学術情報検索における閲覧論文の分析

小林 和央[1]　風間 一洋[1]　吉田 光男[2]　大向 一輝[3]　佐藤 翔[4]

[1] 和歌山大学 [2]豊橋技術科学大学[3] 東京大学[4] 同志社大学

はじめに

背景
- インターネットの普及により，電子化された論文が研究者以外の学生などにも利用されている
  - 既存の検索システムでは出版年や被引用数など順位付け方法が限定
  - 異なる属性のユーザに適した検索結果を提示することが困難

目的とアプローチ
- ユーザの目的に応じた論文の発見を支援する
  - ユーザの論文閲覧行動から複数の指標を計算
  - 論文のトピックや分類を示す検索語を抽出

➢ 各指標の検索結果と検索語の関係を分析する

関連研究

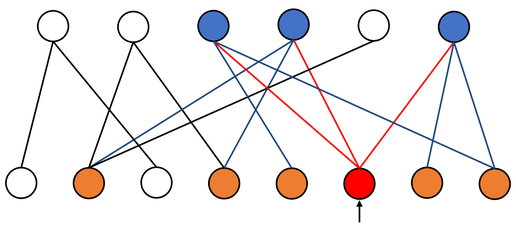
国立国会図書館サーチの検索機能の利用分析[佐藤 2015]
- 検索結果を提示するのみでは必要な情報にたどり着くことは困難

5段階の研究経験のユーザに対して4種類の推薦システムを評価[Hristakeva 2017]
- 研究者や学生によって推薦対象とする論文の範囲が異なる

➢ ユーザの属性に応じた評価方法が必要

論文の評価指標

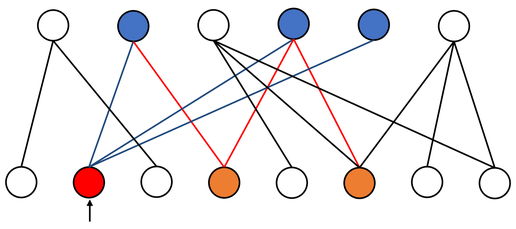
- ユーザ$u_j(j=1,\dots N)$と論文$p_i(i=1,\dots M)$の2部グラフから各指標を求める
- 閲覧人数：多くのユーザに閲覧される論文を重要とする
- 普遍度：ある文献を閲覧したユーザが他に閲覧した文献の和集合が全論文数に占める割合で算出
  - 様々な分野や属性のユーザに閲覧される論文を重要とする



ユーザ$u_j(j=1,\dots N)$　閲覧人数:3

- 異なる分野のユーザに閲覧されると和集合サイズは大きくなる

論文$p_i(i=1,\dots M)$　普遍度: $6/M=0.75$



ユーザ$u_j(j=1,\dots N)$　閲覧人数:3

- 特定分野のユーザに閲覧されると和集合は重なり，サイズは小さくなる

論文$p_i(i=1,\dots M)$　普遍度: $3/M=0.375$

ユーザと論文の2部グラフ

- 2部グラフ上で論文p_iから論文p_jへ2ホップで到達可能なら$r_{i,j}>0$となる
- 論文p_iの普遍度$S_{(p_i)}$は$r_{i,j}>0$の要素数を全論文数Mで割ることで求められる

$$S_{(p_i)}=\frac{|\{r_{i,j}|r_{i,j}>0\}|}{M}\ (\mathrm{j}=1,\dots,M)$$

論文の特性評価手法

論文の特性を評価するために，検索語を用いる

1. アクセスログのリファラより検索クエリを抽出
2. 検索クエリを形態素解析して検索語を抽出
3. 論文ごとに検索語の出現頻度を集計
4. 各検索語のTF-IDF，TF-DFを計算する

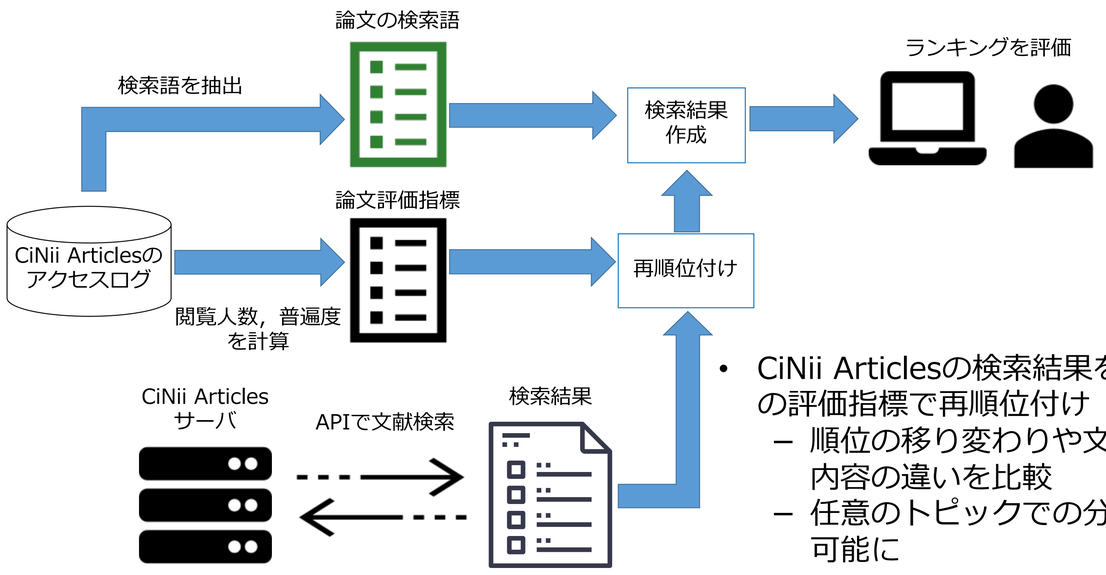
- $TF\text{-}IDF_{(i,j)}=\frac{tf(i,j)}{s(p_i)}*\log\frac{N}{df(t_j)}$　論文$p_i(i=1,\dots N)$,$s(p_i)$：p_iの全単語の出現頻度の和
  - 文書固有で多く検索される単語を抽出→内容指定語

- $TF\text{-}DF_{(i,j)}=\log(tf(i,j)+1)*df(t_j)$
  - 多くの文書で検索される単語を抽出:分類指定語

データセット

- CiNii Articlesのサーバのアクセスログからユーザの閲覧論文を集計
  - 2014年4月1日～2016年3月31日の2年間
  - 閲覧論文数9,106,860件，ユーザ数13,038,381人が対象
- CiNii収録論文の書誌情報約3,750万件分
  - 論文タイトル情報約3,200万件，概要約350万件分を分散表現モデル作成に利用

CiNii Articles再順位付けシステム



- CiNii Articlesの検索結果を論文の評価指標で再順位付け
  - 順位の移り変わりや文献の内容の違いを比較
  - 任意のトピックでの分析を可能に

検索語の比較分析

「推薦システム」での検索結果上位で，内容指定語・分類指定語上位を抽出した

人数	普遍度	題名	内容指定語	分類指定語
1	9	推薦システムのアルゴリズム(1	推薦,神嶌,システム,神嶌敏弘,敏弘,アルゴリズム,serendipity,grouplens,recommender	システム,アルゴリズム,推薦,system,based,知能,人工知能学会,敏弘,content,filtering
2	22	状況依存型ユーザ嗜好モデリングに基づくContext-Aware情報推薦...	推薦,情報,context,嗜好,recommendation,aware,協調フィルタリング,フィルタリング,context awareness,ユーザ	情報,システム,状況,依存,手法,嗜好,ing,推薦,協調,model
3	1	絵本の読み聞かせにおける子どもの好みと絵本の主題との関係性	絵本,読み聞かせ,子ども,好み,幼児,松村,主題,親子,反応,図書館	論文,教育,子ども,分析,研究,幼児,図書館,方法,関係,システム
4	36	協調フィルタリングとコンテンツ分析を利用した観光地推薦手法の検討	推薦,協調フィルタリング,観光地,観光,勇之,樽井,コンテンツ,旅行,システム,情報	論文,分析,システム,情報,観光,大学,計画,経営,特性,手法
5	153	Folksonomy マイニングに基づく Web ページ推薦システム	推薦,協調フィルタリング,folksonomy,システム,web,フィルタリング,マイニング,収集,sns,recommender	論文,システム,web,インターネット,推薦,商品,sns,協調,収集,マイニング
6	11	オノマトペロリ：オノマトペを利用した料理推薦システム	オノマトペ,オノマトペロリ,料理,味覚,推薦,レシピ,擬音語,形容詞,擬態語,知恵美	論文,日本,システム,表現,日本語,検索,利用,料理,人間,味覚
7	3	図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性	推薦,opac,貸出,逸村,図書館,履歴,図書,amazon,池内,svm	教育,システム,評価,情報,図書館,比較,大学,連携,図書,履歴
8	95	Twitter感情分析を用いた感情値可視化とユーザ推薦システム	twitter,感情,分析,推薦,可視化,ユーザ,tweet,加藤,慎一朗,濱川	論文,分析,感情,英語,システム,情報,可視化,加藤,利用,twitter
9	6	図書館の貸出履歴を用いた図書推薦システムの有効性検証	推薦,協調フィルタリング,opac,図書,図書館,貸出,履歴,逸村,有為,池内	システム,研究,図書館,情報,比較,cinii,利用,図書,履歴,推薦
10	2	読み聞かせ時の反応に着目した絵本に対する子どもの好みの取得...	絵本,読み聞かせ,好み,子ども,幼児,反応,松村,幼児教育,遊び,図書館	論文,教育,子ども,日本,幼児,分析,言語,図書館,地域,大学

閲覧人数上位

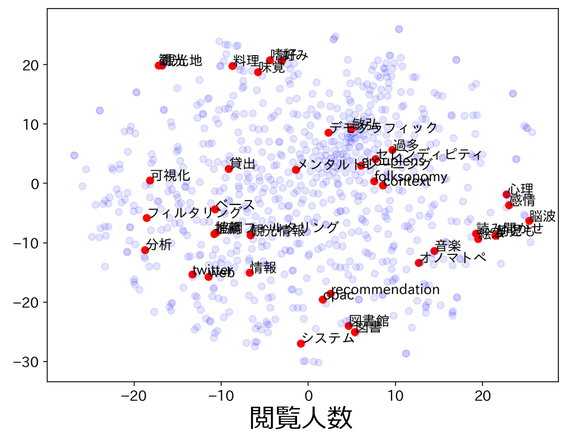
人数	普遍度	題名	内容指定語	分類指定語
3	1	絵本の読み聞かせにおける子どもの好みと絵本の主題との関係性	絵本,読み聞かせ,子ども,好み,幼児,松村,主題,親子,反応,図書館	論文,教育,子ども,分析,研究,幼児,図書館,方法,関係,システム
10	2	読み聞かせ時の反応に着目した絵本に対する子どもの好みの取得方法...	絵本,読み聞かせ,好み,子ども,幼児,反応,松村,幼児教育,遊び,図書館	論文,教育,子ども,日本,幼児,分析,言語,図書館,地域,大学
7	3	図書館の貸出履歴と書誌情報を用いた図書推薦システムの有効性	推薦,opac,貸出,逸村,図書館,履歴,図書,amazon,池内,svm	教育,システム,評価,情報,図書館,比較,大学,連携,図書,履歴
24	4	複数人での旅行における嗜好分析による観光地推薦システムの提案	観光,推薦,旅行,観光地,システム,嗜好,意思決定,集団,web,計画	システム,評価,情報,分析,観光,計画,画像,技術,メディア,集団
87	5	スポーツメンタルトレーニングへの応用を目指した脳波利用の音楽推薦システム	音楽,スポーツ,メンタルトレーニング,メンタル,脳波,推薦,心理,競技,トレーニング,選曲	論文,心理,スポーツ,音楽,評価,システム,方法,トレーニング,脳波,選択
9	6	図書館の貸出履歴を用いた図書推薦システムの有効性検証	推薦,協調フィルタリング,opac,図書,図書館,貸出,履歴,逸村,有為,池内	システム,研究,図書館,情報,比較,cinii,利用,図書,履歴,推薦
43	7	映像コンテキストに基づく多視点映像の視点列推薦	サッカー,映像,スポーツ,推薦,コンテキスト,切り替え,画像,テレビ,カメラ,テレビ番組	スポーツ,研究,システム,評価,画像,効果,映像,比較,サッカー,時間
15	8	高校生の食行動に関する実態報告(第1報)	脳波,音楽,心理,メンタルトレーニング,推薦,機械学習,感性,分類,影響,検討	論文,心理,音楽,学習,影響,効果,システム,行動,分類,脳波
1	9	推薦システムのアルゴリズム(1)	推薦,神嶌,システム,神嶌敏弘,敏弘,アルゴリズム,serendipity,grouplens,recommender	システム,アルゴリズム,推薦,system,based,知能,人工知能学会,敏弘,content,filtering
213	10	介護における声かけに着目した介護職員に対する警告・行動推薦...	介護,掛け,共起,利用者,職員,がけ,介護施設,ライフログ,推薦,事故	論文,介護,文献,行動,言語,報告,サービス,事故,職員,利用者

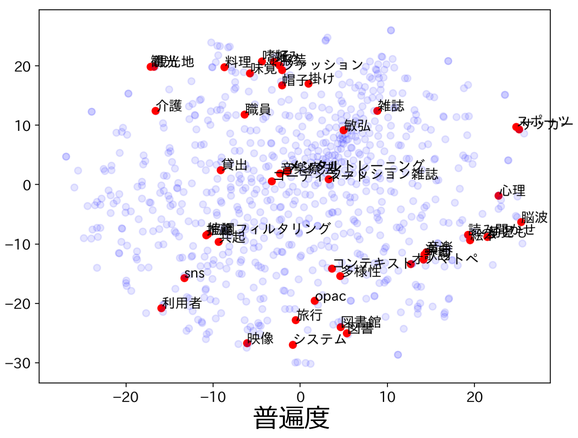
閲覧普遍度上位

- 内容指定語では「観光，音楽，介護」など，内容を示す検索語が出現
- 分類指定語では「教育，心理」 や「研究，分析」など分野や分類を表す語が出現

トピック分布の傾向分析

- 内容指定語上位5件の分散表現をt-SNEを用いて2次元に削減し，プロットした
  - 全論文の題名と概要から分散表現モデルを作成
- 上位15件の論文の検索語を赤色，以下を青色でプロットし分布の違いを分析する





- 閲覧人数上位では「可視化，フィルタリング」などの研究手法が出現
- 普遍度上位では「スポーツ，ファッション」などの一般向けのトピックが出現

分類指定語の頻度分析

閲覧人数上位	頻度		頻度	普遍度上位	頻度		頻度
システム	12	観光	2	論文	8	日本	2
論文	9	日本	2	システム	8	図書館	2
情報	5	評価	2	評価	5	スポーツ	2
推薦	4	図書館	2	音楽	4	画像	2
分析	4	system	1	教育	3	影響	2
based	3	状況	1	研究	3	表現	2
教育	3	依存	1	情報	3	幼児	1
アルゴリズム	2	手法	1	心理	3	観光	1
子ども	2	web	1	子ども	2	比較	1
研究	2	インターネット	1	分析	2	学習	1

分類指定語上位の出現頻度

- 上位15件の論文の分類指定語上位5件の出現頻度を集計，比較した
- 閲覧人数では「アルゴリズム，手法」など研究手法に関する語が出現
- 普遍度では「音楽，心理」などトピックとは異なる分野を表す語が出現

今後の課題と問題点

- ユーザの識別の精度向上
  - IPアドレスとユーザエージェントでは組織的なアクセスを細分化できない
- ユーザの属性を考慮した分析
  - ユーザの属性情報(所属や研究分野など)が存在しない
- 最新の論文では検索クエリが抽出できない
  - HTTPS以降により最新の論文ではリファラが付与されない
  - 手がかり表現を用いて論文の特徴語を抽出する
- 検索結果の絞り込みに検索語を利用する方法を検討する

連絡先

- 小林 和央（和歌山大学システム工学研究科）
- Email: s206099@wakayama-u.ac.jp