5F-3

学習履歴から抽出したキーワードを利用した Web上の学習コンテンツの特定

豊田　哲也†　　鈴木　雅之†　　孫　媛†

†国立情報学研究所

1　はじめに

近年のICTの進歩により，教育現場や個人の学習でICTを活用する事例が数多く提案されている．また，Web上の学習サービスや学習コンテンツの増加，さらにスマートフォン・タブレット端末の高性能化によって，これらを利用した学習アプリ等が手軽に利用可能になっている．このように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，これまで学校教育からのみ把握されてきた学習者の学習特性を多面的に把握可能になると考えられる．学習者が利用するWebコンテンツの中から，学習に関連するコンテンツを特定できれば，個人の学習特性の詳細な分析が可能となり，学習特性に対応した適切なフィードバックによって，より深化した学習を導くことが可能になる．

Web上の学習コンテンツを利用した学習データの解析においては，eラーニングなどの統一されたフォーマットを持つ学習とは異なり，Webコンテンツの種類によってフォーマットが異なることから，データの体系化によって学習コンテンツを同定することが難しい．また，学習者が閲覧したWebコンテンツは，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コンテンツが含まれており，この中にも有益な学習コンテンツがある可能性がある．そこで筆者らは，学習に関するキーワードを前もって確定し，そのキーワードにより学習に関するWebコンテンツを同定する方法を研究している．

この研究の一環として本研究では，Web上の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特定するために必要な学習関連キーワードを抽出する仕組みを提案する．具体的には，オンライン辞書であるWikipediaの記事の見出し語を学習関連キーワードの対象とし，教科書の節単位のような学習単位のキーワードから記事群を抽出し，学習単位との関連度を基にした学習関連キーワードのランキングを作成する仕組みを構築する．

Identification of the Learning Contents on the Web Based on the Keyword Extracted from Learning-Log

†Tetsuya TOYOTA　National Institute of Informatics
†Masayuki SUZUKI　National Institute of Informatics
†Yuan SUN　National Institute of Informatics

2　提案手法

本研究では，Wikipediaを用いて，学習関連キーワードを抽出し，学習内容に特化したキーワード間の関連度と重みづけによってランキングを作成する．Wikipediaを用いる理由は，学習者が意図していない学習に関連するキーワードを抽出するため，幅広い内容を網羅している必要があるためである．

2.1　キーワード抽出

Wikipediaの全記事集合 P の内，学習単位名と一致するWikipediaの該当記事 $P_t \subset P$ を抽出し，P_t が属するカテゴリ集合 $C_{P_t} \subset C$（C はWikipediaの全カテゴリ集合）を得る．次に，P_t に含まれるリンク先の記事集合 $P_l \subset P$，および C_{P_c} に属するすべての記事集合 $P_c \subset P$ として抽出する．次に，$P_c \subset P$ とWikipedia内のリンクで結ばれている記事 $P_{cl} \subset P$ を抽出する．これら P_l, P_c および P_{cl} を合わせた集合を学習関連キーワードの候補記事群 $P_k \subset P$ とする．これらの P_k の各要素と，P_t の関連度を求めることで，学習関連キーワードの順位付けを行う．

2.2　ランキング作成

Wikipediaの記事間の関連度をpf-ibf[2]を基に算出し，学習関連キーワードのランキングを作成する．pf-ibfはWikipediaのように内部リンクで結ばれたネットワーク状のデータにおいて記事間の関連度計算に用いられる手法であり，記事間の距離(pf)と記事間のパスの多さ(ibf)に基づいた計算が行われる．pf-ibfの実際の計算は，記事 P_i と P_j の記事間の全ての経路の長さと P_j の被リンク数で算出され，短い経路と被リンク数が少ない場合に関連度が高くなる．本研究では，P_t と集合 P_k の各要素間の関連度を導出するが，P_k 以外のWikipediaの記事データは除外する．そのため，対象となる P_t および P_k の各要素で作られたネットワークの形状は，非連結成分が存在する．P_t と P_l, P_c は直接リ

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

ンクを持つか，比較的短いパスで相互の記事に到達するが，P_t と P_{cl} 間には経路がない場合が存在する．そこで，経路長の算出を行うためにカテゴリ情報を用いる．Wikipedia のカテゴリは，階層構造を有していて親のカテゴリを辿っていくと，8 つの主要カテゴリに到達する連結した 1 つのネットワーク構造を有している [3]．そこで，P_t と連結していない任意の記事 $P_i \subset P_k$ のカテゴリ $C_{P_i} \subset C$ と C_{P_t} のカテゴリを介した距離を経路長として代用する．図 1 に記事とカテゴリの距離の算出方法を示す．実線は該当記事へのリンクを示しており，破線はカテゴリへのリンクである．P_t と P_{c2} は同一カテゴリにあり，P_{c2} には P_t からのリンクで結ばれていないため，C_{P_t} を介して距離を算出する．P_{c2} と同一カテゴリにある P_{cl} とはカテゴリ $C_{P_t}, C_i, C_{P_{c2}}$ をたどって距離を算出する．

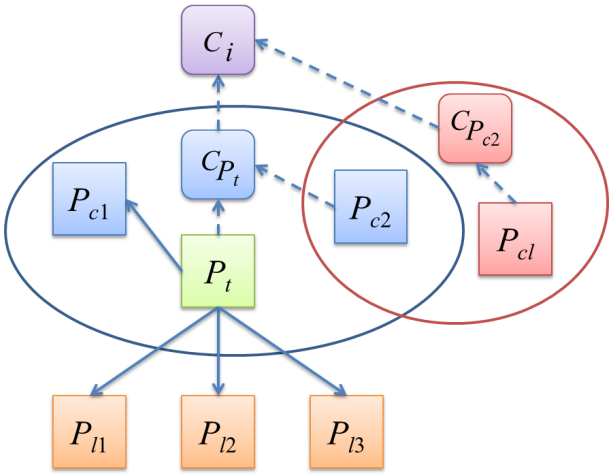

図 1: 記事間距離の概要

次に，pf-ibf で得られた P_k の各要素の関連度に対して，重み付けを行う．これは，ある Web コンテンツが学習に有用であるかどうかは，学習者の知識状態や学力レベルなどによって異なると考えられる．たとえば中学校数学で学習する「因数分解」の Wikipedia の記事には，高校で学習する「複素数」へのリンクが存在するが，中学生にとって「複素数」のキーワードが学習に有用である可能性は低い．そのため，学習者の学習レベルに合わせてキーワードの重み付けが必要になる．そこで，科目/学年ごとの学習指導要領のテキスト情報に対して形態素解析を行い，学習レベルに応じた用語群を抽出し，P_k の各 pf-ibf 値に対してレベルに応じた重み付けを行う．例えば，学習者が中学 1 年の場合，中学 2 年の用語は比較的関連度が高く，高校 3 年の用語は関連度が低いものとして処理する．また，大学生以上の知識を必要とする用語に関しては，より低い値になるよう重み付けを行う．これらの重み付けの処理を行うことで，学習者の学年等に応じた学習関連キーワードのランキング付けが可能となる．

3 結果

上記の手法を利用して，学習関連キーワードの抽出およびランキング付けを行った．実際に抽出したキーワードの詳細については，紙幅の都合上割愛し，発表において提示する．学習者の学習レベルに合わせて提示される学習関連キーワードの順位は高くなるが，学習レベルが低い場合には専門用語が上位には出現せず，より一般的なキーワードや関連性が高いとは言えないキーワードが上位に出現することがわかった．一方で，入力となる科目や学習単位によってもランキングの特徴に大きな変化があり，これは該当する記事のリンク情報の充実度や，分野の記事量の差が影響していると思われる．

4 おわりに

本研究では，学習に関連する Web コンテンツを特定するための学習関連キーワードの導出を行った．Wikipedia の中から学習関連キーワードを導出するために，Wikipedia のカテゴリ構造と内部リンクのデータから，学習単位名の記事との関連度を pf-ibf を基に算出することによって，関連する記事を得るとともに，学習指導要領のデータからキーワードのランキング付けを行った．学習単位名や学習レベルに応じて学習関連キーワードを抽出することが可能となったが，学習レベルが低い場合には精度が落ちることがわかった．今後は，学習関連キーワードの抽出精度の向上と，ブラウザ上で学習関連キーワードを学習者に提示するシステムの開発や，学習関連キーワードを基に特定された Web コンテンツの利用状況を測定することで，学習プロファイルの構築を行う予定である．

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] 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(TOD34), pp. 27–37, 2007.

[3] 豊田哲也, 延原肇. カテゴリ写像に基づく追加学習に対応可能な自己組織化と web ニュース群の動的クラスタリングへの応用. 電気学会論文誌 Sec.C, Vol. 132, No. 8, pp. 1347–1355, 2012.

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.