Exploring Text Highlighting and Copying by Web Users on a Website: A Case Study

Ilan Kirsh
ORCID 0000-0003-0130-8691

Abstract

Selecting text and copying text to the clipboard are common operations of many web users. This study analyzes real usage data from a commercial website to understand what types of textual content users select and copy, for what purposes, and what we can use such user activity data for. This paper distinguishes between two different types of operations: (a) highlighting text by selecting it without copying, for example, to emphasize elements in the text while reading; and (b) copying text to the clipboard, where the preceding selection operation has no distinct function and thus considered part of the copy operation. The paper advocates treating text highlighting and text copying as being part of human-computer dialogues, in which the computer can also gain knowledge about users, their preferences, and their needs, based on their actions. Accordingly, different uses and applications are proposed and discussed, spanning a wide range of areas, including web analytics, web personalization, adaptive websites, text simplification, text summarization, detection of plagiarism, search engine optimization, and reading analysis.

Keywords

Web Analytics, Text Selection, Text Highlighting, Copy & Paste, Human-Computer Interaction.

The Version of Record is available online at: 10.1007/s42979-022-01146-5,

in SN Computer Science 3(4) , 278, Springer, 2022.

1 Introduction

Web analytics tools are widely used to better understand the interests, preferences, needs, and actions of website users. Premium web analytics services track client-side user activity at a very low level and go as far as to record all user mouse movements. However, tracking operations of selecting text and copying text to the clipboard by web users is currently not part of the web analytics toolbox.

This study examines text selecting and copying to the clipboard by users of a website. It also proposes possible practical uses of web usage data of this form, in a wide range of applications. For this study, data of more than one million pageviews with hundreds of thousands of text selection operations have been collected for six months. Interim results, based on three-month data, regarding copying only (excluding selection operations), were presented at the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020) [17]. This paper presents the final results of this study, based on the complete six-month data. It extends the conference paper further in various ways, including by providing additional technical details, examples, and explanations, but the main addition is the analysis of text selection operations, and particularly selection operations that are not followed by copying, i.e., highlighting operations, which are analyzed and discussed in this paper.

Text selection is often considered merely the first step in copying text to the clipboard. However, the vast majority of text selection operations in this study's dataset are not followed by copying. Text selection can be used by online readers also to highlight points while reading, similarly, to highlighting words in a book with a marker pen. Accordingly, this paper distinguishes between two different types of operations: (a) highlighting text by selection without copying; and (b) copying text to the clipboard, where the preceding text selection operation has no distinct function and is therefore considered part of the copy operation.

The main goals of this research were to explore (a) what types of textual content users highlight and copy on a website; (b) what are the possible motivations of web users for highlighting and copying textual content of each one of these types; and (c) how we can use such user activity data on websites for extending web analytics capabilities and for other applications.

The paper is organized as follows: Section 2 reviews related work. Section 3 introduces the dataset that was used in this study and presents first observations regarding text highlighting and text copying in this data set. Section 4 presents examples of textual contents that were highlighted or copied and suggests a taxonomy of highlighting and copy operations into several categories and subcategories. It also discusses possible user motivations to highlight or copy each type of textual content. Section 5 discusses various directions for potential applications and uses that can benefit from data regarding text highlighting and copying. Section 6 concludes this paper and proposes future research directions.

2 Related Work

As various studies show, web analytics is effective in understanding how visitors use websites and can help in improving and optimizing websites and web applications. This has been demonstrated for a wide range

of industries, including, for example, online news [33], online learning [24], e-commerce [9], and digital marketing [4,11]. Web analytics concepts, principles, and methods are described in detail in various books [12,13]. Mouse cursor positions are often used to estimate which areas of the website capture the user's attention. Studies show a correlation between the position of the mouse cursor on the screen and the user's eye gaze [10], with a higher correlation when the user clicks the mouse or moves it [5,28]. Mouse cursor position information is an effective indicator of user's attention in various web applications, including in e-commerce [30], web marketing [34], online surveys [3], task execution [25], and web search [8,10,26,28].

The cumulative attention of all the visitors in different areas of a web page can be visualized by heatmaps [22,2327]. Attention heatmaps are also offered by many commercial web analytics services, which track, record, store, and visualize user mouse activity [18].

Unlike the mouse cursor position and mouse activity in general, data regarding text highlighting and copying are hardly used to gain knowledge. Recent studies proposed and demonstrated the visualization of users' attention on a website using heatmaps of cumulative text highlighting data [1820] and text copying data [18], similarly to the mouse cursor heatmaps, but commercial web analytics services do not offer such heatmaps yet.

Previous studies examined the use of Copy & Paste (C&P) operations by users. Analyzing C&P operations of expert Java programmers in the IBM T. J. Watson Research Center revealed several patterns of using C&P by software developers, such as copying code snippets to use them as templates for new code [14]. A large-scale study of C&P operations of 20,000 Eclipse IDE users analyzed C&P patterns statistically and found various patterns related to IDE users, such as the high frequency of C&P in the same file [2]. Based on repeating C&P patterns of software developers, various tools have been proposed for improving productivity [2,14]. Analysis of all the C&P operations of 15 users on Windows for four weeks revealed differences in the way C&P is used in different applications [32]. Particularly, it was found that web browsers are used more often as the source of C&P (i.e., copying), whereas word processors are used more often as the destination of C&P (i.e., pasting). However, that study has not examined what textual content is copied by users.

Interim results of this study presented an analysis of what is copied to the clipboard by users of a test case website [17]. Following these results, certain types of copy operations have been further explored. Copy operations of single words were found to be associated with word complexity, implying that such usage data can be useful in automatic text simplification [15]. Copy operations of complete sentences were found to be associated with sentence importance, implying possible uses in automatic text summarization [19].

This paper presents further results of this study, regarding text highlighting and text copying by web users, as discussed in section 1.

3 The Dataset

The dataset used in this study consists of activity data of users of the ObjectDB website, collected over six months in 2020. Subsection 3.1 describes the method that was used to collect that data. Subsection 3.2 provides general high-level statistical information about the collected data. Further analysis of the data, including low-level inspection of specific examples, is described in the next sections of this paper.

3.1 Data Collection Method

Figure 1 illustrates the architecture that was used to collect usage data. To track text selecting and copying by the website visitors, a reference to a Tracking Script was embedded in the website pages. As a result, every web page load triggered a request to load the Tracking Script from the Tracking Server. Once loaded, the tracking script recorded text selection and text copying and reported to the Collector component in the Tracking Server, which stored the data in a dedicated database. Following the usual practice of web analytics, and to protect user privacy, all the collected data were anonymized. The database with the collected data was accessed through the Reporter component, which supported retrieval of information regarding the recorded operations based on various parameters (e.g., type, length, and language of the copied content).

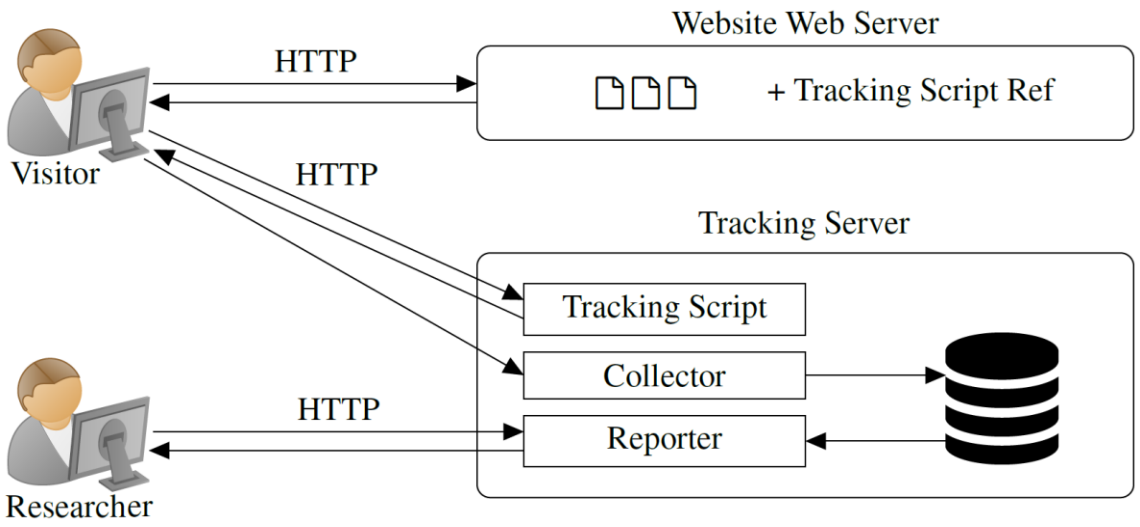

Figure 1: Architecture of the selection and copy tracking system.

Recording selection and copy operations of web users is related to *session-recording*, which is a common practice in modern web analytics of tracking and recording user activity on websites, including mouse movements and keystrokes. This practice raises questions related to personal data protection and user privacy, because of the risks of collecting sensitive personal data intentionally or unintentionally [7]. Session recording does not necessarily require prior user consent under personal data protection regulations, such as GDPR (under certain terms, as explained by the privacy and IT lawyer Arnoud Engelfriet [7]).

Unnecessary personal data should not be collected. If the collected data are completely anonymized, which is a standard practice in web analytics, then they are no longer considered personal data (e.g., according to GDPR). The terms and conditions and the privacy policies of many websites, including the website that was used in this study, explicitly specify that user activity on the website is tracked and recorded and that collected data can be analyzed, and particularly used to identify usage trends and to evaluate ways for improving the service. These activities are compatible with modern privacy protection rules such as the GDPR and the CCPA.

Tracking text selection operations is different from tracking other client-side user actions. Other user actions, such as clicking and moving the mouse, scrolling the page, and also copying text to the clipboard, can be tracked by listening to the relevant JavaScript events in the browser. Text selecting is different, as it does not trigger any JavaScript event. It is possible, however, to get information from JavaScript on which text is selected, if any, at any given point in time. Therefore, a JavaScript timer was used to check routinely for text selections, at a rate of once every tenth of a second. This posed an additional challenge, as single selection operations were detected multiple times with different text strings, due to users extending and shrinking selections. Therefore, detected selections were filtered: selections that were substrings of the preceding or the following selection were discarded.

This paper distinguishes between two different types of operations: highlighting and copying. Copy operations are easily identified by setting an *oncopy* event handler in JavaScript. For each copy operation, the preceding selection operation (i.e., the last selection operation in that pageview before the copy operation) was considered related to the copy operation and therefore was excluded from further analysis. All the other selection operations (after filtering redundant selections as explained above) were classified as highlighting operations. This classification process is not expected to be 100% accurate. For example, if a user started a selection with the intention of copying but eventually did not complete the copy operation (e.g., due to a change of mind), that selection was classified as highlighting. However, since the number of copy operations was considerably smaller than the number of highlighting operations, and also changing mind before copying does not seem to be very common, this classification method fits the purpose of this study.

3.2 General Statistical Information

The examined website contains mainly technical content for programmers. Many of the visits to the website are short. Software developers, who make up most of the traffic to the website, use the website as a learning and knowledge source. They frequently arrive from search engines for short visits, after searching for specific technical solutions and code examples.

Usage data were collected for six months in 2020 from 1,295,221 page-views. In total, 738,577 highlighting operations and 109,525 copy operations were recorded. Examining the distribution of these operations among the pageview shows an interesting difference between highlighting and copying. There were

179,145 pageviews with at least one highlighting operation and 76,148 with at least one copy operation. The ratio between the numbers of highlighting operations and copy operations was 6.74 and the ratio between the number of pageviews with these operations was 2.35. This indicates that highlighting operations tend to be more concentrated. Tables 1 and 2 show more details regarding this difference. Pageviews with a single copy operation are associated with 51.9% of the copy operations, whereas pageviews with a single highlighting operation are associated with only 10.6% of the highlighting operations. 90.6% of the copy operations were in pageviews with up to 4 copy operations, whereas only 31.1% of the highlighting operations were in pageviews with up to 4 highlighting operations. 21.5% of the highlighting operations are in pageviews with at least 20 highlighting operations. Only 0.8% of the copy operations are in pageviews with at least 20 copy operations.

The difference in the distributions of highlighting and copy operations can be explained by the context and purpose of these operations. Copying is often an individual operation that addresses a specific need for specific content at a specific point in time. Highlighting, on the other hand, is often a reading behavior, and therefore, sequences of highlighting operations during reading, by users that use highlighting as a reading assistant tool, are more common than sequences of copy operations.

Table 1: Distribution of highlighting operations among pageviews.

Operations per Pageview	Pageviews	Operations	% of Operations	Accumulated %
None	1,116,076	0	0.0%	0.0%
1	78,608	78,608	10.6%	10.6%
2	25,436	50,872	6.9%	17.5%
3	17,312	51,936	7.0%	24.6%
4	11,979	47,916	6.5%	31.1%
5	9,225	46,125	6.2%	37.3%
6	6,749	40,494	5.5%	42.8%
7	5,235	36,645	5.0%	47.7%
8	4,128	33,024	4.5%	52.2%
9	3,147	28,323	3.8%	56.0%
10	2,557	25,570	3.5%	59.5%
11	2,039	22,429	3.0%	62.5%
12	1,660	19,920	2.7%	65.2%
13	1,431	18,603	2.5%	67.8%
14	1,155	16,170	2.2%	70.0%
15	949	14,235	1.9%	71.9%
16	901	14,416	2.0%	73.8%
17	713	12,121	1.6%	75.5%
18	639	11,502	1.6%	77.0%
19	561	10,659	1.4%	78.5%
At Least 20	4,721	159,009	21.5%	100.0%

Table 2: Distribution of copy operations among pageviews.

Operations per Pageview	Pageviews	Operations	% of Operations	Accumulated %
None	1,219,073	0	0.0%	0.0%
1	56,794	56,794	51.9%	51.9%
2	12,757	25,514	23.3%	75.1%
3	3,766	11,298	10.3%	85.5%
4	1,402	5,608	5.1%	90.6%
5	618	3,090	2.8%	93.4%
6	280	1,680	1.5%	94.9%
7	159	1,113	1.0%	96.0%
8	105	840	0.8%	96.7%
9	79	711	0.6%	97.4%
10	50	500	0.5%	97.8%
11	22	242	0.2%	98.1%
12	19	228	0.2%	98.3%
13	22	286	0.3%	98.5%
14	10	140	0.1%	98.6%
15	16	240	0.2%	98.9%
16	4	64	0.1%	98.9%
17	6	102	0.1%	99.0%
18	6	108	0.1%	99.1%
19	5	95	0.1%	99.2%
At Least 20	28	872	0.8%	100.0%

4 Exploring Types of Text Highlighting and Copying

Users highlight and copy different things for different reasons. This section presents and analyzes real examples of text highlighting and copying. The goal is to classify different types of highlighting and copy operations and to understand possible user motivations for each of these types (although there is no attempt to cover every type of operation and every possible motivation).

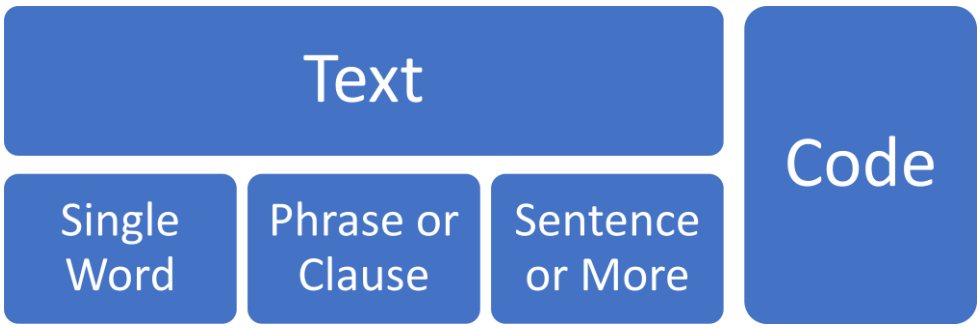

Figure 2: Taxonomy of textual content highlighting and copying (primary categories)

The main taxonomy categories that are used in this paper to analyze text highlighting and copying are illustrated in Figure 2 (further subcategories and special types are discussed later). Software developers,

who make up most of the traffic to the study's website, use the website as a learning and knowledge source. They frequently arrive from search engines for short visits, after searching for specific technical solutions and code examples. When the desired code example is found, relevant fragments are often highlighted and/or copied to the clipboard in order to be pasted and integrated into software projects. The study explores also highlighting and copying of text, and most of the proposed uses and applications are related to highlighting and copying of regular text rather than code. Copying images and other resources, however, are out of the scope of this study.

As shown in Figure 2, text content is divided into three categories by the size of the selected elements: single words, multiword expressions such as phrases and clauses, and full sentences. Each of these categories is associated with different user intentions and purposes, as discussed in the following subsections, and accordingly, related to different potential users and applications.

4.1 Highlighting and Copying Single Words

We start by examining highlighting and copying of single text words. Table 3 presents the most frequently highlighted and copied words (considering only operations related to individual words in regular text, excluding code fragments). The examples in this category are further divided into four subcategories.

Table 3: Examples of frequently highlighted and/or copied words.

Type	Examples with the numbers of highlighting (H) and copying (C) operations
Tutorial	Guestbook (H=20,C=144), guest (H=27,C=79) , GuestListener (H=16,C=44), GuestDao (H=16,C=38)
Concept	JPQL (H=558,C=437), JPA (H=441,C=75), JDO (H=67,C=19)
Complex	persistence (H=358,C=79), transient (H=237,C=141), embeddable (H=114,C=23), explicitly (H=89,C=16), redundant (H=42,C=22)
Other	the (H=699,C=1), database (H=733,C=3), in (H=497,C=1), object (H=487,C=1), objects (H=481,C=0), and (H=372,C=2), a (H=307,C=0), is (H=243,C=1), that (H=246,C=0), following (H=214,C=0), using (H=212,C=0), are (H=190,C=0), an (H=164,C=0), as (H=175,C=0) , not (H=153,C=1), to (H=147,C=0)

The first subcategory consists of real-life words relating to the specific examples used in the website tutorials. The tutorials explain how to develop simple "Guestbook" applications in several environments. Users can follow the step-by-step instructions in these tutorials to create their versions of these applications in their IDEs. The words: "Guestbook", "guest", "GuestListener", and "GuestDao" are the suggested names for the created projects, classes, and files. Users copy these names from the tutorials to their clipboards and

then paste them into the relevant dialog boxes in their IDEs. Note that there are considerably more copy operations than highlighting operations in this subcategory (in contrast to the general frequencies), as using these words in tutorials requires copy & paste rather than highlighting.

The second subcategory consists of acronyms of technical terms. The three examples provided represent key concepts in the website domain knowledge. A possible reason for highlighting and copying these terms is that users need more information about these concepts. Users may select the terms and then search on the internet using the search command on the browser's context menu (usually by right-clicking) or copy the terms into the clipboard and then paste them into an internet search form. Some users may highlight the terms only to emphasize them, without searching, while reading.

The third subcategory consists of complex words. These are regular but less frequent words, which may be less familiar for some non-native English speakers. Users may select or copy these words in order to search for more information (similarly to selecting and copying words of the second subcategory), or translations in online dictionaries. Some users may also highlight them with no additional actions in order to increase concentration while reading. Further research shows that words that are copied frequently are indeed relatively complex [15] and that word copying of these English words is more frequent by users whose preferred language is not English (and particularly frequent by users of foreign languages that are considerably different from English) [21].

The fourth subcategory consists of other ordinary words, including many very simple and frequent words. Words in this subcategory are rarely copied as individual words (unlike the words in the other subcategories), as users do not usually need to search for more information on the internet regarding these words. Some users may highlight words such as 'in', 'and', 'following' and 'not' while they read to emphasize, to increase concentration, or to mark the reading position. Note that the reason that the most frequently highlighted words in this category are simple words is that they are very common on the study's website. Figure 3 presents a heatmap that provides a realistic visualization regarding the frequencies of highlighting individual words in a specific paragraph. The order of colors from indicating the most frequent to the least frequent are red, orange, yellow, green, and blue (see more details in [22]).

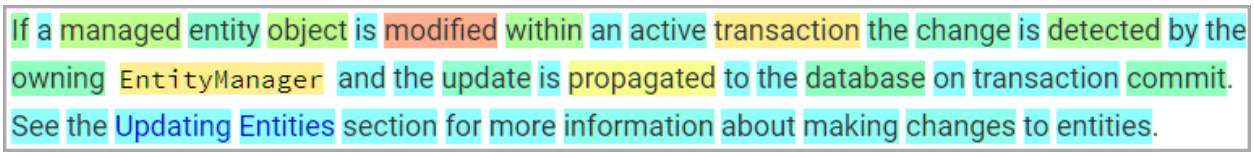

Figure 3: Word highlighting heatmap.

Figure 4 presents another heatmap for the same paragraph, visualizing the frequencies of copying individual words (rather than highlighting). As copying is less frequent than highlighting, the data is more limited, but interestingly when considering copied words rather than highlighted words the focus is moved to words of the first three categories, e.g., the word 'propagated', which may be considered as complex (relative to the other words in the text), and therefore, copied more frequently.

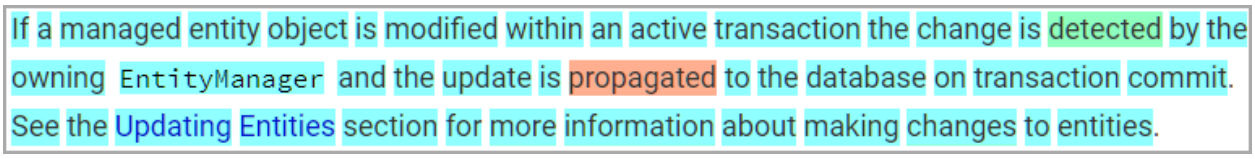

Figure 4: Word copying heatmap.

4.2 Highlighting and Copying Multiword Expressions

In this subsection, we progress from single words to sequences of several words in regular text (excluding code fragments). Table 4 presents the most frequently highlighted and copied multiword expressions (which are not complete sentences) in the dataset.

Table 4: Examples of frequently highlighted and/or copied multiword expressions.

Type	Examples with the numbers of highlighting (H) and copying (C) operations
Phrase	Composite Primary Key (H=120,C=193), Embedded Primary Key (H=74,C=89), The Sequence Strategy (H=113,C=46), Auto Generated Values (H=74,C=38), JPA Named Queries (H=31,C=44), Orphan Removal (H=52,C=23)
Clause	This page covers the following topics: (H=293,C=12), The IDENTITY strategy is very similar to the AUTO strategy: (H=174,C=7), The following query string represents a minimal JPQL query: (H=111,C=8), For example, the following JPQL query: (H=60,C=5), for use when exactly one result object is expected (H=46,C=11), for general use in any other case (H=47,C=2)

The phrases subcategory contains short sequences of two or three words. These are mainly technical terms and concepts, similar to the 'Concepts' subcategory in subsection 04.1. Users may select them and then search on the internet using the search command on the browser's context menu (usually by right-clicking) or copy the terms into the clipboard and then paste them to search on the internet. Some users may highlight these phrases to emphasize them, without searching, while reading.

The clauses subcategory contains longer sequences of words. Usually searching for longer sequences of words on search engines is less effective because such search attempts may be too restrictive. This may explain why these multiword expressions are less frequently copied. However, some of these multiword expressions are frequently highlighted. The top frequently highlighted expressions of this subcategory are clauses that introduce new information (e.g., ending with a colon symbol or including the word 'following') or explain the context (e.g., including words such as 'for use when'). It might be particularly helpful for some users to highlight such context-explaining expressions, in order to emphasize them, and possibly increase concentration while reading.

4.3 Highlighting and Copying Full Sentences

From phrases and clauses, we proceed to full sentences. Table 5 presents some of the frequently highlighted and copied sentences and sequences of sentences in the dataset.

Table 5: Examples of frequently highlighted and/or copied sentences.

Type	Examples with the numbers of highlighting (H) and copying (C) operations
Single	Transient entity fields are fields that do not participate in persistence and their values are never stored in the database. (H=198,C=54)
	The JPA Criteria API provides an alternative way for defining JPA queries, which is mainly useful for building dynamic queries whose exact structure is only known at runtime.(H=179,C=49)
	Changes to detached entity objects are not stored in the database unless modified detached objects are merged back into an EntityManager to become managed again. (H=94,C=14)
Multiple	The IDENTITY strategy also generates an automatic value during commit for every new entity object. The difference is that a separate identity generator is managed per type hierarchy, so generated values are unique only per type hierarchy. (H=523,C=134)
	During a commit the AUTO strategy uses the global number generator to generate a primary key for every new entity object. These generated values are unique at the database level and are never recycled, as explained in the previous section. (H=191,C=37)

Table 5 are single sentences and the others are multiple sentences. It is quite unlikely that users have copied these sentences to use them in searches on the Internet, as such long strings are not effective for searching. These sentences may have been highlighted to emphasize them and assist in reading and understanding key

points in the text. They may have been copied in order to use them in other texts, e.g., as citations or summaries, in documentations, presentations, blogs, websites, answers on forums (such as StackOverflow), or even private communications between colleagues who work on a project together. It is reasonable to expect users to highlight and copy key sentences more frequently than less important sentences, as supported by a preliminary analysis of the examples in this dataset.

4.4 Highlighting and Copying Translated Text

The website under investigation contains only texts in English. Therefore, it was a surprise to find many hundreds of highlighting and copy operations of text in other languages (mainly Spanish, but also French, Portuguese, and other languages). It seems that many users used Google Translate or other translation services to read translations of the pages, and subsequently highlighted and copied translated text as well. Copying individual complex words to search for their meaning (as demonstrated and discussed in subsection 4.1) may be effective for users with moderate or advanced language proficiency, whereas full translation may be needed for users with no (or elementary) language proficiency. Table 6 shows some examples of highlighted and copied translated text from the dataset, which are relatively short and could fit the limited space in the table.

Table 6: Examples of frequently highlighted and/or copied translated text.

Type	Examples with the numbers of highlighting (H) and copying (C) operations
Word	y (H=100,C=0), al (H=24,C=0), de (H=25,C=0), que (H=13,C=0), el (H=10,C=0), la (H=7,C=0), como (H=4,C=0)
Phrase	Requerimientos de plataforma (H=0,C=6) Fiabilidad y Estabilidad (H=0,C=4) Reduce el tiempo de Desarrollo (H=0,C=2) Eficaz en entornos pesados de usuarios multiples (H=0,C=2)
Clause	La siguiente cadena de consulta representa una consulta JPQL mínima: (H=3,C=0)
Sentence	ObjectDB puede administrar bases de datos de varios tamaños de manera eficiente, desde kilobytes hasta terabytes. (H=0,C=2)

Google Translate usage is usually transparent to websites and web applications, as ordinary requests are sent from the browser to the website server, and ordinary responses are sent back from the server to the client. Translations are done by communications between browsers and Google Translate, in which websites are not involved. However, client-side JavaScript code is aware of translated text in the browser (as shown to the users), and therefore, highlighting and copy operations are tracked and recorded with the text as it is shown to the users, including translated text when applicable.

Most of the categories and subcategories of textual content, which are discussed in sections 04.1, 4.24.2, and 4.3 are relevant also for translated text, but not all of them. For example, in the ‘word’ category we see mainly examples from the ‘other’ subcategory. Some words, such as acronyms, are never translated. Complex words are translated (and therefore could be detected if highlighted or copied), but the translation to the user's preferred language eliminates their complexity, so the need to highlight or copy may be significantly reduced.

4.5 Highlighting and Copying Programming Code

All the types of highlighting and copy operations that are discussed above are related to regular text rather than to code fragments. Table 7 shows some of the more commonly highlighted and copied code elements and fragments, which are short enough to be included as examples.

Table 7: Examples of frequently highlighted and/or copied code.

Type	Examples with the numbers of highlighting (H) and copying (C) operations
Word	CURRENT_DATE (H=246,C=535), CURRENT_TIMESTAMP (H=134,C=476), UPPER (H=149,C=215), getSingleResult (H=15,C=76)
Line	@SequenceGenerator(name="seq", initialValue=1, allocationSize=100) (H=3994,C=2387) CriteriaBuilder cb = em.getCriteriaBuilder(); (H=1316,C=690) em.getTransaction().begin(); (H=911,C=604)
Fragment	CriteriaBuilder cb = em.getCriteriaBuilder(); CriteriaQuery<Country> q = cb.createQuery(Country.class); Root<Country> c = q.from(Country.class); q.select(c); (H=1279,C=591)

Code elements and fragments highlighted and copied by users range from single words (query literals, class names, method names, files names, etc.), through lines of codes (e.g., queries), to larger fragments containing complete examples. This reflects the extent of interest and assistance that the users need. In some situations, the focus is on a single word. In other situations, the focus is on a complete example. As with ordinary text, users may highlight code to emphasize and assist in reading and learning. It is reasonable to expect that in most cases code elements and fragments that are copied on the website are then pasted in the users' IDEs and integrated into software projects. Single code words may also be used in further searches.

5 Possible Uses of Highlighting and Copying Data

As shown in the previous sections, highlighting and copying text by website visitors can be tracked, stored, and analyzed. This potential source of web usage data could be useful in various applications. Each one of the following subsections proposes and discusses a potential application that can use such data. The goal of this discussion is to propose novel ideas regarding the use of text highlighting and copying data. Future work should explore each discussed direction thoroughly.

5.1 Understanding the Website Audience's Interest

Knowing the users and understanding how they use the website is a core element of user-centered design and a key to business success. Highlighting and copy operations reveal valuable information about the audience of the website, which is not easily obtained by other means.

The frequencies of highlighting and copying code indicate the importance of specific code fragments to the audience of the website. For example, Table 7 shows a high interest in the date and time literals, CURRENT_DATE and CURRENT_TIMESTAMP. Similarly, the frequencies of highlighting and copying regular text indicate the importance of specific concepts. For example, Table 4 shows that certain types of primary keys (Composite Primary Key and Embedded Primary Key) are of particular interest to many users.

Existing web analytics tools, including premium services that track every user's mouse movement, fall short of providing this precious information. This study advocates the use of text highlighting and copying data as a new source of information in the web analytics toolbox. As discussed in section 2, web analytics is effective in improving and optimizing websites. Extending the boundaries of web analytics to include text highlighting and copying activity of users could make it even more effective.

5.2 Web Personalization and Adaptive Websites

The insights that text highlighting and copying provide about user interest could also be used for personalization. Knowing the user's interest in real-time paves the way to commercial opportunities, such as presenting customized advertisements and special offers, as well as improving user experience by adjusting the website user interface, for example, by presenting new relevant links.

An adaptive website can also change the content that is presented to users based on their highlighting and copy operations. For example, in the context of a software-related website with technical documentation, if a user copies only sample code fragments of a specific programming language on a web page that covers multiple programming languages, content that is relevant to that programming language should be prioritized and presented to this user where available. Such information, regarding the preferences of a user on a web page, is usually absent from standard web server log files, which track requests of pages rather than on-page activity.

5.3 Text Simplification

Shardlow defined text simplification as "the process of modifying natural language to reduce its complexity and improve both readability and understandability" [31]. In lexical text simplification, complex words are replaced with simpler words with similar meanings [31]. This could be done manually or automatically. The identification of complex words can be the first step.

Word complexity is subjective. Information about which words users of a given website consider complex could be very helpful in improving the website content and making it more readable and understandable. Some users move the mouse cursor while reading to mark their reading position, so slowing down or stopping near words might indicate difficulties in reading or understanding [16].

Table 3 shows examples of frequently highlighted and copied words. The third category in that table contains regular words that have been copied by the users of the website. As demonstrated in subsection 4.1 and further analyzed and discussed in follow-up studies [15,21], these are relatively complex words. A plausible explanation as to why users copy complex words rather than simple words, such as "of" and "the", is that they need more information about these words and therefore paste them into online dictionaries or search engines, searching for more information or a translation. Therefore, simplifying these complex words (and further explaining certain complex technical terms in technical websites) might be particularly beneficial to the website's users. Complex words are also frequently selected without copying (i.e., highlighted). These operations may be related to the support that some browsers provide for searching for the currently highlighted text on the internet, and in some settings also for translating the highlighted text (e.g., using browser addons). Using data on highlighting operations in this context could be beneficial when the amount of available data of copy operations alone is insufficient, but it would require distinguishing between highlighting of simple words and highlighting of complex words, possibly based on operations taken by the user immediately after the highlighting operations.

5.4 Tooltips and Glossary

Similar to complex words, users also highlight and copy professional and technical terms in order to search for more information about them, either on the website or externally on the internet. The 'Concept' category in Table 3 and the 'Phrases' category in Table 4 contain terms that express concepts. Unlike complex words, these terms cannot be replaced with simple words to improve readability.

However, there may be other possible techniques to help readers. One option is to underline these terms and display tooltips when the mouse cursor hovers over them. Another option is to present a focused, local glossary of relevant terms beside the main text. Frequently highlighted and copied terms may also be defined and explained on other pages of the website. However, most users arrive from search engines directly to specific pages, missing definitions found on other web pages, so they can benefit from employing such techniques. The frequencies of highlighting and copy operations in the 'Concepts' and 'Phrases' categories may indicate where on the website such assistance is most needed.

5.5 Text Summarization

In the era of information explosion, text summarization is essential in bridging the gap between computer capabilities to store texts and human abilities to read them. A common approach to automatic text summarization is to compose a summary from key sentences, which are extracted from the original text. Various statistical metrics can be used to evaluate the importance of sentences in the original text. The most important sentences, based on the results of this evaluation, are selected, and included in the summary [29]. However, data regarding highlighting and copy operations may provide a different and possibly better indicator of the importance of sentences.

Table 5 shows examples of sentences that were frequently highlighted and copied by users. As discussed in subsection 4.3, users may highlight and copy sentences for various reasons, including for summaries, and it is reasonable to expect that more important sentences would be highlighted and copied more frequently. This is demonstrated by the two highlighting heatmaps in Figure 5 that show that meaningful parts of these paragraphs were highlighted more frequently. The order of colors from indicating the most frequent to the least frequent are red, orange, yellow, green, and blue (see more details in [22]). Clauses that are used merely as external references ("As you can see above", "See the Updating Entities section...") are 'cold' in these heatmaps, while the core textual content is 'hot'.

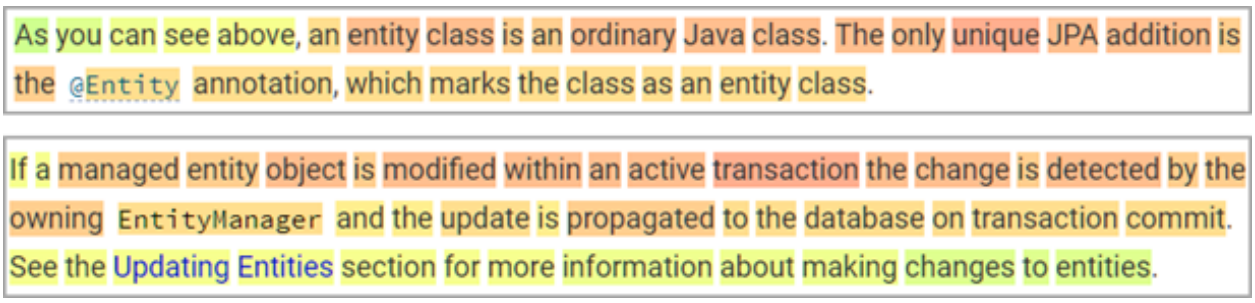

Figure 5: Hot and cold sentences in highlighting heatmaps.

This novel approach is further developed and demonstrated in a follow-up study [19], which shows that users tend to copy important key sentences more frequently and that a good summary can be built from the most frequently copied sentences.

Even if fully automatic text summarization is not used, data of text highlighting and copying can help a human summarizer to make decisions regarding what to include in a summary. Sentences highlighted and copied by users more frequently are likely to be more important to the audience of the website, and therefore, should probably be included in summaries.

5.6 Reference Cards and Tip of the Day

Sentences that are highlighted and copied frequently by users may also emphasize important points. A further examination of the examples of highlighted and copied sentences in Table 5 reveals that there are two different types of sentences in this category.

The first type is summary sentences, such as key definitions and conclusions. For example: “Transient entity fields are fields that do not participate in persistence and their values are never stored in the database.”

The second type includes sentences that provide important details that have to be considered and may be less familiar, e.g., caveats and edge cases. For example: “Changes to detached entity objects are not stored in the database unless modified detached objects are merged back into an EntityManager to become managed again.”

A summary can benefit from both types of key sentences. In addition, sentences of the second type may be selected (manually or automatically) and highlighted using reference cards (or cheat sheets) or as part of a Tip of the Day system.

5.7 Detecting Plagiarism

The vast majority of copy operations are legitimate, e.g., embedding short pieces of copied text as citations with references to the source is usually legal. On the other hand, some copy operations violate copyright rules, e.g., copying and publishing long sections of text, without explicit permission and with no reference to the source is illegal in most circumstances.

Copy operations of isolated words and phrases are usually less concerning in this context. Plagiarism is associated mainly with copying larger sections of text and code (as well as images and other resources, but this study focuses on copying textual content).

Simple Google searches of sentences and lines of code that were copied frequently by the website users reveal that many of them appear on other websites. Some occurrences, such as answers to questions on StackOverflow with proper references to the source, are not only legitimate but even desirable. On the other hand, using large original sections of the website content, without any attribution to the source, are pure plagiarism.

This process of searching for unauthorized uses of the content of a website, based on copy operations data, can be either manual or automatic. An automatic plagiarism detection implementation can track suspicious copy operations, search for the copied text on the internet frequently, and alert the website owner when instances of plagiarism have been found.

5.8 Tracking Reading

Table 3 shows several words that are frequently copied from the website tutorials. ‘Guestbook’, ‘guest’, ‘GuestDao’, and ‘GuestListener’ are the suggested names for the tutorial project and its classes. Users are

expected to copy these names from the tutorial's web pages and paste them into their IDEs. Similarly, users are expected to copy fragments of code from the tutorials and integrate them into their projects.

Tracking these copy operations (possibly in combination with other indicators) may be useful in analyzing user progress in tutorials. It could be used to detect possible breaking points, i.e., points at which many users abandon the tutorials. Some reduction in the number of copy operations throughout a tutorial, due to users' decisions to quit the tutorial, is expected. However, if the copy operations data show an extreme drop at a certain point, it may indicate an issue with that specific section of the tutorial and may require further investigation.

Highlighting operations can also be used to track reading. As discussed above, some users highlight text while they read to emphasize, increase concentration, or mark the reading position. 159,108 highlighting operations in the dataset (20% of the total) are in pageviews with at least 20 highlighting operations. Such data can be used to analyze the reading behaviors of users, including what is read, what is skipped, and what is reread, and also at which speeds. This information may help, for example, in identifying possible obstacles in the text.

5.9 Understanding Language Translation Needs

As discussed in subsection 4.4, some users view websites through Google Translate. The translation provided by Google Translate is normally invisible in the website statistics and web analytics data. The website is accessed ordinarily from the browser, and the translation is done by the browser and Google Translate. In this case study, the copy operations expose a community of users that use Google Translate to translate pages on the website to Spanish.

Examining text highlighting and copying is a simple way to detect the usage of Google Translate on a website. This method has the additional benefit of detecting also specific interests of these users on the website (see subsection 5.1 above).

Decisions on investing in the translation of a website or specific web pages can be affected by data regarding text highlighting and copying by Google Translate users. That data is much more relevant for decisions on translation than the distribution of users by country (which is ordinarily available by web analytics services) because many non-native English speakers do not need translation and may even prefer reading the content in its original language.

5.10 Search Engine Optimization

Data regarding text highlighting and copying can be useful in Search Engine Optimization (SEO), i.e., in making a website perform better in search engine results, and consequently in increasing the traffic to the website (which is usually desired for commercial websites). These operations can indicate which topics are

more popular among users of the website. Investing in new content on these subjects may be a cost-effective way to increase the traffic to the website.

As part of an SEO process, new pages can be created for phrases that users frequently highlight or copy for further searches. A new dedicated page can provide additional content about the subject of a frequently copied phrase, with a title containing that phrase, as well as other optimizations that help search engines to establish the relevancy of the new page to the phrase.

This could attract more visitors from search engines. It may even be possible to catch a rebound of users that leave the website with a phrase in their clipboard to search for more information on the internet. These users may find themselves after that search back on the same website, possibly on a page that was created in this SEO process.

6 Conclusions and Future Work

This study analyzes text highlighting and copying by users of a commercial website.

Selecting text is often the first step in copying text to the clipboard. However, it can also be used by online readers to highlight points while reading, similarly, to highlighting words in a book with a marker pen. The vast majority of text selection operations in this study's dataset are not followed by copying. Accordingly, this paper distinguishes between two different types of operations: (a) highlighting text by selecting it without copying it, for example, to emphasize elements in the text while reading; and (b) copying text to the clipboard, where the preceding selection operation has no distinct function and thus considered part of the copy operation. Users select and copy textual content of various types (e.g., words, multiword expressions, sentences, and code fragments) for various purposes (e.g., searching for more information, using the content, etc.). The paper suggests a taxonomy of text highlighting and text copying into several categories and subcategories. It also discusses possible user motivations to highlight or copy each type of textual content.

Examining the distribution of the recorded operations on the website among pageviews shows an interesting difference between highlighting and copying. The ratio between the numbers of highlighting operations and copy operations was 6.74 and the ratio between the number of pageviews with these operations was 2.35. This indicates that highlighting operations are more frequent than copy operations on that website and that highlighting operations also tend to be more concentrated in pageviews. The difference in the distributions of highlighting and copy operations can be explained by the context and purpose of these operations. Copying is often an individual operation that addresses a specific need for specific content at a specific point in time. Highlighting, on the other hand, is often a reading behavior, and therefore, sequences of highlighting operations during reading, by users that use highlighting as a reading assistant tool, are more common than sequences of copy operations.

Data regarding text highlighting and copying may indicate user attention to particular text parts and may be valuable in various situations. This paper proposes various potential applications and uses that can benefit from such user activity data, in various areas, including web analytics, web personalization, adaptive websites, text simplification, text summarization, detection of plagiarism, search engine optimization, and reading analysis.

Most of the proposed applications focus on specific types of text highlighting and/or copying. Therefore, automatic classification of operation types (possibly based on the taxonomy that this paper proposes), may be needed as the first step in some of these applications.

Work on two of these applications, the use of text copying data in text summarization and the use of text copying data in Complex Word Identification (CWI) and text simplification, has already started with promising initial results, as discussed above. Further work is needed to explore the other proposed uses and to extend the initial work on the two first applications.

This study is based on anonymized web usage data, so the background of users is unknown and their exact motivations in highlighting and copying text could not be verified. Future work may further increase the knowledge about text highlighting and copying on websites by tracking users for which additional information is available, and which can provide further information about their actions and motivations.

Although it is reasonable to expect that the results of this research are not unique for the selected website and can be extrapolated to other websites, due to the study focusing only on one specific technical website, further work on other websites of other types is needed for establishing and generalizing these results.

References

1. Alhlou, F., Asif, S., and Fettman, E. (2016). Google Analytics Breakthrough: From Zero to Business Impact. John Wiley & Sons, USA.
2. Ahmed T., Shang W., and Hassan A. E., (2015). An Empirical Study of the Copy and Paste Behavior during Development. . In Proceedings of the 12th Working Conference on Mining Software Repositories, Florence, Italy, https://doi.org/10.1109/MSR.2015.17
3. Cepeda, C., Rodrigues, J., Dias, M. C., Oliveira, D., Rindlisbacher, D., Cheetham, M., and Gamboa, H. (2018). Mouse tracking measures and movement patterns with application for online surveys. In Holzinger, A., Kieseberg, P., Tjoa, A. M., and Weippl, E., editors, Machine Learning and Knowledge Extraction, pages 28–42, Cham. Springer International Publishing.
4. Chaffey, D. and Patron, M. (2012). From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics. Journal of Direct, Data and Digital Marketing Practice, 14:30–45.

5. Chen, M. C., Anderson, J. R., and Sohn, M. H. (2001). What can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing. In CHI ’01 Extended Abstracts on Human Factors in Computing Systems, CHI EA ’01, page 281–282, New York, NY, USA. Association for Computing Machinery.
6. Dykes, B. (2014). Web Analytics Kick Start Guide: A Primer on the Fundamentals of Digital Analytics. Peachpit, Pearson Education, USA.
7. Gilliam Haije, E. (2018). Are session recording tools a risk to internet privacy? https://mopinion.com/are-session-recording-tools-a-risk-to-internet-privacy/
8. Guo, Q. and Agichtein, E. (2008). Exploring mouse movements for inferring query intent. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’08, page 707–708, New York, NY, USA. Association for Computing Machinery.
9. Hasan, L., Morris, A., and Probets, S. (2009). Using google analytics to evaluate the usability of e-commerce sites. In Proceedings of the 1st International Conference on Human Centered Design, pages 697–706, Berlin, Heidelberg. Springer Berlin Heidelberg.
10. Huang, J., White, R., and Buscher, G. (2012). User see, user point: Gaze and cursor alignment in web search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI’12, page 1341–1350, New York, NY, USA. Association for Computing Machinery.
11. Järvinen, J. and Karjaluoto, H. (2015). The use of web analytics for digital marketing performance measurement. Industrial Marketing Management, 50:117 –127.
12. Kaushik, A. (2007). Web Analytics: An Hour a Day. SYBEX Inc., USA.
13. Kaushik, A. (2010). Web Analytics 2.0. SYBEX Inc., USA.
14. Kim M, Bergman L., Lau T. and Notkin D. (2004). An Ethnographic Study of Copy and Paste Programming Practices in OOPL. In Proceedings of the 2004 International Symposium on Empirical Software Engineering (ISESE 2004), pages 83-92. https://doi.org/10.1109/ISESE.2004.1334896.
15. Kirsh, I. (2020). Automatic complex word identification using implicit feedback from user copy operations. In Proceedings of the 21st International Conference on Web Information Systems Engineering (WISE 2020), pages 155-166. Lecture Notes in Computer Science, forthcoming, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-62008-0_11
16. Kirsh, I. (2020). Using mouse movement heatmaps to visualize user attention to words. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction (NordiCHI 2020), pages 117:1-5, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3419249.3421250

17. Kirsh, I. (2020). What Web Users Copy to the Clipboard on a Website: A Case Study. In Proceedings of the 16th International Conference on Web Information (WEBIST 2020), pages 303–312, INSTICC, SciTePress. https://doi.org/10.5220/0010113203030312

18. Kirsh, I. and Joy, M. (2020). A different web analytics perspective through copy to clipboard heatmaps. In Proceedings of the 20th International Conference on Web Engineering (ICWE 2020), Lecture Notes in Computer Science, vol. 12128, pages 543–546, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-50578-3_41

19. Kirsh, I. and Joy, M. (2020). An HCI approach to extractive text summarization: Selecting key sentences based on user copy operations. In Proceedings of the 22nd HCI International Conference (HCII 2020), Communications in Computer and Information Science, vol. 1293, pages 335-341, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-60700-5_43

20. Kirsh, I. (2021). Visualizing Web Users' Attention to Text with Selection Heatmaps. In Proceedings of the 21st International Conference on Web Engineering (ICWE 2021), Lecture Notes in Computer Science, vol. 12706, pages 517–520, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-030-74296-6_42

21. Kirsh, I. (2021). Word-Copying on a Website as a Word Complexity Indicator and the Relation to Web Users' Preferred Languages. In Proceedings of the 5th Asian CHI Symposium (AsianCHI@CHI 2021), pages 16-20, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3429360.3468172

22. Lamberti, F. and Paravati, G. (2015). Vdhm: Viewport-dom based heat maps as a tool for visually aggregating web users' interaction data from mobile and heterogeneous devices. In Proceedings of the 2015 IEEE International Conference on Mobile Services, MS '15, page 33–40, USA. IEEE Computer Society.

23. Lamberti, F., Paravati, G., Gatteschi, V., and Cannav`o, A. (2017). Supporting web analytics by aggregating user interaction data from heterogeneous devices using viewport-dom-based heat maps. IEEE Transactions on Industrial Informatics, 13:1989 – 1999.

24. Luo, H., Rocco, S., and Schaad, C. (2015). Using google analytics to understand online learning: A case study of a graduate-level online course. In Proceedings of the 2015 International Conference of Educational Innovation through Technology, EITT '15, page 264–268, USA. IEEE Computer Society.

25. Milisavljevic, A., Hamard, K., Petermann, C., Gosselin, B., Dor´e-Mazars, K., and Mancas, M. (2018). Eye and mouse coordination during task: From behaviour to prediction. In International Conference on Human Computer Interaction Theory and Applications, pages 86–93, SciTePress.

26. Navalpakkam, V., Jentzsch, L., Sayres, R., Ravi, S., Ahmed, A., and Smola, A. (2013). Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In Proceedings of the

22nd International Conference on World Wide Web, WWW '13, page 953–964, New York, NY, USA. Association for Computing Machinery.

27. Špakov, O. and Miniotas, D. (2007). Visualization of eye gaze data using heat maps. Elektronika ir Elektrotechnika - Medicine Technology, 115.
28. Rodden, K. and Fu, X. (2007). Exploring how mouse movements relate to eye movements on web search results pages. In Proceedings of ACM SIGIR 2007 Workshop on Web Information Seeking and Interaction, pages 29–32, New York, NY, USA. Association for Computing Machinery.
29. Sajjan, R. and Shinde, M. (2019). A detail survey on automatic text summarization. International Journal of Computer Sciences and Engineering, 7:991–998.
30. Schneider, J., Weinmann, M., vom Brocke, J., and Schneider, C. (2017). Identifying preferences through mouse cursor movements – preliminary evidence. In Proceedings of the 25th European Conference on Information Systems (ECIS), pages 2546–2556, Guimar˜aes, Portugal. Research-in-Progress Papers.
31. Shardlow, M. (2014). A survey of automated text simplification. International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014, 4(1):58–70.
32. Stolee K. T., Elbaum S., and Rothermel G. (2009). Revealing the Copy and Paste Habits of End Users, In Proceedings of the 2009 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pages 59-66. https://doi.org/10.1109/VLHCC.2009.5295296
33. Tandoc, E. C. J. (2015). Why web analytics click. Journalism Studies, 16(6):782–799.
34. Tzafilkou, K., Protogeros, N., and Yakinthos, C. (2014). Mouse tracking for web marketing: Enhancing user experience in web application software by measuring self-efficacy and hesitation levels. International Journal on Strategic Innovative Marketing, 01.