著者検索で得られた大量の論文から同名異人著者を除去する方法

小野寺夏生 onodera@slis.tsukuba.ac.jp †、
岩澤まり子†、辻慶太†、緑川信之†、芳鐘冬樹†、天野晃‡、
大谷裕¶、城山泰彦§、児玉閲†、角田裕之†、山崎静香†
†筑波大学　‡理化学研究所　¶東邦大学　§順天堂大学

抄録　著者名検索で得られる大量の論文を、対象の著者の論文（真論文）とそうでない論文(偽論文)に半自動的に識別する方法を検討した。ソース論文とのアドレス類似度、雑誌間引用関係、発表年差、著者所属国、及び共著関係から検索論文の真偽を予測する手順を導いた。2,595の著者名に対して検索された62.9万件の論文にこの方法を適用した結果、主題分野や所属が広範にわたる大量の文献から、約90%の正解率で真論文を識別できることを確認した。

１．背景と目的

著者名による文献検索や個人研究者レベルでの論文生産分析を行う際に、同姓同名の異なる著者の存在が大きな問題になる[1)]。とりわけ、日本、中国、韓国には非常によく現れる姓が多いため、姓とイニシアルでの著者検索では大量の別人著者論文が混入する[2)]。欧米の著者の場合は問題は軽減されるが、それでも無視し得ないノイズが混在する[3)]。

検索された論文が、意図している著者の論文（真論文）か、同名の別人による論文（偽論文）かを判別するには、検索対象研究者の発表論文リストと照合[4, 5)]する方法、著者所属機関や論文の主題から判断する方法等があるが、大量の論文に対して人手で行うことは非現実的である。論文の研究テーマと助成機関の情報に著者の共著関係の情報を組み合わせたアルゴリズム的方法により、99%の再現率、97%の適合率を達成したという報告[6)]があるが、これは特定の1機関の助成を受けた狭い専門領域を対象としたものである。

本発表では、著者検索で得られた大量の論文（2,600著者に対する63万論文）を真論文と偽論文に半自動的に識別する方法を検討し、その有効性を実証することを試みる。90%以上の正解率を得ることを目標とした。

２．著者判別対象の論文集合

表1に示す6つの分野に属する24の雑誌に、2000年に発表された1,395の論文をソース論文とした。

Web of Science（WoS)により、表1の著者抽出法が「全」の6誌は全著者について、他の18誌は第一著者について検索を行った。検索の著者名は、WoSでソース論文において表記されている通りの綴り（ほとんどの場合名はイニシアルのみ）を用いた。

2,595のソース著者名に対して、1970-2006の期間で約160万の検索論文が得られた。これは1ソース著者あたり600論文以上で、明らかに大量の同名異人著者論文が混入している。今回はこのうち、ソース論文発表年の2000年以前に発表された約62.9万検索論文を調査対象とした。

表１　ソース論文の抽出誌

分野	著者抽出法	雑誌数	ソース論文数	ソース著者数
物性物理学	筆頭のみ	3	175	175
（物性）	全	1	55	182
無機・核化学	筆頭のみ	3	173	173
（無機）	全	1	54	249
電気・電子工学	筆頭のみ	3	170	170
（電気）	全	1	59	209
生化学・分子生物学	筆頭のみ	3	180	180
（生化）	全	1	60	296
生理学(生理)	筆頭のみ	3	178	178
	全	1	58	222
消化器病医学	筆頭のみ	3	174	174
（消化器）	全	1	59	387
合　　計		24	1395	2595

３．真論文判別の方法

3.1 判別に用いる情報項目

(1) 共著者

検索論文とソース論文の間に、検索に用いた著者以外に名の一致する著者がいれば、真論文の可能性が高いと言える。このような検索論文を、以下では複数著者一致論文という。

(2) 著者所属機関アドレス

ソース論文と検索論文のアドレス表記が類似しているほど、真論文である確率は高い。ここでは、全検索論文から取り出した著者所属機関アドレスデータに含まれる単語に出現頻度により重みを付け、この重みを考慮して、

各検索論文とそのソース論文の間のアドレス類似度 Add_Sim を定めた。

(3) ソース論文誌と検索論文誌の引用関係

検索論文の雑誌とソース論文の雑誌の間に引用関係が皆無か僅少であれば、偽論文である可能性が高い。Journal Citation Reports Science Edition の 2004 年版を用いて、2000-2004 の 5 年間に、ソース論文雑誌が検索論文雑誌を引用した回数と検索論文雑誌がソース論文雑誌を引用した回数を数え、その平均の対数を、雑誌間引用強度 log X とした。

(4) 発行年の差

アドレスが異なっていたり引用関係が希薄であっても、検索論文の発表時期がソース論文よりかなり古ければ,著者が移籍したり専門分野を変えたりしている可能性がある。

(5) 著者所属国

中、台、韓、日の著者に特に同名異人が多いので、著者所属国がこの 4 ヶ国のいずれかであることを判別要素の一つとする。

3.2 一次フィルタリング

62.9 万の検索論文から、①アドレス類似度 Add_Sim が 5 未満、②雑誌間引用強度 X が 0、のものを偽論文と見なして除去した。ここでは、多少真論文を逃がすことは犠牲にして、現実的に判別処理が可能な規模の候補論文集合に絞り込むことに重点を置いた。但し、複数著者一致論文は除去対象としない。

3.3 二次フィルタリング

3.3.1 ソース著者以外にソース論文と検索論文の間に一致する著者名がない場合

(1) サンプル検索論文の目視による判別

24 のソース誌ごとに各 100 検索論文を抽出し、それぞれのサンプル検索論文の真偽を各 2 名の判定者（全て本発表の著者）が判定した。2 名の判定者の判定が一致しなかった論文（全サンプル論文の 12.7%）は別の 1 名の判定者の目視により決定した。

(2) ロジスティック回帰による判別モデル

目視判定の結果に基づき、サンプル検索論文が偽論文である確率 p を予測するロジスティック重回帰モデルを分野別に導出した。説明変数は次の 4 つとした：(a)所属機関アドレス類似度 Add_Sim、(b)引用関係強度指標 log X、(c)検索論文とソース論文の間の経過期間 Age、(d)著者所属国によるダミー変数 FEA（ソース論文の著者所属国が日、中、台、韓のいずれかの場合 1、それ以外の場合 0 とする。）

従って回帰式は次のようになる。

$$\ln[p/(1-p)] = \beta_0 + \beta_1 * \text{Add\_Sim} + \beta_2 * \text{logX} + \beta_3 * \text{Age} + \beta_4 * \text{FEA} \quad \cdots [1]$$

回帰分析は、SPSS v.16.0 を用いて行った。得られた重回帰式から予測される p 値が 0.5 未満のとき真論文、0.5 以上のとき偽論文として個々のサンプル検索論文を判別した。

(3) 全検索論文の真偽判別

(2)で得られた回帰モデルを用いて、一次フィルタリングを通過した全ての論文の真偽を判別した。

3.3.2 複数著者一致論文の場合

複数著者一致論文の場合は、ソース論文によって状況が大きく異なっていた。大半のソース論文では、検索論文のほぼすべて（p が 0.99 を越えても）が真論文と見られるが、少数のソース論文（主に日中韓著者の論文）では、もっと p が低くても偽論文になることがある。そこで、$p \geqq 0.5$ の検索論文を 20 以上含むソース著者について、それらの検索論文を個別に見て真偽を判定し、偽と判定したものを除いた。$p \geqq 0.5$ の検索論文が 20 未満のソース著者では、すべてを真論文と見なした。

4．結果と考察

4.1 一次フィルタリング

調査対象の 62.9 万件中 106,163 論文が一次フィルタリングを通過した（16.9%；1 ソース著者あたり 40.9 論文）。通過論文中 39,239 論文(37%)は複数著者一致論文である。

4.2 二次フィルタリングのためのサンプル検索論文の判別

(1) 目視による判別

全てのサンプル論文のうち 69%が真論文と判定された。各分野における真論文率は、物性、無機の 85%以上から生化の 50%弱まで大きな差があったが、これは主に Add_Sim と X の分野間の違いによるもので、同じ Add_Sim と X の領域で比較すると、真論文率に大きな差は見られなかった。

(2) ロジスティック回帰モデルの検証

各分野において、400 のサンプル論文から無作為に 280 を抽出して訓練群論文とし、残り 120 を検証群とした。訓練群に対してロジスティック重回帰を行い、[1]の形の回帰式を得た。結果を**表 2** に示す。Add_Sim と FEA の 2 変数は全ての分野で有意な予測変数であ

り、Log_X も物性と無機の 2 分野を除けば有効である。Age の予測力はこれらに比べてやや弱いが、4 分野で 5%水準有意である。従って、変数の選択は行わなかった。

得られた重回帰式を訓練群と検証群に適用した判別の結果を、人間の目視による判定と比較した（**表 3**）。どの分野も、訓練群、検証群とも 90%前後の正解率が得られた。

表 2　訓練群論文に対するロジスティック重回帰分析の結果

$\ln[p/(1-p)] = \beta_0 + \beta_1 *Add\_Sim + \beta_2 *log\_X + \beta_3 *Age + \beta_4 *FEA$

分野	n	定数 (β_0)	Add_Sim (β_1)	Log_X (β_2)	Age (β_3)	FEA (β_4)
物性	280	3.041 **	−0.5207 **	−0.2300	−0.0824 *	1.691 **
無機	280	1.895	−0.4383 **	−0.0572	−0.0799 *	2.104 **
電気	280	2.200 **	−0.4168 **	−1.5169 **	0.0287	2.804 **
生化	280	6.267 **	−0.4138 **	−1.7206 **	−0.0847 *	1.777 **
生理	280	3.636 **	−0.2782 **	−1.1814 **	−0.0583 *	0.890 *
消化器	280	3.076 **	−0.3804 **	−0.7626 **	0.0028	2.896 **

表 3　ロジスティック重回帰による判別の結果

分野	目視の判定	訓練群の判別 (n=280) 真	偽	正解%	検証群の判別 (n=120) 真	偽	正解%
物性	真	239	6	97.6	100	4	96.2
	偽	17	18	51.4	9	7	43.8
	全体			91.8			89.2
無機	真	232	9	96.3	102	0	100.0
	偽	17	22	56.4	5	13	72.2
	全体			90.7			95.8
電気	真	169	12	93.4	74	3	96.1
	偽	11	88	88.9	8	35	81.4
	全体			91.8			90.8
生化	真	118	13	90.1	50	6	89.3
	偽	14	135	90.6	3	61	95.3
	全体			90.4			92.5
生理	真	176	13	93.1	70	10	87.5
	偽	23	68	74.7	9	31	77.5
	全体			87.1			84.2
消化器	真	167	8	95.4	70	4	94.6
	偽	19	86	81.9	3	43	93.5
	全体			90.4			94.2
全体	真	1101	61	94.8	466	27	94.5
	偽	101	417	80.5	37	190	83.7
	全体			90.4			91.1

(3) 全論文判別のための回帰式の決定

全サンプル論文を用いて再度[1]による回帰式を求め、これを全論文に対する真偽判別式とした。これに基づいて、予測の偽論文率(p_pred)と実測のそれ(p_obs)の関係を**図 1** に示した。両者の一致は良好である。

4.3 二次フィルタリングによる真論文判別

以上の処理による最終的な真偽判定の結果は**表 4** の通りである。一次フィルタリングを通過した 106,163 論文の 85%に当たる 90,217 論文が真論文と判定された。

ソース著者あたりの論文数の平均値は、分野により 23 ～ 58 で、最大 30 年間の活動範囲であることを考えると妥当な大きさと言える。しかし、最大値にはやや過剰と思われるものがあり、同名異人著者の論文が十分除去できなかったソース著者がなお存在することを示唆する。

５．結論

サンプル論文の検証結果から、二次フィルタリングに用いた判別式によって 90%の正解率が得られることが実証された。アドレスの類似度は最も重要な判別要素であり、雑誌間の引用関係強度も判別に有効であった。また、特定の国(日・中・韓・台)の著者に同名異人が極めて多いため、著者がそれらの国に所属するか否かを判別式に組み込むことが重要であった。検索論文とソース論文の間の経過年数は、単独では論文の真偽との間に相関はなかったが、重回帰分析では半数以上の分

野で有意な変数となり、経過年数が長いほど真論文である確率が増大した。

本稿で提案した方法は、主題分野や所属機関・所属国が広範にわたる大量の文献から、一定の精度で「真論文」を選別したいという場合には、有効であると考えられる。

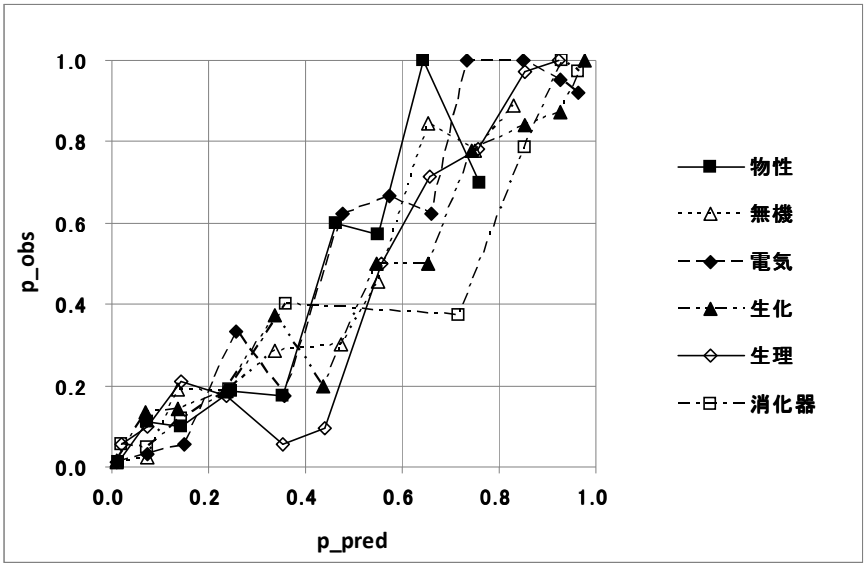

図 1　偽論文確率(p)の予測値対実測値

表 4　最終的な真偽判定結果

分野	ソース著者数	ソース著者のみ一致論文			複数著者一致論文			合計		
		一次フィルタリング通過	二次フィルタリング通過	通過率(%)	一次フィルタリング通過	二次フィルタリング通過	通過率(%)	一次フィルタリング通過	二次フィルタリング通過	通過率(%)
物性	357	8686	7621	87.7	4853	4836	99.6	13539	12457	92.0
無機	422	15206	13951	91.7	9327	9327	100.0	24533	23278	94.9
電気	379	7420	3683	49.6	1824	1619	88.8	9244	5302	57.4
生化	476	12203	7080	58.0	5583	5399	96.7	17786	12479	70.2
生理	400	5380	4308	80.1	3886	3874	99.7	9266	8182	88.3
消化器	561	18029	15114	83.8	13766	13405	97.4	31795	28519	89.7
合計	2595	66924	51757	77.3	39239	38460	98.0	106163	90217	85.0

謝辞

著者らは、本研究の実施に当たり議論に加わって貴重なコメントを寄せられた山崎茂明愛知淑徳大学教授に感謝する。

本研究は、科学研究費補助金基盤研究（B）「論文の引用を支配する要因に関する統計学的研究」の一環として行ったものである。

参考文献

1）Narin, F. Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. Computer Horizons, Inc., 1976, 456p.（p.161）

2）Moed, H. F. Citation analysis in research evaluation. Springer, 2005, 346p.（p.182）

3）Aksnes, D. W. When different persons have an identical author name. How frequent are homonyms? Journal of the American Society for Information Science and Technology. 2008, vol. 59, no. 5, p. 838-841.

4）Rinia, E. J.; van Leeuwen, Th. N.; van Vuren, H. G.; van Raan, A. F. J. Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. Research Policy. 1998, vol. 27, p. 95-107.

5）Moed, H. F. Citation analysis in research evaluation. Springer, 2005, 346p.（p.76-77）

6）Wooding, S.; Wilcox-Jay, K.; Lewison, G.; Grant, J. Co-author inclusion: A novel recursive algorithmic method for dealing with homonyms in bibliometric analysis. Scientometrics. 2006, vol. 66, no. 1, p.11-21.