Keyword extraction method using users' mouse behavior

CHUNYANG HE[1] MASAO TAKAKU[2]

Abstract: Owing to the explosive growth of information, keywords play an essential role in summarizing information and helping search effectively. Existing keyword extraction approaches are mostly focused on the document side, instead of using reading side feedback. In this paper, we proposed a keyword extraction method that incorporates the mouse pointer behavior of the reader when browsing academic papers and conducted an experiment to verify the effectiveness of the proposed method. We developed a mouse tracker to record mouse trajectory, speed, and click behaviors when the participants browsed academic papers. Using a predefined weighting algorithm, a term-weighted ranking was created according to mouse features. We used the term frequency-inverse document frequency (TF-IDF) and TextRank methods as the baseline to compare the effectiveness, and evaluation was performed based on precision, recall, and F-score. The experimental results show that the proposed method outperforms the TextRank algorithm, but there are no significant differences between the proposed method and the TF-IDF algorithm.

Keywords: Academic information, keyword extraction, cursor movement, interactive information retrieval

1. Introduction

With the rapid development of the Internet, academic information is distributed on the web as an electronic form instead of being recorded using books as a medium. Searching for academic papers without keywords is quite a difficult task; keywords can help search engines obtain the most relevant document that a user look for.

However, because it has become increasingly practically impossible to manually assign keywords to documents, automatic keyword extraction technology has recently become popular in the research field of text mining in recent years.

From past approaches based on statistical features, such as term frequency-inverse document frequency (TF-IDF)[1] and BM25[2], to recent machine learning approaches[3][4] that have emerged due to the development of the natural language processing technology, there are many ways to automatically extract keywords. However, these approaches mostly focus on the features in the document itself, and the use of information through behavior analysis of the reader has not yet been considered.

In the field of information retrieval, there are many studies that use users' behavior and feedback to infer their information needs or interests[5]. For example, eye movement data are a type of implicit feedback that efficiently reflects users' intentions and interests, and it is often used for document relevance judgment or users' interest prediction [6][7]. However, the price of the eyeball tracking device is quite expensive, and it is difficult and still not available to collect eye-tracking data in real-world settings. On the other hand, some researchers have conducted an experiments to estimate sight of the user through the movement of the user's mouse, it turns out that the results achieved a high degree of accuracy about 70%[8], this discovery proves that the mouse movement can reflect the user's sight to a considerable extent.

1 University of tsukuba, Tsukuba, Ibaraki 305-0821, Japan
2 University of tsukuba, Tsukuba, Ibaraki 305-0821, Japan

Therefore, in this study, we try to use mouse movement data to replace eye movement data since it can be easily obtained a large amount of data.

This study aims to propose a method that can extract the reader's interest to infer keywords using the users' mouse behavior when they are reading an academic paper and then verify its effectiveness. Considering the academic paper reading situation, readers have to be more focused because there is a great amount of terminology and information that needs to be digested. We inferred that, readers would tend to trace the reading part of the text with pointer more often than the usual reading situation, and the mouse movement and behavior (such as moving slowly or hovering) would be a good indicator to deduce if readers are reading carefully. For this point, we proposed an algorithm to calculate the weight of the corresponding words. In addition, we will explore the impact on proposed method when combining data from multiple users on each single academic paper.

Based on the above, we address the following research questions:

RQ1: Can we use readers' mouse behavior to extract keywords?

RQ2: Can the proposed method be more effective than the baseline methods?

RQ3: Can the proposed method be effective for those who do not use a pointer to trace the reading part of the text?

RQ4: Whether the effectiveness of the proposed method can be improved by combining multiple user data?

2. Related Works

According to several keyword extraction survey papers, keyword extraction technology can be roughly categorized into two types: extraction type and generation type. Furthermore, based on the characteristics of the approaches, they can mainly be classified into four classes, namely, linguistic approaches, statistical approaches, machine learning approaches, and hybrid approaches[8][9]. Statistical approaches are generally based on statistical features derived from the non-linguistic features of

documents, which are quite simple but effective[10]. Linguistic approaches are generally rule-based and derived from linguistic knowledge/features[11]. Machine learning approaches can be divided into supervised and unsupervised learning approaches: supervised learning generally corresponds to keyword generation, whereas unsupervised learning generally corresponds to keyword extraction[12][13]. Hybrid approaches combine each of the above methods or use heuristics such as html tags[14]. Our proposed method uses the reading side's feedback to extract keywords, which can be categorized into a hybrid approach.

Kantor et al.[15] reported that they found that users tend to follow the mouse pointer through the eye while browsing webpages. As a promising candidate for user behavior analysis, mouse trajectories are regarded as important as data that can reflect user's sight and have been used in research such as user interest estimation and search satisfaction estimation. Hijikata et al. extracted the corresponding words that may infer the user's interest from the four mouse behaviors, such as link pointing, link click, tracing, and text selection. By comparing them with the baseline, the results showed that the precision was 1.4 times higher than the baseline[16].

Hijikata et al. showed that the mouse behavior can reflect the user's interest to a certain extent; however, they also reported that users do not strictly trace the line when reading, but simply unconsciously move the mouse pointer to the right, and most movements are performed at a short distance. This finding also pointed to a problem that in the usual reading situation, mouse movement may not reflect the user's sight correctly. Hienert et al. conducted a comparison between queries submitted by users and words extracted by the user's mouse behavior on the search result page. they found that for those terms with a long mouse dwell time often appear in later search sessions which imply that these terms are indicators for user interests[17].

This study mainly uses the features of mouse tracing and text selection based on the research of Hijikata et al. Furthermore, we combine the moving speed and the number of occurrences of terms. The keywords of the study are extracted using the document feature and mouse movement behavior of the reader when they read the academic paper.

3. Methods

3.1 System overview

In this study, we mainly focus on two mouse behaviors: text tracing and text selection. Text tracing features are further subdivided into moving speed and dwell time for each word.

The mouse behavior during users' reading is recorded using a mouse tracker. According to each term's features, such as moving or text selection, a predefined algorithm is used to calculate the weight of the term to create a weighted ranking. The system overview of our method is shown in Figure 1.

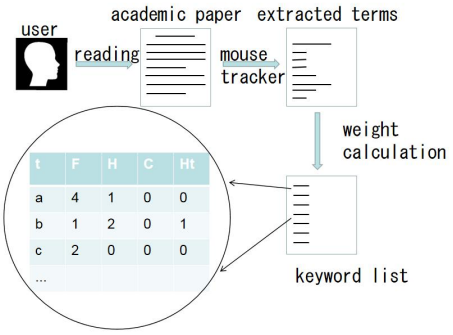

Figure 1: System overview

When a user uses the pointer to hover or move over a term while reading, the mouse tracker records the corresponding term, dwell time, and term length. In addition, because people usually read from left to right, we configured the mouse tracker to avoid recording data when a pointer moves from right to left to avoid collecting noise data.

3.2 Term Weight Calculation

The data recorded by the mouse tracker are processed by removing stop words and stemming. Next, the weight for each term is calculated, and the keyword list is generated. The algorithm for calculating weight W(i) is as follows:

$$tf_{i,j} = \frac{A_i}{\sum_k n_{k,j}}, \quad (1)$$

$$W(\mathrm{i}) = \frac{D_i}{L_i} + tf_{i,j} + p \times C(i), \quad (2)$$

where:

A_i : The number of times the term i is passed by the pointer.

$\sum_k n_{k,j}$: The sum of the number of occurrences of all terms in document j.

$W(i)$: The weight of term i.

L_i : The length of term i.

D_i : The dwell time of term i.

$C(i)$: The number of times the term i has been selected.

p : The coefficient corresponding to text selection behavior which is set to 0.6.

3.3 Combine data from multiple users

Among the top 20 keywords extracted based on the algorithm in Section 3.2, for those keywords that appear repeatedly in the ranking based on different users, we consider further enhancing its weight, according to their frequency of occurrence in the rankings generated by different user data. We define that for those keywords that occur more than 50% of the time which means it occurs in at least half of the rankings generated by

different user's data. Their weights will be increased by the number of their occurrences, for the rest of the keywords, their weights will remain unchanged. Here in after, we refer to proposed method that combines data from multiple users as proposed method*.

4. Experiment

4.1 Experiment environment

To ensure that participants can read as usual, we used the HTML file to construct an academic paper page and make the layout similar to the layout of PDF files and then embedded the code that builds a mouse tracker implemented in JavaScript. An example of an HTML file is shown in Figure 2.

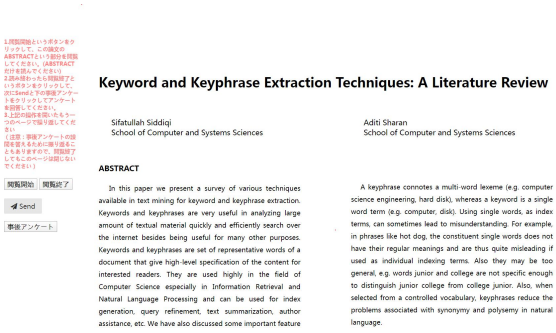

Figure 2: Example of HTML file

The mouse tracker starts recording from the moment the participant presses the button to start browsing and stops recording when presses the button to end browsing. When the participant finishes reading and presses the send button, the recorded data of mouse behavior are transformed to Google form using the Google script. In addition, based on the experiment task design, a mouse tracker is built to record only the abstract part of each paper.

We prepared four pages similar to the example page, which is mentioned above and built an entrance page. We chose four papers from the fields of text mining[9], information seeking behavior[18], recommendation system[19], and information retrieval[20]. These four papers were limited to information science, and we selected papers that were relatively easy to understand so that the potential participants could read smoothly. Figure 3 shows the screenshot of the entrance page.

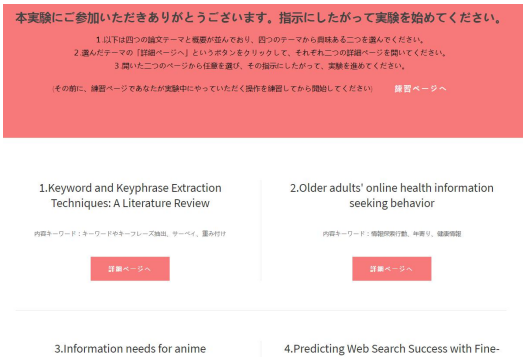

Figure 3: Entrance page

4.2 Participants

We invited 15 students (14 graduate and 1 undergraduate), studying information studies at the University of Tsukuba to be our experiment participants.

To investigate the percentage of users who tend to follow the mouse pointer through the eye while browsing, we conducted a questionnaire survey before the experiment. According to the results of the questionnaire, five students answered that (a) they usually tended to have this habit, three students answered that (b) they occasionally tended to have this habit, four students answered that (c) they did not have such habit, and three students answered that (d) they were not sure. In addition, for the question of how they moved the pointer, 10 students answered using a mouse and 5 answered using a touch pad.

4.3 Procedure

The entire experiment was carried out remotely through ZOOM due to the COVID-19 situation. The experimental process is shown in Figure 4.

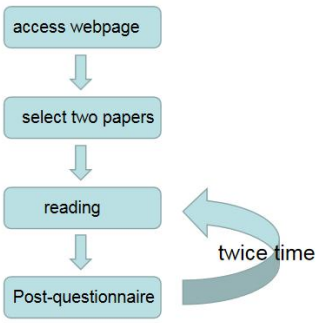

Figure 4: Procedure of the experiment

Firstly, we sent the entrance URL to the participants, they were asked to select two of the four papers as the materials for next reading. Because the reading task needs to be highly focused, in order to reduce the burden on participants and ensure that they could try their best to understand the content, we set a time limitation of 5 min to each reading session. Considering that we expect participants can read a complete content within 5 min, the abstract can be read in about five minutes. Therefore we select the abstract part to be reading material. In addition, we set an instructions as follows to carry out the experiment smoothly:

> "As a class assignment, you need to read the abstracts of two academic papers. When you finish, write down more than 10 words that you think are important."

After they finished reading, they were asked to answer several questions, such as if they understood the content and to write down the words.

4.4 Evaluation Measure

We used the words written by a participant as the correct answer to evaluate the effectiveness of the methods based on precision, recall, and F-score as follows:

$$precision = \frac{|A \cap B|}{|A|}, \quad (3)$$

$$recall = \frac{|A \cap B|}{|B|}, \quad (4)$$

$$F - score = \frac{2 \times precision \times recall}{precesion + recall}, \quad (5)$$

where:

A: Words extracted using the method.

B: Words judged to be important by a participant.

As two famous keyword extraction method, widely used as baseline method, the TF-IDF[1] and TextRank algorithm[21] were used as baselines for comparison with the proposed method. The document frequency of TF-IDF for each word was obtained by manually retrieving the number of matched documents at the ACM Digital Library as a query. Considering that the participants may give more than ten keywords, we choose to use the precision and recall at the top 10 terms to compare the three methods.

5. Results

5.1 Evaluation

Before the analysis, we excluded the data from five participants because they mostly did not move the mouse pointer while reading. In addition, the participants selected only three of the four papers we prepared. Table 1 shows the total effectiveness of each method.

Table 1 shows the effectiveness in total obtained by each method.

Table 1. Total effectiveness of each method

	Precision	Recall	F-score
TextRank	20.5%	19.1%	19.8%
TF-IDF	41.0%	35.7%	38.2%
Proposed method	37.5%	33.8%	35.6%
Proposed method*	**41.6%**	**36.0%**	**38.6%**

Table 2 shows the precision by different threshold cut obtained by each method.

Table 2. Precision by different threshold cut

	Top1	Top3	Top5	Top10	Top20
TextRank	21.1%	22.8%	21.1%	20.5%	18.3%
TF-IDF	**78.9%**	57.9%	**53.7%**	41.0%	25.5%
Proposed method	47.4%	43.9%	37.9%	37.5%	**26.3%**
Proposed method*	**78.9%**	**64.9%**	52.6%	**41.6%**	**26.3%**

As shown in Table 1 and Table 2, the proposed method* achieves the best performance in precision, recall and F-score. With different threshold cut in precision, the proposed method* achieves the best performance in Top1, Top3, Top10 and Top20 while TF-IDF method achieve the best performance in Top 1 and Top5.

The results were analyzed using a two-factor and three-level ANOVA, and LSD was used for the multiple comparison test. Figure 5 shows the F-score of three papers obtained using three methods.

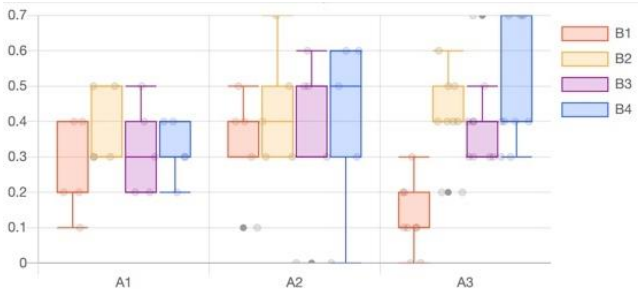

Figure 5: F-score obtained using each method

In Figure 5, B1, B2, B3 and B4 correspond to TextRank, TF-IDF, proposed method and proposed method*. As a result of the analysis of variance, there was a significant difference between the methods ($F(3, 48) = 8.89$, $p < .01$). In addition, according to multiple comparisons, there was a significant difference between the TextRank and the other three methods ($p < .05$). However, no significant difference was found among the TF-IDF method, proposed method and proposed method*.

5.2 Discussion

In this section, we discuss the research questions through the results described in the previous section. For RQ1 (Can we use readers' mouse behavior to extract keywords?), it can be considered that keyword extraction is possible using the proposed method. By calculating the terms weight separately with each of the mouse feature, we find that the dwell time per terms feature ($\frac{D_i}{L_i}$) and the number of times the pointer passed feature ($\mathrm{tf}_{i,j}$) play a major role, which are reached 35.8% and 36.3% precision.

- Dwell time per terms: 35.8%
- Mouse passing: 36.3%
- Text selection: 2.6%

Text selection feature (C_i) only appears in three participants' reading session at five times in total, but the part of terms where it appears, the terms are keyword as the participants write lately. For RQ2 (Can the proposed method be more effective than the baseline method?), according to ANOVA analysis, although the proposed method outperforms the TextRank method, it cannot be confirmed that it outperforms the TF-IDF method.

For RQ3 (Can the proposed method still be effective for those do not use a pointer to trace the reading part of the text?), the answer is negative because the participants who answered that they did not have such habit mostly did not record their data also. We made statistics on the effects of the proposed method on participants of types a, b, and d, the precision of the proposed method on the three types of participants are 40%, 45%, 20%. Two of the three type b participants, although they answered that

they had this habit occasionally in the previous questionnaire, they actually showed this habit frequently during the experiment. On the contrary, mouse tracker couldn't record data of two participants who answered they have this habit frequently in the questionnaire since they barely moved their mouse during the reading session. This finding indicates that the participants' impression of their reading habits may not be accurate.

Meanwhile, we found that the effect of the proposed method on participants who answered that they had such habits (types a and b) is better than types c and d. To determine the relationship between them, we extracted the amount of data recorded by the mouse tracker, which can represent whether the participants frequently traced the text while reading. The distribution of the F-score corresponding to the amount of data recorded by the mouse tracker is shown in Figure 6.

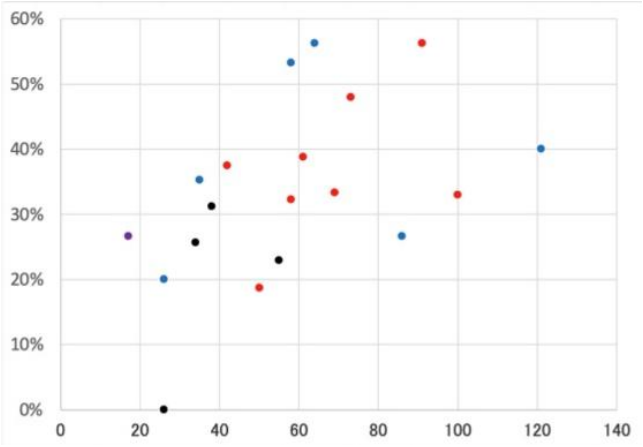

Figure 6: Distribution of amount of data and F-score

In Figure 6, the horizontal axis represents the total number of words recorded by the mouse tracker during each reading session, and the vertical axis represents the F-score. The four colors of red, blue, purple and black represent the four types of participants type a, b, c, and d. We calculate the correlation coefficient of the two data to get r=0.42, which indicates that there have a weak correlation. On the other hand, we can see that the proposed method had a certain effect on participants who did not move the mouse much (but still moved) while reading. We infer that this phenomenon may be because they only moved the mouse when they read the important part. We also considered the impact that may be caused by inadequate amount of data in this experiment. These issues will be addressed in future studies.

For RQ4 (Whether the effectiveness of the proposed method can be improved by combining multiple user data?), as we can see from Table 1and Table 2, combining multiple user data can improve the effectiveness of the proposed method significantly. Also, we expect better performance with larger volumes of user data setting.

6. Conclusions

In this study, we proposed a keyword extraction method for academic papers by using mouse behavior while the user was reading and verified its effectiveness. As a result of the experiment, the proposed method was not significantly different from the existing method. We can see that the proposed method has a relatively poor effect on users who do not have such habit. How to solve this problem will become a focus of further research.

In future work, we can consider setting a threshold to the amount of data recorded by mouse tracker or setting a threshold within a certain period of time to prevent the mouse tracker from recording this type of users, or adjust the proposed method that takes mouse features as optional, to integrate with the baseline method.

Reference

[1] Salton. 1989. Automatic Text Processing, Addison-Wesley Publishing Company.
[2] Robertson et al. 1996. Okapi at TREC4. Proc. of the 4th Text REtrieval Conference. p.73-96.
[3] Boudin et al. 2020. Keyphrase Generation for Scientific Document Retrieval. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL. p.1118-1126.
[4] Meng et al. 2017. Deep Keyphrase Generation.In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. p.582-592.
[5] Huang et al. 2012. Improving Searcher Models Using Mouse Cursor Activity. Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. p.195-204.
[6] Balatsoukas et al. 2012. An Eye-tracking Approach to the Analysis of Relevance Judgments on the Web: The Case of Google Search Engine. Journal of the American Society for Information Science and Technology. Vol.63, No.9. p.1728-1746.
[7] Ajanki et al. 2009. Can eyes reveal interest? Implicit queries from gaze patterns. User Modeling and User Adapted Interaction. Vol.19, No.4. p.307-339.
[8] Bharti et al. 2017. Automatic Keyword Extraction for Text Summarization: A Survey. http://arxiv.org/abs/1704.03242, p.1-12.
[9] Siddiqi et al. 2015. Keyword and keyphrase extraction techniques: a literature review. International Journal of Computer Applications, Vol.109, No.2. p.18-23.
[10] Ramos. 2003. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. p.1-4.
[11] Barzilay et al. 1997. Using lexical chains for text summarization. In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization. p.10-17.
[12] Hong et al. 2012. An Extended Keyword Extraction Method. International Conference on Applied Physics and Industrial Engineering. p.1120-1127.
[13] Mihalcea et al. 2004. TextRank: Bringing order into texts. Proceedings of EMNLP. p.404-411.
[14] Li et al. 2017. A Keyword Extraction Method for Chinese Scientific Abstracts. Proceedings of the 2017 International Conference on Wireless Communications, Networking and Applications. p.133-137.
[15] Kantor et al. 2000. Capturing Human Intelligence in the Net. Comm. of the ACM. Vol.43, No.8, p.112-115.
[16] Hijitaka et al. 2004. Implicit User Profiling for On Demand Relevance Feedback. Proceedings of the 9th international conference on Intelligent user interfaces. ACM. p.198-205.
[17] Hienert et al. 2017. Term-Mouse-Fixations as an Additional Indicator for Topical User Interests in Domain-Specific Search. ICTIR'17. p.249-252.
[18] Huang et al. 2012. Older Adults' Online Health Information Seeking

Behavior. iConference '12: Proceedings of the 2012 iConference. p. 338-345.
[19] Cho et al. 2017. Information Needs for Anime Recommendation: Analyzing Anime Users' Online Forum Queries. Proceedings of 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), p.305-306.
[20] Guo et al. 2012. Predicting Web Search Success with Fine-grained Interaction Data. CIKM '12: Proceedings of the 21st ACM international conference on Information and knowledge management. p.2050-2054.
[21] Barrios et al. 2015. Variations of the Similarity Function of TextRank for Automated Summarization. Argentine Symposium on Artificial Intelligence (ASAI). p.65-72.