協調フィルタリングを用いた論文の共著関係の予測

曽原 寿允† 堀 幸雄‡ 今井 慈郎†‡
†香川大学工学部 ‡香川大学総合情報基盤センター

1. はじめに

昨今の学術論文の出版数は年を追うごとに増加傾向にある．さらに，学問は専門家・細分化が行われ，一人の専門家が学術全体を俯瞰することを困難としている[1]．また，論文における共著ネットワークは単著から共著，単一機関共著から国際共著，機関間共著へと変化しているとの報告がある[2]．そのため，所属機関内外の複数の専門家が集まりそれぞれの得意とする分野を分担し，共同研究を行ったほうがより効率的であると考えられる．

そこで，本研究では，過去の論文情報から著者間の共著関係を抽出・分析し，共著者の可能性を予測する．予測性能を上げるため，大規模データに対応した協調フィルタリングを用いており，先行研究の予測手法との精度の比較を行う．

2. 関連研究

先行研究として Nowell の共著予測の研究[3]がある．Nowell はリンク予測の考え方を共著ネットワークに適応しており，物理学の論文から共著ネットワークを抽出し，ある時点のネットワークから次の時点の共著ネットワークの予測を行った．本研究では，従来の予測モデルで使われている主な予測指標に加えて，協調フィルタリングを用いた予測モデルを使用する．共著リンクの予測には，ノード(著者)，関係(共著)に関する情報の 2 つを用いる．以下に先行研究の各モデルの概要を示す．$v^{(i)}, v^{(j)}$ はそれぞれ研究者ノード，$\Gamma(v^{(i)})$ は $v^{(i)}$ の共著ノード集合を表す．

・共通隣接ノード(common neighbors) 指標

$$\mathrm{common}(v^{(i)}, v^{(j)}) = \left|\Gamma(v^{(i)})\right| \cap \left|\Gamma(v^{(j)})\right|$$

共通隣接ノードは「共著者を共有している研究者が次に研究する可能性が高い」というものである．

・Jaccard 係数

$$\mathrm{Jaccard}(v^{(i)}, v^{(j)}) = \frac{\left|\Gamma(v^{(i)})\right| \cap \left|\Gamma(v^{(j)})\right|}{\left|\Gamma(v^{(i)})\right| \cup \left|\Gamma(v^{(j)})\right|}$$

Jaccard は「共著者の大半を共有する研究者は共著する可能性が高い」というものである．

・Adamic/Adar

$$\mathrm{Adamic/Adar}(v^{(i)}, v^{(j)}) = \sum_{k \in \left|\Gamma(v^{(i)})\right| \cap \left|\Gamma(v^{(j)})\right|} \frac{1}{\log\left|\Gamma(v^{(k)})\right|}$$

Adamic/Adar は「あまり共著していない研究者を共通に持つ研究者と共著する可能性が高い」というものである．

・Katz 指標

$$\mathrm{Katz}(v^{(i)}, v^{(j)}) = \sum_{t=1}^{\infty} \beta^t \left|\mathrm{paths}^{\langle t \rangle}_{v^{(i)}, v^{(j)}}\right|$$

Katz は共通隣接ノードのパス長による一般化を行ったものである．

・優先的選択(preferential attachment) 指標

$$\mathrm{preferential}(v^{(i)}, v^{(j)}) = \left|\Gamma(v^{(i)})\right| \cdot \left|\Gamma(v^{(j)})\right|$$

優先的選択手法は「共著者が多い研究者ほど新たに共著する可能性が高い」というものである．

3. 提案予測方法の概要

3.1 対象データ

本研究で扱う共著ネットワークのデータは表 1 のように表される．

表 1：表による共著関係

	A	B	C	D
A		3	1	
B	3		1	
C	1	1		1
D			1	

英字は著者，数字は共著回数を表している．この表の形式で表されるデータに対して協調フィルタリングを適応し推薦を行う．

3.2 協調フィルタリング

協調フィルタリングとは，多くのユーザの嗜好情報を蓄積し，あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推測を行う方法論である．

取り扱うデータが大規模なデータ集合の場合，データを表としてとらえるとスパースな表となる．協調フィルタリングでは他のユーザのデータを用いて推薦を行うが，その演算の際，他のユーザの値や他のユーザとの類似度が必要となる．しかし，表がスパースな場合，通常の協調フィルタリングでは大規模なデータを扱うことが困難となる．

そこで，本研究では協調フィルタリングを行う際にコンテンツベースの考えをとりいれた．本来の協

Prediction of relation of paper co-authorship based on collaborative filtering
Toshimitsu Sohara†, Yukio Hori‡, Yoshiro Imai†‡
†Kagawa University
‡ Information Technology Center, Kagawa University

調フィルタリングではコンテンツの内容を考慮しなくても，ユーザにとって有意義な結果が出るという長所があるが，大規模なデータを取り扱う場合には，コンテンツの内容を考慮することによってスパースなデータ集合に対応できるものと考えた．

研究者間の類似度を $sim(v^{(i)},v^{(j)})$，研究者 $v^{(i)}$ における予測研究者 $v^{(j)}$ の予測値 $p(i,j)$，共著数を $r(i,j)$ とした時の協調フィルタリングの指標を以下のように表す．

$$p(i,j) = mean\ (v^{(i)}) + \frac{\sum sim\ (v^{(i)},v^{(k)}) * (r(i,k) - mean\ (v^{(k)}))}{\sum sim\ (v^{(i)},v^{(k)})}$$

結果として類似度の高い研究者と共著頻度の高い研究者が推薦されることになる．

3.3 類似度の算出

協調フィルタリングの計算式で必要になるユーザ間の類似度の算出方法について述べる．通常の協調フィルタリングでは，ノード間の類似度の計算にはコサイン類似度や Jaccard 係数が用いられる．コサイン類似度は以下のように定義される．

$$\cos(v^{(i)},v^{(j)}) = \frac{\sum_k (v^{(i)}(k)\cdot v^{(j)}(k))}{\sqrt{\sum_k v^{(i)}(k)^2}\cdot\sqrt{\sum_k v^{(j)}(k)^2}}$$

しかし，これらの指標ではデータがスパースであった場合不都合が多い．具体的には類似度の値が限りなく 0 に近づいてしまう．そこで，共著者間の類似度を共著者間の特徴語ベクトルによって求める方法をとる．これは，著者に付随する特徴語をその著者の特徴語ベクトルとし，そのベクトル値をもとに類似度の計算を行う．この方法はコサイン類似度や Jaccard 係数などのリンクベースの類似度計算ではなく，コンテンツベースの計算となるので，協調フィルタリングに用いているデータがスパースであっても類似度計算の支障を低減できると考えられる．

3.4 著者の特徴語の抽出

著者の特徴語の抽出について述べる．特徴語の元データは論文データのタイトルと抄録を用いる．以下の手順を行い著者の特徴語ベクトルを作成する．

(1) 著者の論文データからタイトルと抄録を抽出
(2) タイトルと抄録を形態素解析し単語に分解
(3) 特定ドメイン内のすべての論文に対して(1)(2)を実行
(4) 形態素解析によって得られた単語に対し，TF-IDF 値を求めて単語の特徴量を算出
(5) 著者に属する単語の TF-IDF 値の中から高得点を選択して著者の特徴語として採用

このような手順によって求められた特徴語ベクトルによりユーザ間の類似度を計算すれば，表のスパース性にとらわれることなく，コンテンツベースの類似度計算を実現することができると考えられる．

4. 評価実験

評価実験として CiNii[4]より取得した論文データを基に前述した手法で共著者の予測を行う．

CiNii で得られる論文データは様々な分野・学会のデータが含まれておりその内容は多岐にわたっている．本研究で提案する手法の類似度の優位性をはかるため，学会をドメインとし学会誌ごとに分類し共著者の予測を行う．また，他の研究者などと全く共著を行っていない著者群や，共著が低頻度の著者群のデータは，今回の実験に対してノイズとなる可能性が高いと判断したため，共著が低頻度のデータは適宜頻度によってカットして実験を行う．

予測精度の評価方法は以下の通りである．

(1)各予測モデルを用いて，ある年の論文集合データを元に予測を実施
(2)その予測したデータと，翌年の論文集合のデータで新たに発生した共著リンクを比較
(3)先行する予測手法の結果と本研究で提案する予測手法の結果を比較することで，各方法による予測精度の評価を実施

5. おわりに

本稿では，研究者の研究活動支援として，論文情報からの共著関係の抽出と予測について，その方法論を中心に述べた．また，従来の予測モデルと比較しつつ，本研究で提案する協調フィルタリングを用いた予測モデルについての概要を示した．

今後の方針として，各指標がどのような特性を持っているか評価を行いたい．また，現在は研究者間のリンクの予測にとどまっているが，推薦されたデータをもとにどのような情報をユーザに推薦すれば，研究者にとってより有意なものとなるかを考慮していきたい．

謝辞

本研究の一部は平成 20 年度香川大学若手研究(萌芽研究)経費によるものです．ここに記して謝意を表します．

【参考文献】

[1]Kajikawa. Y.: Abe. K: and S. Noda:"Filling the gap between researchers studying different materials, different methods: A proposal of structured keywords", Journal of Information Science, Vol.32, No.6, pp.511-524, 2006.

[2] 神田由美子，他: 科学技術指標，文部科学省科学技術政策研究所, 2008.7.

[3]Liben-Nowell, D., Kleinberg, J.: The Link Prediction Problem for Social Networks, in Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM), pp. 556-559,2004.

[4]CiNii(NII 論文情報ナビゲータ),http://ci.nii.ac.jp/