TrustNetwork を用いた論文推薦システムの実装

M2011MM017 平山佳典
指導教員：石崎文雄

1 はじめに

今日，多くの Web サービスが提供されており，ユーザが利用したいサービスや知りたい情報の候補が多く，対象を発見することができない情報過多の問題が発生している．この問題を解決する手法として，ユーザの嗜好に基づいて情報の取捨選択を行う協調フィルタリング [1] が情報検索や推薦システムに多く利用されている．しかし，協調フィルタリングの手法はコールドスタート問題 (Cold Start Problem)[3]，シリング攻撃 (Shilling Attack)，データスパースに弱いという問題点からユーザが的確に情報の推薦を受けるのは困難である．

また，協調フィルタリングの手法はユーザによる評価値を利用した推薦方法がほとんどであり，多くのユーザに評価値が付けられていないアイテムの推薦を行うのは困難である．さらに，本文情報検索システムはユーザが各論文に対して，評価値を付与しているシステムはなく，ユーザによる評価値から推薦を行うことができない．

そこで本研究では TrustNetwork を用いて，ユーザによる評価値が付与されていない論文データを対象に論文推薦システムの実装を行う．本手法では論文の共著関係を基に TrustNetwork を構築していく．共著したことのある著者同士は専門分野が類似しているため，共著したことのある研究者から推薦を行うことでユーザが目的とする論文と類似した検索結果を得られる．そこで論文の共著関係を信頼できるネットワークとして扱い，TrustNetwork を構築していくことで，ユーザが従来の論文検索システムを利用しただけでは検索結果として表れない論文を発見することが可能となる．

2 既存の推薦システムの手法と関連研究

2.1 協調フィルタリングを用いた推薦手法

多くの推薦システムには，協調フィルタリングが用いられており，ユーザによるアイテムの評価値を基にピアソン相関係数 (1) 式やコサイン類似度 (2) 式を利用し，アクティブユーザとの嗜好情報の類似度を計測してアイテムの推薦を行うのが一般的な手法である．

$$sim = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}\sqrt{(x_i - \overline{x})^2}\sum_{i=1}^{n}\sqrt{(y_i - \overline{y})^2}} \quad (1)$$

$$sim = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n}\sqrt{{x_i}^2}\sum_{i=1}^{n}\sqrt{{y_i}^2}} \quad (2)$$

しかし，精度の高い推薦を行うためには，ユーザの嗜好情報を正確に獲得する必要があり，同ユーザが同アイテムに対して異なる評価値を付けるなどといった行為を行うと，評価値にゆらぎが発生するため 推薦精度が低下する問題がある．

また，あるユーザが特定のアイテムだけに対して高い評価を行い，ユーザの嗜好データとは関係なしにそのアイテムを推薦させようとするシリング攻撃 [4] といった問題がある．さらに，推薦システムを初めて利用するユーザは，嗜好情報が存在しないためユーザの嗜好を把握することができないコールドスタート問題も存在する．

2.2 TrustNetwork を用いた推薦手法

Jamali ら [2] は，コールドスタート問題に対応するために，ユーザ同士の繋がりに着目し，信頼できるユーザを探索して，信頼できるユーザ群を生成し，そのユーザ群の評価値を用いて推薦精度を向上させている．協調フィルタリングの手法では，ユーザ間やアイテム間の類似度を基に推薦を行っていたが，この手法では類似度の代わりにユーザの信頼関係を求めて推薦を行っている．Jamali らは e コマースのデータを対象に研究を行い，顧客の商品の購入履歴から信頼関係を発見して TrustNetwork を生成している．本研究では信頼関係を共著関係として TrustNetwork を構築する．図 1 は本研究の論文推薦システムの TrustNetwork の図であり，TrustNetwork の深さを 3 としたときである．図 1 のように従来の検索では発見できなかった論文を TrustNetwork を用いることで発見することができる．また Jamali らの研究では，TrustNetwork の深さを 2 としたときに十分な推薦結果を得ることができ，深さを 3 としたときにノイズが増加し，推薦精度が低下した．

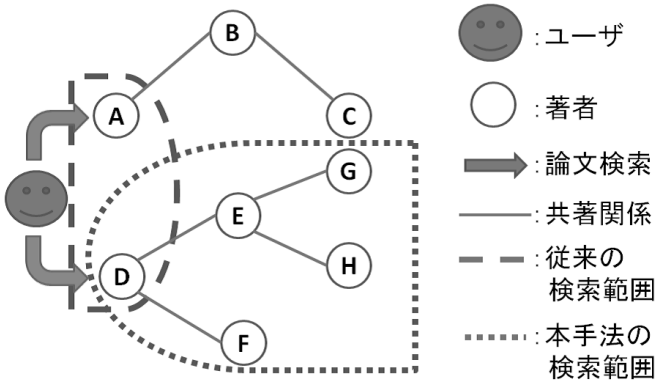

図 1 TrustNetwork の構成

2.3 従来の手法と本手法の相違点

近年では，論文検索を行うことができる Google Scholar や Yahoo!検索 (論文) といった検索エンジンはユーザが入力したキーワードを基に検索を行っている．また，関連著者や研究者，関連刊行物，関連書籍なども同時に検索を行えるが，検索キーワードに対して関連しているものしか出力していないため，ユーザが十分に満足のいく検索結果が得られていない．そこで本研究の手法では，関

連著者を信頼できるネットワークとして，TrustNetwork を生成することで，ユーザが従来のキーワード検索では発見することができなかったユーザの目的に近い論文を発見する．表 1 は，従来の論文検索システムと本手法を用いた論文推薦システムの相違点をまとめたものである．

表 1 手法の相違点

	従来の論文検索	本手法の論文推薦
検索方法	キーワード検索	テクニカルターム + TrustNetwork
検索の深さ	1	1〜2

3 提案する論文推薦手法

本研究では，論文著者の共著関係を利用し TrustNetwork を構築して論文推薦システムの実装を行う．本推薦システムの構成を図 2 に示す．

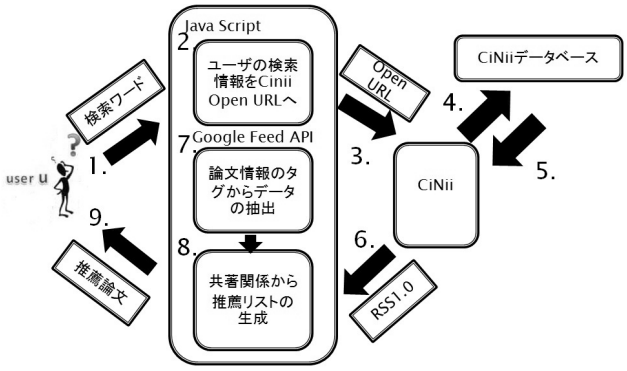

図 2 システムの構成

下記はシステムにおける実行順序である．

1. ユーザによる検索ワード入力
2. ユーザによる検索ワードを基に OpenSearch のクエリを生成する
3. 生成したクエリを基に CiNii にリクエストする
4. CiNii の論文データベースへリクエストを送信
5. CiNii の論文データベースからリクエストを基にデータを受信
6. CiNii Articles の OpenSearch を利用し論文データを RSS1.0 のフォーマットで Java Script へ送信
7. Google Feed API を利用して論文情報の必要なデータを抽出
8. 論文データを基に TrustNetwork を構築し推薦リストの生成
9. ユーザにお薦めの論文リストを送信

3.1 論文推薦システムの流れ

手順 1 でユーザに検索ワードを入力してもらう．この検索ワードとは著者とその著者と関連のあるテクニカルタームの 2 つである．手順 2 では手順 1 で得た検索ワードを基に Java Script を用いて CiNii Articles OpenSearch クエリを生成する．手順 3 と手順 4 で手順 2 で生成したクエリを CiNii Articles を通して CiNii の論文データベースへリクエストの送信を行う．次に，手順 5 で CiNii の論文データベースからその検索ワードに該当した論文データを受け取り，手順 6 で RSS1.0 のフォーマットに論文データを変換して Java Script で受け取る．手順 7 では Google Feed API を用いて手順 6 で受け取った RSS1.0 フォーマットの論文データの必要な部分のデータ抽出を行い，手順 8 で論文の共著関係から TrustNetwork の生成を行った後に，論文の推薦リストを生成する．最後に手順 9 で，ユーザに TrustNetwork を用いて推薦された論文のリストを出力する．

3.2 論文情報

手順 3 から手順 6 では論文情報を取得するために CiNii が提供している API(Application Programming Interface) の OpenURL を利用して論文データを検索と取得をしている．取得する論文データは論文の出版年の新しい順に取得し，推薦される論文も出版年が優先に出力される．表 2 は本研究で取得した論文データを RSS1.0 のフォーマットに変換した RSS フィードのタグ名と内容である．これらを Google Feed API を利用し，必要な論文データの抽出を行っている．

表 2 RSS フィードのタグ名と内容

タグ	内容
dc:creater	著者名
title	論文名
description	抄録
prism:publicationDate	出版年月日
prism:publicationName	刊行物名

4 実験

本研究の論文推薦システムの実装を行う際の，TrustNetwork における深さの影響，同姓同名者の著者による問題点，入力データであるテクニカルタームによる影響，論文の専門分野における推薦出力数の違いについて述べる．

4.1 データセット

CiNiiAPI を用いて実験を行うため，データは CiNii[5] に収録されている論文データを用いる．2011 年 10 月の段階で検索できる論文数は約 1500 万件であり，複数の論文データベースを統合しているため各データベースから論文データの取得を行うことが可能である．

4.2 TrustNetwork の深さ

共著関係の深さ 1〜2 までを TrustNetwork として，実験を行う．深さ 2 を取った理由は Jamali らの e コマースでの TrustNetwork を用いた商品推薦システムにおいて TrustNetwork の深さが 2 のときに最も良い推薦精度を得るといった結果があったためである．そこで本研究の論

文推薦でも TrustNetwork の深さを 1 から 2 の範囲で実験を行った．

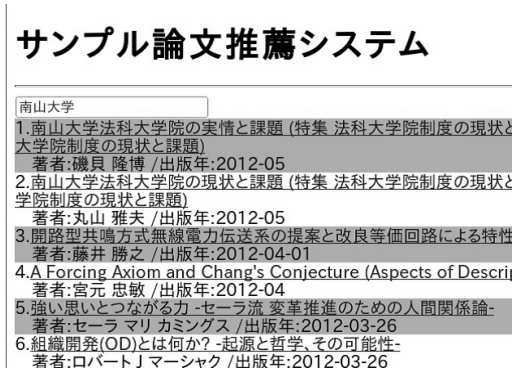

図 3 実装画面

4.3 同姓同名者の著者による推薦の問題点

まず，深さ 1 で著者名を利用して検索を行った結果，同姓同名の著者が存在する場合に上手く検索が行えないことがわかった．ある同姓同名が存在している著者を対象として，検索を行った結果は適合率が 0.54 となった．この結果は，半分は違う著者の論文の推薦がされており，上手く推薦が行われていないことがわかる．そこで，検索クエリに目的とする著者の論文中に含まれていると推測されるテクニカルタームを用いることで，適合率は 1.0 となった．この結果はその目的とした著者を特定したことを表してる．また，著者名のみとテクニカルタームを付与したときを比較した場合，出力された検索結果が 0.5 倍となった．

表 3 深さ 1 における同姓同名者の検索適合率

	適合率	検索一致数	全論文数
著者名	0.54	18	33
著者名 + テクニカルターム	1.0	9	9

本手法では同姓同名者が存在する著者の場合，論文の推薦精度が低下してしまう．そこで本研究では，著者と論文に含まれていると推測されるテクニカルタームを検索ワードに入力を行い，論文の推薦を行う．

4.4 テクニカルターム

本研究の目的は，ユーザが新しい論文を見つけることである．そこで，テクニカルタームを狭義なものや専門性の高いもので行うと十分な結果を得ることができない．そのため，テクニカルタームを上位概念と下位概念に分けて実験を行った．例えば，テクニカルタームをソフトウェアで検索を行う場合上位概念と下位概念の関係は上位概念をソフトウェア工学，下位概念を組込みソフトウェアとなる．表 4 から表 6 は情報工学に着目して検索を行った表である．それぞれ上位概念のテクニカルタームを用いた方が共著関係が多いという結果が得られた．本研究の目的である論文の推薦を行うためには，推薦を行いたい学術分野の上位概念にあたるテクニカルタームを用いることで，ユーザへの論文の推薦量が増加することがわかった．

表 4 テクニカルターム

上位概念	下位概念
ソフトウェア工学	組込みソフトウェア
ソフトウェア工学	アスペクト指向ソフトウェア
データマイニング	相関ルール
通信ネットワーク	高信頼ネットワーク
オペレーションズ・リサーチ	施設配置問題
統計学	多重比較法

表 5 上位概念の共著数

上位概念	共著数
ソフトウェア工学	10
ソフトウェア工学	14
データマイニング	6
通信ネットワーク	3
オペレーションズ・リサーチ	7
統計学	2

表 6 下位概念の共著数

下位概念	共著数
組み込みソフトウェア	4
アスペクト指向ソフトウェア	3
相関ルール	5
高信頼ネットワーク	0
施設配置問題	4
多重比較法	2

4.5 学部・学科ごとの出力結果

テクニカルタームでの実験を行った結果，文系分野と理系分野でのそれぞれの論文に対して共著関係の違いがわかった．表 7 と表 8 は文系分野と理系分野での推薦可能著者数の比較である．推薦可能著者は共著回数が少なくとも 1 回は存在している者を示す．本実験では，文系分野は 39%，理系分野は 100%の確率で推薦可能著者が存在することが確認された．この理由として，文系分野の論文には単著なものが多いことや外国語学部に多く見受けられたのが，論文の翻訳を行うといった翻訳関係であった．よって，文系分野では TrustNetwork を用いて論文の推薦を行うは理系分野と比較すると十分に機能しないことが判明した．

表 7 文系：推薦可能著者数

学部	推薦可能著者数	合計著者数
外国語学	2	9
経済学	3	6
法学	1	6
経営学	1	3
文化学	2	6
哲学	1	3
教育学	2	3
福祉学	1	3

表 8 理系：推薦可能著者数

学部	推薦可能著者数	合計著者数
工学	6	6
医学	6	6
農学	3	3
情報学	6	6
宇宙工学	3	3

5 TrustNetwork の論文推薦結果

理系分野に焦点をあて TrustNetwork を用いて論文推薦を行い，情報工学部から著者 2 名を例として，それぞれ著者 A，著者 B として実験を行った．実験に用いたテクニカルタームはソフトウェア工学である．著者 A との共著者を a_n とし，著者 B との共著者を b_n とした．表 9 と表 10 は著者 A と著者 B におけるそれぞれの深さ 1 と深さ 2 における論文が推薦される量の増加率の表である．教員 A の場合は共著者が 4 名発見され，深さ 1 では 74 本の論文が発見されたが本手法を用いることで，新たに 72 本の論文が発見でき，論文の推薦量が 49%増加した．同様に教員 B の場合でも共著者が 17 名発見され，新たに 411 本の論文が発見でき，論文の推薦量が 1470%増加した．この増加率の違いは，推薦が可能であった共著者数が大きく影響した．本実験では著者 B の方が著者 A よりも共著者数が多く，共著者数が多い方が推薦される論文数が多いことがわかる．

表 9 著者 A における推薦論文増加率

共著者	論文数
a_1	43
a_2	19
a_3	7
a_4	3
合計増加数	72
深さ 1 の論文数	74
増加率	0.49

表 10 著者 B における推薦論文増加率

共著者	論文数	共著者	論文数
b_1	2	b_{10}	1
b_2	26	b_{11}	80
b_3	4	b_{12}	38
b_4	2	b_{13}	11
b_5	11	b_{14}	6
b_6	70	b_{15}	23
b_7	21	b_{16}	7
b_8	35	b_{17}	2
b_9	30	b_{17}	42
合計増加数	411		
深さ 1 の論文数	30		
増加率	14.7		

6 まとめ

本研究では共著関係を TrustNetwork として実験を行った結果，学術分野において本手法の有効性の違いが明らかになった．学部ごとによるテクニカルタームの実験から本手法は理系分野において有効であることがわかった．そして，理系分野の著者 2 名で本手法を用いて実験を行った結果，それぞれ推薦論文量が 49%，1470%増加した．しかし文系分野においては，著者に共著関係がない単著関係が多く存在したために論文の推薦を行うことができなかった．そこで，文系分野においては論文中の単語の頻出度から特徴量の抽出を行い，特徴的な単語を信頼できるタームとして TrustNetwork を構築することで論文の推薦が可能であると考察する．

参考文献

[1] Xiaoyuan Su，Taghi M. Khoshgoftaar,“ A survey of collaborative Filtering techniques，”Journal Advances in Artificial Intelligence，Vol.09，Article No. 4，pp.1-20 (2009).

[2] Mohsen Jamali，Martin Ester,“ Using a Trust Network to Improve Top-N Recommendation，”RecSys ’09 Proceedings of the third ACM conference on Recommender Systems，pp.181-188 (2009).

[3] Andrew I. Schein，Alexandrin Popescul，Lyle H. Ungar，David M. Pennock,“ Methonds and metrics for cold-start recommendations，”SIGIR ’02 Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval，pp.253-260 (2002).

[4] Paulo Messa，Paulo Avesani,“ Trust-aware recommender systems，”RecSys’07 Proceedings of the 2007 ACM conference on Recommender systems，pp.17-24 (2007).

[5] CiNii Articles - 日本の論文を探す - 国立情報学研究所，http://ci.nii.ac.jp/ (accessed 2012.8)