4H-05

学習関連キーワードに基づく Web 上の学習コンテンツの特定

豊田　哲也[†]　　孫　媛[‡]
国立情報学研究所[†‡]

1 はじめに

ICT の進歩により，Web 上の学習サービスや学習コンテンツが充実し，今後も拡大が見込まれている．これらの学習活動によって生成される学習ログデータの複雑化・大規模化に伴って，これらの学習ログを基にした新たな学術領域「Learning Analytics」が注目を集めており，個人の学習特性を解明することが可能になりつつある．学習環境が多様化したことで，学校教育のようなフォーマルな学習以外のインフォーマル学習が注目を集めており[1]，その重要性は今後より大きくなると予想される．特に，個人のインフォーマル学習における学習履歴データを分析することによって，これまで把握困難であった学習者の学習特性などのプロファイルを構築することが可能になると考えられている．

Web 上の学習コンテンツを利用した学習データの解析を実現するためには，Web コンテンツが学習者にとって有益なコンテンツであるかどうかを判断する必要がある．普段の日常生活において閲覧したニュースサイトの記事等，学習することを意図せずに閲覧した Web コンテンツの中に有益な学習コンテンツが含まれている可能性があり，これらの閲覧も学習行動に含まれると考えられる．そのため，閲覧した Web コンテンツが学習に関連するコンテンツであるかどうかを判断するための基準として，我々は学習関連キーワードの導出を行っており，学習項目に関連するキーワードを Wikipedia から抽出する仕組みを提案した[2]．ここでは，学習者の学年レベルに応じて，抽出したキーワードのランキングを学習指導要領から重みづけする仕組みを実現している．

本研究では，抽出した学習関連キーワードの有効性と Web コンテンツを特定するための評価値の妥当性について検証するため，Web コンテンツにどの程度学習関連キーワードが含まれているかを調査し，学習関連キーワードのスコアから Web コンテンツが学習に関連するコンテンツとして特定可能かを分析する．

Identification of the Learning Contents on the Web Based on the Learning Related Keywords
† Tetsuya Toyota
‡ Yuan Sun
National Institute of Informatics (†, ‡)
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

2 学習関連キーワード

Wikipedia を用いて抽出した学習関連キーワードは，学習内容に特化したキーワード間の関連度と学習指導要領を用いた重みづけによってランキングを作成する[2]．入力となるクエリ（「因数分解」や「分数」などの学習項目）から Wikipedia の該当記事を抽出し，該当記事のリンク先の記事集合および該当記事が属する共通のカテゴリ持つ記事集合，さらにそれらのリンク先の記事集合を集め，pf-ibf[3]を基に関連度を算出している．ここで得られた該当記事と抽出記事群の関連度に対して，学習指導要領に出現するキーワード群を学年に応じて重みづけを行い，学習者の学年レベルに応じた学習関連キーワードのランキングを作成している．

分析に用いる Web コンテンツは，Bing Search API を利用して収集する．クエリは学習関連キーワードを生成するために使用した中学校 3 年数学の教科書の索引に記載されているキーワード 10 項目をピックアップして利用する．これらを学習単位名として用い，それぞれ最大 50 件の Web コンテンツを重複がないように得る．得られた 500 件のデータに加え，これらと関連性が低い数学の Web コンテンツ 250 件を加えてデータセットとして用いる．

3 分析

3.1 Web コンテンツの評価値

学習に関連するコンテンツを決定するための方法は，Web コンテンツに含まれるキーワードが学習関連キーワードをどの程度含んでいるかによって決定する．すなわち，学習関連キーワードが全て含まれている Web コンテンツは評価値が最大となる．最終的な Web コンテンツの評価値は，1) Web コンテンツ内に出現した学習関連キーワードのスコアを加算したもの，2) 出現した学習関連キーワードの数を加算したもの，3)学

習関連キーワードのスコアと Web コンテンツ内のキーワードの出現頻度を積算し加算したもの，の 3 つを用いる．これらを評価値が高いものから順にランキングし，上位に出現するコンテンツ数を基に，情報検索の性能評価手法である適合率(precision)と再現率(recall)，およびこれらの調和平均の F 値(F-measure)を用いて評価する．クエリと対応する 50 件の Web コンテンツを基に，適合率はランキング上位に学習コンテンツがどの程度含まれているかを示し，再現率は，学習関連キーワードを基にどの程度のコンテンツが特定できたかを判断することができる．表 1 は 10 件のクエリの平均値と標準偏差である．

表 1：各評価値算出方法による性能評価

	1)	2)	3)
Precision	0.75	0.72	0.68
Recall	0.71	0.70	0.70
F-measure	0.71	0.70	0.65
Precision (Std.)	0.12	0.17	0.26
Recall (Std.)	0.06	0.07	0.07
F-measure (Std.)	0.10	0.09	0.16

どの手法でも再現率は 7 割程度であり，一定の学習コンテンツを特定できていることがわかる．適合率では，1)の Web コンテンツの中に含まれる学習関連キーワードの種類に応じてスコアを加算する方法が最も性能が高いことがわかる．これは，コンテンツ内のキーワードの出現頻度が学習コンテンツの特定に影響しないことを示しており，とくに，出現頻度とスコアを積算した場合は適合率が低くなったことから，学習コンテンツの特定には学習関連キーワードとそのスコアのみでよいことがわかる．

3.2 学習関連キーワードの有効性

学習関連キーワードは一つのクエリに対して数万件が得られるが，それぞれのスコアに応じて特定できるコンテンツの種類も変化すると考えられる．そこで，利用する学習関連キーワードのスコアが高いものを利用した際と，すべてを利用した際の比較を行う．表 2 は，全キーワードと閾値以上のキーワードを利用した際に，上位 100 件以内で特定したコンテンツ数から算出した各評価値である．

表 2　学習関連キーワードの有効性

	全キーワード	閾値以上
Precision	0.75	0.91
Recall	0.71	0.50
F-measure	0.71	0.63

F 値では全てのキーワードを利用した場合の方が高くなるものの，閾値以上のキーワードを利用した場合，再現率は低くなるが，適合率は非常に高くなることが分かった．これは，全てのキーワードを利用した場合に関連する様々な種類のコンテンツを特定することができ，一定の閾値以上のキーワードの場合は，クエリの内容に関連性の高いコンテンツが特定されることを意味している．このため，学習関連キーワードのスコアを閾値で調整することにより，特定する学習コンテンツの内容を調整することが可能であることがわかる．

4 おわりに

本研究では，Web 上の学習コンテンツを特定する方法および提案した学習関連キーワードの有効性，特定した学習コンテンツの評価を行った．Web コンテンツの評価値は，コンテンツ内のキーワードの出現頻度とキーワードスコアの積算によって決定する方法が最も妥当であることがわかった．さらに，学習関連キーワードは，スコアに応じて利用範囲を決定することにより，特定するコンテンツの内容をある程度調整することが可能であることがわかった．また，単一の学習項目に限らず，類似した学習項目のコンテンツを特定することが可能となった．

ただし，コンテンツに含まれるテキストの情報量が多ければ多いほど特定しやすくなるため，画像や動画をメインとするマルチメディアコンテンツに対しては特定が難しい．そのため，Web コンテンツの周辺情報等からテキスト情報を獲得することで，これらのコンテンツに対しても特定が可能になると考えられる．

今後は，学習関連キーワードの抽出精度の向上に加えて，リアルタイムで特定したコンテンツをブラウザ上で提示するシステムの開発を目指す．

参考文献

[1] 山内祐平，”教育工学とインフォーマル学習”，日本教育工学会論文誌，Vol.37，No.3，pp.187-195, 2013.

[2] Tetsuya Toyota, Yuan Sun, “Keyword Extraction for Mining Meaningful Learning-Contents on the Web Using Wikipedia”, 2014 Frontiers in Education Conference (FIE 2014), Madrid, 2014.

[3] 中山浩太郎，原隆浩，西尾章次郎，”Web 事典からのシソーラス辞書構築手法”，情報処理学会論文誌：データベース，Vol.48, No.SIG11(TOD 34), pp.27-37, 2007.