共通関連事項から記事類似度を測定する トピックマップ駆動 wiki の試作

松浦執[*1]・内藤求[*2]・豊田弘巳[*3]
Email: shumats0@gmail.com

*1: 東京学芸大学教育学部基礎自然科学講座
*2: （株）ナレッジシナジー
*3: 町田市立本町田東小学校

◎Key Words Topic Maps, Wiki, Tanimoto Similarity

1. はじめに

近年、児童やその家族、教師や学校を取り巻く状況の変化が激しく、児童が安心して学べる学級を運営することが困難な状況に陥るケースが少なくない。また初任者教員の多くが、学級運営に関する悩みを持つことが報告されている [1]。平成 19 年度からは特別支援教育が通常の学級でも実施されるようになり、教師には、発達障碍への対応など、さらなる知識・技能の習得が求められている。児童の状況をアセスメントし、問題を可視化するとともに、様々な問題やその解決のための知恵を交換しあうことによって、教師自身の気づきを触発するメディアの必要性が高まっている。

本研究では、上述の背景に対する実践研究として、教師の悩みと知恵の集積所となる、教師の wiki, ”せんせい foilo”[2]の構築を試みた。投稿はあまり長くない文章を原則としたが、それでもその内容は多岐に関係する。そこで、様々な観点から記事を特徴付け、記事を多面的に検索可能にする必要がある。このような目的で、情報の整理と検索のための ISO 標準(ISO/IEC 13250)[3]の一つである Topic Maps を用いた。

2. 構築方法

2.1 Topic Maps

Topic Maps は情報の整理と見つけやすさの向上のための indexing 技術であり、また知識のモデル化が必要となる際に有益な技術である。Topic Maps ではあらゆる主題を topic とし、その主題間の関連（association）を様々に定義することによって、多次元的な情報構造を柔軟かつ容易に構築することが可能になる。具体的な情報資源は、各 topic にその出現（occurrence）として結合する [4]。利用者は目的とする主題を検索したり、興味をもつ関連にそって主題間を探索したりする形で検索を進める。Topic Maps は、多様で異質な諸主題や、その間の関連の構築に対して、柔軟な拡張が可能であるという特徴がある。

2.2 開発環境

Topic map の構築および、構築した topic map で駆動する web の開発・運用にはオープンソースの Ontopia を用いた [5]。Topic map は PostgreSQL データベース上に構成、保存した。Ontopia では、Apache Tomcat サーバーをベースに、topic map を tolog クエリ言語で検索してクライアントに出力する。Web ページは主として tolog タグライブラリと java を用いて作成した。

2.3 “せんせい folio”での記事の特徴付け

投稿された記事について関連記事を検索したり、関心のある内容の記事を見つけやすくしたりするためには、記事の特徴づけが必要である。特徴付けの方法として、例えば、記事の文字列を直接解析する方法や、投稿者自身に記事入力とは別に記事の位置づけを依頼する方法などが考えられる。

”せんせい folio”では、記事内容の特徴付けのために全部で 86 の項目を設けた。投稿者が記事を書いた後に、その記事が関連すると思われる項目を選択する構成を採用した。投稿画面を図 1 に示す。投稿者がどの範囲をまでを記事に関連すると認識するかについては、投稿者自身に任せた。この方式を系統的に構築し、かつ拡張や構造の修正を行いやすくするために topic map 駆動 web サービスを用いた。この項目は下述するように、subject topic type の instance 群として本システムの topic map の中に設定した。

図 1 “せんせい folio”の投稿画面。下側が関連事項を選択するエリア。

2.4 “せんせい folio”の topic map

”せんせい folio”の topic map は大きく article topic と subject topic に分かれる。前者は記事一つ一つをトピックとして位置づけるものであり、後者は児童の能力、振る舞い、学校の時空間などのカテゴリーに分けた種々の「観点」をトピックとしたものである。この「観

点」の個々の具体的な topic instance が、前節で述べた、記事の特徴付けのための選択項目に相当する。

Article topic については、さらに図2の4つの topic type に分類し、記事のカテゴリーごとに、「教師の悩み(cares_article)」「教師の知恵(suggestions_article)」「役立つ規則(rules_article)」「教師のとっさの一言(teachers_word_article)」と名付けた。この4区分は必ずしも排他的分類ではなく、書かれた記事は複数の区分の内容を持つことがあり得る。初期段階の設計でこれらの区分を設定したのは、分類整理のためというよりも、それらの観点から投稿する動機付けのためという機能指向の分類である。実際に記事を投稿すると、article topic instance が生成され、タイトルや記事本文はその instance に結合された occurrence として記録される。

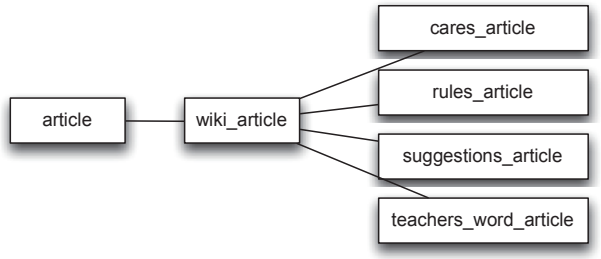

図2　Article topic type の階層

一方、記事を書いた後に、記事を特徴づけるための選択項目に対応するのが、subject topic type の instances である。Subject topic type の分類を図3に示し、そのうちの competence type の instance を例として表1に示した。記事に関連づける subject topic は、article_subject と article_situation の2つに分類した。

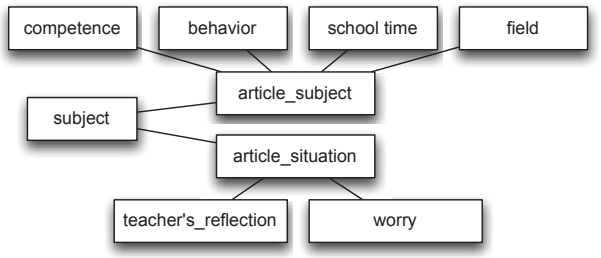

図3　Subject topic type の階層

このうち article_subject は記事の内容を特徴づけるものである。Competence type は児童の能力についての観点であり、behavior は児童の問題のある振る舞い方、school time は登校から下校に至るまでの学校での特徴的な場面、field は学校や家庭その他の空間的な場の分類を示す。

もう一つの article_situation は記事の内容を直接特徴づけるのではなく、記事を書いている先生がどのような思いでその問題を考えているか、不安を抱えている主体は誰であるか、という内容の topic type である。

記事文書を occurrence に持つ各 article topic instance は、投稿者によって選択された subject topic instance と関連づけられる。この関連タイプを Article_related_with association type と名付けた。各記事(article topic instance)は Article_related_with association によって、subject topic instance のある集合と関連づけられる。この集合が、関連づけられた記事を特徴づけるものと考えた。

2.5　記事の類似度測定

表1　Competence type のサブタイプとインスタンス

Category	Competency types	Indications
Physical strength	Posture	Standing and sitting postures
	Group gymnastic skill	Ball games, group games
	Individual gymnastic skill	Run, jump, apparatus gymnastics
	Health	Likes and dislikes in food, illness
Intelligence	Reading skill	Reading aloud, comprehension, Kanji
	Writing skill	Writing letters, figure, sentences
	Arithmetic skill	Arithmetic
	Logical thinking	Vocabulary, comprehension, thinking, expressing
Willingness	Expressiveness	Smiling, laughing
	Perseverance	Tenacity
	Will for living	Positive thinking, not depressed
	Will to act	Positive attitude
Practice	Cooperation with friends	Play with friends
	Cooperation in work	Act in cooperation
	Roles in daily life	Day duty, activity in charge
	Rules in daily life	Following rules
Communication	Listening to	Looking at, listening to others
	Assessment of situation	Behave according to the situation
	Expression	Tell others what the child feels or thinks
	Sympathy	Guess what others feel

“せんせい folio”では、subject topic およびその階層をもとに、選択した subject topic に関連づけられた article topic instance をリストする。これは subject topic によるナビゲーションである。

さらに、2つの記事に関連づけられた subject topic の和集合に対する積集合の割合を、2つの記事の類似度として定義した。この類似度は Tanimoto Similarity と呼ばれる[6]。即ち、ある article topic instance a に subject topic instance の集合 A が関連づけられており、別の article topic instance b に subject topic instance の集合 B が関連づけられているとするとき、記事 a と b の Tanimoto Similarity を次のように書くことができる。

$$T_{ab} = \frac{|A \cap B|}{|A \cup B|} .$$

この値により記事の類似度を評価した。

図4は単独の記事表示のページの例である。左側の

垂直ナビゲーションは subject topic の分類であり、その subject topic に関連づけられた記事のリストページへのリンクである。上部の水平ナビゲーションは、登録利用者ごとの記事の新規作成、更新、リストページなどへのリンク、および全体の記事の類似関係表示ページなどへのリンクである。中央のコンテンツ・エリアでは、単独の記事内容と、類似する記事のリスト、記事へのフィードバック入力、その記事に関連づけられた subject topic instance のリストを示した。類似記事として表示するための基準値の設定については以下に議論する。

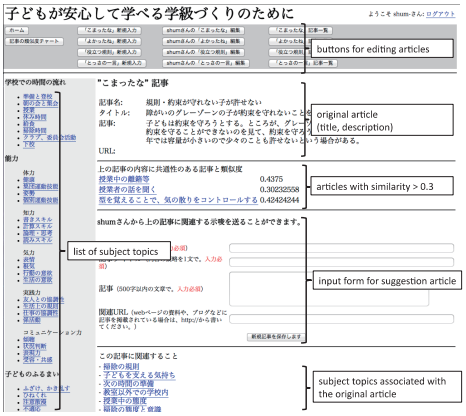

図 4　単独の記事の表示と、類似する記事リスト。

3.　結果と考察

図 5 は投稿数が 102 件の時点での、記事に関連づけられた subject topic instance の数（以下関連項目数と表記する）のヒストグラムを示したものである。選択肢となる subject topic instance の総数は 86 アイテムである。実際の関連項目数は 3 件が最も多く、ついで 4 件、2 件であった。関連項目数がこの程度の小数であると、記事が特定の subject topic によって特徴づけられてしまうことになり、記事の多面性や、関係のひろがりにおける特徴付けがないことになる。総数で 86 ある subject topic instance のなかには、同じカテゴリー内で排他的な関係に近いものもある。しかし、投稿されている事象は様々な場面や能力と関連を持ち得るものが少なくない。投稿者に、よりひろく関連づけを求めることが必要である。

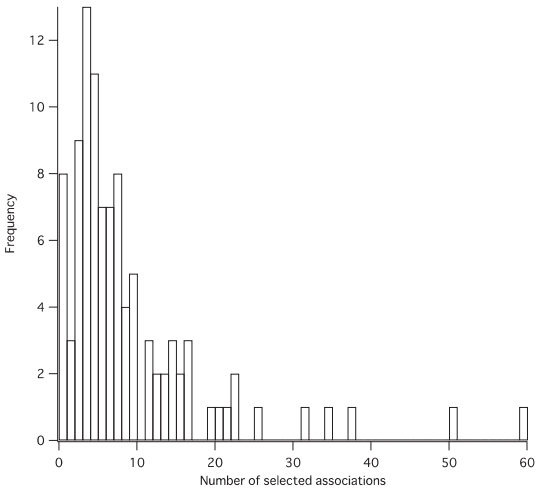

図 5　投稿記事に関連づけられた subject topic の数の分布

図 6 は、すべての記事のペアについて Tanimoto 類似度を計算し、類似度の値のヒストグラムを示したものである。縦軸は頻度の対数値を表示した。類似度が 0 とは、共通の関連づけられた subject topic を持たない記事ペアであることを意味する。類似度 0 のペアが全体の 64.7%にのぼり、類似度が 0.1 から 0.3 の範囲のペアが全体の 33.5%であった。類似度の頻度分布は、図に見られるように、類似度の大きいものほど、ほぼ対数的にペア数が減少していた。

この頻度分布から、類似記事としてリストするための類似度の基準値も低い値に設定せざるを得なかった。現状では暫定的に類似度 0.3 以上の記事をリストしている。類似度 0.3 以上でリストされるのは、全体として 5.6%の記事ペアに過ぎない。

記事ごとの関連数に対して、２つの記事を取り上げたとき、実際どれだけの関連が共通であるかを調べるため、次のようなプロットを試みた。記事のペアごとに、それぞれの関連数の積を横軸にとり、そのペアに共通する関連の数を縦軸にとってプロットした。このプロットが図 7 である。

一般に、各記事の関連項目数が多いほど共通する関連項目も増加するはずである。選ばれる関連項目に偏りがなければ、２つの記事 i, j に共通の関連項目の数 C_{ij}（一致数）は、各記事の関連項目数 m_i、m_j の積 $m_i \cdot m_j$ に比例すると考えられる。さらに、共通関連項目数は、項目の総数（“せんせい foilo” の場合には 86 アイテム）M に反比例すると推測できる。図 7 には M^{-1} の値の勾配を持つ直線を書き入れ、実際のプロットの回帰直線を点線で書き入れた。回帰直線の勾配は、およそ 1/67 となり、$M^{-1} = 1/86$ よりも大きな値であった。

回帰直線の勾配が $M^{-1} = 1/86$ よりも大きいことから、実際の投稿記事では選択される関連項目に偏りがあることが示唆された。

さらに、プロットの分布は $m_i \cdot m_j$ の値が小さい領域

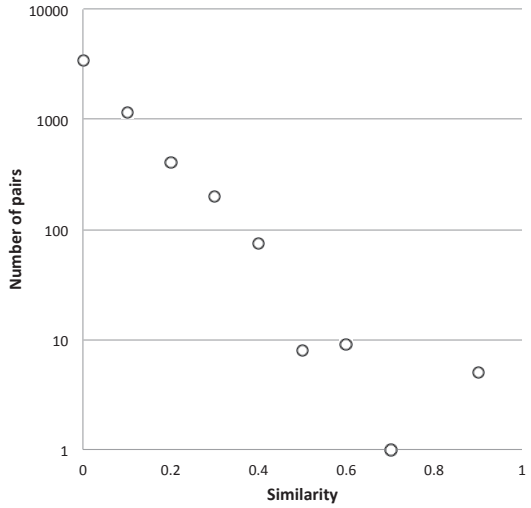

図 6　記事のペアの類似度のヒストグラムを片対数プロットした図

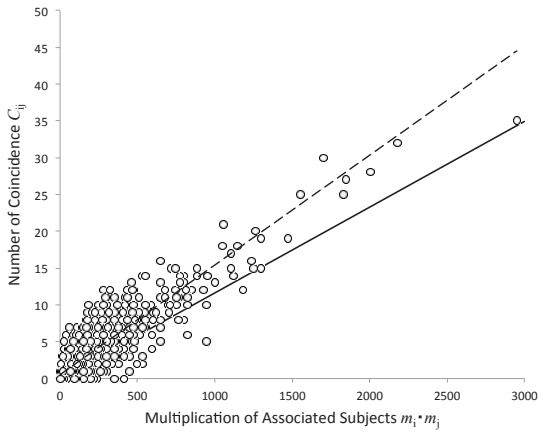

図 7　各記事 i, j のペアについての、記事ごとの関連項目数の積 $m_i \cdot m_j$ に対する、両記事に共通する関連の数 C_{ij}。直線は、勾配 $M^{-1} = 1/86$ を示す。ただし M は関連項目の全数。点線はプロットから最小二乗法により求めた回帰直線を示す。回帰直線の勾配は、およそ 1/67 である。

に集中している。これは図 5, 6 に見られた選択数の少なさに対応する。このことから、選ばれた選択肢の範囲が狭く、特定の話題で類似記事が集まってしまうことが示唆された。

表 2 は図 3 に分類した 5 種類の subject topic type に属する subject instance の数と、実際に各 instance に関連づけられた記事の平均数を示す。例えば、子どもの振る舞い behavior topic type には 7 つの instance が存在するが、各 instance には平均として 6.4 本の記事が関連づけられていたということを意味する。

表 2　記事に関連づけられた subject topic instance の平均数

Subject type	Subject sub-type	Number of instances	Mean number of associated articles
Article _subject	Behavior	7	6.4
	Competence	35	8.1
	School_time	26	8.5
	Field	8	14.8
Article _situation	Teacher's_reflection	5	19.4
	Worry	3	31.0

Article_subject についての平均関連記事数は、Field topic type が最も多い。すなわち、どのような場に関連する記事かということについては、比較的には特徴付けが積極的に行われていることを意味する。

一方、Article_situation については、平均関連記事数が Article_subject に比べて大きくなっている。こちらは項目の排他性も小さく、記事の具体的内容によらずに選択することができる。このため、関連事項の選択において、平均的に Article_subject topic よりも Article_situation topic の選択をしやすいという傾向が生じていたと推測される。

記事の類似性を特徴づけるためには、記事の固有性を適切に表現する必要がある。一つの記事は、投稿者の視点として、あるスケールの粒度を持つものと考えられる。この粒度よりも微視的な粒度で特徴づければ、多面的な特徴が表現できる。また、2 つの記事を比べるとき、微視的粒度での特徴づけの組み合わせにおける共通性の度合いとして類似性を定義できる。一方、個々の記事の差異を超えた粒度での特徴付けでは、類似性の有無を抽出するデータが得られない。

現状の、投稿者による subject topic instance 選択の仕方では、Article_situation topic の選択の比重が高くなってしまい、その結果、類似性を持つ記事の抽出という意味では十分な効果を発揮するに至っていないと思われる。

特に、選択数が 3, 4 アイテムと小数の場合で、そのうち 1, 2 アイテムが Article_situation topic instance である場合には、それらの一致による類似度への寄与が高くなり、粒度の小さい Article_subject topic instance による細かな特徴付けが類似度の評価に寄与しがたい。

これらのことから 2 つの課題が考えられる。第 1 は、特徴付けの粒度に応じて、類似度計算への寄与率を調節することである。第 2 は粒度の小さい Article_subject topic instance をより積極的に、かつより低い負担で選択できるように工夫する必要がある。そのための、馴染みやすく、また関心を引きやすいインタフェースを制作する必要がある。

4.　おわりに

現場の教師の悩みや知恵を集め、関心のある主題にまつわる投稿を見つけやすくし、また関連する記事を検索しながら、閲覧者あるいは投稿者自身がなんらかの気づきを得やすい wiki サイトを目指して、Topic Maps 技術を利用した wiki のパイロットケースを作成した。投稿記事の特徴づけをするために、投稿者自身に記事に関連する事項の組を作成してもらい、これに基づいて記事の類似度を測定した。

現状では、記事に対する特徴付けのための、小さい粒度の関連数が少ないことにより、類似度の値が低い記事ペアが多くなっている。より気づきの得られやすい検出を実現するためには、まず記事をよりよく特徴付けすることが必要条件である。記事を書くだけでなく、これを特徴づけることに関心が持てて、かつその負担を低くすることが課題である。

謝辞

本研究は科研費（24501042）の助成を受けている。

参考文献

(1) 西田晋, http://www.edu.city.kyoto.jp/sogokyoiku/kenkyu/outlines/h21/pdf/541.pdf.
(2) "せんせい folio". http://tm.u-gakugei.ac.jp/ca/k/.
(3) JTC 1/SC 34, http://www.itscj.ipsj.or.jp/sc34/.
(4) Ontopia code google.: http://code.google.com/p/ontopia/.
(5) Pepper, S.: The TAO of Topic Maps, http://www.ontopia.net/topicmaps/materials/tao.html#d0e632.
(6) Rogers, D. J., Tanimoto, T. T.: A Computer Program for Classifying Plants, Science **132**: 1115-1118 (1960).