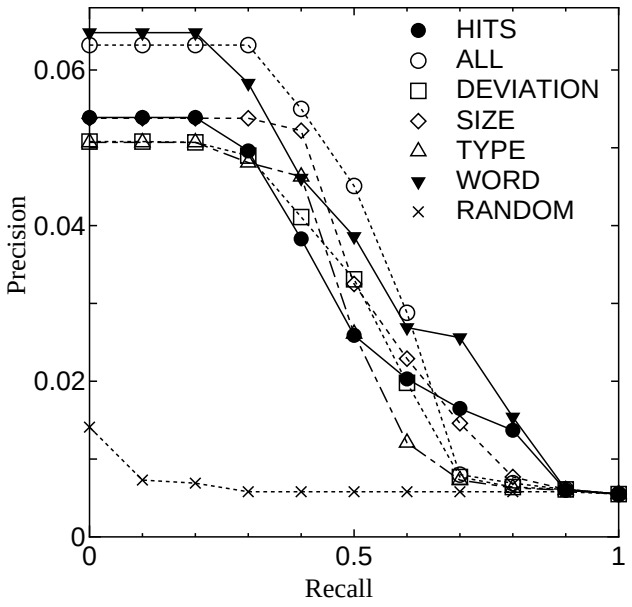

図 1: サーベイ論文検出精度の比較

ハブ値が高くない論文でも，他の分野と比較するとハブ値が大きくなっていた．このような傾向は，Cohn らの指摘する HITS アルゴリズムの問題点，すなわち「対象となる論文集合が複数の研究分野から構成されている場合，参照･被参照関係が最も密な分野のハブとオーソリティ値が，疎な分野よりも相対的に非常に大きくなる傾向にある.」と一致する．実際，文献 [6] によれば，一論文あたりの被参照数は，物理学が 7.3 回，計算機科学では 1.8 回と，非常に開きがあり，このような研究分野毎の被参照数の違いが結果に反映されてしまったと思われる．この点を改善する一つの方法は，Cohn らが行っているように，何らかの方法で論文集合を分野毎に分類しておき，分野毎にオーソリティ値とハブ値を算出することである．もう一つの方法は，論文データベースからあらゆるサーベイ論文を検出するという実験の設定を見直すことである．実際にユーザがデータベースからサーベイ論文を探す場合を考えると，ユーザがデータベース中のすべてのサーベイ論文を探すといった状況は想定しにくく，むしろ，ユーザがキーワードを入力して論文検索し，その結果の中にサーベイ論文があれば検出するといった実験設定の方が，より現実的であり自然である．キーワード検索の結果はある程度内容の類似した論文が集められていると考えられるため，検索結果からサーベイ論文を検出するというタスクは，Cohn らの手法における，主成分毎にハブ値とオーソリティ値を算出することに相当すると考えられる．結果として，このような設定で実験を行うことで，Cohn らが指摘する HITS アルゴリズムの問題が回避できると考えられる．

5 おわりに

本研究では，サーベイ論文の自動検出を行うために，サーベイ論文の持ついくつかの特徴を考慮した HITS アルゴリズムの計算方法を提案し，多言語論文データベースを用いて実験を行った．その結果，HITS アルゴリズムはサーベイ論文の検出に有用であり，さらに提案方法は，HITS アルゴリズムを改善できることが分かった．しかし，物理学のように被参照数の多い研究分野の論文は，他の分野の論文よりもハブ値やオーソリティ値が相対的に高くなる傾向にあり，これが物理学以外の分野におけるサーベイ論文検出精度の低下の要因になっている．

この点を改善するには，2 つの方法が考えられる．一つはあらかじめ論文集合を分野毎に分けて提案方法を適用するという方法である．もう一つは，論文データベースからすべてのサーベイ論文を検出するという実験設定を見直すことである．前者に関しては，例えば Cohn らの手法が解決方法の一つとして挙げられる．後者に関しては，まずキーワード検索を行い，次に検索結果の中からサーベイ論文を検出するといった，ユーザの検索手順に基づいた実験設定が考えられる．キーワード検索の結果は，ある程度内容の類似した論文が集められていると考えられるため，このような設定で実験を行うことで，Cohn らが指摘する HITS アルゴリズムの問題が回避できると考えられる．

今後は，ここで述べた HITS アルゴリズムの問題点の改良方法と，より適切なシステムの評価方法について検討を行う．

なお，本研究で実験に用いた多言語論文データベースは http://presri.pi.titech.ac.jp:8000 から利用可能である．

謝辞

本研究の一部は科学研究費補助金 (特別研究員奨励費) の援助を受けて行われたものである．

参考文献

[1] Chakrabarti, S., Dom, B.E., Gibson, D., Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A.S., "Mining the Link Structure of the World Wide Web," *IEEE Computer*, Vol.32, No.8, pp.60–67, 1999.

[2] Cohn, D. and Chang, H., "Learning to Probabilistically Identify Authoritative Documents," *Seventeenth International Conference on Machine Learning*, http://www.andrew.cmu.edu/~huan/, 2000.

[3] Gilbert, G. and Sparck Jones, K., "Statistical Bases of Relevance Assessment for the 'ideal' Information Retrieval Test Collection,", BL R&D Report 5481, 1979.

[4] 難波英嗣，奥村学，"論文間の参照情報を考慮したサーベイ論文作成支援システムの開発," 自然言語処理，Vol. 6，No. 5，pp.43–62，1999.

[5] 難波英嗣，奥村学，"WWW 上の多言語論文データを用いたサーベイ支援システムの開発," 情報処理学会 第 64 回全国大会，http://oku-gw.pi.titech.ac.jp/~nanba/study/ipsj2002.ps.gz, 2002.

[6] 根岸正光，山崎茂明，"研究評価 - 研究者・研究機関・大学におけるガイドライン -," 丸善株式会社，2001.