2S-7

サーベイ論文の解析結果を用いた文献分類手法の提案

福井達也† 中村健二† 小柳 滋†
†立命館大学

1 はじめに

社会ニーズの多様化と学問体系の細分化に伴い，様々な研究分野の統廃合や研究分野間の協業による学際化が進んでいる．このような研究分野の大きな流れを把握するための一助として，サーベイ論文が執筆されている．このサーベイ論文により分野が整理され，研究領域の概観の把握や近年の新たな研究分野出現動向なども効率的に把握できる．しかし，サーベイ論文はその論文が執筆された時点までの情報を整理したものであり，最新の研究動向を把握するためには独自に調査する必要がある．

本研究ではサーベイ論文を解析し，論文を整理分類するための情報を取得することで，サーベイ論文の執筆時以降を含む，最新の研究動向の把握を支援するシステムを提案する．

2 関連研究

研究領域を概観するための取り組みとして，論文を整理・分類する研究が行われている．

論文を整理・分類する研究では，論文間の内容の類似度に基づきクラスタリングを行う手法 [1] や論文の引用情報を用いて算出した類似度に基づいてクラスタリングする手法 [2] があり，これらは多数の論文を類似したトピックで纏めることが可能である．しかし，クラスタリングでは，作成されたクラスタがどのような内容を示すものなのかを推定しなければならないという課題がある．一方，あらかじめ分類対象カテゴリを決定した状態で文書を分類する研究 [3][4] も提案されているが，これらの手法では，分類対象のカテゴリの決定とそのカテゴリの元となる教師データを人手で準備しなければならないという課題がある．

3 提案手法

3.1 研究の概要

本研究では, サーベイ論文から得た情報に基づき論文を分類し，研究動向を可視化する手法を提案する．本研究においてサーベイ論文を用いる理由は，サーベイ論文では研究分野について章節ごとに整理分類して記述されており，分類先のカテゴリおよび各カテゴリの学習データを同時に取得できると考えたためである．これにより既存研究における分類先カテゴリの生成についての課題や，分類されたカテゴリの内容把握についての課題を補うことができる．

3.2 提案手法の流れと実装

本節ではサーベイ論文を用いた論文分類の処理の詳細を説明する．処理の流れを図 1 に示す．

処理 1 サーベイ論文が引用している論文群を取得

本処理ではサーベイ論文が引用している論文を取得する．論文の取得はサーベイ論文中の参考文献部分から各論文のタイトルを得ることで行う．この処理で得られた論文群を $DQ = \{dq_1, dq_2 ... dq_i\}$ と定義する．

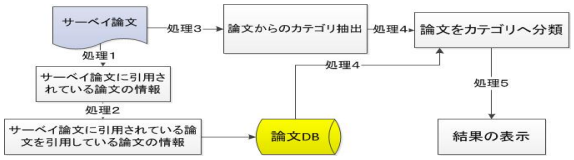

図 1: 処理の流れ

処理 2 分類対象となる論文群を取得

本処理では処理 1 で得た DQ を引用している文献を GoogleScholar[1]を用いてそれぞれ 50 件づつ取得する．この処理で得られる論文群を $DT = \{dt_1, dt_2 ... dt_j\}$ と定義する．

処理 3 分類カテゴリの生成

本処理ではサーベイ論文を章節ごとにブロック単位に分割する．次に各ブロックで論文の引用があるものをカテゴリとして抽出する．ここで, 抽出したカテゴリを $C = \{c_1, c_2 ... c_k\}$ とし，c_k はその文書中で引用されている論文 dq_i を含む集合とする．なお，本研究では正確にカテゴリを生成するために手動でこの処理を行う．

処理 4 論文の分類

本処理では論文と各カテゴリとの関連度を求め，関連度が閾値以上になったカテゴリへ論文を分類する．論文と各カテゴリとの関連度を求めるためには式 1 を用いる．

$$Relation(dt_j, c_k) = \frac{c_k \cap Quote(dt_j)}{|c_k|} * \log \frac{|C|}{C \cap CQuote(dt_j)} \quad (1)$$

式 1 にて，$Quote(dt_j)$ は論文 dt_j が引用している論文群，$CQuote(dt_j)$ は論文 dt_j が引用している論文を含むカテゴリ群を示す．

処理 5 研究動向の可視化

本処理では研究カテゴリ間の関係や，カテゴリの盛衰の明確化を目的とした可視化を行う．

4 評価実験

本実験では論文のカテゴリ分類と可視化の評価実験を行う．実験には推薦システムに関するサーベイ論文 [5] を用いる. サーベイ論文を解析して得られた情報として DQ は 112 件，DT は 4,684 件（重複なし 2,869 件），取得カテゴリ件数 $|C|$ は 13 であった．

A Paper Classification Method Based on Analysing Surveys
†Tatsuya FUKUI †Kenji NAKAMURA †Shigeru OYANAGI
†Ritsumeikan University

[1] http://scholar.google.co.jp/

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.

表 1: 各カテゴリごとの分類論文数と分類精度

カテゴリ No.	1	2	3	4	5	6	7
論文数	229	145	516	944	472	654	57
正分類数	189	90	359	789	429	474	37
分類精度 (%)	82.5	62.0	69.5	83.5	90.1	72.5	65.0
カテゴリ No.	8	9	10	11	12	13	全体
論文数	176	0	316	138	286	321	4,254
正分類数	149	0	205	92	236	251	3,298
分類精度 (%)	84.7	0	64.9	66.7	82.5	78.2	77.5

4.1 実験 1：論文のカテゴリへの分類

4.1.1 実験内容

サーベイ論文を解析して得られた情報を基に実際に各カテゴリへ論文を分類を行い，論文の分類先の正誤判定を評価する．なお，本実験では論文を分類する基準となる閾値は 0.1 とした．

評価については，サーベイ論文中の各カテゴリの文書と分類対象の論文の内容を比較し，人手で判断する．なお，正誤判定はそれぞれのカテゴリごとに行った．

4.1.2 結果と考察

表 1 にカテゴリごとの重複あり分類論文数と分類精度を示す．分類精度の最高値は 90.1%，最低値は 62,0%となった．精度の低かったカテゴリ 2 は推薦システムの基本的な考え方について述べられており，カテゴリの内容と異なる分野の情報が多く含まれていることが精度が低くなった要因であると考えられる．一方，最も精度がよかったカテゴリ 5 は Hybrid Recommendation について述べられており，この手法は多様な手法と関連があることが高精度となった理由と考えられる．以上の結果より学際的なカテゴリは精度は高精度で，専門的なカテゴリは精度が低下する傾向があることが分かる．

4.2 実験 2：論文とカテゴリの関係の可視化

4.2.1 実験内容

本実験では論文とカテゴリの関係を可視化する．論文中でのカテゴリのトピックと可視化結果を比較することで，得られた知見について考察する．

時系列ごとのカテゴリと論文の関係の可視化手法を次に示す．まず，年表はサーベイ論文執筆年の 10 年前より始める．次に，式 1 で求めた関連度を用いて，論文を執筆年次中で最も関連の強いカテゴリの位置に配置する．なお，10 年前以前に執筆された論文は 10 年前時点に配置する．そして，カテゴリはサーベイ論文執筆時を軸に前後 3 年ごとに配置し，カテゴリと関連のある論文をそれぞれ振り分けた色の直線で結ぶ．

4.2.2 結果と考察

すべての論文をカテゴリと結んだものを図 2，特徴が見られた個々のカテゴリに注目して描画した図を図 3，図 4 に示す．図 3 は Collaborative Filtering(カテゴリ 4) のトピックについて表現した図である．この図 3 を確認するとカテゴリ 4 はある特定のカテゴリと強く関連があり，2002 年度を区切りに関連がカテゴリ 2 からカテゴリ 13 へと推移しているということが読み取れる．次に，図 4 は Hybrid Recommendation(カテゴリ 5) について表現した図である．図 4 を確認するとカテゴリ 5 は特定のカテゴリと関連しているのではなく幅広いカテゴリと関連があり，比較的最近盛んになった研究であること読み取れる．

これらの結果より本研究では研究カテゴリ間の関係や，カテゴリの盛衰を明確化できていることがわかる．

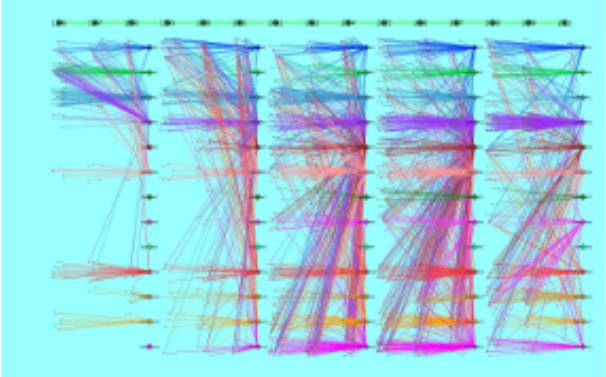

図 2: 全体の可視化

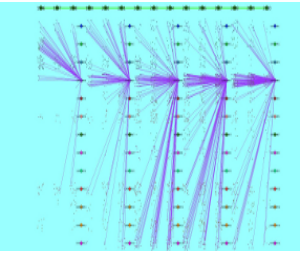

図 3: カテゴリ 4 に注目した可視化

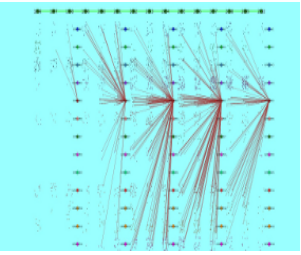

図 4: カテゴリ 5 に注目した可視化

5 おわりに

本稿ではサーベイ論文を用いた論文の分類と研究カテゴリの時系列的変化を可視化する手法を提案した．実験結果より，内容の明確なカテゴリに論文を分類する方法について新たな可能性を見出すことができた．しかし，論文のフォーマットは著者によって異なるため，論文の形式によっては本手法を適応できないという課題を残した．今後は多様な論文で実験を重ねることで本手法の汎用性を高めることが必要である．

参考文献

[1] 柏木裕恵，高田雅美，佐々木和貴：アブストラクトを用いた論文分類システムの設計と実装，数理モデル化と問題解決研究会研究報告，情報処理学会，Vol.2006, No.95，pp.33-36 (2006).

[2] 江藤正己：論文の構成単位に基づいた共引用関係の尺度，情報処理学会論文誌 (データベース)，情報処理学会，Vol.49，No.SIG7，pp.1-15 (2008).

[3] 榊剛史，松尾豊，石塚満：制約付きクラスタリングを用いた論文分類，人工知能学会全国大会論文集，人工知能学会，Vol.20，No.1A1-1，pp.1-4 (2006).

[4] Sebastiani, F.:Machine Learning in Automated Text Categorization, ACM Computing Surveys, ACM, Vol.34, No.1, pp.1-47 (2002).

[5] Adomavicius, C.:Toward the Next Generation of Recommender Systems - A Survey of the State-of-the-Art and Possible Extensions, IEEE Transactions on Knowledge and Data Engineering, IEEE, Vol.17, No.6, pp.734-749 (2005).

Copyright ©2011 Information Processing Society of Japan.
All Rights Reserved.