

# 学習分野を考慮したインターネット検索の提案

下倉 雅行<sup>1,2,a)</sup> 大国 航<sup>1,b)</sup> 島袋 舞子<sup>1,c)</sup> 兼宗 進<sup>1,d)</sup> 村上 晴美<sup>2,e)</sup>

**概要:** 電子書籍の普及が進んでおり、端末では国語辞典などで用語の意味を表示する機能が提供されている。電子教科書において、用語の意味にとどまらず、学習内容に関連する Web 上の情報源にアクセスしてより深い学習を行うことが有効と考えられる。しかし、Web 検索の結果をすべて表示すると、どの検索結果が正しいのかを判定することが、学習を始めようとする学習者には困難であると予想される。そこで電子教科書に適した Web 検索手法を検討し試作を行った。検索したい語が含まれる電子教科書の段落と、検索結果のスニペットを分野判定する。それぞれの分野が合致するページのみを提供することで、学習中の文脈に対応した Web 検索が可能になる。さらに分野判定の精度を向上させるための手法を検討した。

**キーワード:** 電子教科書, 電子書籍, 情報検索, 分野判定

## A Proposal of Field Judgment Function in Web-Searching for Digital Textbooks

SHIMOKURA MASAYUKI<sup>1,2,a)</sup> OGUNI WATARU<sup>1,b)</sup> SHIMABUKU MAIKO<sup>1,c)</sup> KANEMUNE SUSUMU<sup>1,d)</sup>  
MURAKAMI HARUMI<sup>2,e)</sup>

**Abstract:** The worldwide popularity of digital books continues to increase. One useful function is that they display word meanings with an embedded dictionary function. Such web-searching functions are becoming very effective and more popular for learners. However, students often have difficulty selecting correct meaning from multiple meanings on the screen. This may result in misunderstandings and confusions. Therefore, our presentation enables context-based web-searching in digital textbooks by proposing a new classification function that determines a word's fields from the textbook in which it appears and displays search results limited to specific fields.

**Keywords:** digital textbook, digital book, information retrieval, field judgment

### 1. はじめに

近年、小学校から高等学校まで電子教科書の利用が始まっている。タブレット端末などで容易に閲覧でき、さら

に動画などのコンテンツを含むことが可能である。また、電子書籍には国語辞書や英和辞書などの辞書により、言葉を調べる機能 (図 1) が付属している。これは言葉の意味が出てくるものであり、専門的な用語に関する答えを得られるものではない。電子教科書では、通常の国語辞書や英和辞書などから得られない用語も出てくることもあり、それらの用語については Web 検索を利用することとなる。Web 検索を利用するとすれば、ブラウザを別途立ち上げ、検索サイトで調べたい単語を検索することとなる。学習途上の学習者にとっては、この未知の用語を検索する、という行為はかなり敷居が高く、検索結果が得られたとしても、どれが正しいのか理解できず、全く違う分野の言葉を見て

<sup>1</sup> 大阪電気通信大学  
Osaka Electro-Communication University, 1130-70, Kiyotaki, Shijonawate, Osaka 575-0063, Japan

<sup>2</sup> 大阪市立大学  
Osaka City University, 3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585, Japan

a) masayuki@karekinada.com

b) ht12a009@oecu.jp

c) shimabuku.m@gmail.com

d) kanemune@acm.org

e) harumi@media.osaka-cu.ac.jp

しまうこともありうる。学習するための情報の難易度となじみのない言葉の意味の理解を支援するシステム [1] もあるが、wikipedia に限定しており、一般的な Web ページには対応していない。

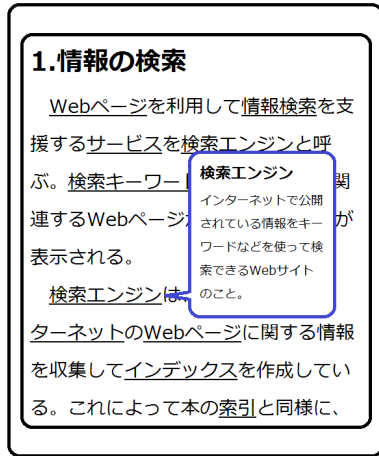


図 1 辞典機能を備えた電子書籍の画面例

検索支援として、(1) ブックマーク共有により検索支援するもの [2]、(2) 検索ログを利用したニーズを抽出 [3]、(3) HTML タグによるクラスタリング手法 [4]、(4) 小規模カテゴリ構造を構築する検索支援 [5] などが存在している。しかし、学校教育の場では、(1)(2) はブックマークがない状態や検索ログがない状態も考慮しなければならないため、利用が困難となる。また、(3)(4) は Web ページ自体の分類となり、教科書の分類と合致するとは限らない。

そこで、我々は、電子教科書から直接検索でき、図 2 のように検索結果と電子教科書分野が同じであるものを選別できる機能を考案した。さらに、精度を高めるために隣接分野というものを検討し導入した。ここでは、この検索機能の実装および評価について説明する。

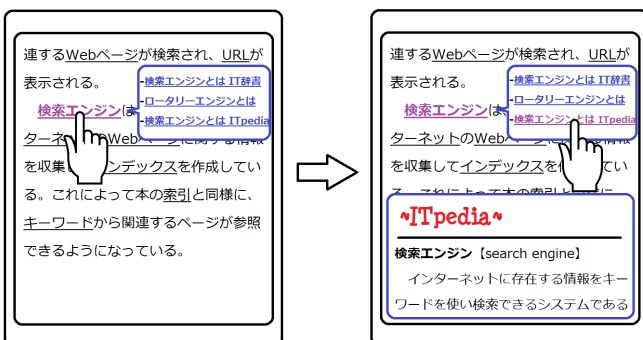


図 2 分野判定検索機能付電子教科書

## 2. 電子教科書に求められる機能

電子教科書に求められる機能は、日本教育情報化振興会 (JAPET) が 2010 年 11 月に「デジタル教科書の位置づけと機能」[6] という資料の中でデジタル教科書に必要な機能としてまとめている。これを表にしたものが表 1 である。

表 1 デジタル教科書に必要な機能

No.	機能	概要
1	マルチメディア提示機能	教科書に即した内容を、音声、写真、映像、アニメーションなどを利用してわかりやすく提示する
2	教材カスタマイズ機能	教師の指導方法や学習者の特性に応じて、教材を選択または編集できる
3	関連教材・資料へのリンク機能 (内部・外部)	学習内容に関連する教材・資料等にアクセスできる。
4	内容の自動更新機能	統計値、地図、地名、制度などを常に最新の状態に保つ。
5	学習習熟度に応じた演習問題提供機能	教師の指示で個々の学習者に適した演習問題を設定できる。
6	学習者データの収集・分析機能	演習問題等の実施結果を収集し、教師が分析しやすいかたちで提示する。
7	学習者の入力インタフェース機能 (認識機能)	手書きによる筆算や計算の途中過程も認識できる。
8	コミュニケーション機能	学校内の児童生徒間コミュニケーション、学外とのコミュニケーション、宿題や課題の提出、家庭とのコミュニケーション
9	各種ツール機能	資料まとめ・発表、資料の協働製作、音楽、美術等の創作、辞書、事典

電子教科書が活用される場面として、同じ JAPET の資料の学習者用のデジタル教科書活用場面例の中の 1 つに「(3) 個々にあるいはグループで主体的な、探求的な学習」とある。また、この具体的な例として「インターネットなどによる調べ学習」が挙げられている。しかしながら、求められる機能の中には、この調べ学習に利用できる機能が挙げられていない。

学習者が電子教科書を利用して予習や復習をする際にわからない単語があると、そこで辞書やインターネットを利用して調べることとなる。辞書に載っている単語のみであ

れば問題はないが、辞書にない単語を Web の検索エンジンを利用して調べるとなると、検索結果が正しいかどうかの判別もできない。そこで、現在挙げられている電子教科書に求められる機能に加えて、検索をサポートする機能が必要であると考えられる。

ここでいう検索をサポートする機能は、教科書の各節や項の関連する分野に検索結果を限定し、大幅に異なった分野を出さないようにする機能（ここでは分野判定検索機能と呼ぶ）である、と考えている。

### 3. 分野判定検索機能の設計と実装

#### 3.1 分野判定検索機能の概要

Web 検索結果は、どのような結果が得られるかはわからない。教科書に載っている用語であっても、期待したものと大幅に異なる結果が得られることがよくある。そこで、検索結果の分野を判定し、検索元の教科書の分野が合致したものののみを提供することで、この大幅に異なる結果のために起こりうる、間違えた学習を減らすことが可能になると考えた。流れとしては次のような形になる。

- (1) 電子教科書上で意味のわからない語句をクリック（またはタッチ）
- (2) 語句が含まれる段落（または節や項）から分野を判定
- (3) 語句をキーワードとして Web 検索にかけて得られた結果の一部（スニペット部分）を分野判定
- (4) 元々の語句の分野と同じ分野の検索結果のみ提供

#### 3.2 分野判定検索機能の設計

分野判定検索機能を実現するための流れを整理する（図 3）。

- (1) 分野判定のための辞書（分野判定辞書）の作成
- (2) 電子教科書の本文を段落ごとに分野判定
- (3) 電子教科書内でのキーワードの選択
- (4) 検索エンジンを利用した検索
- (5) 検索結果の抜粋部分（スニペット）を分野判定
- (6) 分野が合致する結果のみ表示

分野判定辞書を作成するためには、単語と分野の対応を示すデータが必要である。今回提案する機能では、国立国会図書館件名標目表 [7](NDLSH) を利用した。NDLSH には、件名標目、標目よみ、ID、同義語、上位語、下位語、関連語、注記、分類記号 (NDLC)、分類記号 (NDC9)、参照 (LCSH)、参照 (BSH4)、出典 (BSH4)、出典、編集履歴、作成日、最終更新日が含まれる。このうち、目録を検索する手がかりとして用意されている件名標目、分野を特定する手がかりとなり得る同義語と出典、分類記号 (NDC9) を利用することとした。分野を判定したい文章も単語に分割して利用することから、件名標目、同義語、出典は単語に分割して利用する。ここで得られた単語の出現頻度 TF (Term Frequency)[8] をカウントし、重み付けを行えるようにす

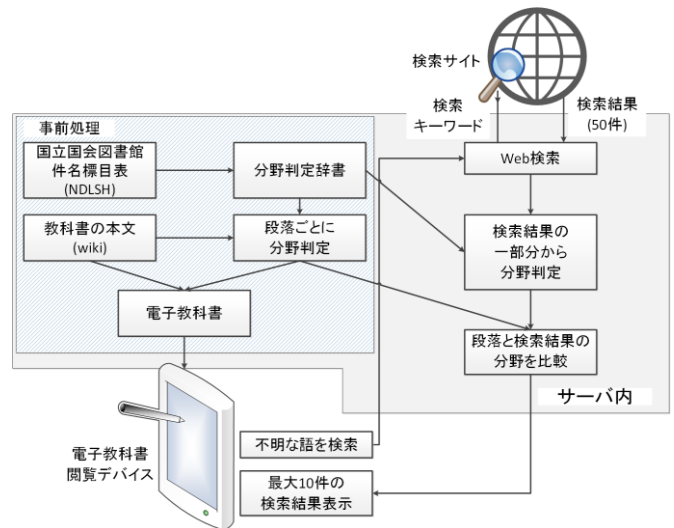


図 3 分野判定検索機能の流れ

る。これにより得られた辞書を用いて、分野を判定する。検索エンジンへのアクセスには、検索エンジンが提供する API を利用した。今回は Google Custom Search API[9] を利用する。検索結果の分野判定には、検索結果の Web ページから抜粋された文章であるスニペットを利用する。電子教科書の本文の、キーワードの含まれる段落を分野判定した結果と、スニペットを分野判定した結果が合致するもののみを提示することで、学習内容に近い検索結果だけを提示することが可能になる。

#### 3.3 分野判定検索機能の実装

分野判定辞書は、国立国会図書館件名標目表 (NDLSH) として提供されているタブ区切りファイルを元に作成した。作成手順は次の通りである。

- (1) NDLSH のタブ区切りファイルから” (ダブルクォーテーション) を削除する。
- (2) 件名標目、同義語、分類記号 (NDC9)、出典を抜き出す。(表 2)
- (3) -を含む行(細目)と分類記号 (NDC9) を持たない行を削除する。(表 3)
- (4) 1 つの行に複数の分類記号 (NDC9) を持つ場合は分割して 1 行 1 分類記号にする。(表 4)
- (5) 分類記号 (NDC9) のうち、上位 2 桁 (大分類と中分類) を取り出し、これを分類記号とする。(表 5)
- (6) 件名標目、同義語、出典を KAKASI(kanji kana simple inverter)[11] で単語に分割する。(表 6)
- (7) 分割された語のうち、数字のみ、記号のみ、ひらがなのみ、アルファベット 2 文字以下、ストップワード

(MySQL Full-Text Stopwords[10]) を除外し、語と分類記号のペアを作成する。(表 7)

- (8) 作成されたペアで同じものの個数をカウントする。(表 8)
- (9) 個数の平方根を取り、四捨五入したものを重みとする。この重みが 1 以下のものは除外。(表 9)
- (10) 同じ語で複数の分類記号を持つ場合、10 未満の場合のみまとめ、10 以上のものは語そのものを除外し、最終的に、語 (tab) 分類記号 1/重み 1, 分類記号 2/重み 2,... という形で保存する。これを分野判定辞書とする。(表 10)
- (11) 分類記号ごとに含まれる語数を数えたものを保存する。(表 11)

表 2 データの抽出

件名標目, 同義語, 出展	分類記号 (NDC9)
アフガニスタン,Afghanistan, アジア高原の旅 / アーノルド・トインビー 著 ; 黒沢英二 訳	292.71
アメリカ合衆国--歴史--1783-1809,United States--History--1783-1815, アメリカの歴史. 2 / メアリー・ベス・ノートン 他著	253.04
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	299.4;452.24
包装 (細目)	
破壊活動,Subversive activities, 政治学辞典	316.74;317.81

表 3 利用できないデータの削除

件名標目, 同義語, 出展	分類記号 (NDC9)
アフガニスタン,Afghanistan, アジア高原の旅 / アーノルド・トインビー 著 ; 黒沢英二 訳	292.71
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	299.4;452.24
破壊活動,Subversive activities, 政治学辞典	316.74;317.81

表 4 1 行 1 分類記号 (NDC9) への変換

件名標目, 同義語, 出展	分類記号 (NDC9)
アフガニスタン,Afghanistan, アジア高原の旅 / アーノルド・トインビー 著 ; 黒沢英二 訳	292.71
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	299.4
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	452.24
破壊活動,Subversive activities, 政治学辞典	316.74
破壊活動,Subversive activities, 政治学辞典	317.81

次に、文章の分野判定方法は次の通りである。

- (1) 分野判定対象の文章を KAKASI で語に分割する。(図 4)

表 5 分類記号 (NDC9) から大分類中分類の抜き出し

件名標目, 同義語, 出展	分類記号
アフガニスタン,Afghanistan, アジア高原の旅 / アーノルド・トインビー 著 ; 黒沢英二 訳	29
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	29
インド洋,Indian Ocean, 印度洋 / 柴田賢一 著	45
破壊活動,Subversive activities, 政治学辞典	31
破壊活動,Subversive activities, 政治学辞典	31

表 6 単語への分割

分割された語	分類記号
アフガニスタン,Afghanistan, アジア 高原 の 旅 / アーノルド ・ トインビー 著 ; 黒沢 英二 訳	29
インド 洋,Indian Ocean, 印度洋 / 柴田 賢一 著	29
インド 洋,Indian Ocean, 印度洋 / 柴田 賢一 著	45
破壊活動,Subversive activities, 政治学 辞典	31
破壊活動,Subversive activities, 政治学 辞典	31

表 7 辞書に利用する単語の選定

語	分類記号	語	分類記号
アフガニスタン	29	Indian	45
Afghanistan	29	Ocean	45
アジア	29	印度	45
高原	29	柴田	45
アーノルド	29	賢一	45
トインビー	29	破壊活動	31
黒沢	29	Subversive	31
英二	29	activities	31
インド	29	政治学	31
Indian	29	辞典	31
Ocean	29	破壊活動	31
印度	29	Subversive	31
洋	29	activities	31
柴田	29	政治学	31
賢一	29	辞典	31
インド	45		

- (2) 分割して得られた語が分野判定辞書にあれば、分類記号と重みを取り出す。(図 5)
- (3) 分類記号に含まれる語数が多い場合には出現頻度が小さいものを利用しないようにするために分類記号ごとの語数の平方根/重みが 15 より大きい場合は利用しない。(図 6)
- (4) 15 以下の場合には分類ごとの語数が大きい場合には重みとして小さくするために、重み/分類ごとの語数の四乗根を新たな重みとして分類記号のスコアとして加算

表 8 語と分類記号の出現回数

出現回数	語	分類記号
26	アフガニスタン	29
10	Afghanistan	29
103	アジア	29
6	高原	29
1	アーノルド	29
1	トインビー	29
3	黒沢	29
1	英二	29
21	インド	29
16	Indian	29
10	Ocean	29
4	印度	29
1	柴田	29
1	賢一	29
8	インド	45
9	Indian	45
12	Ocean	45
6	印度	45
2	柴田	45
1	賢一	45
5	破壊活動	31
4	Subversive	31
3	activities	31
26	政治学	31
10	辞典	31

表 9 出現数によるデータ選別

出現回数の平方根	語	分類記号
5	アフガニスタン	29
3	Afghanistan	29
10	アジア	29
2	高原	29
2	黒沢	29
5	インド	29
4	Indian	29
3	Ocean	29
2	印度	29
3	インド	45
3	Indian	45
3	Ocean	45
2	印度	45
2	破壊活動	31
2	Subversive	31
2	activities	31
5	政治学	31
3	辞典	31

表 10 分野判定辞書 (語, 分類記号/重み) の例

語	分類記号/重み
アフガニスタン	29/5
Afghanistan	29/3
高原	29/2
黒沢	29/2
インド	29/5,45/3
Indian	29/4,45/3
Ocean	29/3,45/3
印度	29/2,45/2
破壊活動	31/2
Subversive	31/2
activities	31/2
政治学	31/5
辞典	31/3

表 11 分類記号ごとの語の出現数例

出現数	分類記号
9	29
5	31
4	45

植物辞典によると、胡麻の栽培は東インドが原産であり、そこから中国、インド、アフリカに広がっていった。



KAKASIで分割

植物辞典によると、胡麻の栽培は東インドが原産であり、そこから中国、インド、アフリカに広がっていった。

図 4 KAKASI による分割例

植物辞典によると、胡麻の栽培は東インドが原産であり、そこから中国、インド、アフリカに広がっていった。

分野判定辞書からマッチした語を取り出す  
辞典 31/3  
インド 29/5,45/3  
インド 29/5,45/3

図 5 辞書とのマッチング

分類記号 31 出現数 5  
分類記号 29 出現数 9  
分類記号 45 出現数 4

辞典 31/3  $\sqrt{5/3} \leq 15$  分野として利用  
インド 29/5  $\sqrt{9/5} \leq 15$  分野として利用  
インド 45/3  $\sqrt{4/3} \leq 15$  分野として利用

図 6 分類記号のスコアとして採用するか判断

する。(図 7)

- (5) 分類記号のスコアを持っている場合にこの文章の分野とする。  
教科書の段落ごとに事前に分野判定を行い、段落番号と

分類記号を保存する。

分野判定検索機能は、電子教科書の本文からのリンクで渡されたキーワードと段落番号から、キーワードを検索

辞典	31/3	31のscoreに3/49を加算
インド	29/5	29のscoreに5/45を加算
インド	45/3	45のscoreに3/44を加算
最終的に		
		31のscore=3/49
		29のscore=10/45
		45のscore=6/44

図 7 スコアの計算

エンジンに渡す。検索エンジンから得られた結果は 50 件までを利用する。検索結果には、Web ページのタイトル、URL、スニペットなど含まれる。このうち、スニペットのみを分野判定し、得られた分類記号と分類記号のスコアを、教科書の段落と一致したものを一致度のスコアとして加算し、順位付けに利用する。最終的に順位の高いものから最大 10 件を分野判定検索の結果として提供する。

### 3.4 隣接分野の利用

ここまでで実装した分野判定検索機能を利用し、予備実験を行った。この予備実験では、小学校 6 年生向けの理科の教科書を利用した。予備実験のうちの一つとして食物連鎖というキーワードで検索し、得られた結果を確認した。この結果では、教科書の該当段落の内容と検索結果のスニペットがほぼ同じ内容であったが、分野判定の結果が異なった。分野判定結果は、教科書の該当段落は植物学、スニペットは動物学や医学・薬学などになった。そのため表示されていないことに気がついた。そこで、我々は隣接する分野というものが存在するのではないかと考えた。

最初、分類記号 (NDC9) のうち、大分類を利用することを考えた。しかし、あまりにも範囲が大きくなりすぎるために今回は利用しないこととした。

そして、我々が着目したのは、分野判定辞書を作成中に得られた、語と分類記号の対 (表 12) である。同じ語が複数の分類記号にまたがっている場合には、それらが近い分野となるのではないかと考えた。ここから 2 つずつの組み合わせを表 13 のように作成し、分類記号の組み合わせの出現回数を調べた。ただし、一つの語が 10 種以上の分類記号にまたがっている場合には利用しないこととした。この組み合わせが多い順番に表 14 のように並べ、これを隣接分野候補とした。この隣接分野候補に合致した場合には、分野判定検索時のスコアの計算を  $1/2$  と設定した。

表 12 語と複数分類記号

語	分類記号
アフガニスタン	22,29,31,38,39,46,59,92
インド	02,12,16,22,28,29,33,36,38,45,46
破壊活動	31,32
戦記	22,38,39,46,78
攻撃	22,31,39

表 13 分類記号の組み合わせ

組み合わせ			
22:29	22:31	22:38	22:39
22:46	22:59	22:92	29:31
29:38	29:39	29:46	29:59
29:92	31:38	31:39	31:46
31:59	31:92	38:39	38:46
38:59	38:92	31:32	22:38
22:39	22:46	22:78	38:39
38:46	38:78	39:46	39:78
46:78	22:31	22:39	31:39

表 14 隣接分類記号候補

回数	組み合わせ	回数	組み合わせ
3	22:39	1	29:59
3	31:39	1	29:92
2	22:31	1	31:32
2	22:38	1	31:38
2	22:46	1	31:46
2	38:39	1	31:59
2	38:46	1	31:92
1	22:29	1	38:59
1	22:59	1	38:92
1	22:78	1	38:78
1	22:92	1	39:46
1	29:31	1	39:78
1	29:38	1	46:78
1	29:46		

## 4. 分野判定検索機能の評価実験

分野判定検索機能を利用して、教科情報の中の科目「社会と情報」の教科書から検索を行った。キーワードは、教科書で設定されている索引語 341 個から選んだ。選定基準は、通常の Google 検索の結果に正しいものが上位 3 件に 1 つ以下のみのものとした。ここから 22 個のキーワードを選定し、Google 検索と分野判定検索機能とを比較した。22 個のキーワードは表 15 である。

まず最初に隣接分野候補のうち、上位何件まで利用すればよいかを調べた。この 22 個のキーワードと教科書の文章を利用してそれぞれ検索した結果から、比較的正しい結果が上位に来るような隣接分野候補の順位について調べた。隣接分野候補を利用しない、上位 10 位まで、上位 20 位まで、と順に調べ、最終的に 170 位まで 10 位刻みで調べてみた。隣接分野候補を幅広く利用すると、元の文章から離れたものも検索結果の候補として上がってきた。また、少なすぎるとあまり効果が見られなかった。今回利用した 22 個のキーワードでは 50 位まで利用することで検索結果の改善が見られた。そのため、ここでの評価には隣接分野候補を 50 位まで利用することとする。

今回利用した 22 個のキーワードを、隣接分野候補の 50 位まで利用して検索した結果、Google に比べて少し改善さ



表 15 キーワード候補

キーワード	意味
TB	容量を示すテラバイト
PB	容量を示すペタバイト
MOV	動画形式
WAVE	音声形式
CMC	コンピュータを介したコミュニケーション
AP	無線 LAN のアクセスポイント
WWW	World Wide Web の略
アクセス	セキュリティに関するアクセス権限
アニメーション	動画の意味でのアニメーション
家出サイト	家出に関する情報をやりとりする掲示板など
カット	動画編集時のカット
国名	ドメインの国名
検索	Web 検索
残像	動画におけるコマとコマの間の残像
自殺サイト	自殺情報のやりとりをする掲示板など
タイムライン	動画編集時のタイムライン
セル	表計算ソフトの一つの箱
ロゴタイプ	文字を装飾や図案化したもの
ドリー	動画撮影時に使われる追従用の車
パン	写真や動画撮影時のカメラの左右の動き
バイト	情報量の単位
flame	掲示板などで発生する言い争い

れたもの、またはほぼ等価であるものがほとんどであった。しかし、ロゴタイプ、バイト、WWW の 3 つのキーワードについては、大幅な改善が見られた。この 3 つのキーワードについて授業で利用できるかどうかを検証するべく、中学校技術科または高等学校情報科を担当している先生 3 名に、結果について次のように確認してもらった。

- (1) 3 つのキーワードごとに、Google 検索で得られた各 10 個の検索結果について、それぞれ教科書の文章と比較して内容が正しいかどうか。
- (2) 3 つのキーワードごとに、分野判定検索機能を利用して得られた各 10 個の検索結果について、それぞれ教科書の文章と比較して内容が正しいかどうか。

確認してもらった結果についてまとめる。各表で A-C は回答者を表す。まず Google の検索結果についての回答は表 16 である。次に分野判定検索機能を利用した検索結果についての回答は表 17 である。

表 16 Google 検索結果が正しい数

キーワード	A	B	C
ロゴタイプ	1	1	2
バイト	1	1	1
WWW	1	1	1

検索の評価に利用される適合率  $P$  は、システムが提示した結果のうち、正解の比率を表す。正解数を  $tp$ 、不正解数を  $fp$  とした場合、適合率は式 1 として表すことができる。Google 検索と分野判定検索機能について、各教員の回答

表 17 分野判定検索機能の結果が正しい数

キーワード	A	B	C
ロゴタイプ	1	6	6
バイト	4	5	5
WWW	1	1	1

からそれぞれ適合率を計算し、平均をまとめたものを表 18 にまとめた。

$$P = \frac{tp}{tp + fp} \quad (1)$$

表 18 適合率の比較

キーワード	Google	分野判定
ロゴタイプ	0.13	0.43
バイト	0.10	0.47
WWW	0.10	0.10

表 18 から、これら 3 つのキーワードのうち、ロゴタイプとバイトについては大幅に改善されている。しかし、WWW については、Google と同程度という結果が得られた。それぞれのキーワードについて考察する。

ロゴタイプは競走馬に同じ名前がある。そのため、Google 検索ではこのロゴタイプという馬に関する情報が上位に現れてしまい、今回のようにデザイン関連のロゴタイプは下位に隠れていた。それらを分野判定検索機能では引き上げることに成功している。

次にバイトについては、Google 検索ではアルバイト情報が上位にくる。これは一般的にアルバイト情報を検索することが多いためと推測される。これも下位に隠れていたサイトを引き上げることができた。

最後に WWW は、Google 検索では WWW を含むもの、ということでアドレス (URL) に WWW が含まれるもののうち、より検索結果が上位に来るように設定されたサイトが表示されていた。分野判定検索機能では定義をより上位に移動させることに成功はしている、しかし、全体として語としては分野が正しくなっているが、教科書の内容とは異なるものが多くなってしまった。Google 検索では、50 件中 49 件が英語のサイトであり、日本語の教科書の文章と英語とでうまく適合できなかったのではないかと推測される。高校教員からの意見として、英語のサイトは授業では使いにくい、という意見も出ていた。

単純なキーワード検索のみであれば Google は非常に優秀な結果を返す。実際、341 個の索引語のうち 280 個程度の索引語では関連するサイトが上位に 2 つ以上表示されていた。しかし、教科書の文脈を考慮する必要がある場合には、Google のランキングで 11 位以降にあるサイトが有用になることがあり、分野判定検索機能ではこの 11 位以降に隠れてしまったサイトを抜き出すことに成功している、と言える。

分野判定検索機能でも適合したサイトを取り出せないものもある。今回とりあげた 22 個の索引語のうち、19 個については Google 検索に比べて改善が見られなかった、と述べたが、これら以外の 40 個程度の索引語でも同様に改善が見られなかった。例えば、計算システムや通信システムのように会社名の一部として使われている場合、会社の事業説明が関連する語を含むために分野が一致してしまうが、内容が異なっている。インタビューやドキュメントは、芸能人のインタビューやドキュメンタリー番組なども内容を表す文章に含まれる語により分野が一致するが、内容が異なる。このように分野判定では判別できない例も存在している。

## 5. おわりに

今回提案した、文脈を考慮した検索である分野判定検索機能は、特定の語について効果があることがわかり、さらに学校の授業でも利用できることがわかった。しかし、分野判定としては一致するが、本来調べたい内容と異なるサイトが存在し、そのために正しい結果が得られないこともわかった。また、中学校や高等学校には英語で書かれている場合には、内容が一致していても使えないことがあることもわかった。

今回の分野判定検索機能では、処理の簡略化のために段落の分野判定時に得られたスコアを利用しておらず、隣接分野候補についても出現回数を考慮していない。スニペットの分野判定を行う際にこれらのスコアを利用することにより、順位がどう変化し、正しい結果が得られるかを検証する必要がある。

今後の課題として、分野が一致するが内容として正しくないサイトや、教科書の内容に対して難易度が高いサイトなどについて、多くの例を集めて対処方法を考えると共に、他の判定方法を組み合わせることにより、精度が向上できないかを検証する。また、正しい結果の網羅率に相当する再現率は、50 件の候補から検証する必要があるため、今回は見送った。今後、再現率についても調査する必要がある。

今回は教科情報の社会と情報に限定して調査したが、他の教科、科目についても検証していくことも考えている。

## 参考文献

- [1] 西原陽子, 砂山渡, 谷内田正彦: Web ページの難易度と学習順序に基づく情報理解支援システム, 電子情報通信学会論文誌, Vol.J89-D, No.9, pp.1963-1975, 2006.
- [2] 森幹彦, 山田誠二: ブックマークエージェント: ブックマークの共有による情報検索の支援, 電子情報通信学会論文誌, Vol.J83-D1, No.5, pp.487-494, 2000.
- [3] 大久保雅且, 杉崎正之, 井上孝史, 田中一男: WWW 検索ログに基づく情報ニーズの抽出, 情報処理学会論文誌, Vol.39, No.7, pp.2250-2258, 1998.
- [4] 折原大, 内海彰: HTML タグを用いた Web ページのクラスタリング手法, 情報処理学会論文誌, No.49, Vol.8,

- pp.2910-2921, 2000.
- [5] 仲川こころ, 高田喜朗, 関浩之, 可変なカテゴリ構造を用いた文書検索支援手法, 情報処理学会論文誌 Vol.42, No.10, pp.2441-2453, 2001.
- [6] 一般社団法人日本教育情報化振興会: デジタル教科書の位置づけと機能, <http://www.japet.or.jp/jog7cjygu-403/>, 2010.
- [7] 国立国会図書館: 書誌データ作成ツール. <http://www.ndl.go.jp/jp/library/data/bunruikenmei.html>
- [8] Christopher D.Manning, Prabhakar Raghavan, Hinrich Schutze 著, 岩野和生, 黒川利明, 濱田誠司, 村上明子 訳: 情報検索の基礎, 共立出版, 2012.
- [9] Google Inc.: Google Developers Custom Search, <https://developers.google.com/custom-search/>, 2013.
- [10] Oracle Corporation: MySQL 5.1 Reference Manual, <http://dev.mysql.com/doc/refman/5.1/en/fulltext-stopwords.html>
- [11] KAKASI project: KAKASI - 漢字→かな (ローマ字) 変換プログラム, <http://kakasi.namazu.org/index.html.ja>, 2014.