

コミュニティマイニングにおける Web 引用解析と文献引用解析の比較

野村早恵子^{†,††} 三木 武[†] 石田 亨^{†,††}

Comparative Study of Web Citation Analysis and Bibliographical Citation
Analysis in Community Mining

Saeko NOMURA^{†,††}, Takeru MIKI[†], and Toru ISHIDA^{†,††}

あらまし 研究者の Web サイトを互いに結ぶ Web 上のリンク構造がどのように学術コミュニティを表現しているかを発見するため、本研究では Web 引用解析を行った。Web 引用解析は、文献引用解析を Web に適用したもので、Web サイト間の共引用関係及び直接引用関係を調べるものである。Web リンクのセマンティクスは文献引用のそれよりも異種混交としているため、Web リンクがどの程度正確に学術コミュニティ内の紐帯をとらえられるかが課題である。この目的のために、本研究では計算機科学研究者によって制作された 3,000 余りの Web サイトと 8,000 本余りの論文を調査した。その中から、Web・文献の両方で被引用数の高い上位 200 人の研究者を対象として、Web 著者共引用解析及び Web 相互引用解析を行った。その結果、1) Web 著者共引用解析が、文献著者共引用解析により得られるクラスタのスーパーセットとなる強い傾向を得た。また、2) Web 相互引用関係に対する共著関係の割合は、Web 引用関係に対するそれよりも有意に大きいことが分かった。

キーワード Web コミュニティマイニング, Web リンク解析, 書誌引用解析, 知的紐帯, 社会認知的紐帯

1. ま え が き

大量の Web データから識別される Web 構造を解明する多くの研究がなされている。Broder らは、この分野において認識されている研究意義を五つのカテゴリーに分類した上で、“Web コンテンツ生成における社会学の理解”の研究成果はほとんどないとしている [1]。本研究の最終的な目的は、人々が Web サイトのリンク構造を通してどのように社会ネットワーク [13] を形成しているかを解明することにある。しかし、Web のリンクの意味合いは文献の引用に比べあいまいである [3]。これまで、書誌学の共引用解析によって、Web の中から知識ネットワークを抽出することを目的とした研究 [4], [9] や、また当解析手法を Web の知識ネットワーク抽出に適用する際の注意点を議論した研究 [10] 等は報告されている。しかし、個人の Web サイトに適用することにより、人の社会ネットワーク

を抽出しようと試みる研究成果は、我々の知る限りない。第 1 段階として、Web の引用解析が、Web 上の社会ネットワーク分析の手法として適切であるかを検討する必要がある。

本論文では、Web の引用解析を学術コミュニティに対して適用する。近年、特に計算機科学分野の研究者は、個人の Web サイトを積極的に制作し、プロフィールや研究論文を公開している [6], [8]。そのため、解析のための Web リンクを十分に取得でき、また、学術コミュニティには文献引用解析が適用できるので、その結果と比較すれば Web 引用解析の正確さを容易に評価することができると考えた。

先行研究 [12], [14] における文献共引用解析は、知的業績の類似度などを測るものであり、得られる紐帯は知的紐帯 (intellectual tie) と呼ばれる。一方、文献共著解析 [7] は、研究者の協働関係、つまり社会認知的紐帯 (sociocognitive tie) を抽出する手法である。また、最近 White ら [15] は、文献引用が共著ネットワークと同様、研究者の社会認知的紐帯を反映していることを確認している。

そこで、本論文では、文献における引用解析に対応する手法として、Web 著者共引用解析 (Web author

[†] 京都大学大学院情報学研究科社会情報学専攻, 京都市
Department of Social Informatics, Kyoto University, Yoshida
Honmachi, Sakyo-ku, Kyoto-shi, 606-8501 Japan

^{††} 科学技術振興機構 CREST デジタルシティプロジェクト, 京都市
Japan Science and Technology Corporation, CREST Digital
City Project, Kyoto-shi, 606-8501 Japan

cocitation analysis) と Web 直接引用解析 (Web intercitation analysis) を導入する．その上で以下のよ
うに，文献解析との比較を行う．

- 研究者の個人 Web サイト間の共引用を文献の共引用と比較し，Web 著者共引用解析が学術コミュニティにおける知的紐帯を抽出するかを確認する．

- 研究者の個人 Web サイト間の直接引用を文献の直接引用と比較し，Web の直接引用が社会認知的紐帯を抽出するかどうかを確認する．

2. 文献引用解析

文献引用解析 [2] は，学術コミュニティにおける研究者やその著作の価値を測る手法であり，主に二つが提案されている．共引用解析 [12]，及び直接引用解析 [15] である．前者は関連の深い学術領域やホットトピックを発見するものであり，一方後者は共著ネットワーク解析 [7] のように研究者間の社会的な関係を抽出したり，コミュニティ内のオーソリティを発見するものである．

2.1 著者共引用解析

著者共引用解析 [14] は共引用解析の [12] 一種で，解析の単位が各文献ではなく，1 人の研究者による著作文献集合とされる．ここで文献の著者は主著者，すなわち単一のまたは第 1 の著者のことを指す．文献 a が，著者 A_1 の書いた文献と著者 A_2 の書いた文献を引用しているとき，著者対 (A_1, A_2) は a によって共引用されているという． (A_1, A_2) を共引用している文献の数が，著者 A_1, A_2 間の関連の強さを測る尺度である．著者共引用解析は，以下の 3 プロセスからなる．

1. 著者の選択: n 人の著者，つまり集合 $A = A_1, A_2, \dots, A_n$ を解析の対象として選ぶ．一般的に，解析結果の信頼性を高めるため，被引用数の最小しきい値を定め，被引用数がそのしきい値以上であるような著者を選ぶ．

2. 共被引用数の算出: あらゆる著者対 $P = (A_i, A_j) \mid A_i, A_j \in A, i \neq j$ の共被引用数をデータから計算する．結果は共引用行列 C_{cc} として表される． C_{cc} は，節を著者，枝を共被引用数で重み付けされた著者対とするネットワークの隣接行列である．

$$C_{cc} = (c_{ij}); c_{ij} = c_{ji} = \text{CocitationCount}(A_i, A_j)$$

3. 主成分分析: 主成分分析を用いて，著者をクラスタに分割する． $n \times n$ の共引用行列の行をサンプル，列を変数として，バリマックス回転により主成分を計

算する．ある変数 (著者) が最も強く相関する主成分が，その著者が属するクラスタになる．

2.2 直接引用解析

一般的に，文献間の引用関係は社会認知的紐帯を解析するものではなく，学術ネットワークの中で重要な文献や著者を発見するものである．しかし White ら [15] は最近，著者間の引用は，学術コミュニティにおける社会認知的紐帯をも表現しているに違いないという研究課題に基づき，ある組織における引用に対して時系列分析を行った．まず，16 人の研究者からなる学際研究グループにおいて，面識の度合やコミュニケーション頻度を測るため，社会ネットワーク分析の手法に基づいた質問紙調査を行った．次いで，上記の社会関係の要因と文献引用数との間の相関係数を求めた．その結果，組織の歴史が深化し，研究者間の社会的認知度合が強まるにつれ，引用頻度も高まることが明らかとなった．更に，著者への引用関係が，本の編集者・貢献者関係のような，社会認知的紐帯を表現していることも確認された．

3. データ

3.1 データソース

計算機科学研究者によって制作された，文献の引用データと個人 Web サイトを収集し解析した．データは，*CiteSeer* [5] を用い，2001 年 9 月に “Computer Science Directory” に収録されていた論文の全リストを取得した．*CiteSeer* では，各論文に HPSearch [16] が提供する著者の Web サイトの URL が付随している．ロボットによってこれらのリンクをたどり，第 1 著者の Web サイトを収集した (表 1)．8,811 論文から取り出された 5,220 人の研究者のうち，3,878 人が個人 Web ページをもっていた．したがって，この 3,878 人を解析の対象とした．

表 1 データの各種統計
Table 1 Data statistics.

データソース	Computer Science Directory, <i>CiteSeer</i>
研究領域	計算機科学・工学
データ取得日	2001 年 11 月
収集した論文数	8,811
研究者数 (論文の単一/第 1 著者)	5,220
Web サイトをもつ研究者数	3,878
総 Web ページ数	273,404
研究者間の Web リンク数	6,263
研究者間の文献引用数	27,840
研究者間の共著関係数	4,852

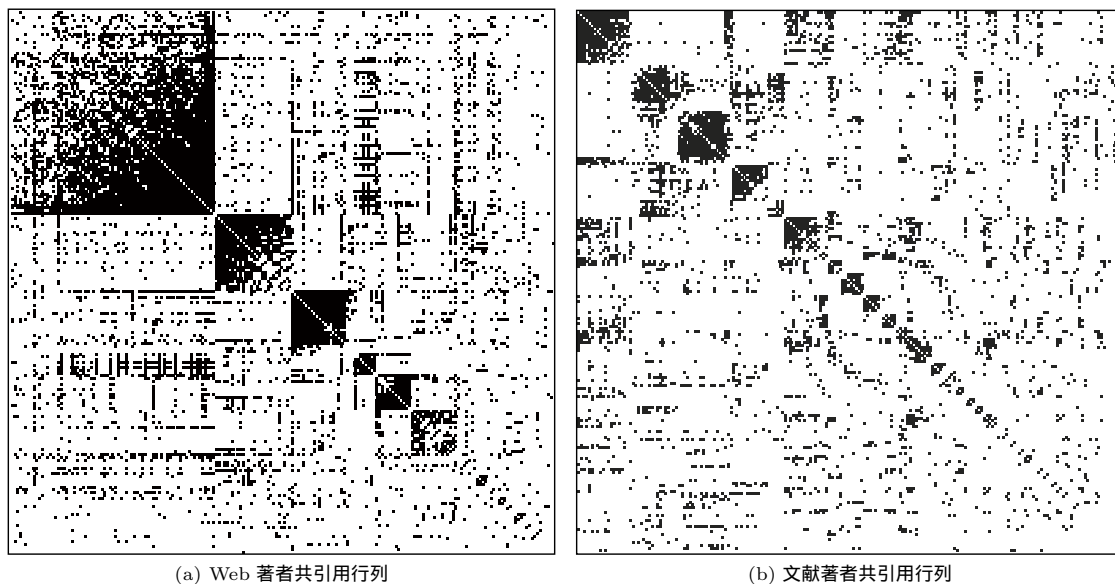


図 1 クラスタと因子負荷量でソートされた Web と文献における著者共引用行列 ($n = 200$)

Fig. 1 Web and bibliographical author cocitation matrix ($n = 200$).

3.2 データの単位

Web 引用解析をする際には、Web サイトを一つの単位として取り扱った。具体的には、Web サイトを共通の接頭辞をもつ URL の集合と定義した。例えば、以下の URL は、<http://webscience.edu/~smith/>を接頭辞としてもっており、一つの Web サイトの一部とみなした。

- <http://webscience.edu/~smith/cv.html>
- <http://webscience.edu/~smith/course/2002Nov.html>

1 人の研究者が複数の Web サイトを所有している場合は、それらを単一 Web サイトとみなした。サイト A からサイト B へのリンク数は、それらの間に複数のリンクが存在していても、0 か 1 に正規化した。

3.3 データの修正と著者の同定

CiteSeer では、論文のタイトルや著者の名前は自動テキスト解析によって抽出されており、研究者の Web サイトの URL はサーチエンジン結果から推測されたものである [5]。このため、*CiteSeer* には誤ったデータが含まれている。そこで、“Computer Science Directory” に分類されていた著者名と URL を、2001 年 11 月にすべて人手でチェック修正を加えた。

著者の名前は、先行研究 [7] にならい、“T. White” などのように「名の頭文字 + 姓」に正規化して区別し

た。名の頭文字と姓は同じだが、名やミドルネームが異なる人については、人手で区別した。

4. Web 著者共引用解析による知的紐帯の発見

4.1 Web 著者共引用解析

White らの手法 [14] に基づいた著者共引用解析を Web に適用した。White らは、より顕著なコミュニティの構造を把握するため、著作論文が非常に頻繁に引用されている研究者を解析対象としている。本研究では、文献著者共引用解析の結果と Web 著者共引用解析の結果を比較するために、文献の被引用数と Web の被引用数の積を計算し、そのスコアの上位 200 人の研究者 (200 人の Web サイト間のリンク数: 794, 文献引用数: 1,365) を選択した。

次に、Web と文献の被引用数を解析し、研究者をクラスタに分類した。生成されたクラスタについては、まず Web や文献のクラスタが何の研究分野を表しているかを調べ、更に Web 著者共引用クラスタと文献著者共引用クラスタの対応関係を調べた。

4.2 著者共引用クラスタが表現するコミュニティ

図 1 は、200 人の研究者をクラスタごとに集めた Web と文献の著者共引用行列である。著者 A_i は第 i 行と第 i 列で表される。著者らはまず属するクラスタ

表 2 Web・文献著者共引用解析におけるクラスタが表す研究分野
Table 2 Fields represented by clusters on Web and bibliographical author cocitation analysis.

(a) Web 著者共引用クラスタ

(b) 文献著者共引用クラスタ

Cluster Number	Categories in CiteSeer Computer Science Directory														No. of Researchers					
	Agents	Applications	Architecture	Artificial Intelligence	Compression	Databases	Hardware	HumanComputerInteraction	InformationRetrieval	MachineLearning	Networking	OperatingSystems	Programming	Security		SoftwareEngineering	Theory	WorldWideWeb		
1-	28	0	4	69	1	9	4	12	39	63	8	5	8	7	16	31	39	75		
2+	7	0	29	4	0	7	25	0	4	4	14	11	89	18	25	54	7	28		
3+	9	0	0	0	0	9	30	0	4	0	26	17	30	70	48	52	39	23		
4-	63	0	0	75	0	25	38	25	13	13	13	13	13	50	50	50	50	8		
5-	8	0	0	23	8	62	15	0	38	0	8	23	23	8	8	31	46	13		
6+	6	0	41	0	0	12	47	0	18	0	59	47	29	29	24	12	0	17		
8+	0	0	0	0	50	0	0	0	0	100	0	0	0	0	0	0	0	2		
8-	0	0	0	100	0	50	0	0	0	0	50	0	50	50	50	0	0	2		
10+	0	0	0	25	0	25	0	25	0	0	25	50	50	50	25	4	23	4		
13-	0	0	0	50	50	0	0	0	50	50	0	0	0	0	0	0	50	2		
15-	0	0	0	33	0	0	0	0	67	0	0	0	100	0	0	0	0	3		
other	4	0	30	13	0	13	39	13	4	13	22	30	39	13	30	17	4	23		
																			Total	200

+/-: Sign of factor loadings. Number represents %, except for *No. of Researchers*.

Cluster Number	Categories in CiteSeer Computer Science Directory														No. of Researchers					
	Agents	Applications	Architecture	Artificial Intelligence	Compression	Databases	Hardware	HumanComputerInteraction	InformationRetrieval	MachineLearning	Networking	OperatingSystems	Programming	Security		SoftwareEngineering	Theory	WorldWideWeb		
1+	10	0	5	80	0	10	5	23	80	75	0	0	15	5	10	25	45	20		
1-	18	0	27	0	0	18	36	0	0	27	27	100	55	27	73	27	11			
2+	17	0	0	0	0	17	0	0	0	0	0	67	50	17	67	33	6			
2-	0	0	65	0	0	0	82	12	18	0	53	65	53	12	53	0	0	17		
3+	0	0	0	33	0	33	0	33	100	0	0	0	67	0	33	67	33	3		
4+	15	0	0	0	0	46	0	8	0	31	15	23	92	46	38	31	13			
4-	0	0	17	0	17	17	0	0	0	17	83	0	0	33	0	0	6			
5+	69	0	0	85	0	31	8	31	62	23	15	8	8	0	23	15	85	13		
5-	50	0	0	50	0	0	0	0	100	0	0	50	0	0	0	50	2			
7+	0	0	20	0	60	20	40	0	0	80	20	40	40	5	40	40	5			
7-	0	0	14	0	0	14	14	0	100	0	0	29	14	14	0	7				
9+	33	0	17	33	0	17	0	0	100	0	0	0	17	67	33	6				
10-	40	0	0	60	0	0	0	20	100	0	0	0	40	40	40	5				
11-	0	0	0	25	50	0	0	75	75	0	0	0	0	25	50	4				
12+	67	0	0	100	0	0	0	0	67	0	0	0	67	33	33	3				
13+	17	0	0	50	0	67	0	0	0	17	0	0	17	100	0	6				
14+	0	0	0	0	0	20	0	20	20	0	0	0	80	60	20	5				
15+	0	0	0	0	0	0	0	0	50	0	50	100	50	0	50	2				
other	14	0	11	39	6	14	12	3	11	23	12	16	21	14	20	30	17	66		
																			Total	200

順に、次に因子負荷量の高い順にソートされている。色の付いたセルは、Web・文献上で共引用されている著者対である。この共引用行列では、同じクラスタに属する研究者グループが対角線上に並んでいるが、それは彼らが互いに密な共引用関係で結ばれていることを表している。Web 著者共引用解析で生成された行列では六つの明確なクラスタが現れ、一方文献著者共引用行列ではより多くの小さなクラスタが現れた。

抽出された Web・文献のクラスタの特徴を、CiteSeer の Computer Science Directory での論文分類に基づいて調べた。表 2 には、2 人以上の著者を含むクラスタが並んでいる。主成分に対する因子負荷量が正である著者は、負である著者とは別のクラスを構成する。表では、クラスタ番号の右の “+” / “-” 記号が因子負荷量の符号を表している。また、各セルの値は右端列の “No. of Researchers” を除き、各クラスタに属する研究者のうち、列が示すカテゴリーに論文が収録されている者の割合 (%) である。百分率を累積しても 100%にならないのは、ある著者が書いた文献が複数のディレクトリにリストされていることがあるためである。

表 2(a) の “1-” クラスタでは、69%の著者が “Artificial Intelligence” ディレクトリに分類され、63%の著者が “Machine Learning” に分類されている。“2+” では 89%が “Programming” に、“3+” では 70%が

“Security” に、それぞれ分類されている。三つのクラスタが特定の分野を 100%表すほかは、“10+”のように、50%程度の割合で複数の関連分野にまたがって分布しているクラスタが多い。しかし、Web 著者共引用解析は、特定の研究領域を抽出することに成功している。

文献著者共引用解析 (表 2(b)) では、“1+” では 80%の研究者が “Artificial Intelligence” に、80%が “Information Retrieval” に、そして 75%が “Machine Learning” に分類されている。“1-” では、すべてのメンバが “Programming” に論文が収録されている。また、特定の研究分野を 100%表すクラスタが半数あり、その他も約 70%以上という高い割合で特定の分野を表している。Web と比べより細かなクラスタが抽出でき、かつ各クラスタの精度が高いことから、文献著者共引用クラスタは Web に比べてよりめいりょうに、どの研究者が同じ研究コミュニティに属するかを示している。

文献に関しては、共引用は文献の内容に即して厳密に選ばれた引用によって生じるため、抽出される研究分野は、より小さな、特化された専門領域になる。一方 Web のリンクがもつセマンティクスは、文献の引用ほど厳密ではない。実際我々のデータでは、「AI 研究者へのリンク集」のようなコンテンツによって生じた Web 著者共引用が存在していた。そのため Web 著

者共引用解析は、より一般的な研究分野、ないし「トピック」を生成するといえる。

4.3 Web-文献著者共引用クラスタの関係

Web 著者共引用から抽出されたクラスタと文献著者共引用から抽出されたクラスタの関係を明らかにするため、特定の文献共引用クラスタに属する研究者が Web 共引用クラスタのどこに属しているかを調べた(表 3)。表のセル内の値は、ある文献クラスタに所属する研究者のうち、特定の Web クラスタに所属する者の割合である。

ほとんどの文献著者共引用クラスタで、そのクラスタに属する研究者はある特定の Web 著者共引用クラスタに属している。例えば、文献クラスタ“1+”に含まれる研究者の 95%が、Web クラスタ“1-”に含まれている。また、六つの文献クラスタで、100%のメンバが一つの Web クラスタに属している。

次に、Web クラスタの 60%を占める大クラスタ“1-”、“2+”、“3+”の特徴を表 2 と対比させて検討する。Web クラスタ“1-”は、“Artificial Intelligence” “Information Retrieval” “Machine Learning” を説明する文献クラスタにより構成されており、Web クラスタ“2+”は、“Security” 及び “Programming” “Theory” を説明する文献クラスタが集まっている。また、“3+”は、“Security” 及び “Software Engineering” にかかわる文献クラスタにより構成される。また、それ以下のクラスタでも、文献における関連分野がまとめられていることが分かる。つまり、

各 Web クラスタは、文献クラスタの特徴を用いることで、大きな枠組みの研究コミュニティとして説明することができる。このことから、表 3 は、各文献著者共引用クラスタが、ある Web 著者共引用クラスタのサブセットになる強い傾向を確認している。

5. Web 直接引用解析による社会認知的紐帯の発見

5.1 Web 直接引用解析

引用は、第一義的に研究内容によって生じるものである。しかし協働者間の関係においては、彼らは互いに密に引用しあっており、引用は社会認知的紐帯を含むと推測される。本研究では、大量のリンク情報から社会認知的紐帯を抽出するため、相互引用を解析の対象とした。Web 相互引用とは研究者の Web サイトが互いにリンクしていることを意味し、文献相互引用とは論文などが互いに引用し合っていることを意味する。加えて、相互引用ネットワークという言葉をも、節が研究者を、枝が研究者間の相互引用を表すような無向ネットワークの意味で用いることとする。本章では、文献と Web における相互引用ネットワークを分析し、それらを比較する。

5.2 Web 及び文献の相互引用ネットワーク間の関係

2 人の研究者が互いの Web サイトをリンクし合っているならば、互いに社会認知的紐帯の関係をもっている可能性がより高くなると仮定した。この仮定を検討するために、用いる社会認知的紐帯のデータとして、文献解析において抽出しやすく、また協働関係として明確な共著関係に注目した。そして、比較するデータとして、被引用頻度の高い上位 200 人の研究者を節とする 3 種のグラフ(それぞれ枝は Web 相互引用、文献相互引用、そして文献共著)を用いた。これらのグラフの隣接行列は 200 × 200 の対象行列である。Web の相互引用行列がそもそも 2 値(0 または 1)をとるので、比較のためにすべての行列を 2 値にした。すなわち、行列の (i, j) 成分は、著者 A_i と A_j の間に紐帯が存在するか否かを表す。

まず、行列の全要素をサンプルとして、行列間の相関係数を測ったところ、文献相互引用行列と文献共著行列の相関は 0.44、Web 相互引用行列と文献共著行列の相関は 0.27 となった。この結果は、強くはないが相関が存在していることを示している。

次に文献に関して、相互に引用している研究者が、

表 3 Web・文献著者共引用におけるクラスタ間の関係
Table 3 Cluster relationships between Web and bibliographical author cocitation.

		Web Cluster													Total	
		1-	2+	3+	4-	5-	6+	8+	8-	10+	13-	15-	other			
Bibl. Cluster	1+	95	0	0	0	0	0	0	0	0	0	0	0	0	5	20
	1-	0	64	18	0	0	0	0	0	0	0	0	0	0	18	11
	2+	0	50	17	0	0	0	0	0	0	33	0	0	0	0	6
	2-	12	24	0	0	0	35	0	0	0	0	0	0	0	29	17
	3+	100	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	4+	0	0	100	0	0	0	0	0	0	0	0	0	0	0	13
	4-	0	100	0	0	0	0	0	0	0	0	0	0	0	0	6
	5+	62	0	0	23	15	0	0	0	0	0	0	0	0	0	13
	5-	50	0	0	0	0	0	0	0	0	0	0	0	0	50	2
	7+	0	0	0	0	60	20	0	0	0	0	0	0	0	20	5
	7-	0	0	14	0	0	57	14	0	0	0	0	0	0	14	7
	9+	100	0	0	0	0	0	0	0	0	0	0	0	0	0	6
	10-	100	0	0	0	0	0	0	0	0	0	0	0	0	0	5
	11-	75	0	0	0	0	0	0	0	0	0	25	0	0	0	4
	12+	100	0	0	0	0	0	0	0	0	0	0	0	0	0	3
	13+	0	0	0	33	33	0	0	17	0	0	0	0	0	17	6
14+	20	0	40	20	0	0	0	0	0	0	0	0	0	20	5	
15+	50	0	50	0	0	0	0	0	0	0	0	0	0	0	2	
other	35	12	5	3	9	9	2	2	3	2	5	15	66	66		
Total	75	28	23	8	13	17	2	2	4	2	3	23	200			

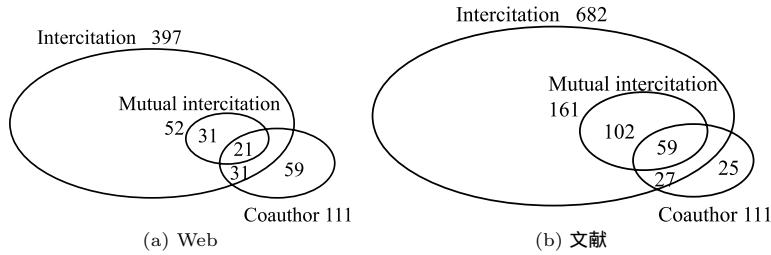


図 2 Web・文献における引用, 相互引用, 共著関係数を表すベン図
 Fig. 2 Venn diagrams with the number of intercitation, mutual intercitation and coauthorship.

単なる引用よりも深い社会認知的関係をもっているかどうかを調査した。図 2 (b) は、200 人の研究者からなる 19,900 通りの研究者対について、引用関係, 相互引用関係, 共著関係にある対の数をベン図にしたものである。図では引用している研究者対のうち共著関係にあるものは 12.6%だが、これを相互引用ネットワークに限定すると、共著関係の比率は 36.6%に増している。つまりより多くの紐帯が、社会認知的関係の情報をもっているといえる ($Z = 9.68, p < .001$)。この結果は White ら [15] の、引用が社会認知的紐帯を含むという結果を確認するものである。

更に、Web においても文献同様の関係が現れるかを調べた。文献の引用が第 1 に知識的要因に基づいているのに比べ、Web のリンクは異種混交で、その性質を知るのが難しいのは確かである。しかし我々は、少なくとも研究者同士では、Web の相互引用は文献同様に、一般の引用に比べてより社会的な関係を表していると仮説を立てた。Web に関して、引用, 相互引用, 文献共著関係にある研究者対を数えた結果を図 2 (a) に示す。引用している研究者対のうち共著者の割合は 13.1%であったが、相互引用している対では 40.4%であった ($Z = 5.63, p < .001$)。Web 相互引用は、文献と同様、社会認知的紐帯を含んでいることが確認できた。

上記の議論を補うため、Web と文献の相互引用, 及び共著関係をベン図にしたものが図 3 である。Web と文献の両方で相互引用している研究者の 92%が共著関係にあった。引用の相互性が、リンクを社会認知的なものへとフィルタリングする効果が、再度確認された。

我々は、更に Web や文献において相互引用をしながら、共著関係になかったそれぞれ約 6 割の研究者同士の関係性を手作業で調査した。本研究は、CiteSeer のディレクトリに収集された文献のみを解析対象とし

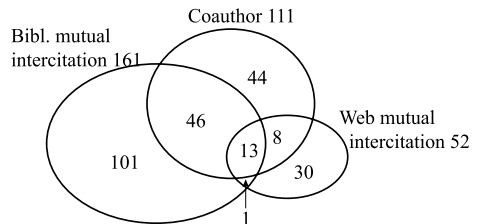


図 3 Web・文献相互引用及び共著関係のベン図
 Fig. 3 Venn diagram of Web and bibliographical mutual intercitation and coauthorship.

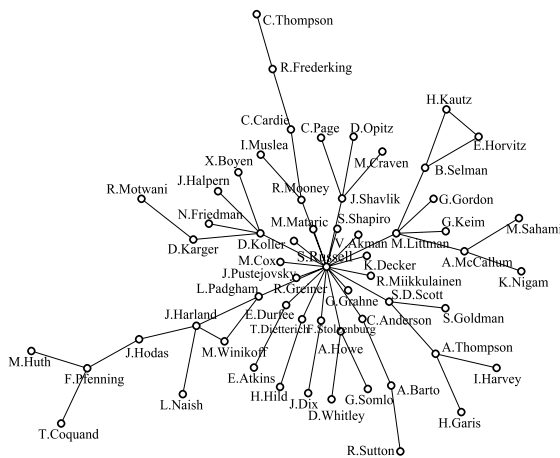
ていたため、その文献集合で共著者関係として抽出できなかった著者対であっても、分析対象外の文献で共著関係にある場合はある。ここでは、そうしたケースを除いた結果を示す。

文献では、102 対のすべてにおいて、社会認知的紐帯であるかは特定できないものの、特定の専門研究に関する知識的要因に基づいた引用関係が確認された。

一方、Web では、所属研究分野の研究者リンク集において相互的にリンクが張られているケース (16.6%)、著名な教科書の執筆研究者のサイトが他の研究者の授業シラバスサイトからリンクされており、かつその有名な研究者が自分の属する研究分野の研究者リンク集をもっているケース (37.5%) と、社会認知的紐帯とは特定できないものが約 6 割を占めた。残りの約 4 割は、友人関係 (16.7%)、研究会や論文誌特集号の編集委員共同主催者 (25.0%)、同じプロジェクトに所属している研究者同士 (4.2%) と、明らかに社会認知的紐帯であった。このことから、Web の相互引用が、共著関係に限らず様々な社会認知紐帯抽出へのフィルタリング効果をもつことが更に確認できた。

5.3 相互引用ネットワークが表現するコミュニティ

最後に、研究者 Web サイト間の相互引用からどのようなコミュニティが抽出されたのかを明らかにする



Web Intercitation Component (#authors: 531)

Component 1 (#authors: 59)

32%	S. Russell	2%	A. McCallum
9%	M. Littman	2%	A. Howe
7%	L. Padgham	2%	C. Cardie
6%	D. Koller	2%	C. Anderson
6%	J. Harland	2%	B. Selman
5%	R. Mooney	1%	T. Dietterich
5%	Stephen. D. Scott	1%	R. Frederking
4%	J. Shavlik	1%	F. Stolzenburg
3%	J. Hodas	1%	E. Durfee
2%	F. Pfenning	1%	D. Karger
2%	A. Thompson	1%	A. Barto

著者名の左の%は、媒介中心性を表す $((n-1)(n-2)/2)$ 。

図 4 Web 相互引用ネットワークの第 1 連結成分における媒介中心性 ($n = 59$)

Fig. 4 Between of Web intercitation network: Connected component 1 ($n = 59$).

ため、Web 及び文献の相互引用ネットワークを構成し、共著ネットワークと比較した。

まず、ネットワークを連結成分に分割した。結果、共著ネットワークでは 2,199 の節 (全体の 67%) が、文献相互引用ネットワークでは 1,027 節 (68%) が、最大の連結成分として抽出された。この傾向は、大規模な共著ネットワークの解析に関する先行研究に一致する。一方 Web 相互引用ネットワークでは、いくつかの中規模な連結成分が得られ、最大連結成分の大きさも 59 節 (10%) であった。

次に、Web 相互引用によってコミュニティを相互接続している研究者を発見しようと試みた。共著解析研究では、わずか数人のかぎとなる人物が、多数の最短共著経路が通過する要の位置を占めコミュニティを接続する役割を果たし、またネットワークの中でただ 1 人が 60% 程度の最短経路を握っているとされる [7]。この媒介中心性 (betweenness) は、ネットワークの中で誰が大きな影響力をもち、情報の流れをコントロールしているかを示す。

Web 相互引用ネットワークにおける各研究者の媒介中心性とその上限に対する比率を計算した。図 4 は、最大連結成分として抽出された “Artificial Intelligence” コミュニティにおける、上位 20 名の媒介中心性の値を示したものである。連結成分は、全節対間最短経路を描画する PFNETs [11] によって可視化している。Web 上の社会認知的リンクが、S. Russell 氏を中心に

広がっている様子を示している。このような研究者は、Web コンテンツが制作される過程において、研究分野の中心として多くのリンクを集めつつ、自らも積極的に他の重要人物にリンクすることによって、Web 上の学術コミュニティを形成する中心的役割を果たしているといえる。

6. むすび

本研究では、CiteSeer の Computer Science Directory から、3,878 人の研究者の個人 Web サイトと彼らの 8,811 本の論文を収集し、Web と文献の両方で被引用頻度の高い上位 200 人の研究者を対象に解析を行った。文献データには著者共引用解析、共著ネットワーク解析、引用解析を適用し、Web データには著者共引用解析、引用解析を適用した。そして、Web と文献の解析結果を互いに比較した。

解析の結果、以下の 2 点の知見を得た。

- Web 著者共引用解析が抽出するクラスタは、文献著者共引用解析のように、学術コミュニティにおける知的紐帯を反映するものである。Web のクラスタは、文献著者共引用クラスタのスーパーセットとなる強い傾向を示す。各 Web クラスタは、学術コミュニティにおけるより広い研究分野を表すものである。

- Web・文献の相互引用関係と、文献共著関係との関係を測った結果、相互引用ネットワーク中の共著関係の割合は、引用ネットワーク中の割合に比べて十

分に大きい。また、共著関係でなくとも論文誌の編集委員同士などが抽出できることから、Web の相互引用は研究者間の社会認知的紐帯抽出としてのフィルタリング効果が高いといえる。

本研究は、Web 引用解析が社会ネットワークを抽出する上で効果的であることを示している。Web 引用解析は、研究者だけでなく、膨大な Web データに基づいて様々な集団の社会ネットワークを分析できる可能性を秘めている。一方文献引用解析に比べると、Web のリンクがもつセマンティクスは多様であるため、精細な解析を行うのは難しい。Web 引用解析の有用性を増すためには、Web リンクのセマンティクスを限定する必要がある。我々は現在、本研究の次のステップとして、Web ページの内容解析を行い、リンクのセマンティクスを限定する枠組みに基づいた Web リンクオントロジーを生成し、より詳細な Web 引用解析の可能性を追求している。そのアプリケーションとしては、様々な社会ネットワークをリンクオントロジーから自動検索できる検索エンジンを考えている。

謝辞 本研究は科学技術振興事業団戦略的基礎研究推進事業「デジタルシティのユニバーサルデザイン」プロジェクトから援助を受けて行われた。有益なコメントを頂いた京都大学の小山聡氏、*CiteSeer* のデータ訂正に助力頂いた研究室メンバに深く感謝します。

文 献

- [1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," Proc. 9th WWW Conference, pp.309–320, 2000.
- [2] E. Garfield, Citation Indexing: Its theory and application in Science, Wiley, New York, 1979.
- [3] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, vol.46, no.5, pp.604–632, 1999.
- [4] R. Larson, "Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace," Proc. 59th Annual Meeting of the Am. Soc. Inf. Sci., vol.33, pp.71–78, 1996.
- [5] S. Lawrence, C.L. Giles, and K. Bollacker, "Digital libraries and autonomous citation indexing," Computer, vol.32, no.6, pp.67–71, 1999.
- [6] S. Lawrence, "Online invisible?" Nature, vol.411, no.6837, p.521, 2001.
- [7] M.E.J. Newman, "Who is the best connected scientist?: A study of scientific coauthorship networks," Phys. Rev. E, vol.64, no.016131, 2001.
- [8] 野村早恵子, 石田 亨, 正木幸子, 横澤 誠, 篠原 健, "インターネットにおけるアイデンティティの国際比較," 信学論 (D-I), vol.J84-D-I, no.2, pp.222–231, Feb. 2001.
- [9] J. Pitkow and P. Pirolli, "Life, death, and lawfulness on the electronic frontier," Proc. CHI 97, pp.22–27, 1997.
- [10] C. Prime, E. Bassecoulard, and M. Zitt, "Co-citations and co-sitations: A cautionary view on an analogy," Scientometircs, vol.54, no.2, pp.291–308, 2002.
- [11] R.W. Schvaneveldt, F.T. Durso, and W.D. Donald, "Network structures in proximity data," The Psychology of Learning and Motivation, vol.24, pp.249–284, 1989.
- [12] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," J. Am. Soc. Inf. Sci., vol.24, no.4, pp.265–269, 1973.
- [13] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, UK, 1994.
- [14] H.D. White and B.C. Griffith, "Author cocitation: A literature measure of intellectual structure," J. Am. Soc. Inf. Sci., vol.32, no.3, pp.163–172, 1981.
- [15] H.D. White, B. Wellman, and N. Nazar, "Does citation reflect social structure?: longitudinal evidence from an interdisciplinary research group," J. Am. Soc. Inf. Sci. Tech., vol.55, no.2, pp.111–126, 2004.
- [16] <http://hpsearch.uni-trier.de/>
(平成 15 年 4 月 9 日受付, 8 月 15 日再受付)



野村早恵子 (正員)

平 12 京大大学院情報学研究科社会情報学専攻修士課程, 平 15 同大学院同研究科博士後期課程研究者指導認定退学。博士 (情報学)。現在, UCSD Dept. of Cognitive Science 客員研究員。CMC 解析に興味をもつ。



三木 武

平 14 京大・工・情報卒, 現在, 同大学院情報学研究科社会情報学専攻修士課程在学中。Web コミュニティマイニング, セマンティック Web に興味をもつ。



石田 亨 (正員)

昭 51 京大・工・情報卒, 昭 53 同大学院修士課程了。同年日本電信電話公社電気通信研究所入所。現在, 京都大学大学院情報学研究科教授。工博。人工知能, コミュニケーション, 社会情報システムに興味をもつ。