6M-01

嗜好抽出を目的とした電子書籍へのアノテーションの分析

鈴木　啓史†　　松村　敦‡　　宇陀　則彦‡

筑波大学 情報学群 知識情報・図書館学類†　　筑波大学 図書館情報メディア系‡

1. はじめに

インターネットの普及にともない、誰もが大量の情報を入手できるようになった一方で、その中から自分が欲しい情報を自力で探すことが難しくなった。そのため、ユーザの嗜好に基づいて必要なアイテムを推定し提示する情報推薦システムが注目されている。

情報推薦システムにおいて、ユーザの好みに合ったアイテムを推薦するためには嗜好抽出が重要な技術となっている[1]。嗜好を抽出する方法の 1 つに、コンテンツに付与された情報を利用する方法がある[2]。しかし、この方法では、ユーザがコンテンツに多数の情報を付与するため、ノイズとなる情報が生成されてしまうという問題がある。これに対して、コンテンツの一部をそのまま嗜好として利用する方法がある。この方法では、精度の高い推薦が実現できることが真野らの楽曲推薦システムにより明らかになっている[3]。そこで本研究では、コンテンツの一部をそのまま嗜好として利用する方法を電子書籍の推薦に適用することを考えた。コンテンツの一部を取り出す方法として読書中に行うアノテーションに着目し、電子書籍で自然に読書をしながら嗜好抽出することを目指す。

本研究の目的は、電子書籍の推薦における嗜好抽出手法として、電子書籍へのアノテーションが利用可能かどうかを実験的に検証することである。

2. 関連研究

松岡らは論文概要の文書内の重要な箇所に赤・青・緑の 3 色を使って下線を付与することが出来るシステムを開発し、アノテーションの分析を行った[5]。その結果、ユーザが引いた下線には文章内の重要語を多く含むことが明らかになった。しかし、秋山らの研究[4]によると、なぜアノテーションをするのかに関しては「後から読むときにすぐに分かるようにするため」「ページに書かれている内容について理解しやすくするため」「文章の内容が難しかったため」「重要な箇所だったため」など、さまざまな理由が挙げられる。本研究では、自然なアノテーションからの嗜好抽出を目指している。したがって、松岡らが実験で指定した重要な箇所だけではなく、ユーザのこれまでの経験に基づいてアノテーションを行ってもらい、その理由によって嗜好抽出の精度に差が出るかどうかを検証していく。

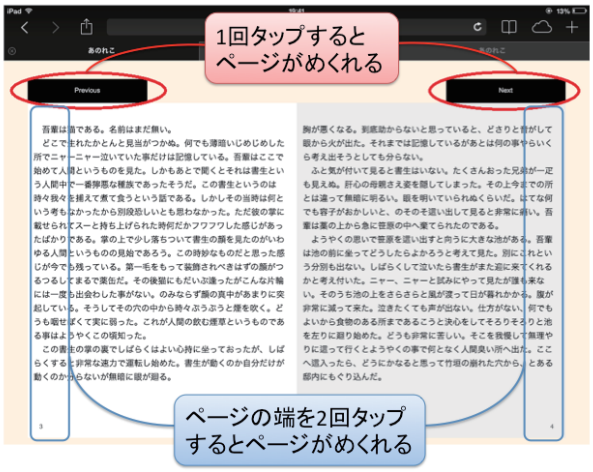

図 1 実験システム「あのれこ」

3. 実験システム

実際に電子書籍へのアノテーションを収集し実験で利用するために、アノテーションを記録・保存できる Web アプリケーション「あのれこ」を開発した。図 1 はあのれこの画面である。HTML5, Javascript, jQuery を利用して開発した。「あのれこ」の機能は「ページめくり機能」と「アノテーション機能」の 2 種類がある。ページめくり機能は、現在閲覧しているページから次のページや前のページに移動する時に使用する機能である。ページめくりの方法は 2 つある。1 つ目の方法は、画面右上と左上に表示されている「Next」と「Previous」のボタンを 1 回タップする方法である。2 つ目の方法はページの端を 2 回タップする方法である。アノテーション機能は図 2 に示したように、アノテーションをしたい文字列を指定してボタンをタップすることによって、指定した文字列を記録することができる機能である。

Analysis of annotations to the e-book for preference extraction
†Keishi Suzuki, College of Knowledge and Library Sciences, School of Informatics, University of Tsukuba
‡Atsushi Matsumura, Norihiko Uda, Faculty of Library, Information and Media Science, University of Tsukuba

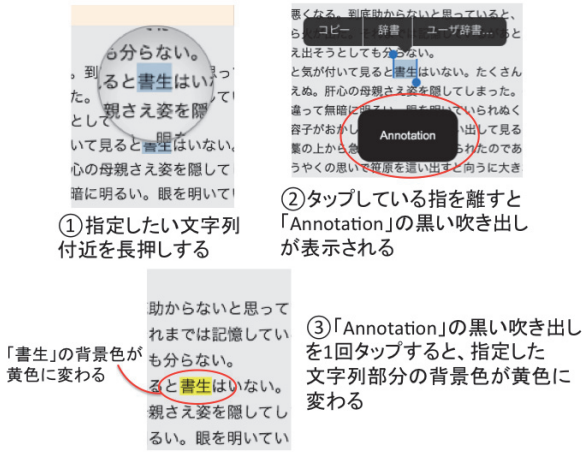

図 2 アノテーション機能の使い方

4. 実験

嗜好抽出手法としてアノテーションが利用できるかを検証するため、書籍に書き込みを行った経験のある 12 名の大学生を対象に実験を行った。まず、「あのれこ」を使用してそれぞれ 2 冊の電子書籍に対してアノテーションをしてもらった。その後、アノテーションされた文字列中の単語、本文全体からランダムに抽出された単語、tfidf により抽出された単語それぞれに対して興味のある単語を実験参加者に判定してもらった。また、実験参加者が付与した各アノテーションについて、アノテーションをした理由を尋ねた。

5. 実験結果

表 1 にアノテーション、ランダム、tfidf により抽出された総単語数、興味関心のある単語の数とその精度を示す。これにより、アノテーションから単語を抽出する手法の方が、ランダムや tfidf による単語抽出手法よりも、興味関心のある単語を多く抽出できることがわかった。なお、ランダムと tfidf の総単語数は各実験参加者のアノテーションによる抽出総単語数に揃えたため、同数となっている。

一方、付与されたアノテーションの長さを見ると、全 424 アノテーションのうち、1 単語のアノテーションは 50 個だけであり、2 単語以上のテキストを対象に付与されることが多かった。したがって、興味関心のある単語をそのまま抽出することは難しいことが明らかになった。

表 2 にアノテーションの理由による興味関心のある単語の数と精度の分布を示す。6 つの理由の精度で多重分析を行った結果、どの理由においても有意な差がみられなかった。よって、アノテーションの理由によって興味関心のある単語の抽出数に影響が出ないことが示唆された。

表 1 興味関心のある単語の抽出数と精度

	アノテーション	ランダム	tfidf
総単語数	3500	3500	3500
興味語数	441	217	228
精度	13%	6%	7%

表 2 アノテーションの理由による興味語数と精度

アノテーションの理由	後で読む	内容理解	興味関心	重要	疑問	その他
抽出単語数	478	1347	676	952	38	9
興味語数	58	160	129	91	3	0
精度	12%	12%	19%	10%	8%	0%

6. おわりに

以上の結果から、アノテーションによって興味関心のある単語を抽出できる可能性があると考えられる。しかし、実験参加者が付与したアノテーションの多くは 2 語以上のテキストであり、どの単語に興味関心を持ったかを推定することが難しいという問題が残っている。推薦システムとして運用していくためには、興味関心のある単語を適切に抽出する必要がある。今後の課題は、システムのアノテーション機能の改良、キーワード単位で興味関心のある箇所にアノテーションをする実験の実施が挙げられる。

参考文献

[1] 土方嘉徳. 嗜好抽出と情報推薦技術. 情報処理, 2007, Vol.48, No.9, p.957-965.
[2] 田上道士, 山場久昭, 高塚佳代子, 岡崎直宣, 冨田重幸. ユーザが抱く印象を用いた個人志向の表現 -Folksonomy を利用したユーザの印象の推測方法-. 宮崎大學工學部紀要, 2014, Vol.43, p.263-272, http://hdl.handle.net/10458/5011, (参照 2014/12/25).
[3] 真野洋平, 高田俊弘, 齋藤洋典. 時間区間アノテーションの集約に基づくダイジェストを対象とする楽曲推薦システム. 情報処理学会研究報告, 2014, Vol.31, No.70, p.1-6.
[4] 秋山博紀. 安村通晃. アノテーション付加による知識共有型電子書籍の提案, 情報処理学会研究報告, 2011, Vol.142, No.13, p.1-8.
[5] 松岡有希, 坂本竜基, 中田豊久, 伊藤禎宣, 武田英明. 論文概要に対する色付きアンダーライン付与システムの運用・分析. 電子情報通信学会第 17 回データ工学ワークショップ, 2006.