第 23 回情報知識学フォーラム予稿

新書本を用いた学問発見支援手法の提案
A method to support for finding academic disciplines using *Shinsyo*, nonfiction paperback books

清水 花菜子 [1*], 高久 雅生 [2]
Kanako SHIMIZU[1*] , Masao TAKAKU[2]

[1*] 筑波大学 情報学群 知識情報・図書館学類
College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: s1511514@klis.tsukuba.ac.jp
[2] 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: masao@slis.tsukuba.ac.jp
* 連絡先著者 Corresponding Author

大学等における学部選択や科学コミュニケーションの文脈において，学問に興味を促し，興味のある学問分野を発見する必要がある．そのため，興味のある学問分野を発見するための手法を提案する．本研究では，多様な学問の基礎的な内容を扱っており，手にされやすいという特性をもつ新書本を用い，利用者が選択した新書本に類似する学問分野を，利用者に対し，興味のある学問分野として提示する．新書本と学問分野の類似度を算出する手法として，BM25 による単語の重みを利用し，コサイン類似度で新書本に対する学問分野の類似度を算出する手法 1 と，BM25 で新書本と学問分野の類似度を算出する手法 2 を検討した．結果として，手法 2 より手法 1 のほうが有効であることが示された．

Several situations such as student's decision making on the application to university and citizen's participation in science communication require to find interesting academic disciplines. We propose a method for finding academic disciplines using *Shinsyo*, nonfiction paperback books. *Shinsyo* covers several subjects and is easy to read for beginners. We implemented and evaluated two methods of calculating similarity between *Shinsyo* and academic disciplines. First method calculates BM25 term weights both on *Shinsyo* and academic disciplines, and then calculates the cosine similarity between them. Second method calculates ranking scores based on BM25 using *Shinsyo* as a query and academic disciplines as a document. As a result, we showed that the first method was better than second one.

キーワード: 新書本, 学問発見, BM25, コサイン類似度
keywords: *Shinsyo*, finding academic displine, BM25, Cosine similarity

1 はじめに

現在の日本において，大学入学以前から学部選択を迫られることが多い．文部科学省によれば,「現在，多くの大学においては，学科，更には専攻，コースというように，募集単位を細分化した上で，それぞれの募集単位ごとに入学者選抜を実施し，学生を受け入れている」[1] と述べられている．

そのため，入学希望者は入学以前から，その学科や専攻，さらにはその学科等が対象とする学問に興味をもち，興味のある学問分野を発見する必要がある．

また近年，科学コミュニケーションを目的としたイベントが活発になっている．ところが，イベントの参加者は,「科学・技術への高関与層」が主体となっており,「科学・技術への低関与層」はほとんど参加しておらず,「科学・技術への低関与層」にアプローチをする必要性があると加納ら [2] は述べている．

そのため，科学コミュニケーションの文脈においても，特に，学問への興味関心が低い人々を中心に，

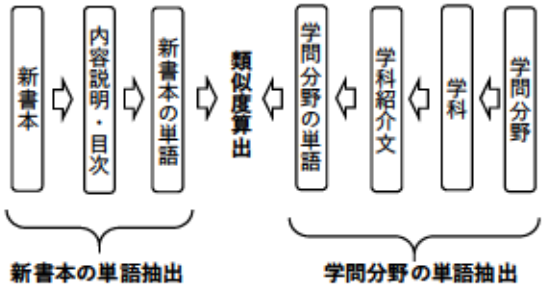

図 1: 提案手法の概要

学問に興味をもち，興味のある学問分野を発見できるようにする必要があると言える．

そこで本研究では，学問への興味を促し，興味のある学問分野を発見するための学問発見支援手法を提案する．

2 関連研究

利用者に対して，学問への興味を促し，学問分野を推薦するシステムは数多く存在する．

例えば，京都大学学際融合教育研究推進センターによる Your Schola[3] がある．これは，「1,757 人の研究者からのアンケート結果を元に，12 の領域の中から回答者の適正に近いと考えられる学問を提示する診断ツール」[3] である．また，夢ナビ 興味が湧く学問発見サイト [4] は，関心のあるキーワードを選択することで学問が提示される．

3 提案手法

3.1 概要

本研究における提案システムは，利用者が興味のある新書本を選択することで，利用者に対して興味のある学問分野を提示する．図 1 に提案手法の概要を示す．提案手法は「新書本の単語抽出」,「学問分野の単語抽出」,「類似度算出」の 3 ステップで構成されている．

3.2 新書本

提案システムの選択対象に新書本を選出した理由を述べる．

新書本に関連する研究はいくつか行われている．例えば，大学図書館と新書本についての調査・考察から，新書本について「学問の導入的読書材として，新書本のもつ役割はけっして小さくない」,「もっとも，ハードカバー本と比べ，同一主題であれば手にされる可能性も，また，通読される可能性もそれぞれ高い」[5] と述べており，新書本には学問研究の基礎的内容を扱っており，手にされやすいという特性があると言える．また，4 つの新書本シリーズの主題区分を調査した研究がある [6]．主題区分の分布は偏っているものの，どの主題区分にも一定数の新書本が該当し，多様な学問領域をカバーしているという特性があると言える．

以上の新書本の特性が，学問に興味を促し，興味のある学問分野を発見するという目的に適していると考え，提案システムの選択対象として，新書本を選出した．

3.3 対象データ

新書マップの「主な新書シリーズ」[7] に該当する図書を新書本として定めた．新書本に関する文書として，2012 年度版「BOOK」データベース [8] のタイトル，内容説明，目次を用いた．対象とする新書本は 793 冊である．

また，学問分野は，学校基本調査における学科系統分類表 [9] を基にした．学科系統分類表は 10 の大区分，76 の中区分，3,363 の小区分からなっている．学科系統分類表の小区分名は学科名に対応するため，学科系統分類表を基にした学問分野と，学科について記述された学科紹介文は対応づけることができる．学問分野に関する文書として用いる学科紹介文には，スクレイピングによって収集した逆引き大学辞典の学科紹介文 [10] を利用した．

本研究においては原則，10 以上の学科紹介文をもつ中区分を学問分野として採用した．なお，「大区分名+その他」となっている中区分については，中区分に属する小区分の中から 10 以上の学科紹介文をもつものを学問分野として採用した．また，複数の中区分に属する小区分のうち，小区分自体に対応する学科紹介文が 10 以上あるものについては，独立した学問分野として扱った．採用した学問分野数は 54 である．

採用する 54 の学問分野と対応した，実際に用いる文書数は，延べ 2,418 文書，異なり 2,128 文書である．

3.4 単語抽出

新書本に関する文書，学問分野に関する文書から単語を抽出する．形態素解析器には Mecab，辞書には mecab-ipadic-neologd[11] を用いた．学問分野に関する文書から抽出された単語数が延べ 177,145 単語，異なり 9,208 単語，また，新書本に関する文書から抽出された単語が延べ 72,174 単語，異なり 15,332 単語であった．さらに，学問分野に関する文書と新書本に関する文書で一致する単語数については，延べ 190,396 単語，異なり 4,021 単語となった．

3.5 類似度算出手法

類似度算出手法として，2 つの手法を検討する．

3.5.1 手法 1

手法 1 では，新書本に関する文書内の単語と学問分野に関する文書内の単語を BM25[12][13] によって重みづけし，重みを用い，コサイン類似度で新書本に関する文書と学問分野に関する文書の類似度，すなわち，新書本と学問分野の類似度を算出する．以下の式により，単語の重みづけを行った．

$$score(D,T) = IDF \cdot \frac{(k_1+1)tf}{tf + k_1 \cdot K} \tag{1}$$

ここで，IDF は単語 T の逆文書頻度，tf は単語 T の文書 D での頻度である．また，$K = k_1((1-b) + b \cdot \frac{dl}{avdl})$ であり，文書長の正規化を行っている．dl は該当文書 D の単語数，$avdl$ は文書 D 全体の単語数の平均である．なお，k_1，b はパラメータであり，$k_1 = 1.2$，$b = 0.75$ を用いている．

続いて，以下の式により，新書本と学問分野間の類似度算出を行った．

$$cos(S,A) = \frac{\sum_{T \subset S \cap A}(score_{S,T} \cdot score_{A,T})}{\sqrt{\sum_{T \subset S} score_{S,T}^2} \cdot \sqrt{\sum_{T \subset A} score_{A,T}^2}} \tag{2}$$

ここで，S は新書本，A は学問分野，T は出現単語，$score$ は (1) 式によって算出された重みである．

3.5.2 手法 2

手法 2 では，新書本に関する文書をクエリ，学問分野に関する文書を検索対象とし，BM25[14] によって，クエリと検索対象の類似度，すなわち，新書本と学問分野の類似度を算出する．以下の式により，新書本と学問分野間の類似度算出を行った．

$$score(D,Q) = \sum_{T \subset Q} IDF \cdot \frac{(k_1+1)tf}{K + tf} \cdot \frac{(k_3+1)qtf}{k_3 + qtf} \tag{3}$$

ここで，IDF は単語 T の逆文書頻度，tf は単語 T の文書 D での頻度，qtf は単語 T のクエリ Q での頻度である．また，$K = k_1((1-b) + b \cdot \frac{dl}{avdl})$ であり，文書長の正規化を行っている．dl は該当文書 D の単語数，$avdl$ は文書 D 全体の単語数の平均である．なお，k_1，k_3，b はパラメータであり，$k_1 = 1.2$，$k_3 = 1000$，$b = 0.75$ を用いている．

4 評価実験

手法の有効性を検証するために，人手による適合判定を行った．判定は，「本システムでは，興味のある新書本を選択することで，興味のある学問分野を推薦する．以下の新書本を選択した結果から推薦される学問分野として，以下の学問分野は適切か.」という基準で行った．また，判定者数は大学教員 1 名，学部生 2 名の計 3 名で行った．判定用データには，新書本 20 冊に対して，54 の学問分野を用いた．また，判定時の情報源としては，新書本のタイトル，内容説明，目次を利用した．さらに，適合レベルは高適合，適合，部分適合，不適合の 4 段階とした．

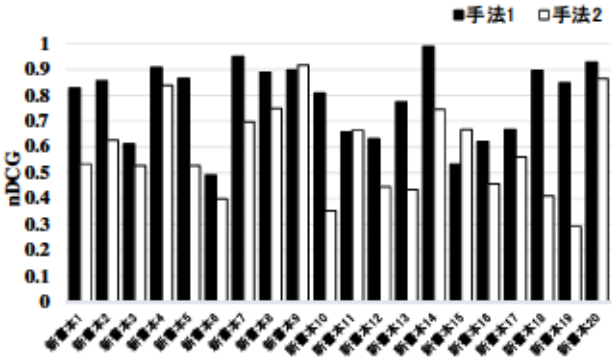

図 2: 新書本別 nDCG 値

3 名の判定結果の一致度を算出するために，ケンドールの一致度係数 w[16] を用いた．ケンドールの一致度係数は 2 名以上の判定者による判定結果の一致度を示すものであり，0 から 1 までの値をとる．3 名の判定結果の一致度係数は $w = 0.71$ であった．

続いて，3 名の判定結果から正解データを決定した．正解データは，判定結果が 3 名間で一致する場合には，3 名の一致する判定結果，2 名間で一致する場合には，2 名の一致する判定結果，3 名間で不一致の場合には，中間の判定結果とした．

以上の正解データを用い，評価実験を行った．評価指標には nDCG[15] を用いる．本研究では，54 学問に対する評価値 $nDCG_{54}$ を以下の式により算出した．

$$DCG_{54} = \sum_{rank=1}^{54} \frac{rel_{rank}}{\log_2(rank+1)} \tag{4}$$

$$IDCG_{54} = \sum_{rank=1}^{54} \frac{rel_{rank}}{\log_2(rank+1)} \tag{5}$$

$$nDCG_{54} = \frac{DCG_{54}}{IDCG_{54}} \tag{6}$$

また，評価値の利得は以下の通りである．

$$rel_{rank} = \begin{cases} 0(\text{不適合}) \\ 1(\text{部分適合}) \\ 2(\text{適合}) \\ 3(\text{高適合}) \end{cases} \tag{7}$$

5 評価結果と考察

図 2 は評価実験に用いた新書本 20 冊の新書本別 nDCG 値である．手法 1 による値と手法 2 による値を比較している．

手法 1 の nDCG 値の平均が 0.78 に対し，手法 2 の nDCG 値の平均が 0.59 となった．また，個別の

新書本に対しても，新書本 9，新書本 11，新書本 15 以外の 17 冊については，手法 1 の nDCG 値が手法 2 の nDCG 値を上回っている．そのため，手法 2 より手法 1 のほうが有効であると言える．

正解データにおいて，新書本に対して高適合または適合であるとされた学問分野のうち，手法 1 におけるランキングで上位 10 位までに含まれないものは 9 組あった．これら 9 組において，新書本と学問分野間で単語が一致しにくい理由を，目視で単語を確認することによって考察する．新書本に関する単語には「ルネサンス」「世田谷区長」「リフォーム詐欺」といった具体的な単語が多い一方，学問分野に関する単語には「文化」「行政」「社会問題」といった抽象的な単語が多いことが，単語を一致しにくくする原因であるとわかった．また，新書本に関する文書としては，新書本のタイトル，内容説明，目次のみを扱っており，得られる単語数が少ないことも単語が一致しにくい原因であることがわかった．これらは今後の課題であると言える．

また，評価実験に用いた新書本 20 冊において，手法 1 と手法 2，いずれに対しても nDCG 値が 0.5 を下回った新書本 6「帝国ホテルの流儀」について考察する．この新書本は，適合判定において，高適合となる学問分野が存在せず，IDCG が 3.56 であった．さらに，評価実験に用いた新書本 20 冊中において，高適合となる学問分野が存在しない新書本が 10 冊存在したことからも，一部の新書本に対しては，正解となる学問分野が存在しないことがある．

6 おわりに

新書本を用いた，興味のある学問分野を発見するための手法を提案した．類似度算出手法としては，BM25 による単語の重みを利用し，コサイン類似度で新書本に対する学問分野の類似度を算出する手法 1 が有効であった．

今後は，手法 1 を基にシステム実装を行い，利用者による評価を行う．また，新書本と学問分野間で単語が一致しにくいという課題についても，検討を行う．

参考文献

[1] 大学審議会: 「大学入試の改善について (答申) (平成 12 年 11 月 22 日大学審議会) 」, 文部科学省, 2000. http://www.mext.go.jp/b_menu/shingi/old_chukyo/old_daigaku_index/toushin/1315961.htm (2018 年 9 月 30 日参照)

[2] 加納圭; 水町衣里; 岩崎琢哉; 磯部洋明; 川人よし恵; 前波晴彦: 「サイエンスカフェ参加者のセグメンテーションとターゲティング:「科学・技術への関与」 という観点から」, 科学技術コミュニケーション, Vol. 13, pp. 3–16, 2013.

[3] 京都大学 学際融合教育研究推進センター: 「ナビスコラ:学問分野診断＆相関図」, NaviSchola: ナビスコラ. https://navischola.app (2018 年 9 月 17 日参照)

[4] 夢ナビ編集部: 「夢ナビ 大学教授がキミを学問の世界へナビゲート」. http://yumenavi.info/index_sp.aspx (2018 年 6 月 3 日参照)

[5] 吉田昭: 「大学図書館と新書本」, 大学図書館研究, No. 38, pp. 60–66, 1991.

[6] 今村成夫: 「新書本の主題範囲」, 大正大學研究紀要, No. 99, pp. 326–313, 2014.

[7] 新書マッププレス: 新書マップ: 知の窓口, 日経 BP 社, 2004.

[8] 日外アソシエーツ: 「BOOK」データベース, 2012 年 1 月-12 月, 2015, (CD-ROM).

[9] 文部科学省: 「学科系統分類表」, http://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2018/03/27/1388724_4.pdf (2018 年 6 月 8 日参照)

[10] 廣告社: 「大学検索」, 逆引き大学辞典. https://www.gyakubiki.net (2018 年 6 月 27 日参照)

[11] “mecab-ipadic-NEologd : Neologism dictionary for MeCab”. https://github.com/neologd/mecab-ipadic-neologd (2018 年 7 月 12 日参照)

[12] Robertson, S. E; Walker, S; Jones, S; Hancock-Beaulieu, M. M.; Gatford, M: “Okapi at TREC-3”, In Proceedings of the 3rd Text REtrieval Conference, 1994.

[13] Robertson, S; Zaragoza, H: “The Probabilistic Relevance Framework: BM25 and Beyond”, Information Retrieval, Vol. 3, No. 4, pp. 333–389, 2009.

[14] Robertson, S. E; S. Walkery; M. Beaulieuz: “Okapi at trec-7: Automatic ad hoc, filtering, vlc and interactive”, In Proceedings of the Seventh Text REtrieval Conference, 1998.

[15] Jarvelin, Kalervo; Jaana, Kekalainen: “Cumulated gain-based evaluation of IR techniques”, ACM Transactions on Information Systems (TOIS), Vol. 20, No. 4, pp. 422–446, 2002.

[16] Kendall, M. G; Gibbons, J. D: Rank correlation methods, 5th ed, Edward Arnold, 1990.