

Web 検索による余談抽出に基づく穴埋め問題生成システム

大貫颯[†] 北山大輔[†]

[†] 工学院大学情報学部

1 はじめに

現在、書店では受験対策や定期テスト対策、模試対策に関係する書籍が数多くならんでいる。しかし参考書は必要な知識を書いているだけのものが多い。一方、問題集に関しては単元別に問題が書いてあり、解答や解説が詳しく書いてあるものもある。しかし、問題集は学生自らが学習意欲を持っていなければ解かない。すなわち、これらは学生自らが楽しく学習したくなるような問題ではないといえる。問題を実際に解いたとしてもただ問題を解いているだけなのであれば学習内容が印象に残りににくく学習効果が得られない。そこで本研究では、学習内容に関係はするが、本筋ではない余談に着目する。授業中などに学生の興味を引くために行われる余談は、印象に残りやすく、学習内容が余談に紐付けられて記憶されることも多いと考える。上記を踏まえ、我々は学習意欲の向上を目的とした、Web 検索を用いた余談抽出手法を提案し、余談に基づく穴埋め問題生成システムを開発する。具体的には、余談抽出としてまず学習しようとしている単元における重要語を抽出する。その重要語を用いて Web 検索を行う。検索結果において学習しようとしている内容と一番類似性の低いものを抽出する。そのことにより、内容に関連付けられているが教科書や参考書には載っていない内容を抽出することができる。

提案手法の概念図を図 1 に示す。図 1 の右部における各ページをつなぐ線に関しては類似度の値を示しており、線の太さが太いほど類似度の値が大きいことを表している。

2 関連研究

学生が学習内容を記憶する場面に関する研究を紹介する。井戸坂 は一人で学習するワークシートより、グループで学習するゲームに人気が集まる傾向がある

A Generation System of Fill-in-the-blank Questions based on Extraction Attractive Topic using Web Search
Hayate OHNUKI[†] and Daisuke KITAYAMA[†]

[†] Information Science, Kogakuin University
163-8677, Tokyo, Japan

{j114022@ns, kitayama@cc}.kogakuin.ac.jp

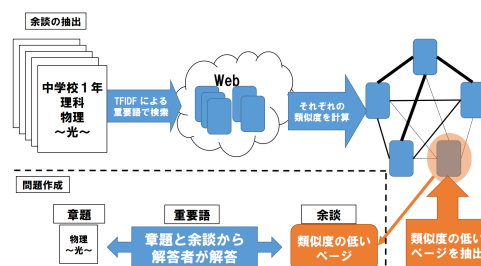


図 1: システムの全体構造

と主張している [1]。また小森は直接実験や観察などで体感することにより中期的な記憶がつきやすいと主張している [2]。本研究では学生が学習内容を記憶ためには学習内容に関連する余談や付加知識などがあることにより学習内容を印象付けることができ記憶に残りやすくなるという仮説を立てた。このような、余談と学習に関する研究は、我々の知る限り行われておらず、学習内容に対する補足内容の関連度と記憶の定着に関して明らかにすることも、本研究の目的の 1 つである。

3 余談に基づく穴埋め問題生成システム

3.1 学習対象データ

本稿では中学校の理科の内容を対象とする学習内容とした。中学校の学習内容を学年と分野に分割し 12 単元での学習内容においてシステムを構築する。数学を除外しているが、本システムはテキストベースの処理となるため、数学では数式や図形が多く重要語が判断しにくいと考えたからである。今回使用するデータは市販されている中学理科の参考書を OCR により読み取り、テキストデータ化したものを用いた。参考書によっては学年別になっていないものもあるが、そのような場合著者らが判断し、分類した。そのため今回は各参考書ごとの節を 1 つの授業と仮定して分析を行った。

今回使用した対象は旺文社中学総合的研究理科、学研パーフェクトコース理科、塾講師が公開！中学入試塾技 100 理科の 3 種類である。この OCR の結果には誤認識や脱字が多く発生する。しかし本研究では単語

レベルでの統計処理であるため、問題にならないと判断した。

3.2 重要語の判定

今回の対象データから単語を抽出するために形態素解析器の MeCab を使用した。

節別に分かち書きされたテキストデータを文章ごとに分割し重要語を判定する。判定方法としては TFIDF 法を用いて行う。1つの授業ごとに分割された節のテキスト i に対する単語 j の出現回数を $TF_{i,j}$ 、単語 j が出現する節の数を DF_j 、すべての節の数を $|D|$ とした時、単語のテキスト内における重みを、式 1 に定義する。各節ごとに、TFIDF 値上位 3 単語を重要語として用いる。

$$weight_{i,j} = TF_{i,j} \times IDF_j \quad (1)$$

$$IDF_j = \log_2\left(\frac{|D|}{DF_j}\right) \quad (2)$$

3.3 Web 検索を用いた余談ページの抽出

3.2 節で判定した重要語を Web 検索エンジンに入力し、Web ページ集合を得る。本稿では、Web 検索エンジンは Google を使用する。今回使用する重要語は TFIDF で重要語と判断された上位 3 単語を用いて Web 検索を行う。単語ごとに検索結果の上位 15 ページのテキストデータを取得する。この取得した Web ページを用いて学習内容に関連した余談となる Web ページを出力する。このテキストデータすべてを TFIDF 法における文章とし文書ベクトルを生成する。この文章ベクトルを用いて、各ページの類似度を測定する。類似度の計算方法はコサイン類似度を使用する。類似度は 15 ページ全体に対し各ページ類似度を出力させる。この中で全体に全体の文章に対してもっとも似ていない文章を出力させる。Web ページ検索における上位 15 ページはある程度学習内容に関連したページが出てくることが予想される。しかしこの文章全体におけるもっとも類似度の低いページは内容に直接的には関連はしていないが間接的に関連することが期待される。これを余談内容の Web ページとして抽出する。

最も学習内容に関連した余談を判断する方法としてコサイン類似度を用いて階層クラスタリングを用いる。階層的クラスタリングのデンドログラムにおいて階層が高いところから出発し次の分岐点までの距離長い方の分岐している道をたどる。この方法を最下点になるまで行う。また最後 2 ページの分岐になった場合は最後の分岐から分かれた 2 ページにおいてランダムに余談のある Web ページとして出力を行う。

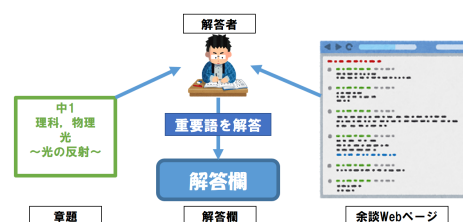


図 2: 生成した問題

3.4 余談に基づく穴埋め問題生成

今回作成する問題の想定としては授業後に復習する際のものを想定して作成する。問題の形式は章題と余談となる Web ページを表示する。解答者は章題と余談となる Web ページの内容から共通している重要語を解答する形式である。生成した問題の略図を図 2 に示す。

例えば、節の内容を中学校 1 年理科、物理、光の性質の反射と屈折に指定した際には重要語は“屈折角”、“反射角”、“反射の法則”と判別した。問題を生成したとき“反射角”ではビリヤードを用いたの反射の説明の Web ページが出題された。このページは本研究での余談に当てはまると考えることができる。しかし、“屈折角”の時は Web 上の百科事典の内容が出力されるなど、本研究の余談にはあまりそぐわないものも出力された。

4 おわりに

学習者が自ら学習するための問題生成システムを構築した。今回のシステムは学習する際学習した内容を印象付けるために余談という学習の本筋とは異なるものを提示することで学習者の意欲を上げようとするものであった。このシステムでは余談となる Web ページが入力した節の内容とさほど変わらないことがあったので検討が必要である。

参考文献

- [1] 井戸坂 幸男, 兼宗 進, 久野 靖, 中学校におけるコンピュータを使わない情報教育 (アンプラグド) の評価情報処理学会論文誌, No. 1, Vol.13, pp.49-56 (1988)
- [2] 小森章史, 実感を伴った理解を図る理科学習指導, 北海道立教育研究所附属理科教育センター, 研究紀要 第 25 号, pp.70-72, (2013)