

電子書籍立ち読み支援のためのレビューを利用した興味文抽出に関する研究

A Study on Extraction of Sentences Arousing Interest Based on Reviews for Supporting to Browse Electronic Books

1DS07206K 村井聡一 MURAI Soichi

1. はじめに

近年、書籍の電子化が進んでおり、膨大な電子書籍の中から、読者にとって価値のある書籍を、効率的に発見するシステムの必要性が高まっている。これまで、この目的のために、検索や推薦による書籍の絞り込みに関する研究が行われてきた[1, 2]。しかし、読者は絞り込まれた複数の本に対して、何らかの手法で個々の本の価値を判断し、「読む」か「読まない」か、を判断する必要がある。我々は、その判断の際に、現実の書店で行われているような、主体的な立ち読みが有効だと考えた。本研究では、電子書籍を対象とした、小説の立ち読み支援手法を提案する。利用者が、読む価値のある小説を選別する際に、利用者が興味を引くと思われる文章を提示することで、効率的な立ち読みを支援する。

2. 興味を喚起するキーワード

本手法では、「利用者の興味を喚起する単語を含む文章は、利用者の興味を喚起する文章である」という仮定に基づき、読者が興味を引くと思われるキーワードを自動的に抽出し、それに基づいて文章の興味喚起度を求める。

利用者の興味を引く単語を見つけるために Web 上のレビューを利用する。レビューには、読者が書籍を読んだ感想が含まれており、そこには読書中に興味を喚起された場面や人物に関する単語も含まれている。レビューワーが興味を感じた対象は、他の読者の興味を引くことがあると考え、書籍とレビューに出現する単語を図1のように分類した。『レビューが対象とする書籍中の文章は、他の読者の興味を引く可能性が高い』と仮定し、書籍の本文とその書籍のレビューの両方に出現する、名詞、動詞、形容詞をキーワードの候補とする。キーワード候補に関して、対象とする書籍のレビューにおける出現数が多く、他の書籍のレビューの出現数が少ない単語は、その書籍の特徴的な単語であると考え、キーワードとする。

		書籍本文に出現する単語	
		する	しない
レビューに出現する単語	する	興味を喚起する単語 (キーワード)	・背景情報 ・レビュー特有の表現
	しない	興味を喚起しない単語	

図1 キーワード候補となる単語

3. 登場人物に代わるキーワード

物語の主要な登場人物名は出現頻度が高く、キーワードとして抽出されることが多い。しかし、登場人物名自体は読者の興味を喚起しないと考えられる。そこで、登場人物名はキーワードにはならないが、登場人物と密接にかかわる単語は興味を喚起する可能性が高いと考え、登場人物名の代わりにキーワードとして利用する。登場人物名と関連する単語には、物語全体を通して関連する単語と、ある場面においてのみ関連する単語、の二種類が考えられる。

レビューにおいて、登場人物と同一の文に出現する度合いが高い単語を、物語全体を通じた人物像に関するキーワードとし、登場人物の代わりに用いる。

書籍本文において、登場人物と同じ場面に出現する度合いが低く、登場人物が出現しない場面に出現する頻度が低い単語を、ある場面に特徴的なキーワードとして用いる。このキーワードは、他のキーワードとは一緒に用いず、特徴的な場面中の文章を提示するために用いる。

なお、登場人物の識別は人手で行うことを前提とする。

4. 文章の興味喚起度の計算

抽出したキーワードに基づいて、文章が利用者の興味を喚起する程度を推定する。 i 文目を中心とした $i-1 \leq i \leq i+w$ の範囲を一つの文章とする。 i 文目の理解を補う情報は前文にある、と考え、 $i-1$ 文目も文章に入れた。文章の興味喚起度は、文章に含まれる文に出現するキーワードの重みを足し合わせた値に、文章中に出現するキーワードの種類をかけ合わせることで決定する。 i 文目から離れて出現するキーワードの重みを軽くして扱うためにハニング窓関数を用いた[3]。

5. 評価

実験により提案手法の評価を行った。複数の被験者に、提案手法によって提示した文章と、ランダムに提示した文章に関して、それぞれの文章が「書籍の特徴が表れているか」について3段階で評価してもらった。実験結果により、提案手法の有効性が統計的に示された。

[1] 見並史彬, 小林幹門, 伊藤孝行, “概念辞書を利用した目的指向書籍推薦システムの試作”, 日本ソフトウェア科学会第24回大会(2007年度)論文集

[2] 丸川雄三, 阿辺川武, “横断的連想検索サービス「想一IMAG-INE」データベース連携が拓く新たな可能性”, 情報管理, Vol. 53(2010), No. 4, pp. 198-204

[3] 黒橋禎夫, 白木伸征, 長尾眞, “出現密度分布を用いた語の重要説明箇所の特定”, 情報処理学会論文誌, Vol. 38, No. 4, pp. 845-854, 1997