論文閲覧時のポインター行動を用いた特徴語抽出手法に関する研究

筑波大学
人間総合科学学術院人間総合科学研究群
情報学学位プログラム

2022年3月

賀　純陽

論文閲覧時のポインター行動を用いた特徴語抽出手法に関する研究

Keyword extraction method using users' mouse behavior

氏名：賀　純陽

HE CHUNYANG

インターネットの普及に伴い、近年ウェブ情報資源が爆発的な発展を遂げた。その中で、情報を整理し要約するためのキーワードは重要な役割を果たしている。キーワード抽出や生成技術が進歩する中、かつての文書内の特徴を用いた手法から近年の機械学習を用いた手法まで、多くの手法は文書自体に着目しているものの、読む側のフィードバックを活用する方法はあまりない。本論文の目的は学術論文を閲覧する時の読み手のポインター行動を用いて、論文のキーワードを抽出する手法を提案し、その抽出の有効性を検証することである。実験参加者が論文を閲覧する時のポインター軌跡、速度、クリック特徴をマウストラッカーによって記録し、各特徴量によって重み付きランキングを作成する。ベースライン手法では TF-IDF 手法と TextRank 手法を用い、精度、再現率と F-スコアに基づいて本手法の有効性についての検証と考察を行う。評価実験の結果として、提案手法は TextRank 手法よりよい効果を得られ、TF-IDF 手法と比べて若干よい効果を得られたことから提案手法の有効性を示すことができた。

Owing to the explosive growth of information, keywords play an essential role in summarizing information and helping search effectively. Most of the existing keyword extraction approaches merely focus on the document-centric information, without well incorporating users' reading behaviors, such as the mouse-related information. In this thesis, 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. Specifically, we developed a mouse tracker to record mouse trajectory, speed, and click behaviors during the participants' reading process of academic papers. Using a predefined weighting algorithm, a term-weighted ranking was proposed by incorporating mouse-related features. We used the term frequency-inverse document frequency (TF-IDF) and TextRank methods as the baseline methods to compare the effectiveness. The evaluation was performed in terms of precision, recall, and F-score. Based on an in-depth comparison with the baseline methods, the experimental results show that the proposed method is able to achieve a better performance, which demonstrates its effectiveness.

主研究指導教員：高久　雅生

副研究指導教員：于　海涛

目 次

図 目 次

表 目 次

第1章　はじめに

本章では研究の背景と目的について説明をする。まず、1.1 節はキーワードの重要性、自動キーワード抽出技術における課題とインタラクティブなユーザ行動によるユーザの分析に関する研究背景を紹介する。1.2 節は研究目的を説明する。1.3 節は本論文の構成について述べる。

1.1　研究背景

ウェブの急速な発展に伴って、学術情報もかつての書籍を媒体にした記録から電子的な形としてウェブで流通するようになった。無数の学術情報資源から素早く自分が欲しい論文情報を手に入れるため、検索システムにとって文書を要約できるキーワードは大きな役割を果たしている。

1.1.1　キーワードの重要性

キーワードは、1 つのワードまたは複数のフレーズのシーケンスとして、文書の本質的な内容を凝縮した形で表すものと見られている。しかし、キーワードは検索システムにおいて検索、索引など様々な用途に役立つ一方、ほとんどの文書にはキーワードが割り当てられていない。2013 年度のデータ分析世界大会「KDD CUP」で Microsoft Academic が提供した論文のうち、およそ 75 %の論文はキーワードが割り当てられていない [1]。ネットワークの発達に伴って情報資源が爆発的に増える状況に対してキーワードは、大量の情報を効率的にフィルタリングするために必須のツールであり、文書ごとに適切なキーワードを割り当てることは検索効率の向上に非常に役立つ。

キーワードを割り当てる従来の方法では、主に図書館や情報検索分野の専門家が固有の分類基準に基づいて手作業で割り当てるか、もしくは著者の判断で作成するかがほとんどであるが [2]、情報が爆発的に増えることで人手で文書をキーワードに要約することはまず大きな労力的な負担になり、人手ですべてのドキュメントに適切なキーワードを付与することがますます現実的に不可能となりつつあるため、近年ではテキストマイニング分野で自動文書要約やキーワード抽出技術が盛んになっている。

自動キーワード抽出は、人手を介さずにテキスト文書から自動的に核となる内容を表している単語やフレーズを見つけ出すプロセスである。従来手法ではかつての統計的な手法に

図 1.1: 眼球追跡装置によって作成したヒートマップの一例 [7]

基づいた TF-IDF[3]、BM25 手法 [4] などから近年の自然言語処理技術の発展に伴ってホットスポットになった機械学習手法 [5, 6] まで様々あるが、これらの手法はほとんど文書自体の情報に着目しているものの、読む側の行動分析による情報の活用はあまり考慮されていない。読む側の行動分析による情報の活用に関する例の一つとして、ヒートマップを用いる手法がある [7]。ヒートマップとは、コンテンツのどこがよく読まれているか (熟読率)、どこまで読まれているか (読了率)、どこがクリックされているのか (クリック)、ユーザーのマウスの動き (マウスムーブ) などのデータを色の濃淡で表現した可視化グラフである [8]。図 1.1 では眼球追跡装置によって作成したヒートマップの一例を示す。

眼球追跡装置によって作成したヒートマップを用いて読者の熟読したエリアを可視化することで、読者の関心を抽出することができる。このように、読む側の立場に立ち、読者らが文書のどの部分に興味を示し、関心を持つかを推測することで文書のキーワードを抽出することは新しい視点を提供し、より網羅的な学術情報検索支援を提供できるため有意義だと考えられる。

1.1.2 インタラクティブなユーザ行動からのデータ抽出

一方、情報検索分野ではユーザの行動分析を行うことでユーザの意図を推測することがよくあり、多くの研究では、ユーザの行動やフィードバックからユーザのニーズや興味を予測してきた [9, 10]。インタラクティブな情報検索分野において、ユーザのフィードバックはそれぞれ明示的なフィードバック (explicit feedback) と暗黙的なフィードバック (implicit feedback) の二種類に分けられる [12]。

明示的なフィードバックを用いる手法ではアンケートや採点などの形で興味や適合性判定についてユーザに回答してもらうのが一般的である [13, 14]。明示的な手法のメリットは直接ユーザから情報が提供され、かなり信頼性のあるフィードバック情報が手に入るのに対し、その評価の手順が煩わしくかつユーザへかける負担が大きいため、大規模なユーザデータ分析には適用しにくいというデメリットも存在する。

暗黙的なフィードバックを用いる手法では直接にユーザから答えを求めるのではなく、ユーザが情報探索過程において示す様々な行動からユーザの意図を汲み取り、ユーザの興味またはドキュメントの適合性の予測に用いる。この種類の手法のメリットは、より低コストで、大量に、またユーザに負担をかけずに収集できるため、大規模のユーザデータ分析が可能になる [15]。

暗黙的なフィードバックの内容としては具体的に、ページごとの閲覧時間、閲覧履歴、入力クエリまたはユーザの視線などが挙げられる。その中で、眼球運動データはユーザの意図や興味を反映している暗黙的なフィードバックの一種であり、文書適合判定、ユーザ分析やユーザの興味を推測するなどの研究でよく使われている [10, 11, 16]。しかし、眼球追跡装置の精度が向上しているにもかかわらず、現段階で高精度のものはまだ非常に高価で、またリアルの世界で大規模にユーザからデータをとることが困難などの問題点もあり、実用が難しい。一方、研究者らはユーザが操作したマウスポインターの軌跡などを用いて視線を予測する実験を行うことで、マウスポインターの動的な動きが、正確に任意の時点での視線ポイントを予測できるということを明らかにしてきた [17]。そのため、近年では視線と似た効果を持ち、かつ安価で取得しやすいマウス動作のデータを眼球運動データの代わりとする考えが出た。ここでマウス動作とはマウスの移動軌跡、クリック、スクロールなどユーザがマウスを使用するときに示すあらゆる動作を指す。

1.2 研究目的

本研究の目的は読者が学術論文を読む時に示すポインター行動から読者の関心を抽出することを試み、キーワードを推測する手法を提案し、その有効性を検証することである。ここでポインター行動とはユーザがマウスやタッチパッドを使用するときのポインターの移動軌跡、ホバリング、テキストをドラッグするなどの動作を指す。マウス動作との区別は、ポインター行動はタッチパッドを使用するときのポインターも含む点が異なり、本研究ではポインター行動を研究の対象とする。

学術論文を読むときの状況を考えると、読者は大量の専門用語や情報を消化する必要があるため、より集中しなければならない。そのため、読者は一般的な閲覧シチュエーションよりもポインターでなぞり読みしがちであると考える。この時のポインター行動はゆっくり移動する、ホバリング、クリックするなど読者がじっくり読んでいるかどうかを推測する

ための良い指標になるのではないかと考え、該当する単語の重要度を推測する手法を提案する。そのうえで、同じ学術論文ごとに複数のユーザが読む場合、複数ユーザのそれぞれのポインター行動によって論文のキーワードを推測し、共通するところを見つけることでより良いキーワードを抽出する手法も試みる。こうした手法を以下では複数ユーザからの集約に基づく手法と呼ぶ。

以上を踏まえ、本研究のリサーチクエスチョンでは以下の四つを設定する：

RQ1: ポインター行動から読者の関心を持つ部分を抽出することによって文書のキーワードを推測できるか？

RQ2: 提案手法はベースライン手法より良い効果を得られるか？

RQ3: ポインターでなぞり読みを行わないユーザに対して、複数ユーザからの集約による総合的な特徴語抽出手法は有効であるか？

RQ4: ユーザデータの集合が増えることによって、複数ユーザからの集約による総合的な特徴語抽出手法の効果は良くなるか？

なお、本研究では英語の学術論文を対象とし、その他の言語の論文閲覧は対象としない、なぜなら、英語の学術論文が全世界で広く利用され、応用可能性が高いこと、さらに、単語の抽出が比較的容易であることから英語の学術論文を対象に定めた。

1.3 本論文の構成

本論文の構成は以下の通りである。

第 1 章では、研究背景と研究目的を説明した。第 2 章では、キーワード抽出手法とマウス動作の追跡についての関連研究を紹介する。第 3 章では提案手法の仕組みについて詳しく述べる。第 4 章、第 5 章は、評価実験の結果と考察について論じる。そして第 6 章は、本論文で議論した内容についての総括である。

第2章　関連研究

本章では、関連研究について紹介する。2.1 節はキーワード抽出技術の種類や特徴に基づいて、キーワード抽出手法に関する先行研究を紹介する。2.2 節は学術論文におけるキーワード抽出に関する先行研究を紹介する。2.3 節はマウス動作に基づいたユーザ分析に関する先行研究を紹介する。

2.1　キーワード抽出手法

キーワード抽出手法は大まかに抽出型と生成型という二種類がある。抽出型の手法は原文からそのまま重要だと思われる単語を抽出してくるタイプの手法である。生成型の手法は近年自然言語処理の発展に伴って出現した、主に機械学習の教師なしモデルを用いて文書の内容を単語や重要文に要約するタイプの手法である。この二つのタイプの手法のもとで各手法の特徴によってさらに統計的な特徴を用いた手法 (statistical approaches)、言語的な特徴を用いた手法 (linguistic approaches)、機械学習を用いた手法 (machine learning approaches) と混合的な手法 (hybrid approaches) という主に四つの種類に分けられる [18, 19]。図 2.1 では上記四つの手法の種類を表す。

統計的な特徴を用いた手法では主に統計的な特徴とコーパスに依存し、シンプルで効率良いという長所がある [20]。言語的な特徴を用いた手法では基本的に特定のルールを用いた言語的特徴を用いてテキストの中のキーワードを検出する [21]。機械学習を用いた手法は教師ありと教師なし学習に分けることができ、教師あり学習は生成型、教師なし学習は抽出型手法に対応するのが一般的である [22, 23]。最後に混合的な手法は先ほど述べた三つの手法を組み合わせた手法である [24]。

本研究で提案する手法は本文からそのまま重要語を抽出するため抽出型の手法に属すると考えられるが、主流の四つのタイプの特徴と違い、ユーザによる暗黙的なフィードバックを分析することでユーザが重要だと思うキーワードの抽出を試みる。

2.2　学術論文における特徴語抽出

学術論文は出版者の版権などの問題で直接全文にアクセスできるものが少ないため、学術論文を対象にした自動的特徴語抽出手法は論文の抄録、アブストラクト、タイトルと引用関係などを対象にする研究が存在する。

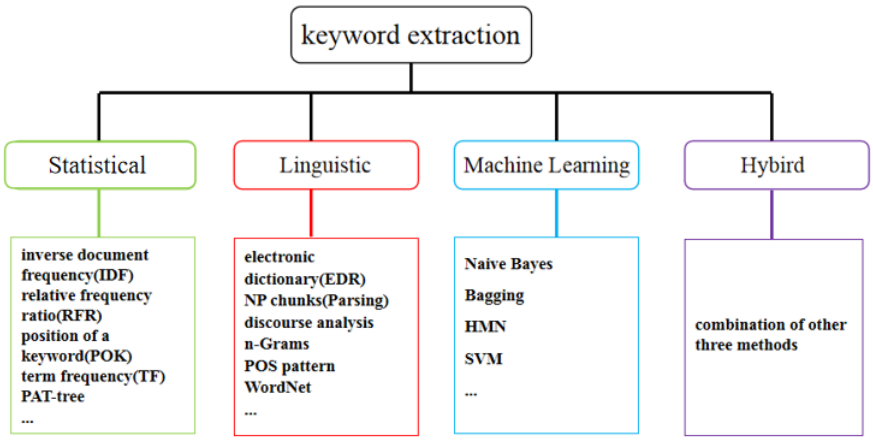


図 2.1: キーワード抽出手法の種類 [18]

原田らは抄録文のうち、文の構造的な特徴や文中に出現する特別な表現に基づいて、主題概念を表している単語を抽出することを試みた [25]。

Blank らは論文のメタデータに着目し、論文の引用関係グラフのネットワークを用いて論文のキーワード推薦を作成した [1]。

Ma らは論文のキーワードは常にタイトルと関連する傾向があるという仮説に基づいて、論文タイトルに着目し、重みつきハイパーグラフを構築することで論文のキーワードを抽出する提案をした [26]。

中川らは学術論文コーパスを対象に専門用語の抽出手法を検討した。論文の主旨はコンテンツにおける連続する名詞が重要語である可能性が高いという考えに着目し、単一の名詞 N に連接する単一の名詞の頻度の統計量を利用する N の重みつけ方法を提案した [27]。

このように、既存研究は文書内の特徴やメタデータに注目している研究が多数存在しているのに対して、ユーザ側の行動分析に基づいた論文重要語の抽出手法に関する例が見受けられない。したがって、本研究は読者閲覧時のポインター行動に基づいて文書に対する特徴語抽出手法を提案する。

2.3 マウス動作によるユーザ分析

ユーザのマウス動作を用いてユーザの分析やプロファイリングを行う研究は多く存在する。従来の研究ではマウス動作のデータがユーザの意図を予測するのに価値のあるデータであることを証明してきた。Kantor らはユーザがウェブページをブラウジングするときに、マウスポインターで視線を追っていく傾向があると報告している [28]。Huang らはマウスポインターの軌跡を用いてユーザの注目箇所を予測する実験を行い、70 %の精度を得た [17]。ユーザの行動分析に役立つ暗黙的なフィードバックとして、マウスポインターの行動はユーザの視線を反映できるデータとして重要視され、ユーザの興味推測、検索意図や満足度の

推測などの研究に使われてきた。

Guo らは情報探索行動を対象に、検索結果ページにおけるユーザのマウス軌跡を用いてユーザが行う検索意図を推測し、さらに、ユーザの探索のタイプはナビゲーション探索か情報型探索かを区別することを試みた [29]。

Huang らはユーザの検索結果ページにおけるマウスのホバリングやスクロールは検索結果ページがユーザによって閲覧されたかどうかを推測するためのよい特徴であると報告し、この二つの特徴を検索モデルに取り入れることでユーザが検索結果に対する満足度を予測できると述べた [30]。

Hienert らは学術情報検索エンジンの検索結果ページにおけるユーザのマウス動作によって抽出した単語と再入力クエリとの比較で、ユーザが関心を持つ単語に対する平均的なマウスポインターの停留時間は、ほかの単語に対する停留時間よりおよそ二倍長いと報告した [31]。

これまでは主に単一のユーザごとにユーザ分析を行った研究を紹介したが、以下は複数ユーザのポインターデータの集約に基づいた情報の活用に関する研究も紹介する。Ageev らは情報収集型探索タスクを対象に、ユーザが検索エンジンからたどった先のページにおけるマウスの移動とスクロール動作を用いて、複数のユーザが最も見ている部分を抽出し、ドキュメントのスニペットの生成手法を改善する手法を提案した [32]。

Hijikata らはユーザが任意にウェブページを閲覧しているときに示すリンクポインティング、リンククリック、なぞり読み、テキスト選択などという四つのマウス動作から対応するユーザの興味語を抽出することを試みた。結果としてベースラインと比べて有効性が 1.4 倍向上したとの結果が出た [33]。しかし、リンクポインティングとクリックがもっとも効いている特徴であるが、文書閲覧の場合ではリンクが存在せず、適用しにくい。また、Hijikata らはユーザ実験の時、なぞり読みという特徴について、ユーザは実際厳密に文字の行の上をなぞり読みしていないことが多いと報告した。その結果、なぞり読みは出現頻度が最も多く出現する特徴であるが、四つの特徴の中では最も精度が低い。これはタスクや検索シチュエーションの違いによるもので、より細かい粒度の制約条件が必要であると考えられる。

本研究と既存研究との違いは以下の三つにある。

- 本研究は学術論文閲覧を対象にする。
- 本研究はより細かい粒度でポインター行動を扱う。
- 本研究は複数ユーザからの集約効果を導入し、その有効性を評価する。

本研究は英語の学術論文を対象に、読者が論文内容を閲覧するときに示すポインター行動に基づき、なぞり読み、ポインターが通過した回数とテキスト選択特徴を用いる。加えてなぞり読み特徴は閲覧速度と停留時間などより細かい粒度の指標を用いて扱う。以上の三つのポインター行動の特徴を用いて論文内容の特徴語抽出を試みる。さらに、複数ユー

ザからの集約という概念を導入し、論文を読むときに複数ユーザが関心を持つ語ほど特徴的な語であるという仮説に基づき、読む側からの新しい視点を提供することで、より網羅的な学術情報検索支援を試みる。

第3章　提案手法

本章では、英語の学術論文閲覧における読者のポインター行動を対象に、文章内にある特徴などを使わず、かつ事前の学習用データを必要としない文章の特徴語抽出手法を提案する。まず、3.1 節では、手法の全体像について説明する。次に 3.2 節ではマウストラッカーの実装について説明をする。3.3 節では、特徴語の重み付けアルゴリズムについて述べ、3.4 節では複数ユーザからの集約について説明する。

3.1　手法の全体像

本研究ではまずユーザが関心を持っている内容を閲覧する時に示すいくつかのポインターの行動から、なぞり読み (ポインターで文字列を追っていく動作)、ポインターが通過した回数 (ポインターが単語の上を通過する回数) とテキスト選択 (ポインターでテキストをドラッグする動作) という三つの行動特徴に着目する。このうち、なぞり読み特徴は単語ごとのポインターの移動速度として扱う。移動速度を扱う理由は、移動速度は読者が熟読しているかどうかを反映でき、移動速度の遅い単語ほど重要な語である可能性が高いと判断したためである。

ユーザの閲覧中のポインター行動はマウストラッカーによって記録し、各特徴量によって単語単位の重みつきランキングを作成する。手法の全体像を図 3.1 に示す。

3.2　マウストラッカーの実装

閲覧中のポインター行動を記録するために、JavaScript を用いて独自のマウストラッカー機能を実装した。

ユーザのブラウザ上の操作イベント (mousemove,mouseup,mousedown) を DOM (Document Object Model) インターフェースで検出する。マウストラッカーはユーザ閲覧時のポインターが通過した単語の文字列、単語を通過する平均的なポインターの移動速度とテキスト選択された部分を記録する。具体的には、mousemove(fn) でポインターが動くたびにテキストノードの検出を行う。このとき、テキストノードは一文字単位の範囲から構成される。この検出には caretRangeFromPoint() メソッド (Chrome に適用、IE の場合は createTextRange()、Firefox の場合は caretPositionFromPoint()) を用いる。このメソッドは任意

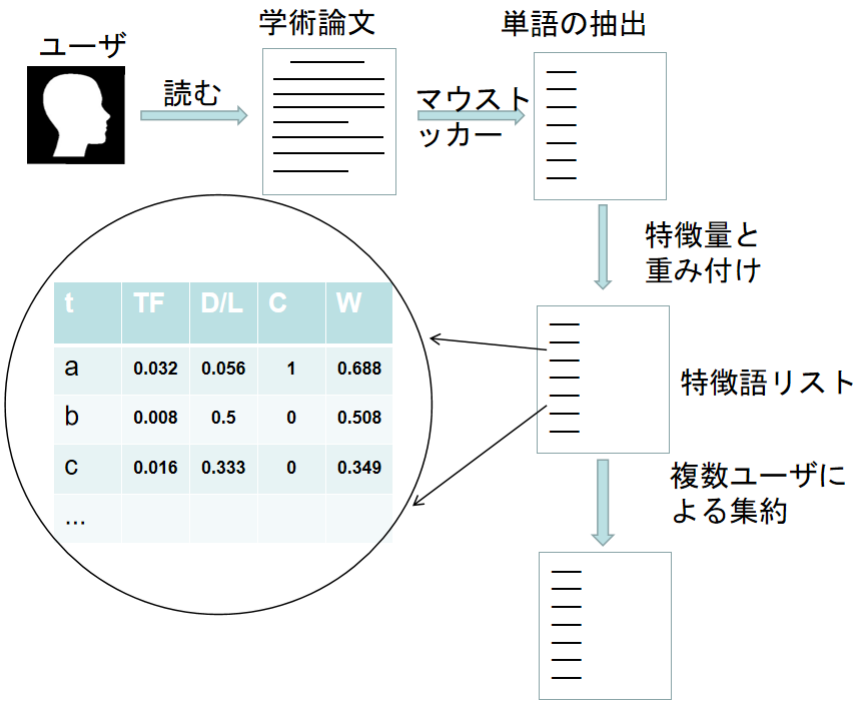


図 3.1: 手法の全体像

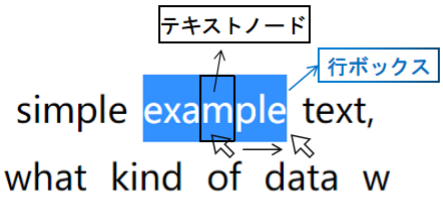


図 3.2: テキストノードと単語の行ボックス範囲の例

の時点のポインター座標を指定すると、その座標の位置にあるテキストノードが属するタグにあるすべてのテキストと該当するテキストノードのテキストの中の位置情報を返す。位置情報に基づいて該当するテキストノードから出発して、その文字のそれぞれ左と右の空白の区切りまで一文字ずつ組み合わせることによって、そのテキストノードに対応する単語の文字列を認識することができる。テキストノードに対応する単語の文字列を認識するソースコードを表 3.1 に示す。

単語ごとの認識範囲はその単語の行ボックスで決める。テキストノードと単語の行ボックス範囲の例を図 3.2 に示す。

マウストラッカーが記録する情報には 3 つの情報がある。それらは停留時間、ポインターの移動距離、平均的な移動速度である。ポインターが該当する単語の行ボックスの範囲を離れると該当する単語の文字列を記録する。また、離れたときのタイムスタンプ (t_2) と入っ

表 3.1: テキストノードに対応する単語の文字列を認識するソースコード (Chrome の場合)

```
1       function getWordUnderCursor(event) {
2                var range, textNode, offset;
3                 if (document.caretRangeFromPoint) {
4                    range = document.caretRangeFromPoint(event.clientX, event.
                         clientY);
5                    textNode = range.startContainer;
6                    offset = range.startOffset;
7                }
8                var data = textNode.data,
9                    i = offset,
10                   begin,
11                   end;
12               while (i > 0 && data[i] !== " ") { --i; };
13               begin = i;
14               i = offset;
15               while (i < data.length && data[i] !== " "&&data[i]!=',’&&data[i
                     ]!='.') { ++i; };
16               end = i;
17               return data.substring(begin, end);
18           }
```

たときのタイムスタンプ (t_1) をもとに、その差分 (t_2-t_1) を該当する単語の停留時間と見なす。ポインターが該当する単語の範囲にいる間は 0.01 秒おきに移動した距離を計算式 $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$ によって計算する。X_1, Y_1 と X_2, Y_2 はそれぞれ 0.01 秒間隔の前後におけるポインターの横軸と縦軸の座標を表す。ポインターが該当する単語の範囲にいる間に 0.01 秒おきに移動した距離を足し合わせた値を該当する単語に対するポインターの移動距離と見なす。該当する単語に対するポインターの移動距離 (L) と該当する単語の停留時間 (D) があったとき、$\frac{L}{D}$ を単語の平均的な移動速度と見なし、記録する。

テキスト選択は mousedown と mouseup イベントによって検出を行う。これらのイベントの発生したタイミングを記録し、mousedown の生起した語と mouseup の生起した語の間で移動した単語をテキスト選択したものと見なす。

This:0.078 is:0.083 a:0.079 mousedown simple:0.068 example:0.057 mouseup text:0.098

図 3.3: マウストラッカーの出力形式の例

マウストラッカーの出力形式の例を図 3.3 に示す。「This is a simple example text」という文書に対し、ポインターが語を通過した後に該当する単語と単語に対応する移動速度が出力される。この例では、「This」に対して 0.078、「is」に対して 0.083 という速度が出力され

る。「simple」と「example」は mousedown と mouseup イベントが生起した語であるため、テキスト選択された語と見なす。

また、人の閲覧習慣は基本的に左から右へという方向が一般的であるため、行を切り替えるときのノイズデータを除外するためにポインターが右から左へと移動するときに語の記録はしないように設定した。

3.2.1 マウストラッカーの信頼性

構築したマウストラッカーはいくつかの制限がある。これらの制限は予備実験を行って確認した。これらの制限を超える範囲については、提案手法では扱わないこととした。

一つ目は単語の行間設定が記録の精度に影響を与える。上の行と下の行の距離が小さい場合 (line-height 属性が 10px 以下) はノイズデータが多く発生する。

二つ目は言語の特徴に依存する。日本語や中国語など語の間に空白がない特徴を持つ言語やアラビア語のような右から左へ書く言語には適用できない。

三つ目は行を切り替えるときのノイズデータを除去するために、右から左へポインターを動かすときは語の記録をしないように設定したため、まれに記録が取れないケースがある。具体的には語の範囲を出たときの座標がその直前の座標よりも左になっていると、記録できない。

四つ目はブラウザによって互換性問題が存在し、現時点では Chrome, Firefox と IE においてのみ動作検証を行っており、これら以外のブラウザでは動作を確認していない。

3.2.2 記録したデータの取得

記録した生データは埋め込まれた Google Script によって Google Form に転送し、別の Java プログラムを用いて処理し、特徴語リストを作成する。

3.3 重み付けアルゴリズム

マウストラッカーが記録したデータはまずストップワード (自然言語をコンピュータで処理するにあたって、一般的である等の理由で、処理対象外とする単語のこと、例えば a,of,the など) 除去やステミング (例えると「swims」「swimming」「swimmer」などの複数の語形を語幹である「swim」でマッチングを行う) の前処理を行った後、単語単位のデータセットを生成し、各特徴量によって重み付けをする。

特徴語 i の重みは以下の式で計算する。なお、$tf_{i,j}$ は Salton の定義 [3] を基に、ポインター行動に合うように定義しなおしたものである。3.2 式のうち、第 1 項はその単語に対するポインターの平均的な速度の逆数、第 2 項はその単語のポインター通過回数の文書内の全単

語数に対する比率、第 3 項はテキスト選択回数を示す。すべての項は特徴的な語ほど高い値となることを想定してモデル化した。

$$tf_{i,j} = \frac{A_i}{\sum_k n_{k,j}} \tag{3.1}$$

$$W(i) = \alpha \frac{D_i}{L_i} + (1 - \alpha) tf_{i,j} + p * C(i) \tag{3.2}$$

それぞれの記号が表す意味は以下の通り。

α : なぞり読みとホバリングという二つの特徴が重み付けに与える影響を表す係数、テストデータによってチューニング最適化を行う。

A_i : ある特徴語 i をポインターが通過した回数。

$\sum_k n_{k,j}$: 文書 j 内のすべての単語の出現回数の和。

$W(i)$: 特徴語 i の重み。

L_i : ポインターがある特徴語 i の行ボックス範囲内で移動した距離。

D_i : ポインターがある特徴語 i での停留時間。

$C(i)$: 特徴語 i がテキスト選択された回数。

p: テキスト選択特徴を表す係数。

j: 対象文書 j を表す。

3.4 複数ユーザからの集約

この節では一つの文書に対して複数のユーザが閲覧し、それぞれのポインター行動によって抽出した特徴語リストを集計し、重複回数の高い単語 (つまり、複数のユーザが関心を持つ単語) ほど、その単語は重要語であるという仮説に基づき、以下の 2 種類の特徴語リストを生成する手法を提案する。一つはユーザごとにリランキング操作を行う手法、もう一つは文書に対する総合的な特徴語リストを作成する手法である。

3.4.1 ユーザごとにリランキング

複数の読者が同じ論文を読む場合、それぞれの理解や興味が異なり、ポインター行動によって作成した特徴語ランキングもそれぞれ異なるが、上記の重み計算アルゴリズムによって各ユーザに対して作成した特徴語のランキングの上位 20 語のうち、全ユーザが作成した特徴語の集合において半分以上で重複する単語があれば、その単語の重複回数に基づいて降順でリランクを行い、そのほかの特徴語の順序はそのまま保持する。図 3.4 は複数ユーザのランキングからの集約手順を表す。

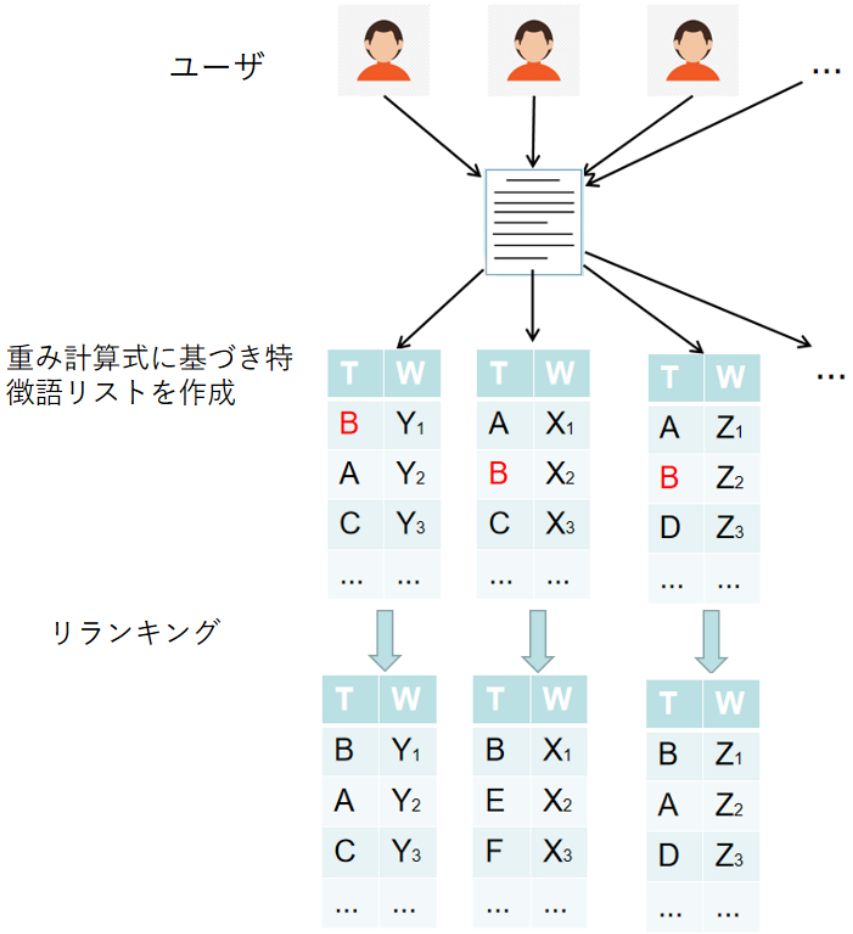


図 3.4: ユーザごとに特徴語リストを作成

3.4.2 文書の総合的な特徴語リスト作成

複数ユーザそれぞれの特徴語リストに共通して出現する語がより特徴的な語であるとしたとき、ポインター行動を取得したユーザ数が増えることによって総合的な特徴語リストの変化が安定すると考え、複数のユーザの特徴語リストに基づく文書の総合的な特徴語リストを作成する。

文書に対する総合的な特徴語リストを作成するために、一つの文書に対して、ユーザそれぞれのデータによって作成した特徴語リストの集合から上位 20 位までの特徴語を取り出し、すべての特徴語リスト全体における出現回数の降順に基づいて特徴語のランキングを作成する。同じ出現回数の単語は全ての特徴語のランキングにおける順位の平均値を取った値で上下関係を決めることで総合的な特徴語リストを作成する。

図 3.5 は文書の総合的な特徴語リスト作成手順を表す。

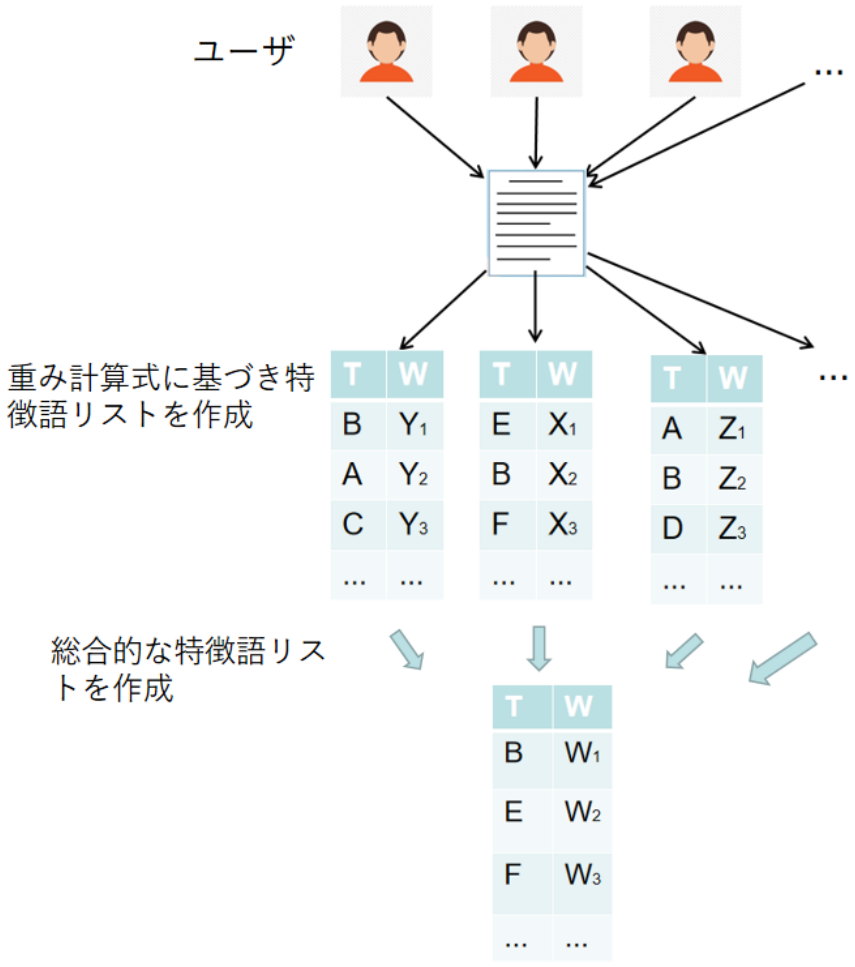


図 3.5: 文書の総合的な特徴語リスト作成

第4章　評価実験

本章では提案手法の有効性を評価するための実験について説明する。まず実験の流れについて述べ、そのあと事前調査アンケート、実験環境の構築について説明を行い、最後に評価基準と実験の結果について述べる。

4.1　実験の手順

実験参加者は事前調査アンケートに回答した 25 人を実験参加者として招待した。また、新型コロナウイルスの影響で、なるべく対面での接触を避けるため、評価実験はオンラインで行うこととし、実験過程は zoom により記録を行い、録画をした。実験のプロセスは以下の図 4.1 に示す。

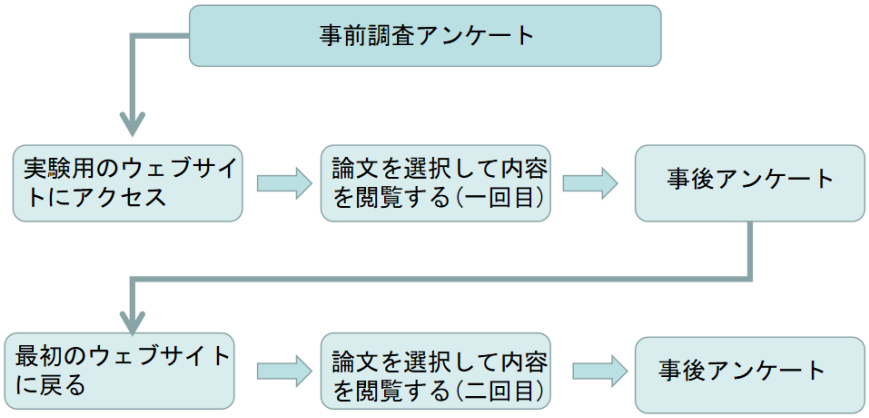


図 4.1: 実験の流れ

まず実験参加者に実験用のホームページにアクセスさせ、四つの論文から二つを選んでもらう。選んだ二つの論文から一つを選択して閲覧する。閲覧が終わったら事後アンケートに回答する。事後アンケートが終われば残りのもう一つの論文を閲覧し、閲覧が終わったら先ほどの手順で事後アンケートに回答すれば実験終了となる。実験用のホームページは図 4.2 に示す。

実験参加者が閲覧中に高度な集中力を保ってもらうため、負担をあまりかけないように閲覧時間を一本の論文ごとに 5 分という制限を設定した。そのため、5 分間で完結する内容

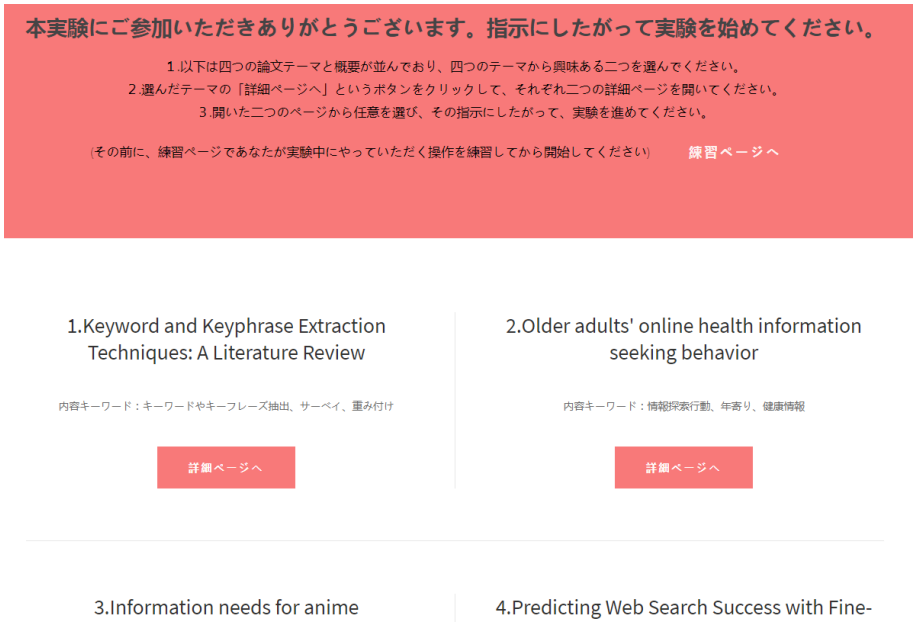


図 4.2: 実験用ホームページ

としては論文の要旨部分が最適と考え、それぞれ選んだ論文の要旨部分を閲覧するタスクを設定した。また、実験参加者に実験時に zoom の画面共有をしながら閲覧するように指示し、閲覧中に別のタブを開くことやわからなかった単語の意味を調べることは禁止した。

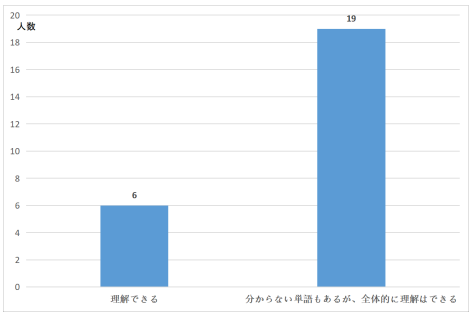
さらに、実験を円滑に行えるよう、以下のようにインストラクションを設定する：

大学授業の課題として、とある二本の論文を読むことになりました。読み終わった後、小テストがあります。テスト内容としては、読んだ部分についてあなたが思うキーワードを十個またはそれ以上並べてもらいます。

それぞれの閲覧時間は 5 分間、閲覧が終わったら事後アンケートに回答してもらう。事後アンケートは選んだ論文の番号、読んだ内容についての理解度、実験参加者が重要だと思うキーワードの入力と、閲覧中に意味がわからなかった単語の四つの質問からなっている。

4.2 実験参加者

学部生 5 名 (男性 3 名、女性 2 名、年齢 21 歳から 24 歳)、大学院博士前期課程 18 名 (男性 12 名、女性 6 名、年齢 22 歳から 27 歳)、博士後期課程 2 名 (男性 1 名、女性 1 名、年齢 27 歳から 30 歳) 合計 25 人がアンケートに回答した。実験参加者は情報学学位プログラムを専攻している学生がメインで、ほかには社会工学、システム情報工学、数理物質科学、生命地球科学などであった。



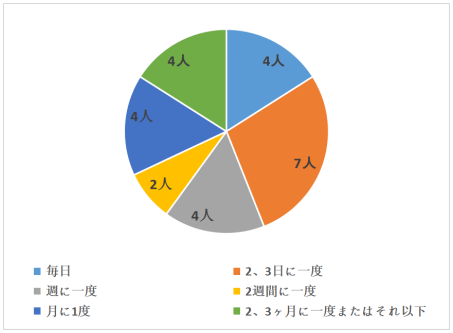


図 4.3: 英語論文に対する理解度に対する結果　図 4.4: 普段論文を読む頻度に対する結果

4.3　事前アンケート調査

学術論文を閲覧する場合、ポインターでなぞり読みする習慣を持つか、習慣を持っている読者の割合などを調査するため、また今回の実験は英語論文を閲覧すると想定しているため、英語力も含めて閲覧習慣などを調査し、かつ、フィルタリング目的で Google Form を用いて事前アンケート調査を行った。

4.3.1　アンケート項目と結果

調査アンケートの項目を表 4.1 に示す。その中で項目 8 は最も重要で、なぞり読みの習慣についてたずねた。項目 5 と 7 はフィルタリング項目であり、普段英語論文を読まないもしくは読めない参加者、またはスマホ、タブレットと紙媒体だけで論文を閲覧する参加者は本実験から除外される。

アンケートの結果として、25 名の回答者は全員普段パソコンで論文を読み、かつ英論文に対する理解度は 4 レベル（分からない単語もあるが、全体的に理解はできる）以上であるため、フィルタリングを通過し、実験参加者とした。

普段論文を読む頻度や英語論文に対する理解度についての調査結果を図 4.3、図 4.4 に示す。

ポインターでなぞり読みする習慣について、頻繁にあると答えたのは 8 人、たまにあると答えた人は 10 人、習慣がないと答えたのは 4 人、自分でもわからないと答えたのは 3 人であった。また、ポインターを動かすためにマウスを使うと答えたのは 19 人で、タッチパネルを使うと答えたのは 6 人であった。

4.4　実験環境の構築

実験環境は読者が普段通りに閲覧しているような、リアルな閲覧環境を作るため、PDF 形式のレイアウトのように HTML ファイルで学術論文ページを構築し、3.2 節で説明した

表 4.1: 事前調査アンケートの質問項目

事前調査アンケート		
質問番号	質問項目	選択肢
1	メールアドレス	回答者入力
2	性別	男性/女性/その他
3	所属組織	回答者入力
4	普段論文を読む頻度	毎日/2，3 日に一度/週に一度/2 週に一度/月に一度/それ以下
5	普段論文を読む環境	パソコン/スマホ/タブレット/紙媒体 (複数選択可)
6	ポインターを動かす主な手段	マウス/タッチパネル
7	英語論文に対する理解度	理解できる/分からない単語もあるが、全体的に理解はできる/分からない部分が多く、全体の内容を把握できない場合が多い/理解できない/英語論文を読むことはない
8	論文を読むときはポインターでなぞり読みする習慣はあるか	頻繁に/たまに/ない/わからない
9	実験可能日	回答者入力

マウストラッカーを埋め込んだ。論文詳細ページの一例は以下の図 4.5 のように示す。

マウストラッカーは実験参加者が閲覧開始というボタンを押す瞬間から記録できるようになり、閲覧終了ボタンを押すと記録が停止する。今回の評価実験の需要に応じて、ノイズデータの生成を減らす目的で二つの設定をした。一つ目はマウストラッカーが記録する部分は論文の要旨部分に限定すると設定した。二つ目は行間設定をデフォルトの 1.5 倍 (30px) に設定した。

マウストラッカーが記録した閲覧中のマウスポインターの行動データは実験参加者が send ボタンを押す時に事前に設置した Google Script によって Google Form まで転送される。

以上のような論文詳細ページを四つ用意し、実験入口となるホームページを構築した。閲覧を円滑に行えるように、それぞれ情報科学分野における text mining、information seeking behavior、recommendation system、interactive information retrieval の四つの主題から比較的に読みやすく、かつハードルが高い英単語のない四つの論文 [19, 34, 35, 36] を選定して、評価実験の閲覧材料にした。この四つの論文の主題はそれぞれアニメ推薦、キーワード抽出のサーベイ論文、高齢者における健康情報探索活動、マウス動作によるユーザ分析である。

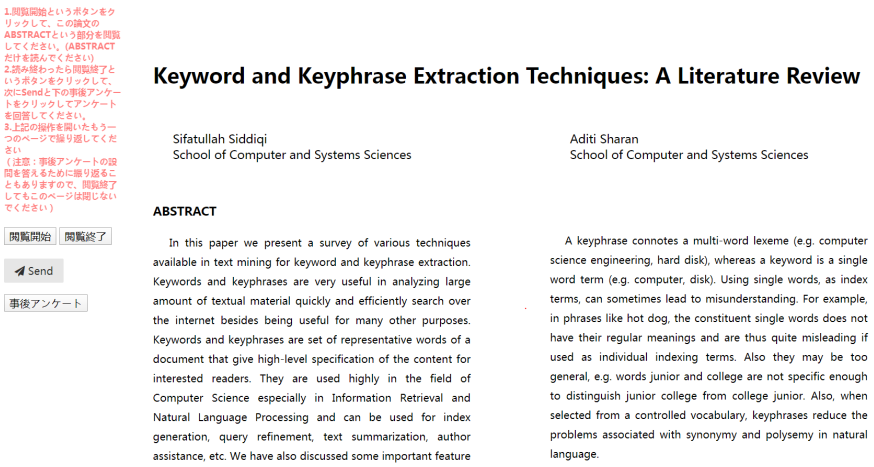


図 4.5: 論文詳細ページの一例

4.5 評価指標

このセクションでは、有効性を検証するための指標について説明する。4.5.1 節は評価に用いる正解データについて説明し、4.5.2 節は評価の基準を説明する。4.5.3 節は比較対象となるベースライン手法について説明する。

4.5.1 正解データ

評価に用いる正解データは二種類に分けて評価を行う。まずは実験参加者単体ごとに提案手法の有効性を評価するための正解データである。それぞれの実験参加者の論文に対する理解や関心ポイントが異なるため、論文に対して重要だと思うキーワードリストも実験参加者によって異なる。したがって実験参加者が事後アンケートで判断した重要だと思うキーワードリストを正解データとして、その実験参加者のポインター行動によって作成した特徴語リストを評価する。

次に、集約後の提案手法による論文に対する総合的な特徴語リストを評価するための正解データは、それぞれ実験参加者が判断した重要だと思う単語リストの集合全体から、単語の重複回数を基準に上位 10 位まで抽出し、正解リストと見なす。図 4.6 では論文の正解リストの集合の一例を表す。横軸は単語を表し、縦軸は実験参加者が判断した正解リストの集合における各単語の重複回数である。重複回数に基づいて上位 10 位までの単語を正解リストとして取っているが、実際には上位 10 位と同じ重複回数の単語が存在する場合は上位 10 位と同じ重複回数の単語を全部正解として扱うことにする。

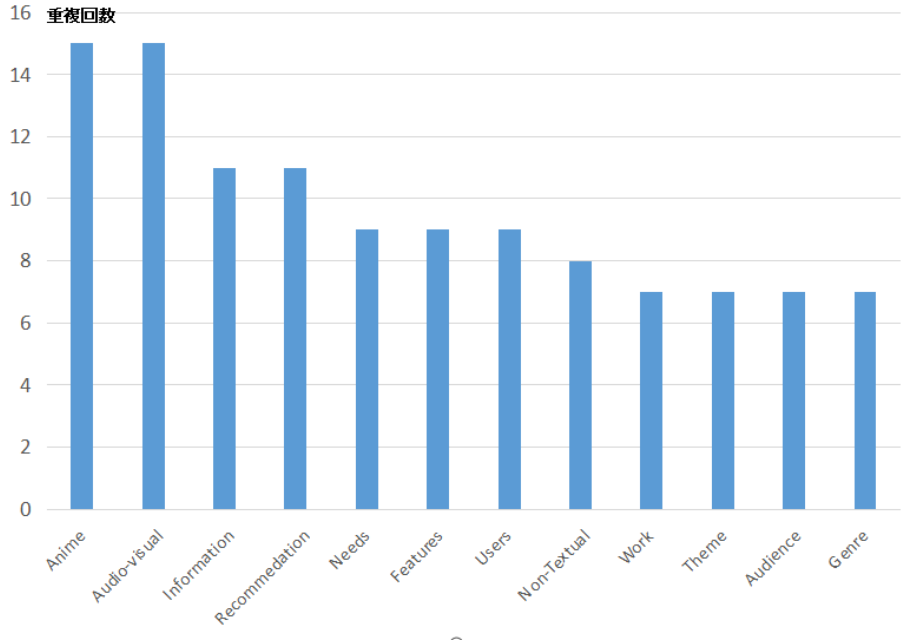


図 4.6: 論文 3 の正解リストの集合

4.5.2 評価基準

それぞれの正解リストを用い、有効性を評価するための指標として、精度、再現率と F-score を用いる。それぞれ以下の式で表す：

$$precision = \frac{|A \cap B|}{|A|} \tag{4.1}$$

$$recall = \frac{|A \cap B|}{|B|} \tag{4.2}$$

$$F\text{-}score = \frac{2 * precision * recall}{precision + recall} \tag{4.3}$$

A: 手法によって抽出した単語
B: 正解リスト

4.5.3 ベースライン手法

提案手法と比較するためのベースラインとして、キーワード抽出分野においてよくベースラインとして使われる TF-IDF 手法 [3] と TextRank 手法 [37] を採用する。

今回の閲覧材料に使用した四本の学術論文は ACM デジタルライブラリーからダウンロードしたものであり、クローラーを使って自動的にアクセスすることが禁止されているため、TF-IDF 手法の IDF 計算は人手で単語ごとにクエリとして ACM サーチエンジンに入力し

て検索し、サーチエンジンからヒット件数に応じて得られた結果を計算することで IDF 結果を得た。

精度と再現率の計算について、提案手法が抽出する単語の数が限られているため、かつ研究目的の一つであるより網羅的な学術検索支援のため、今回は評価指標において上位 20 位までの精度も追加し、評価しやすいように三つの手法による上位 20 位と 10 位までの単語で評価を行う。

4.6 評価結果

まず、4.6.1 節は係数の最適化について説明をする。4.6.2 節は、正解は実験参加者が重要だと思う単語リストである場合に提案手法とベースライン手法が得た結果を述べる。4.6.3 節は、正解は実験参加者が重要だと思う単語リストの集合を集計した場合に提案手法とベースライン手法が得た結果を述べる。4.6.4 節は実験参加者数の増加が集約後の提案手法による総合的な特徴語抽出手法に与える効果について検討する。

また、今回の評価実験で 7 人の実験参加者はデータが取れなかったため（閲覧中にほぼポインターを動かしていなかった)、この 7 人を除外して分析を行った。また、準備した四つの論文は三つしか選ばれていなかったため、残りの選ばれなかった論文は除外する。

4.6.1 係数の調整と最適化

実験データセットから 4 分の 1 のデータを切りだし、訓練データとして係数の調整と最適化を行う。また、係数の調整は重み係数 α について行う。

テキスト選択特徴は出現回数が低く、今回の実験では 4 人の閲覧セッションしか現れなかった。しかし、現れた閲覧セッションでテキスト選択された語はすべて正解の単語であった。対応する重み係数 p は実数値の 0.6 と設定した。なぜなら、予備実験より重み付けアルゴリズム (3.2 式) の第 1 項と第 2 項は主にそれぞれ 0.1 から 0.5 までの値として算出されることが多く、これら第 1 項と第 2 項よりも強調して重み付けがされるよう 0.6 を設定した。この 0.6 という値により、テキスト選択された単語の重みは上位に位置づけられることができる。

3.2 式の係数 α をそれぞれ違う値で手法の効果を検証した結果、α は 0.05 の値で最も良い効果を得られた (表 4.2)。α は 0.5 以上の値で手法効果の変化はなかった。このセッティングにおいて第 2 項の特徴 (ポインターが通過した回数) はアルゴリズムに対する影響が小さく、ランキングの順序はほぼ第 1 項の特徴 (なぞり読み) で決まる。0.5 から 0.05 の間は少し上昇が見られ、このセッティングにおいて第 2 項の特徴のアルゴリズムに対する影響が上昇し、0.05 の値で第 1 項と第 2 項の特徴がアルゴリズムに及ぼす影響はほぼ同じであった。

0.05 からさらに値が小さくなると手法の効果が下がる。このセッティングにおいて第 2 項の特徴がアルゴリズムに及ぼす影響が強く、ランキングの順序はほぼ第 2 項で決まる。

表 4.2: 係数 α の最適化

α	精度
0.8	37.5 %
0.7	37.5 %
0.6	37.5 %
0.5	37.5 %
0.4	38.0 %
0.3	38.0 %
0.2	38.0 %
0.1	39.0 %
0.075	39.0 %
0.05	41.2 %
0.035	39.5 %
0.025	38.5 %
0.015	37.5 %
0.01	36.5 %

4.6.2 実験参加者が重要だと思う単語リストを正解データとした場合

実験参加者が重要だと思う単語リストを正解データとした場合に、提案手法とベースライン手法が得た上位 20 位と上位 10 位までの精度、再現率と F 値を表 4.3、4.4 のように示す。提案手法 (集約) は 3.4.1 節で述べたリランキング操作を行った後の提案手法を指す。

表 4.3: 上位 20 位までの精度、再現率と F-score

	精度	再現率	F-score
TextRank	20.5 %	19.1 %	19.8 %
TF-IDF	27.1 %	25.7 %	26.4 %
提案手法	28.8 %	27.3 %	28.0 %
提案手法 (集約)	28.9 %	27.4 %	28.1 %

表 4.3、4.4 が示している通り、提案手法は TextRank 手法より良い数値を得られ、TF-IDF 手法より若干良い数値を得られた。リランキングによる効果は上位 10 位までの精度において上昇が少し見られ、上位 20 位までの精度においては効果の上昇がほぼ見られなかった。

表 4.4: 上位 10 位までの精度、再現率と F-score

	精度	再現率	F-score
TextRank	24.0 %	22.3 %	23.1 %
TF-IDF	40.3 %	38.7 %	39.5 %
提案手法	39.8 %	38.1 %	38.9 %
提案手法 (集約)	41.8 %	39.8 %	40.8 %

それぞれ上位 10 位と上位 20 位までの精度を用いて論文と手法の二要因混合分散分析を行った。図 4.7 では上位 20 位までの精度を用いた分散分析の結果を示す。

図 4.7 において、横軸 A1、A2、A3 はそれぞれ三つの論文に対応し、B1、B2、B3 と B4 はそれぞれ TextRank 手法、TF-IDF 手法、提案手法と提案手法 (集約) を表す。分散分析の結果として、手法間には有意な差が見られ ($F\ (3, 72) = 11.07, p<.01$)、LSD 法を用いて多重比較を行った結果、TextRank 手法と提案手法もしくは提案手法 (集約) の間で有意な差が見られ、TextRank 手法と TF-IDF 手法の間にも有意な差が見られ ($p<.05$)、TF-IDF 手法と提案手法、提案手法 (集約) の間では有意な差が見られなかった。また、論文間にも有意な差が見られ ($F\ (2, 24) = 3.01, p<.10$)、LSD 法を用いて多重比較を行った結果、論文 1 と論文 2 の間にも有意な差が見られ ($A1<A2, p<.05$)、論文 1 と論文 3、論文 2 と論文 3 の間では有意な差は見られなかった。

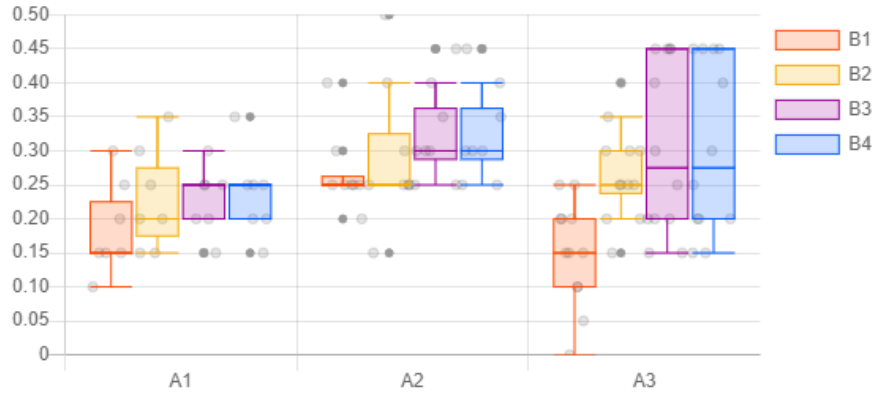


図 4.7: 上位 20 位までの精度分布 (A: 論文 B: 手法)

4.6.3 実験参加者が重要だと思う単語リストを集計して正解データとした場合

実験参加者が重要だと思う単語リストの集合を集計して正解データとした場合に、各手法にて得られる効果を表 4.5、4.6、4.7、4.8 のように示す。

各表は、提案手法は三つの論文における性能が TextRank 手法より良いことを示している。

表 4.5: 論文 1 における三つの手法の効果

	上位 20 位までの精度	上位 10 位までの精度	上位 10 位までの再現率
提案手法	25.0 %	40.0 %	36.4 %
TextRank	15.0 %	10.0 %	9.1 %
TF-IDF	25.0 %	50.0 %	45.5 %

表 4.6: 論文 2 における三つの手法の効果

	上位 20 位までの精度	上位 10 位までの精度	上位 10 位までの再現率
提案手法	35.0 %	60.0 %	37.5 %
TextRank	35.0 %	50.0 %	31.3 %
TF-IDF	45.0 %	60.0 %	37.5 %

論文 1 では TF-IDF 手法が 50 %（上位 10 位まで）の精度を得ているのに対し、提案手法は 40 %の精度を得ていた。論文 2 では提案手法が TF-IDF 手法と同じく 60 %の精度を得ていることがわかる。論文 3 では TF-IDF 手法が 50 %の精度を得ていることに対し、提案手法は 70 %の精度を得ていた。三つの論文における効果の平均値から見ると、提案手法は TextRank 手法よりはるかに良い性能を見せ、TF-IDF 手法と比べて比較的よい性能を持っていることが分かった。

4.6.4 実験参加者数の増加による変化

集約後の提案手法による総合的な特徴語抽出手法において、実験参加者のデータセットの増加による変化を調べるために、それぞれの論文ごとに実験参加者人数の三分の一ずつを分割し、提案手法が異なる実験参加者数のもとで与える効果を調査した。表 4.9 は実験参加者数の三つの段階ごとに得られる効果を表す。

表が示している通り、三つの論文における提案手法の効果を平均値で取ると、実験参加者三分の一から実験参加者三分の二までになると、手法の精度は 7 ポイントの上昇が見られ、実験参加者三分の二から実験参加者全員までに変わると、2.4 ポイントの上昇があった。提案手法は全体的に実験参加者の人数が増えることによって効果が上がる傾向があると考えられる。ただ、一定人数になると、上昇の効果が安定すると推測できる。

表 4.7: 論文 3 における三つの手法の効果

	上位 20 位までの精度	上位 10 位までの精度	上位 10 位までの再現率
提案手法	45.0 %	70.0 %	58.3 %
TextRank	10.0 %	10.0 %	8.3 %
TF-IDF	30.0 %	50.0 %	41.7 %

表 4.8: 論文 1 から 3 までの評価指標の平均値

	上位 20 位までの精度	上位 10 位までの精度	上位 10 位までの再現率
提案手法	35.0 %	56.7 %	44.1 %
TextRank	20.0 %	23.3 %	16.2 %
TF-IDF	33.3 %	53.3 %	41.6 %

表 4.9: 異なる実験参加者数による変化 (上位 10 位までの精度)

	実験参加者三分の一	実験参加者三分の二	実験参加者全員
論文 1	40 %	35 %	40 %
論文 2	50 %	60 %	60 %
論文 3	52 %	68 %	70 %
平均値	47.3 %	54.3 %	56.7 %

第5章　考察

本章では、評価実験の結果に基づいて、第1章で述べたリサーチクエスチョンに回答し、次に提案手法の有効性、重み計算に用いる特徴、または提案手法の制限と改善点について考察を行う。

5.1　リサーチクエスチョンに対する回答

RQ1: ポインターから読者の関心を抽出することによって文書のキーワードを推測できるか？

表4.4で示している通り、提案手法は41.8 %の精度を得られ (上位10位までの精度)、分散分析をした結果としてTextRank手法と有意な差が見られ、TF-IDF手法と同程度の効果を得られたということからポインター行動を用いて読者の関心を抽出することによって文書のキーワードを抽出できるということが言える。

RQ2: 提案手法はベースライン手法よりよい効果を得られるか？

表4.8が示しているように、文書に対する総合的な特徴語リストにおいて、上位20位までの精度でTextRank手法で得られた20 %とTF-IDF手法で得られた33.3 %に対し、提案手法は35 %の精度を得られた。上位10位までの精度でTextRank手法で得られた23.3 %とTF-IDF手法で得られた53.3 %に対し、提案手法は56.7 %の精度を得られた。提案手法の効果はベースライン手法のTextRank手法より良い効果が得られ、TF-IDF手法と比較して若干良い効果を得られた。

RQ3: ポインターでなぞり読みを行わないユーザに対して、複数ユーザからの集約による総合的な特徴語抽出手法は有効であるか？

実験参加者全員25人のうち、7人は閲覧中ほぼポインターを動かさなかったため、データの入手はできなかった。

図5.1はこの7人がアンケートで書いた重要だと思う単語リストを対象に、複数ユーザからの集約によって作成した総合的な特徴語リストを用いて評価した結果を表す。結果として、平均的に42.9 %の精度 (上位10位まで) を得られたことから有効であると判断できる。すなわち、ポインター行動を入手できないユーザに対しても、他のユーザから得られた総合的な特徴語リストを用いることにより、一定の効果を持つキーワード抽出が可能である。

RQ4: ユーザデータの集合が増えることによって、複数ユーザからの集約による総合的な

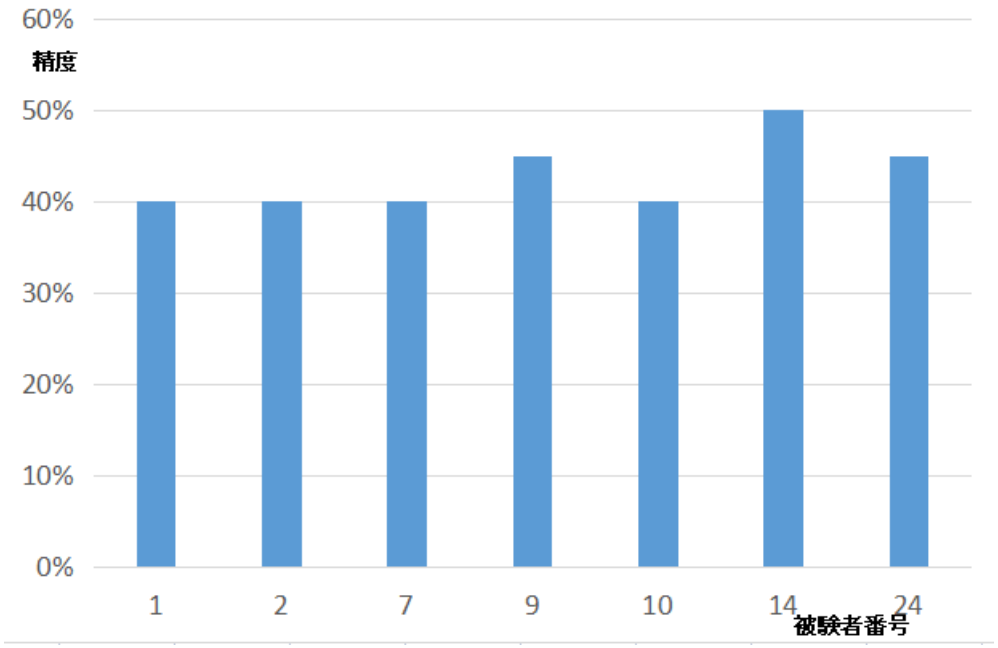


図 5.1: なぞり読みを行わないユーザに対して提案手法が得られた効果 (上位 10 位までの精度)

特徴語抽出手法の効果はよくなるか？

表 4.9 が示しているように、実験参加者の数が三分の一から実験参加者三分の二までに変わると、全体的に 7 ポイントの精度上昇が見られ、実験参加者三分の二から実験参加者全員までになると、全体的に 2.4 ポイントの精度上昇が見られた。提案手法はユーザデータの集合が増えることによって精度がよくなる傾向が見られるが、一定数のユーザデータセットが蓄積すれば提案手法の精度の上昇は緩やかになるという傾向を見せている。

5.2 重み計算に影響を与える特徴

重み付けアルゴリズムにある三つの特徴について、それぞれ重み計算において重みに対する影響を調査するために、それぞれの特徴を単独で用いて重み計算をした結果、第 1 項の特徴 (なぞり読み) と第 2 項の特徴 (ポインターが通過した回数) はどちらも重み計算において重要な役割を果たしていることが分かった。三つの特徴をそれぞれ単独で重み計算を行ったときに得られた平均的な上位 10 位までの精度を以下に示す。

- なぞり読み:36.7 %
- ポインターが通過した回数:36.4 %
- テキスト選択：2.6 %

図 5.2 は三つの特徴を単独で計算したときにで得られた精度の分布を表す。

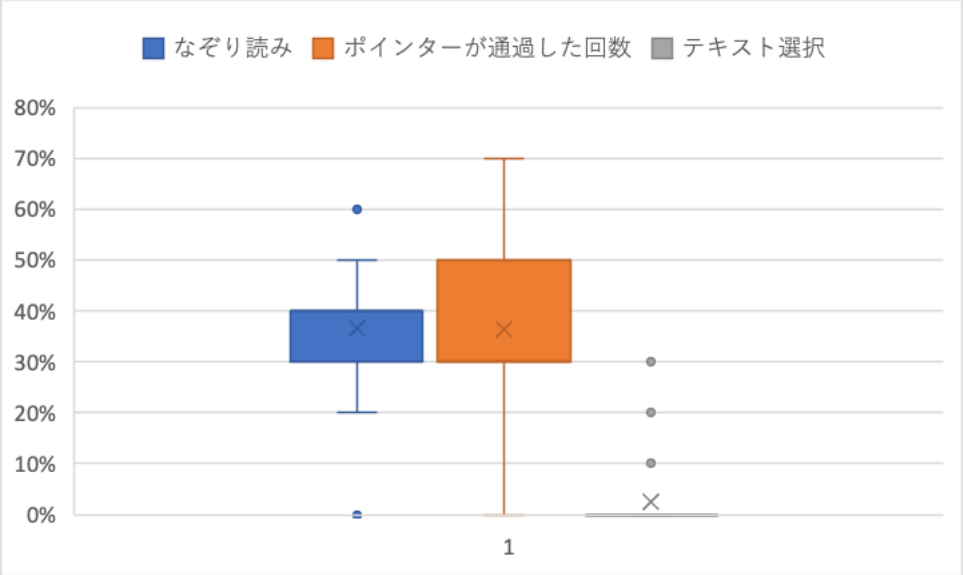


図 5.2: 各特徴の単独計算で得られた精度

第 1 項の特徴 (なぞり読み) と第 2 項の特徴 (ポインターが通過した回数) は単独で重み計算を行った結果としては得られた精度がほぼ同じ水準であった。テキスト選択特徴は出現頻度が低いため、単独で計算したときは得られた精度が低いが、テキスト選択された単語での正解率が高いため重み計算において必要とされる特徴と考えられる。

また、上記の結果は 4.6.1 節の重み計算 α の調整と最適化の結果とも一致し、その理由は係数 α が最適化された値で第 1 項の特徴と第 2 項の特徴が全体的な重み計算アルゴリズムに及ぼす影響はほぼ同じためである。

5.3 ユーザの個性が提案手法に与える影響

ユーザの個性が提案手法に与える影響を調査するために、事前調査アンケートにおけるポインターでなぞり読みする習慣の回答に基づいて実験参加者を分類し、それぞれのタイプの実験参加者が提案手法によって得られる効果を調査した。

5.3.1 なぞり読みが頻繁にあるタイプ

事前調査アンケートに回答した 25 人のうち、ポインターでなぞり読みする習慣について頻繁にある（以下、タイプ a で示す）と答えたのは 8 人であるが、この 8 人のうち、1 人は実験中ポインターを動かさなかったため分析から除外した。それぞれ集約前の提案手法によって得られた上位 10 位までの精度を表 5.1 で示す。

表 5.1: タイプ a の実験参加者で提案手法が得られた効果

実験参加者番号	なぞり読みする習慣	論文 1	論文 2	論文 3	平均値
3	a	30 %	40 %	未選択	30 %
5	a	50 %	未選択	70 %	60 %
8	a	40 %	未選択	30 %	35 %
12	a	30 %	未選択	40 %	35 %
17	a	30 %	未選択	40 %	30 %
20	a	40 %	50 %	未選択	45 %
25	a	未選択	50 %	50 %	50 %
平均値					43.6 %

表 5.2: タイプ b の実験参加者で提案手法が得られた効果

実験参加者番号	なぞり読みする習慣	論文 1	論文 2	論文 3	平均値
4	b	未選択	60 %	40 %	50 %
6	b	未選択	50 %	40 %	45 %
13	b	未選択	50 %	30 %	40 %
16	b	未選択	60 %	60 %	60 %
19	b	未選択	50 %	60 %	55 %
21	b	40 %	未選択	40 %	40 %
23	b	未選択	45 %	55 %	50 %
平均値					48.6 %

5.3.2 なぞり読みがたまにあるタイプ

事前調査アンケートを回答した 25 人のうち、ポインターでなぞり読みする習慣についてたまにある（以下、タイプ b で示す）と答えたのは 10 人であるが、この 10 人のうち、3 人は実験中ポインターを動かさなかったため分析から除外した。それぞれ集約前の提案手法によって得られた上位 10 位までの精度を表 5.2 で示す。

5.3.3 なぞり読みをしない、自分でもわからないタイプ

事前調査アンケートを回答した 25 人のうち、ポインターでなぞり読みする習慣についてない、（以下、タイプ c で示す）自分でもよくわからない（以下、タイプ d で示す）と答えたのはそれぞれ 4 人と 3 人であるが、タイプ c の実験参加者は 3 人、タイプ d の実験参加者

は 1 人が実験中ポインターを動かさなかったため分析から除外した。それぞれ集約前の提案手法によって得られた上位 10 位までの精度を表 5.3 で示す。

表 5.3: タイプ c と d の実験参加者で提案手法が得られた効果

実験参加者番号	なぞり読みする習慣	論文 1	論文 2	論文 3	平均値
2	c	未選択	データ取れず	40 %	40 %
11	d	未選択	0 %	30 %	15 %
15	d	20 %	未選択	30 %	25 %
22	d	30 %	30 %	未選択	30 %
平均値					27.5 %

以上の三つの表が示している通り、提案手法はタイプ a とタイプ b の実験参加者での効果はタイプ c とタイプ d での効果よりはるかに良いことがわかる。タイプ b の実験参加者による提案手法の効果はタイプ a の実験参加者による提案手法の効果よりも高いという結果は二つの要因があると考えられる。一つ目の要因はタイプ b の実験参加者のうち、3 人がマウストラッカーによって記録されたデータの数は非常に高く、実験時の zoom 録画をチェックしたところ、この 3 人はなぞり読みする習慣についてたまにあると答えたが、実際に閲覧したときは頻繁になぞり読みをすることが分かった。二つ目の要因としてはたまになぞり読みをするユーザは自分が気になっている部分、もしくは重要だと思う部分でしかなぞり読みをしないという可能性が高いと考えられる。したがって、このようなタイプのユーザによって提案手法で抽出した特徴語リストの精度が高いと推測できる。

また、データが取れなかったタイプ c と d の実験参加者のうち、二人が論文閲覧中にポインターを動かさなかったが、事後アンケートの重要だと思う単語を入力するという質問に回答するときに、振り返って論文を読むときは逆になぞり読みをする現象が起きたことが zoom 録画によって観察された。これは与えるタスクの違いによっての影響なのではないかと推測できる。

5.4 マウストラッカーの記録回数による影響

ユーザが閲覧中に示すポインターの動作頻度と提案手法との関係を明らかにするために、マウストラッカーが記録したデータ量を抽出し、ポインターが一回単語を通過したことを一回の記録としたとき、マウストラッカーの記録回数を、1 人のユーザの閲覧中のポインターの動作頻度とみなすことができる。それぞれの実験参加者によるマウストラッカーの記録回数と対応する提案手法で得られた効果の関係を図 5.3 に示す。図の縦軸は提案手法で得られた上位 10 位までの F-score を表し、横軸は実験参加者それぞれの閲覧時におけるマウストラッカーによる記録回数を表す。マウストラッカーの記録回数が大きければ大きいほど閲

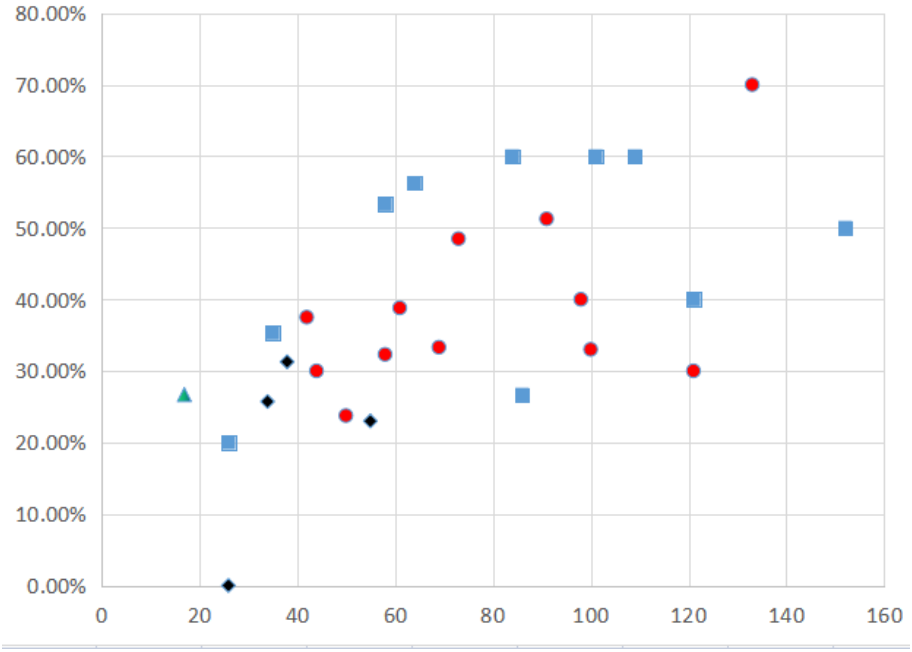


図 5.3: 記録回数と対応する F-score

覧時のポインター行動が活発に行われたと考えられる。図の中で赤い点、青い点、緑の点と黒い点はそれぞれタイプ a、タイプ b、タイプ c、タイプ d の実験参加者を表す。

記録回数と F-score について相関分析を行った結果、やや強い相関があることがわかった (相関係数 r=0.609)。提案手法は閲覧中にポインター行動が頻繁に行われるユーザでの効果がより良い傾向が見られる。

5.5 提案手法の改善点と展望

本節では提案手法が存在するいくつかの問題点と改善点について説明を行い、今後の課題について検討する。

5.5.1 今回の評価実験のセッティング

今回の実験では事前に選定した四つの学術論文の要旨を対象に行われた。実験参加者は 5 分以内で選んだ論文の要旨を読み終わるという設定であったが、要旨のような短い文書に対して、ユーザの個性の影響で提案手法にて抽出される特徴語の数が少ない場合があるという問題が存在した。

また、今回の実験で TF-IDF 手法は高い性能を示していたが、これはおそらく高度に情報がまとまった論文要旨を対象にしたことによる影響と推測する。

今後は実験のセッティングについて、全文の閲覧に基づいた場合の提案手法の効果も検討したいと考える。

5.5.2 提案手法が誤って抽出した単語

集約する前の提案手法によって抽出した特徴語リストのうち、実験参加者が事後アンケートでわからないと回答した単語を誤って特徴語として抽出したケースが 3 件あった。これらの単語を誤って特徴語として抽出した原因は実験参加者が閲覧中に自分がわからない単語に遭遇したとき、理解しようとするか戸惑っていることによってポインターはその単語の停留時間が高くなり、したがって重み計算で高い重みを得られたと考えられる。一方、このような単語は複数ユーザからの集約によって抽出した特徴語の集合においては重複回数が低いため、この問題は集約後の提案手法によって解決できた。

5.5.3 TF-IDF 手法で抽出できなかった単語

TF-IDF 手法によって抽出した上位 20 位までの単語と提案手法によって抽出した上位 20 位までの単語を比較すると、主題がアニメ推薦と高齢者健康情報探索活動である論文 2 と論文 3 では user, needs, information、主題がキーワード抽出サーベイである論文 1 では feature などの正解単語を TF-IDF 手法によって抽出できなかったが、提案手法は抽出できた。上記の単語はコーパスにおいてよく使われる単語であるため IDF 値が低くなり、かつ文書の中では出現頻度である TF 値も高くないことから TF-IDF 手法における重み計算が小さくなると思われる。それに対して提案手法では文書内の特徴を用いず読者の理解や関心に基づいた文書の特徴語を抽出しているため、このような問題を避けることができる。

一方で、5.5.1 節でも述べたように、提案手法は短い文書を対象とするときは、頻繁にポインターでなぞり読みする習慣を持たないユーザによって抽出される特徴語の数が不足しているケースに対して、文書内の特徴を用いるベースライン手法ではこの問題を解決することができる。したがってベースライン手法はある程度提案手法を補えるのではないかと考える。今後の課題としてうまくベースライン手法と組み合わせることでより良い提案手法の効果が得られることが期待できる。

第6章　おわりに

本研究では読者が学術論文を閲覧するときに示すポインター行動から学術論文の特徴語リストの抽出を試みた。抽出した特徴語をそれぞれ複数ユーザによる集約後と集約前に基づいて有効性を検証した。評価実験の結果により、集約前の提案手法は TextRank 手法より有意な差が見られ、TF-IDF 手法とは有意な差が見られなかった。複数ユーザからの集約後の提案手法は精度 56.7 %の効果（上位 10 位まで）が得られ、ベースライン手法の精度 53.3 %と 23.3 %と比べて比較的に良い結果を得られた。この結果は提案手法に対する改善や実験参加者のデータセットの拡張によって一定程度に性能向上が期待され、提案手法の有効性を示すことができた。

ポインターでなぞり読みする習慣について、提案手法の効果は頻繁にあるユーザとたまにあるユーザでの性能はなぞり読みする習慣を持たないユーザよりよいということがわかった。また複数ユーザからの集約によってなぞり読みする習慣を持たないユーザからの影響を避けることができた。マウストラッカーの記録回数と提案手法の効果との関係を検証したところ、記録回数と提案手法の効果はやや強い正の相関があることがわかり (r=0.609)、記録回数の多いユーザでの提案手法の効果が良いという傾向が見られた。

一方、ベースライン手法単独で抽出できなかった単語が提案手法によって抽出できる例があること、さらに、文章内特徴を用いることで提案手法の改善にもつながることから、今後の課題として文書内の特徴とポインター行動の特徴を組み合わせることで提案手法の改善を目指していきたい。

謝辞

この論文の執筆にあたり、多くの方々からたくさんのご支援いただいたことに感謝の意を表します。

まずは指導教員の高久雅生先生には、いろいろお世話になりました。最初の研究の着想から、関連文献の調査、手法の実装、評価実験の設計、最後の論文執筆まで多くのご指導をいただきました。心から感謝申し上げます。

また、同研究室のみなさまには多くのアドバイスやコメントをいただきまして、誠にありがとうございました。

次に博士前期課程の授業を担当される先生方や論文審査するときにたくさんのアドバイスをいただいた先生方には深く感謝申し上げます。

最後に、日常生活で支えてくださった家族に深くお礼を申し上げます。

日本でのこの二年半の留学生活は私の人生の貴重な経験となりました。これからも大切にしていきたいと思います。

参考文献

[1] Blank et al, Leveraging the citation graph to recommend keywords, RecSys ’13, 2013, p.359-362.

[2] Rose et al, Automatic Keyword Extraction from Individual Documents, Text Mining: Applications and Theory, 2010, p.1-20.

[3] Salton, G, Automatic Text Processing, Addison-Wesley Publishing Company, 1989.

[4] Robertson et al, Okapi at TREC4, Proc. of the 4th Text REtrieval Conference, 1996, p.73-96.

[5] Boudin et al, Keyphrase Generation for Scientific Document Retrieval, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, p.1118-1126.

[6] Meng et al, Deep Keyphrase Generation,In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, p.582-592.

[7] Hienert et al, Reading Protocol: Understanding what has been Read in Interactive Information Retrieval Tasks, Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, 2019, p.73-81.

[8] ユーザー行動分析がひと目でわかる「ヒートマップ」とは？. 2020-03-13, https://www.kwm.co.jp/blog/heatmap/, (accessed 2022-01-21)

[9] Liu, F. et al, Personalized Web Search by Mapping User Queries to Categories, Proceedings of the Eleventh International Conference on Information and Knowledge Management, 2002, p558– 565.

[10] Ajanki et al, Can eyes reveal interest? Implicit queries from gaze patterns, User Modeling and User Adapted Interaction, Vol.19, No.4, 2009, p.307-339.

[11] 高久雅生, 江草由佳, 寺井仁, 齋藤ひとみ, 三輪眞木子, 神門典子, タスク種別とユーザ特性の違いが Web 情報探索行動に与える影響: 眼球運動データおよび閲覧行動ログを用いた分析, 情報知識学会誌, Vol.20, 2010, p.249.

[12] Sugimoto, M, User Modeling and Adaptive Interaction in Information Gathering Systems, Journalof Japanese Society for Artifificial Intelligence, Vol. 14, No. 1, 1999, pp. 25–32.

[13] Lang, K, NewsWeeder: Learning to Filter NetNews, Proc. of ICML ' 95 , 1994, pp. 331–339 .

[14] Smyth, B, P, A Personalized Television Listings Service, Comm. of the ACM, Vol. 43, No. 8, 2000, pp. 107–111.

[15] Morita, M, Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval, Proc. of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 272–281.

[16] Balatsoukas et al, 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, 2012, p.1728-1746.

[17] Huang et al, User see, user point: gaze and cursor alignment in web search, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, p.1341-1350.

[18] Bharti et al, Automatic Keyword Extraction for Text Summarization: A Survey, http://arxiv.org/abs/1704.03242, 2017, p.1-12.

[19] Siddiqi et al, Keyword and keyphrase extraction techniques: a literature review, International Journal of Computer Applications, Vol.109, No.2, 2015, p.18-23.

[20] Ramos, Using tf-idf to determine word relevance in document queries, Proceedings of the first instructional conference on machine learning, 2003, p.1-4.

[21] Barzilay et al, Using lexical chains for text summarization, In Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, 1997, p.10-17.

[22] Hong et al, An Extended Keyword Extraction Method, International Conference on Applied Physics and Industrial Engineering, 2012, p.1120-1127.

[23] Mihalcea et al, TextRank: Bringing order into texts, Proceedings of EMNLP, 2004, p.404-411.

[24] Li et al, A Keyword Extraction Method for Chinese Scientific Abstracts. Proceedings of the 2017 International Conference on Wireless Communications, Networking and Applications, 2017, p.133-137.

[25] 原田隆史，細野公男，野美山浩, 諸橋 正幸, 抄録からのキーワード自動抽出，Information Processing Society of Japan，SIG Notes 31(8), 1994, p55-61.

[26] MA et al, Keywords Extraction Algorithm Based on Weighted Hypergraph Random Walk, Acta Electronica Sinica, 46(6), 2018, 1410-1414.

[27] 中川裕志，湯本紘彰，森辰則, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, 10 巻 1 号, 2003, p. 27-45.

[28] Kantor et al, Capturing Human Intelligence in the Net, Comm. of the ACM, Vol.43, No.8, 2000, p.112-115.

[29] Guo, Q et al, Exploring Mouse Movements for Inferring Query Intent, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008, p707–708.

[30] Huang et al, Improving Searcher Models Using Mouse Cursor Activity, Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2012, p.195-204.

[31] Hienert et al, Term-Mouse-Fixations as an Additional Indicator for Topical User Interests in Domain-Specific Search, ICTIR ’17, 2017, p.249-252.

[32] Ageev, M. et al, Improving Search Result Summaries by Using Searcher Behavior Data. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013, p.13–22.

[33] Hijikata et al, Implicit User Profiling for On Demand Relevance Feedback, Proceedings of the 9th international conference on Intelligent user interfaces, ACM, 2004, p.198-205.

[34] Cho, H et al, Information Needs for Anime Recommendation: Analyzing Anime Users’ Online Forum Queries, Proceedings of 2017 ACM/IEEE Joint Conference on Digital Libraries(JCDL), 2017), p.305-306.

[35] Huang et al, Older Adults’ Online Health Information Seeking Behavior, iConference ’12: Proceedings of the 2012 iConference, 2012, p.338-345.

[36] Guo, Q et al, Predicting Web Search Success with Fine-grained Interaction Data. CIKM ’12: Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, p.2050-2054.

[37] Barrios, F, Variations of the Similarity Function of TextRank for Automated Summarization, Argentine Symposium on Artificial Intelligence (ASAI), 2015, p.65-72.