Temporal Information Searching Behaviour and Strategies

Hideo Joho[*,a], Adam Jatowt[b], Roi Blanco[c]

[a]Research Center for Knowledge Communities, Faculty of Library, Information and Media Science, University of Tsukuba, Japan
[b]Department of Social Informatics, Graduate School of Informatics, Kyoto University, Japan
[c]Yahoo Labs, Barcelona, Spain

Abstract

Temporal aspects have been receiving a great deal of interest in Information Retrieval and related fields. Although previous studies have proposed, designed and implemented temporal-aware systems and solutions, understanding of people's temporal information searching behaviour is still limited. This paper reports the findings of a user study that explored temporal information searching behaviour and strategies in a laboratory setting. Information needs were grouped into three temporal classes (Past, Recency, and Future) to systematically study their characteristics. The main findings of our experiment are as follows. 1) It is intuitive for people to augment topical keywords with temporal expressions such as *history*, *recent*, or *future* as a tactic of temporal search. 2) However, such queries produce mixed results and the success of query reformulations appears to depend on topics to a large extent. 3) Search engine interfaces should detect temporal information needs to trigger the display of temporal search options. 4) Finding a relevant wikipedia page or similar summary page is a popular starting point of past information needs. 5) Current search engines do a good job for information needs related to recent events, but more work is needed for past and future tasks. 6) Participants found it most difficult to find future information. Searching for domain experts was a key tactic in Future search, and file types of relevant documents are different from other temporal classes. Overall, the comparison of search across temporal classes indicated that Future search was the most difficult and the least successful followed by the search for the Past and then for Recency information. This paper discusses the implications of these findings on the design of future temporal IR systems.

Key words: Temporal Information Retrieval, Information Searching Behaviour, Search strategies, User Study

1. Introduction

Predicting the future and remembering the past are very common cognitive processes of humans. Same as in real life, we should expect many search activities to be of a strong temporal character. Previous studies [19, 33] have confirmed this fact and elucidated a relatively high number of search intents and queries that center on information associated to particular temporal scopes such as future or past. Regarding future-oriented search, users often need to know more about planned events, forecasted trends, possible scenarios, speculations, predictions and so on. This kind of information can effectively help them to be better prepared for events to come or tasks to be undertaken. Imagine a user who wants to purchase shares or futures of a particular company, and another user who plans a visit in Kyoto. Any predictions or speculations about the company's future, its plans or forecasts and any information about forthcoming cultural events in Kyoto would be valuable in these scenarios. Similarly, the study of the past helps us to explain the present, to learn the background of the current course of actions or to form opinions on any trends and changes occurring over time. In general, we should be able to easily imagine a multitude of reasons on why someone would need to find future or past-related information.

However the core interest of society centers on the present or, at least, on the near past. The Web, for example, abounds in rather up-to-date information that is mainly about "now" [12, 32]. The information about more distant

[*]Corresponding author

Email addresses: hideo@slis.tsukuba.ac.jp (Hideo Joho), adam@dl.kuis.kyoto-u.ac.jp (Adam Jatowt), roi@yahoo-inc.com (Roi Blanco)

past or future tends to be buried in the wealth of data on the current topics and events. Intuitively, it may be difficult for an average user to extract content that is related to the future and to the past, especially, to distant future and past, for arbitrary queries. Such information is often scattered across many documents and expressed in a large number of different ways and is thus likely difficult to be extracted, merged, processed and understood by searchers. We would then expect the state-of-the-art search engines to offer some kind of support for those users who try to seek for time-related information. The academic community has already started developing methods for enabling the temporal search [5, 9, 20, 16, 21, 22, 27, 31]. What is still missing is the knowledge of how searchers actually filter out present-related information when they have search intents related strongly to either the past of the future. Empowered with such knowledge it will be easier to reason about the required level of support searchers should receive or about any temporal mechanisms that ought to be implemented for enabling effective temporal search. In this work we aim to fill in this gap and shed light on the actual behavior of searchers who wish to find past or future-related information. In particular, we conduct controlled settings experiments with 30 participants who are asked to perform searches on variety of topics on the Web to find information related to particular time scopes. We then analyze their behaviour as well as feedback regarding the tasks and their difficulties. We report a large number of observations that have not been known to the community and which could have considerable implications on the design of temporal search mechanisms and search interfaces for facilitating retrieval of time-related information.

The reminder of the paper is organised as follows. Section 2 provides the literature survey in temporal information retrieval. Section describes the design of user study we performed to capture information seeking behaviour of temporal search. Section presents the results of the user study. Section 5 discusses the implications of these findings on the design of future temporal search engines, followed by a conclusion and the outline of future directions.

2. Related Work

Temporal Information Retrieval (e.g., [1, 5, 7, 8, 9, 17, 19, 20, 22, 30, 31, 33]) has increasingly been gaining much interest in the IR community. This subarea of information retrieval focuses on temporal aspects of search, treating time as crucial facet for determining document relevance. Prior work mainly focused on either estimating temporal features of documents [16, 34, 35], temporal aspects of queries [7, 20, 31, 33] or on matching temporalities of queries with ones of documents for realizing effective time-aware retrieval [3, 23, 21, 27].

According to a study performed on the AOL query dataset [33] about 1.5% of queries are explicit temporal queries, that is, they contain an explicit temporal expression. Examples of such queries are: “Poland 1940s”, “Olympics 2016” or “most popular songs 2000s”. A subsequent study [7] revised this number to about 1.21% queries after excluding some false positive temporal expressions (e.g., “Excel 2007”, “Honda civic 2004”). Searchers also issue implicit temporal queries that are related to time despite lacking apparent temporal expressions (e.g., “Einstein childhood”, “WWII major battles”, “USA debt size”, “Rio de Janeiro Olympics”) [20].

Metzler et al. [31] concluded that about 7% of queries have a certain temporal intent. Considering the popularity of Web search, this rate translates to a remarkable number of unique searches. Some methods have been already proposed to automatically classify queries into different temporal classes with the underlying aim of improving search results once query temporal intent is known (see, for example, [9, 20, 22]). For a more extensive overview of existing approaches the reader may refer to [1, 8].

Given the relatively large amount of temporal queries, search engines should handle them in a way which appropriately considers particular temporal classes to prevent potential mismatch like returning past-related information for a query with obvious future-focused intent. In addition, search engines could offer support for finding information from particular temporal class (e.g., by query recommendation, by elucidating temporal aspects of snippets). For this to be effective, however, a prior comprehensive analysis of the way in which users search for time-sensitive information is needed.

Although previous works focused on ranking documents for temporal queries and on estimating temporal intents underlying user queries, little has been done to uncover the actual search patterns and behaviour of users who seek information of temporal character. The only previous study in this regard that we are aware of was conducted as an online questionnaire involving 110 users [19]. This study suggested that although many users search for information about recent events and current state, a good proportion of them also seek for information about past as well as future events. The authors have also reported that the temporal search was done mainly in office and less at home, as well

	Temporal class					
	Atemporal	Future	Past	Recency	$\chi^2(2)$	p
Temporal search practice	1.5 (0.7)	2.6 (1.3)	2.4 (1.2)	1.7 (0.8)	183.7526	$\leq .001$

Table 1: Mean and standard deviation of temporal search practice scores (1: High, 7: Low). N = 30.

as, they confirmed that users are not entirely satisfied with current search engine results. Nearly 24% respondents agreed that they could not conceptualize a suitable query for their searches, while about 35% had problems with finding relevant search results. Unlike that work, in this paper we report on a more detailed study that has been carried with realistic search tasks and, more importantly, conducted in controlled settings rather than employing an online questionnaire. We introduce a set of realistic search tasks, each with different variants depending on the temporal scope.

Recently, Kato et al. [24] studied cognitive search intents in web search although they did not approach the temporal search needs. Cognitive information intents are characterized by search requirements related to document characteristics in contrast to requirements on document topics. For example, users may search for comprehensible documents, objective documents or those containing concrete information. Temporal information needs can be considered as another type of cognitive search intents.

Several search strategies and tactic models have been proposed for general search, and examined by researchers [4, 13, 14, 28, 37, 38]. In principle, a search strategy is the largest unit in information seeking process to determine "overall plan for, or approach to, a whole search session" [36, p. 34]. Keyword searching and directory browsing are an example of search strategies. Search tactics are, on the other hand, a more specific unit including broadening or narrowing search terms, for example. However, the boundary between strategies and tactics is not always clear. In this work, we mostly focus on a level of search strategies that people employed in temporal search tasks, because such strategies allow us to understand intentions behind their search behaviour. Although the descriptions of specific tactical behaviors are included where possible, the detail analysis on tactics is beyond the scope of this work.

To summarise, little work has been done so far to clarify what strategies, patterns or heuristics users employ when searching information related to a particular class of temporal search needs as well as to elucidate what kind of problems they encounter. This work aims to fill in these gaps and serve as guidance for subsequent studies of user behaviour and automatic approaches within temporal information retrieval.

3. Experimental Design

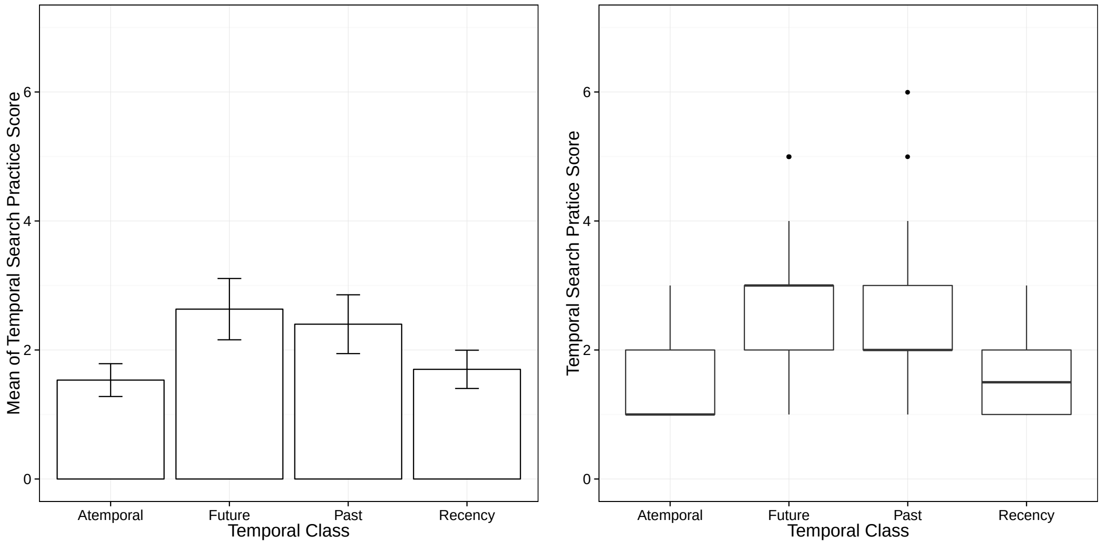
This section describes the design of user study in detail. It should be noted that the entire experiment was conducted in one week in Feb/March 2014 in Tokyo, Japan. The information presented here contains translation from the descriptions originally presented in Japanese.

3.1. Participants

30 people were recruited for our study by a third party agent - 15 female and 15 male. The mean age of the participants was 21.5 with standard deviation of 1.1. Most participants were university students; 27 were undergraduate and 3 were postgraduate students. The subject of their current degree program widely varied, including Architecture, Cultural Studies, Mechanical Engineering, Laws, Computer Science, Liberal Arts, Economics, Commercial Science, Politics, Literature, Life Science, Arts, Cognitive Science, History, Marketing, Education, Linguistics, and Social Science.

The entry questionnaire established that participants had on average 10 years of search engine experience, and 80% (24) used search engines several times everyday, 13% (4) used a couple of times a day, and 6% (2) used a couple of times a week. 80% (24) reported that they can mostly find for what they were looking for.

As a part of entry questionnaire, we asked participants about their practice of temporal search. Participants were requested to indicate a level of agreement (7-points Likert-like scale, Strongly Agreed: 1; Strongly Disagreed: 7) to the following statement: *I have searched for [the meaning or explanation] of a given topic*, for the case of atemporal information needs. The hard bracket part of the statement was replaced with "the origin or history" (Past), "latest



(a) Mean and Confidence Interval (CI)

(b) Median and Quartiles

Figure 1: Plots for temporal search practice scores. (1: High, 7: Low). $N = 30$.

information" (Recency), and "future plan or direction" (Future), respectively. The result is shown in Table 1. It can be seen that participants reported that they have more experience in Atemporal search and Recency search than Future or Past search. Figure 1 plots a bar chart for the confidence interval of mean values and box plot for the median and quartile values. These figures further highlight the difference between Atemporal and Recency group and Future and Past group.

Friedman's ANOVA shows that the difference among the temporal class is significant. Post-hoc pairwise Wilcoxon tests with Bonferroni correction show that the difference between the following pairs was significant with medium effect size: Atemporal-Future ($p = .0046, r = -.3181$), Atemporal-Past ($p = .0037, r = -.3327$), and Future-Recency ($p = .0049, r = -.3241$). Therefore, participants have more experience on Atemporal and Recency search than Future search. Participants have also more experience on Atemporal than Past, but the difference between Recency and Past was not significant, potentially due to a relatively large variance in the Past class.

In summary, our participants were mostly university students with varied background who have used search engines since their youth. They were mostly confident about their search capability. Some of them did not have much experience of searching for past or future information compared to atemporal and recency information.

3.2. Search Topics

We prepared a set of search topics based on NTCIR-11 Temporalia Test Collection [17, 18, 29]. Temporalia[1] aims to build a test collection that allows researchers to investigate temporal aspects of Information Access technologies. The challenge is composed of two subtasks: Temporal Query Intent Classification (TQIC) and Temporal Information Retrieval (TIR). Topics for the TIR subtask were designed to have a common title and description, along with a search question for each of four temporal classes (atemporal, past, recency, and future). These temporal class-specific questions allowed us to investigate people's information searching behaviour in a structured way.

The dry run data[2] of NTCIR-11 Temporalia TIR subtask has 15 topics and every topic has 4 temporal search questions. Of those, we selected 6 topics so that the topical diversity was maximised (see Table 2 for the list of topics used in our study). Furthermore, since TIR topics were mainly designed for system evaluation, we needed to make

[1] `https://sites.google.com/site/ntcirtemporalia/` (Last Accessed: 17/09/2014).

[2] NTCIR tracks tend to have two phrases: dry run and formal run. The former is used for training purpose while the latter is for testing.

Background Information
You have started regularly cooking since you enrolled to your university. Then you became more aware of ingredients used for your cooking than before. In particular, you started to pay attention to the use of genetically modified organism (GMO) in foods. On the other hand, you were not sure if GMO was really harmful. Therefore, you started to look into the issues of GMO so that it can help your food selection in the future. When you started to investigate GMO, it turned out that genetical modification has been applied to not only vegetables but also animals such as fish. This made you realise that the technology of genetical modification had much wider applications than you had initially thought.

Information Need (Past)
You would like to know how people reacted to products that contained genetically modified organism. In particular, please find web pages that describe people reactions when the technology of genetical modification was first appeared.

Information Need (Recent)
You would like to know the latest information about the issues of the development and application of genetic modification technology. Please find web pages that describe new subjects that only recent technology is able to apply to and its associated problems or concerns.

Information Need (Future)
You would like to know what will happen to genetic modification in the future. In particular, you would like to find out if genetically modified organism are indeed an ideal way to achieve sustainable foods in the future.

Figure 2: Simulated work task situation for Topic 014 (Genetically Modified Organisms).

some modifications to apply them to a user study. We followed the basic principle of simulated work task situation [6]: the title and description of the original topic description were augmented with a realistic contextual story regarding university students (e.g., report writing, buying a new smartphone, cooking, etc). Furthermore, temporal search questions were also augmented with information needs and indicative search instructions. Figure 2 shows a modified topic description. It has been suggested that these modifications can increase a level of engagement of participants in laboratory-based user studies [6]. We decided to omit atemporal search questions from our topic description to keep experimental complexity at a manageable level.

Note that the original topics for NTCIR-11 Temporalia TIR subtasks have been developed following real users' interests. A group of volunteers have manually formulated search queries that were interesting to them. This sometimes caused the difference between the content of subtopics of the same topic that pertain to different temporal classes. Furthermore, at certain cases it was impossible to formulate exactly the same search question for all the temporal classes. Nevertheless, each subtopic is strongly related to its parent topic.

We also note that while the distinction between the recency and future is usually clear, this may not be the case with the border between the past and the recency. Naturally, such a border can be fuzzy and topic-dependent. Similarly to the settings of NTCIR-11 Temporalia, we rely on participants' own judgments to determine where the "past" ends and where the "recency" begins.

3.3. Search Task and Corpus

The search task given to participants was to find as many relevant documents as possible within 10 minutes. Participants were allowed to use any resources to find relevant information on the Web. The only restriction we set was to prohibit to concur with their friends or acquaintances using social network systems. When they found a relevant web page, they were asked to highlight relevant parts in the page using the Evernote Web clipper.[3] This allowed us to

[3] `https://evernote.com/webclipper/` (Last Accessed: 17/09/2014)

Table 2: Main titles and temporal questions of NTCIR-11 Temporalia Dry Run dataset.

ID	Title	Class	Subtopics
003	Android Phone	Future	What is the roadmap of Android phones in the future?
		Past	What kind of security problems have been reported in Android-based phones?
		Recency	What are the functionalities available in recent Android phones?
006	Fashion Trends	Future	What is the forecast for 2015's fashion trends?
		Past	What designers won a fashion award in the past?
		Recency	What type of fashion is lately trendy?
011	Oprah Winfrey	Future	What are her hopes about future activities and career?
		Past	What has she done in order to help people in the world?
		Recency	What are her latest activities after she finished the Oprah Winfrey Show?
012	Abenomics	Future	What changes is the application of Abenomics policies expected to bring to Japan?
		Past	What was the state of Japan's economy before Abenomics?
		Recency	What is the latest performance of the Abenomics process?
013	Biodiversity	Future	What are expectations about the biodiversity in the future?
		Past	What were the previous global efforts in preserving biodiversity?
		Recency	What are the current efforts to protect biodiversity?
014	Genetically Modified Organisms	Future	Are genetically modified organisms ideal for sustainable consumption?
		Past	What were the response of the people to genetically modified organism products so far?
		Recency	What are the present usage and issues about genetically modified organisms?
Training	Obesity in US	Recency	What is the current statistics related to obesity in US?

Table 3: Rotation of experimental conditions, where 003-014 are topic IDs, and P, R, and F are temporal classes.

	Search session					
	1	2	3	4	5	6
User1	003P	006R	011F	012R	013F	014P
User2	003R	006F	011P	012F	013P	014R
User3	003F	006P	011R	012P	013R	014F
User4	014P	013R	012F	011R	006F	003P
User5	014R	013F	012P	011F	006P	003R
User6	014F	013P	012R	011P	006R	003F

revisit not only the web pages participants identified as relevant but also the actual relevant texts for post-hoc closer examination.

3.4. Rotation

We had three sets of variables to consider in the rotation table of experimental conditions. One was search topics with a level of six, and another was temporal classes with a level of three. Since it was infeasible to carry out an experiment that could consider all the combinations of these variables in a counter-balanced manner, we had to prioritise the variables.

The first priority went to temporal classes, and we created three rotations using greco-latin square table (P-R-F, R-F-P, F-P-F) where P, R and F stand for past, recency and future temporal classes, respectively. A lower priority was set on search topics since they tended to be different from one other. Table 3 shows the final rotation table used in our experiment for six users, where each participant performed six search tasks. The rotation shown in Table 3 is repeated five times for all the 30 participants. They are not necessarily perfectly balanced, but all participants performed a search task on six different topics, and two tasks per temporal class. Therefore, we consider that this rotation design was acceptable for the aim of our investigation.

3.5. Protocols

We carried out a laboratory-based experiment in the following manner. Participants were directed to a webpage where a recorded movie was shown to them to explain an overview of the experiment. They were asked to sign a consent form when they agreed to participate in the experiment, followed by an entry questionnaire to capture their background information. Then, another recorded movie was shown to them to explain how to read the search topic, how to highlight relevant information within a web page, how to save web pages, among other experimental instructions. This was followed by a training session where participants were asked to find at least one relevant page for a training topic, and highlight the relevant part in the page. Participants were given opportunities to raise any questions during and after the training session.

Then, the first search topic was distributed to participants with a mark on the temporal class to search in the session. Note that our topic description file contained all temporal information needs (as shown in Figure 2), although no topic was used twice by the same participant. This was to ensure that participants focused on a particular temporal class in their search and did not broaden the interpretation of information needs to other temporal classes. Before the search began, participants were asked to indicate their level of familiarity to the given topic.

Participants were then given 10 minutes to search and find as many relevant documents as possible. When the time was up, they were asked to answer a set of post-task questionnaires to indicate their perceptions on the search task. This was repeated 6 times. When all tasks were completed, participants were asked to fill in an exit questionnaire to provide the description of their search strategies across three temporal classes. A break was given every 45 to 60 minutes during the experiment. Participants were paid at the end of the experiment.

It should be noted that relevance judgements in our study were binary and based on participants' perceived relevance. We did not have or use official relevance assessment files (like those employed in test collection-based system evaluation). This was because the document collections used in our experiment (the entire Web) was larger than that of NTCIR-11 Temporalia (News and blog archives in English). However, an experimenter was monitoring the site during the experiment for technical support, and we did not observe any obvious misconduct (e.g., saving clearly off-topic pages on purpose). In addition, we went through all the saved pages, and, again, no obvious off-topic page was identified. Therefore, we consider that pages saved by participants were reliable for our analysis.

3.6. Data Collection

The main sources of our data analyses were client-side log data and questionnaires. A web browser plugin was installed to record major actions (e.g., clicks, form input, tab operations) on the browser with their timestamps. We also collected a cache of web pages which participants thought to be relevant to search tasks (and thus, saved). Participants were also asked to highlight relevant parts inside web pages. Questionnaires had a mixture of closed questions and open questions. Most closed questions used a 7-point Likert-like scale to capture participants' level of agreement with a certain statement, where Score 1 indicated the strongest agreement while Score 7 indicated the strongest disagreement with a statement.

4. Results

This section presents the results of our analyses on temporal information searching behaviour based on 180 search sessions conducted by participants. Unlike hypotheses-driven studies, we start this section by looking at overall performance and participants' summative descriptions of their temporal search strategies and difficulties encountered

	Temporal class					
	Future	Past	Recency	$F(2, 58)$	p	η^2
Number of saved pages (Perceived relevance)	2.4 (1.6)	3.3 (1.6)	3.8 (2.1)	17.81	$\leq .001$	.17
Task experience (Lower score is more positive)	30.9 (8.6)	26.4 (7.9)	24.7 (7.4)	12.450	$\leq .001$	.16

Table 4: Overall performance and perception of temporal search tasks. $N = 60$.

	Temporal class					
	Future	Past	Recency	$F(2, 58)$	p	η^2
Pre-task topical familiarity (Lower indicates a higher familiarity)	4.8 (1.8)	4.9 (1.7)	4.3 (1.7)	2.917	.061	.05

Table 5: Mean and standard deviation of pre-task topical familiarity (1: High, 7: Low). $N = 60$.

during the experiment of six search sessions. Participants' prominent experience reported in the exit questionnaire was used to determine focal points in subsequent analyses using system logs, accessed web pages, and structured questionnaires. Finally, it should be noted that all supporting figures are available in the Appendix for reference.

4.1. Overall Search Performance

To gain an overall picture of three temporal search classes, we looked at the average number of web pages saved by participants, and average of the sum of task perception scores collected at the end of individual search sessions. The former data should be seen as an overall task performance, while the latter should be seen as an overall task experience. The detail of task experience will be discussed in Section 4.3.

The results of these overall performances are shown in Table 4. The columns of Table 4 consist of mean and standard deviation values of three classes, F-statistics, p-values, and effect size (η^2) of one-way ANOVA test.

First, participants saved more web pages in the Recency tasks than in the Future and Past tasks. The number of saved pages was the lowest in Future tasks. One-way ANOVA shows that the difference among the three classes is significant and that the effect size is large. Post-hoc pairwise t-tests with Bonferroni correction show that the difference between Future and Past ($p = .027$), and Future and Recency ($p \leq .001$) are significant. Effect size of the former pair is medium ($d = -.54$) and that of the latter pair is close to large ($d = -.76$). The difference between Past and Recency was not significant. Therefore, participants in Future search tasks were able to find significantly fewer number of web pages than the other two classes. In other words, Future search tasks were less successful than Past and Recency tasks.

Second, a similar trend was observed from the task experience scores. Again, Past and Recency had a lower score than Future, suggesting that participants' search experience were less positive in Future search. One-way ANOVA test shows that the difference among the three classes is significant with large effect size. Post-hoc pairwise t-tests with Bonferroni correction show that the difference between Future and Past ($p = .008$), and Future and Recency ($p \leq .001$) are significant. Effect size of the former pair is medium ($d = .54$) and that of the latter pair is close to large ($d = .77$). However, the difference between Past and Recency was not significant. Two-way ANOVA test shows that there is no significant interaction effect with topics ($F(6, 81) = 1.994, p = .076$). Therefore, participants perceived that they struggled with Future search than the other two classes. This is in line with the result of saved pages.

Third, when we looked at interaction plots of these two overall performance scores (i.e., Figure 5c and 6c), we noticed that the variance of some topics had a different pattern from other topics. For example, in Figure 5c (Number of saved pages), Topic 13 had a larger difference across three classes than the other topics. On the other hand, in Figure 6c (Task experience), Topic 12 had a smaller difference among three classes than the others. Therefore, we looked at the level of pre-task topical familiarity per class which are reported in Table 5. As can be seen, Recency had a lower score than Future and Past, but their difference was not significant. On the other hand, the effect of topics on their perception was found to be significant ($F = 15.249, p \leq .001$). Figure 7c indicates Topic 11 on Oprah Winfrey was found to be particularly unfamiliar to the participants of this study. Nevertheless, two-ways ANOVA did not show a significant interaction effect between temporal classes and topics. It means that although one topic was particularly unfamiliar to participants, its effect was common to all temporal classes.

With this overall picture in mind, the following sections explore the search strategies and behaviour employed by participants to perform three temporal search tasks.

4.2. Reported Search strategies

In the exit questionnaire, we asked participants to describe what would be the best search strategies for each of temporal classes based on the six search sessions they performed. We requested to follow the template below to encourage participants to describe their strategies ideas: "I think the system should facilitate *[...]* strategies since *[...]* approach worked well for my search" or "I think the system should enable us to perform *[...]* strategies since I found it difficult to take *[...]* approach in my search". In other words, we encouraged participants to describe their best strategies or to formulate their system needs based on success experience and difficulty experience during the experiment. Reported strategies were then manually grouped by the authors to identify common or popular methods and issues. It should be noted that participants were not necessarily aware of advanced search options and resources, and thus, there could already be a facility available to achieve what participants wanted in the comments. Nevertheless, it is possible that such options are not widespread across people's search practices.

4.2.1. Common strategies and Difficulties

Two strategies were common (i.e., frequently mentioned) to most participants across the temporal classes. The first tactic was to augment topical terms with *temporal control expressions* (TCEs). A TCE can be explicitly temporal such as "latest", but can also be less explicit such as "outcome". Whatever the level of explicitness, the intention behind the use of TCEs is to control search results to elicit the information in a targeted temporal class from others. For example, "latest information GMO" was a popular query in Topic 014, where "latest information" was an added temporal control expression. The expressions employed varied across temporal classes (and can be artifact of topic descriptions), but most participants used this approach at least once during the experiment. Commonly reported expressions were summarised in Table 8. However, participants reported mixed results with this tactic: some stated that it worked well while others stated otherwise. A common comment in an unsuccessful case was that a web page contained the word "latest" but it was published several years ago, and thus, it did not have recent data.

The second common tactic was to use a search engine's option to specify a particular time span in search results. However, most participants described such tactic as a "wish list" of search engines. Indeed, many participants stated that "it would be great if we could sort the results based on publishing dates or filter out based on dates" or similar ideas. Only two participants clearly stated that they used the time-based search options to control their search results. This suggests that, as expected, searchers would like to control their search using time-based operations, but they rarely knew the availability of search options in the current search engines, or rarely used them for temporal information needs in practice. Moreover, controlling the timestamps of returned documents cannot be used for finding information about the future. It is also ineffective when the information need refers to events from more distant past unless the search is carried over longitudinal document collections such as news or web archives rather than on the open Web.

4.2.2. Past Search strategies and Difficulties

Past information needs in our study tended to seek for an origin, historical development, past major incidents of a given topic.

One common tactic reported by participants in past information needs was to find a manually curated summary page of a search topic in Wikipedia or an official web site. One participant stated that "I first got an overview from Wikipedia, then searched for detail information by following the links or by using keywords in the Wikipedia page". Although this tactic can be useful for any temporal classes, it was more often reported in the past information needs than others. Further analysis of participants' feedback suggests that one of the reasons for this tactic was due to a difficulty in finding past information on the Web. Several participants reported that past information was often buried by recency information in search results. Therefore, one needed to submit a specific keyword to gather past information on the topic. The summary page was often regarded as a source to find such specific keywords. Overall, participants appeared to struggle on past information needs when they were not able to find the summary page at an early stage of search.

4.2.3. Recency Search strategies and Difficulties

Recency information needs in our study tended to seek for the current status, latest findings, recent breakthroughs of a given topic.

Participants feedback suggests that recency search was perceived to be more straightforward than the other two temporal classes. Several participants stated that current search engines were sufficient to find relevant documents for recency information needs. Many of them also reported that adding TCEs such as "latest information" or "current circumstances" to topical keywords were effective in many topics. However, there were cases where such TCEs were used in returned documents with a wrong year such as "latest information in 1999". Therefore, just like in other classes, TCEs can be harmful in recency information needs.

Browsing was more often mentioned as a tactic than in the other two classes. The main sources of browsing were official home pages and news websites. One participant described that "when the keyword like *current* did not work, it was effective to find and browse domain-specific news sites". When a topic was centered around a named entity, their home page often showed the latest news. On the other hand, if a topic was about recent social issues then it was likely that they have been featured by news sites. Therefore, participants appeared to find browsing those pages was complementary to keyword search in recency information needs.

4.2.4. Future Search strategies and Difficulties

Future-related information needs in our study tended to be related to potential technologies, prediction of changes, future development and application, personal plans and future trends.

It was evident from their comments that participants found it most difficult to satisfy future information needs. One factor of the difficulty is that temporal expressions were less likely to associate with future-related information than other classes. One participant stated that "the keyword like *prediction* did not often allow me to find detail future-related information." Therefore, TCE could have been least effective in future information needs. This meant that participants had to find more specific topical terms to retrieve future-related information. The same problem occurred in past information needs, however in that case participants were able to find specific terms in the summary pages which usually did not exist for future-related information. Another factor frequently mentioned by participants was the lack of credibility or reliability of future-related information. Many of our subjects were aware that credibility and reliability of information were crucial for future information needs. However, they were not fully cognizant of how to confirm these qualities in search results.

In such circumstances the following is some of the major strategies taken or suggested by participants for future search. One popular tactic participants reported to overcome the credibility problem was to find experts on the topic, and locate future information from their sources (e.g., commentaries, blogs, reports). Since information from domain experts was often published in PDF files rather than HTML documents, some participants reported that focusing on a particular file type was another tactic in future-related information seeking. A couple of participants expressed a need to control search results based on a presence of figures and tables in documents, since they could help them to judge the credibility of information.

Another tactic suggested by several participants was to gain an idea of future-related keywords from the recent information of the same topic. For example, one participant reflected that "It felt that future-related information were more likely to be found from statistical prediction and academic papers, but those documents were not easy to understand. Therefore, I felt that some supporting mechanism to deepen my understanding of a given topic up till now was needed, to find future-related information.". In other words, by reading the latest information on the topic, one might be able to infer keywords that can elicit future information in search results. This *cross temporal class information seeking* (i.e., past information for recent information, recent information for future information) was mentioned for other classes, but it was more prominent in future-related information needs. This again highlights the difficulty of finding future-related information compared to other classes.

4.3. Task Perceptions

Based on the range of strategies and difficulties reported by participants, we examined if their perceptions of search sessions were in line with the findings. This is basically a breakdown of the total task experience scores we analysed in Section 4.1. Post-task questionnaires were administered to capture participants' perceptions of various aspects of search tasks, when they were completed. They include 1) clarity of information needs, 2) ease of first

	Temporal class					
	Future	Past	Recency	$F(2,58)$	p	η^2
(1) Clarity of information needs	2.9 (1.5)	2.3 (1.2)	2.5 (1.3)	3.871	.026	.06
(2) Ease of first query formulation	2.5 (1.1)	2.3 (1.1)	2.2 (1.0)	1.471	.238	.02
(3) Ease of subsequent query formulation	3.7 (1.5)	3.4 (1.5)	3.2 (1.5)	2.100	.132	.03
(4) Match to expected search results	4.5 (1.7)	3.6 (1.6)	3.2 (1.5)	10.540	$\leq .001$	.18
(5) Ease of SERP triage	4.3 (1.7)	3.5 (1.5)	3.0 (1.4)	9.800	$\leq .001$	.16
(6) Ease of finding relevant information	4.4 (1.7)	3.6 (1.6)	3.0 (1.4)	13.540	$\leq .001$	.20
(7) Confidence in relevance judgements	4.2 (1.6)	3.3 (1.6)	3.3 (1.5)	5.718	.005	.09
(8) Ease of search strategy decision making	4.5 (1.5)	4.5 (1.4)	4.4 (1.6)	0.286	.752	.00

Table 6: Mean and standard deviation of task perception scores (1: High, 7: Low). $N = 60$.

query formulation, 3) ease of subsequent query formulation, 4) match to expected search results, 5) ease of triage on search engine result pages (SERPs), 6) ease of finding relevant information from visited documents, 7) confidence in relevance judgements, and finally, 8) ease of search strategies decision-making (i.e., querying or clicking). These aspects can be found in a typical search process model proposed by literatures (e.g., [15]), although our questions were more elaborated than such a search model.

The results of task perceptions are shown in Table 6. The third row indicates the level of certainty about what kind of information they were asked to find. A noticeable difference was observed between Future and Past classes. One-way ANOVA test shows that some of the differences among the three classes are significant with medium effect size. Post-hoc pairwise t-tests with Bonferroni correction show that the difference between Future and Past is significant ($p = .049$) and effect size is closer to medium ($d = .44$). Therefore, it is possible that participants in Future tasks were, relatively speaking, less clear about what sort of information they were looking for than in Past tasks. However, since the mean score was below 3 in all temporal classes, we can judge that participants had a clear understanding of information needs in most cases.

The fourth and fifth row of Table 6 show the results of ease of 1st query formulation and subsequent query formulations, respectively. Participants perceived that formulating subsequent queries was generally more difficult than the first queries. Furthermore, the mean score for Future search tasks was higher than Past and Recent, but the difference among them is small. One-way ANOVA test did not find significant difference among the classes in the first query nor subsequent queries. Therefore, temporal classes did not seem to have significant effects on participants' perceptions regarding query formulation and reformulation. In Section 4.4, we will analyse their query formulation behaviour in more detail using log data.

Next, we examined to what extent participants found search results as expected. The sixth row of Table 6 (denoted by (4)) shows the results. Future tasks had a higher score than Past and Recency tasks and the difference is larger than the one in the previous results. One-way ANOVA test confirms that the differences among the temporal classes were significant with a large effect size. The post-hoc t-tests show that the difference between Future and Past ($p = .0044, d = .52$), and Future and Recency ($p \leq .001, d = .84$) were both significant with medium and large effect, respectively. Therefore, participants often found search results of Future tasks different from their original expectations.

We obtained a similar result for the ease of search results triage, which is a process of deciding which document in the list to visit to access the full-text. The result is shown in the seventh row of Table 6. Following previous trends, the mean score of Future tasks were higher than the Past and Recency tasks. One-way ANOVA test shows that the differences among the temporal classes were significant. Post-hoc t-tests indicate that the difference between Future and Past ($p = .024, d = .49$), and Future and Recency ($p \leq .001, d = .78$) were both significant with medium and large effect, respectively.

The eighth row of Table 6 (denoted as (6)) shows the result about the ease of finding relevant information from retrieved documents. Recall that we asked participants to highlight the most relevant part in a web page when they save it as a relevant document. Therefore, they engaged in identifying relevant information within a page, not just judging the relevance. We observed a similar pattern to previous results, where the score of Future tasks was higher than other two classes. One-way ANOVA test shows that the differences among the temporal classes were significant with large effect size. Post-hoc t-tests confirm that the difference between Future and Past ($p = .0026, d = .51$), and Future and Recency ($p \leq .001, d = .94$) were both significant with medium and large effect, respectively. Therefore, participants in Future tasks tended to find it more difficult to locate relevant information in retrieved documents than in tasks of the other classes.

We asked participants to include their level of confidence in the relevance judgements they saved. The ninth row of Table 6 shows the result about it. Again, the score of Future tasks was higher than Past and Recency tasks, suggesting that they were less confident about the relevance judgments in Future class than other two classes. One-way ANOVA test shows that the difference among the three classes is significant with small to medium effect size. Post-hoc t-tests confirm that the difference between Future and Past ($p = .012, d = .52$), and Future and Recency ($p = .012, d = .54$) were both significant with medium effect.

Finally, we look at the ease of search strategies decision making. In particular, we asked about the ease of making the choice whether to submit a new query or to view the next result page. The result is shown in the bottom row of Table 6. We did not observe a large difference among the temporal classes. ANOVA test also shows that there was not statistical significance among the classes. Therefore, the temporal class did not have a strong impact on the way participants made search tactic decision.

4.3.1. Summary of task perceptions

This section allowed us to elaborate what aspects of search process were strongly affected by temporal classes, based on participants' perceptions of search tasks. The results show that perceptions of query re/formulation or strategic decision making were not strongly affected by temporal classes, whereas triage of search results, extraction of relevant information, and confidence in relevance judgements were strongly affected, particularly in Future tasks.

Furthermore, although we did not find significant difference, the perceptions on Recency tasks were often more positive than Past tasks. These relatively consistent trends can be observed by the supporting figures (Figure 9 - 17) in the Appendix. The figures in the Appendix also allow us to observe that Topic 012 (Abenomics) often had a small difference among the three classes, while other topics mostly showed a varied level of differences among them, regarding the task perceptions. Therefore, the effect of topics seems to be limited.

The next section will investigate the client-side query logs based on the findings we obtained so far.

4.4. Log Data Analyses

We looked at search logs to gain further details of participants' strategies and behaviour on temporal information searching. In particular, we analysed queries formulated, and the domain and types of web pages accessed by participants during the search sessions.

4.4.1. Query formulation

As for participants query formulation behaviour, we looked at five aspects such as the average number of queries submitted in a search session, the ratio of zero clicked queries, average length of queries by terms, total number of unique terms used in a search session (i.e., search vocabulary), and finally, the ratio of temporal control expressions (TCEs) in search vocabulary. A zero click query is a query where a participant did not click any document from the search result.

	Temporal class					
	Future	Past	Recency	$F(2,58)$	p	η^2
(1) Number of queries	8.1 (4.7)	8.5 (4.9)	6.1 (3.6)	6.583	≤.001	.09
(2) Ratio of zero-click queries in (1)	.22 (.21)	.23 (.19)	.21 (.19)	.330	.720	.01
(3) Query length (terms)	2.1 (0.7)	2.2 (0.8)	2.4 (1.0)	1.226	.298	.03
(4) Search vocabulary size	7.9 (4.6)	8.9 (4.4)	7.2 (3.7)	4.556	.015	.05
(5) Ratio of temporal control expressions in (4)	.49 (.21)	.21 (.19)	.33 (.23)	25.499	≤.001	.34

Table 7: Query formulation behaviour. $N = 60$ (except (3)).

The results of these analyses are shown in Table 7. First, the third row of the table (denoted as (1)) shows that participants submitted on average the largest number of queries to Past tasks, followed by Future and Recency tasks. The difference between Past/Future and Recency appears to be relatively large with some outliers (Figure 16). One-way ANOVA test shows that the difference among the three classes is significant with medium effect. Post-hoc pairwise t-tests with Bonferroni correction show that the difference between Future and Recency ($p = .049, d = .47$) and Past and Recency ($p = .011, d = .55$) are significant with medium effect size. Therefore, participants submitted fewer number of queries in Recency tasks than the other two classes.

Second, we looked at the ratio of zero-click queries since this can indicate the percentage of poorly formulated queries. Contrary to our expectation, the difference of zero-click queries was not so high among the three classes. One-way ANOVA test also shows that the difference is not significant. A similar insignificant difference was observed in the average length of queries submitted by participants. Supporting figures of these results can be found in Figure 18 and 19.

Third, we looked at the range of vocabularies employed to complete a search session. Unlike navigational tasks, finding many relevant documents often requires to reformulate initial queries with various terms. Therefore, looking at search vocabulary can help us understand the level of diversity required to progress search tasks. The result of search vocabulary is shown in the 6th row of Table 7, and the ratio of TCEs in the vocabulary is shown in the bottom row of the table. The average size of search vocabulary in Past tasks was slightly larger than the other two classes, like the number of queries. One-way ANOVA test shows that the difference among the three classes is significant with effect close to medium. However, post-hoc pairwise t-tests with Bonferroni correction did not show any pairs of classes to be significantly different. This led us to look at interaction effect with topics. Two-ways ANOVA test shows that there is a significant interaction effect between the temporal class and topics ($F(7, 79) = 3.345, p = .004$). Interaction plot in Figure 20 illustrates complex patterns of their interaction effect: the size of search vocabulary cannot be explained by the temporal class alone, and there could be several topics that affected the difference of the mean values.

The result of TCE ratio in search vocabulary was more prominent. In particular, we observed that the TCE ratio in Future tasks was more than the double of Past tasks. One-way ANOVA test shows that the difference among the three classes was significant with large effect. Post-hoc pairwise t-tests with Bonferroni correction show that the difference between all pairs is significant with medium to large effect (Future-Past: $p \leq .001, d = 1.44$, Future-Recency: $p \leq .001, d = .75$, Past-Recency: $p = .0072, d = -.56$). However, like the search vocabulary, two-ways ANOVA test shows that there is a significant interaction effect with topics ($F(7, 79) = 11.130, p \leq .001$). Interaction plot in Figure 17 illustrates that the ratio of Future tasks is consistently higher than the other two classes except Topic 012. This suggests that although there was a significant interaction effect, it is safe to say that participants used TCEs in Future tasks much more frequently than Past or Recency tasks in most cases.

We then examined the most common temporal control expressions used in queries. Table 8 lists the most frequent TCEs in different temporal classes. We have manually grouped them into three classes: *explicit temporal expressions*, *explicit temporal category markers* and *implicit temporal category markers*. *Explicit temporal expressions* denote a precise point in time or period and thus can be directly anchored on timeline without need for any further information or processing [1, 8, 30]. They can belong to different levels of granularity but due to the character of the search topics used in our study the most common granularity was a year or decade. Explicit temporal expressions are usually contrasted with *implicit* and *relative temporal expressions* [1, 30], which also point to particular time points or periods. Relative temporal expressions, however, need an absolute reference such as document timestamp of other explicit temporal expressions occurring in nearby text in order to be correctly anchored on a timeline. We did not find any of such expressions in the query logs. *Explicit temporal category markers* directly indicate a general temporal class Past, Recency and Future but cannot be positioned on timeline due to their impreciseness. On the other hand, *implicit temporal category markers* bear certain semantic meaning other than the temporal class indication. However, implicitly, the temporal class can be inferred with a high degree of accuracy.

4.4.2. Accessed pages

In order to gain a deeper insight into the characteristics of accessed resources we manually grouped them into 10 categories: ORG, NEWS, NEWS2, BLOG, MAG, COM, KB, FORUM, SNS, SHOP, and MISC. See Table 9 for the description of these categories.

First, we looked at the top 10 most frequently visited domains across the three temporal classes. Table 10 shows the results along with the category labels. The most noticeable characteristics in the table is the access to Wikipedia.

Table 8: Popular TCEs observed and reported in Past, Recency and Future search tasks. *Italic* terms are those which were explicitly identified by participants in the exit questionnaires.

	Temporal Control Expressions		
	Explicit temporal expressions	Explicit temporal category markers	Implicit temporal category markers
Past	1990, 1990s, 90s, 2000	*past*, *history*, at that time	origin, cause, background, effect, emergence, *progress*, *achievement*, *outcome*
Recency	2014	*now*, *latest* (information), *current*, present	effect, *trend*, new, outcome, timeline, *circumstance*, *numbers*, *data*
Future	2013, 2014, 2015, 2016 after 2030	*future*, in the future	*prediction*, change, *prospect*, influence, *plan*, trend, development, *schedule*, *possibility*, *impact*

Table 9: Categories of accessed pages.

Category	Description
BLOG	Personal blogs, group blogs, home pages, etc.
FORUM	Bulletin boards, CQA sites
NEWS	Major general news sites
NEWS2	Domain specific news sites
MAG	Online magazine sites
ORG	Governments, associations, universities, NGOs, etc.
KB	Knowledge Base, databases, curated sites, etc.
SNS	Social network systems, Facebook, twitter, etc.
COM	Commercial companies websites
SHOP	E-commerce shops
MISC	Other resources

Table 10: Top 10 most frequently visited domains. Wikipedia is highlighted. See Table 9 for the description of category keys.

	Future (N = 1157)		Past (N = 1093)		Recency (N = 1194)
61	[BLOG] oprah.com	197	[KB] *ja.wikipedia.org*	101	[ORG] wwf.or.jp
58	[KB] *ja.wikipedia.org*	58	[ORG] wwf.or.jp	88	[ORG] biodic.go.jp
41	[ORG] wwf.or.jp	21	[NEWS2] macs.mainichi.co.jp	50	[KB] *ja.wikipedia.org*
36	[ORG] biodic.go.jp	21	[ORG] biodic.go.jp	48	[BLOG] oprah.com
33	[KB] matome.naver.jp	20	[MAG] diamond.jp	40	[KB] matome.naver.jp
33	[ORG] env.go.jp	18	[KB] matome.naver.jp	35	[MAG] wedge.ismedia.jp
21	[NEWS] jp.reuters.com	17	[NEWS2] fashion-press.net	30	[NEWS]nikkeibp.co.jp
21	[FORUM] android-group.jp	15	[NEWS2] fashion-gp.com	30	[COM] monsanto.co.jp
18	[ORG] mhlw.go.jp	13	[ORG] satoyama-initiative.org	21	[COM] au.kddi.com
18	[ORG] maff.go.jp	12	[NEWS] yomiuri.co.jp	16	[MISC] allabout.co.jp

Although all classes had frequent access to Wikipedia, Past tasks (18%) had access more than twice as frequent as Future (5.0%) or Recency (7.4%) tasks. This indicates that participants' reliance on Wikipedia for Past tasks was much stronger than the other two tasks. Apart from that, there seems to be more commonality than difference among the three classes, sharing several domains in the top 10.

Figure 3 shows the distribution of all categories between the three temporal classes. An overall trend we observed

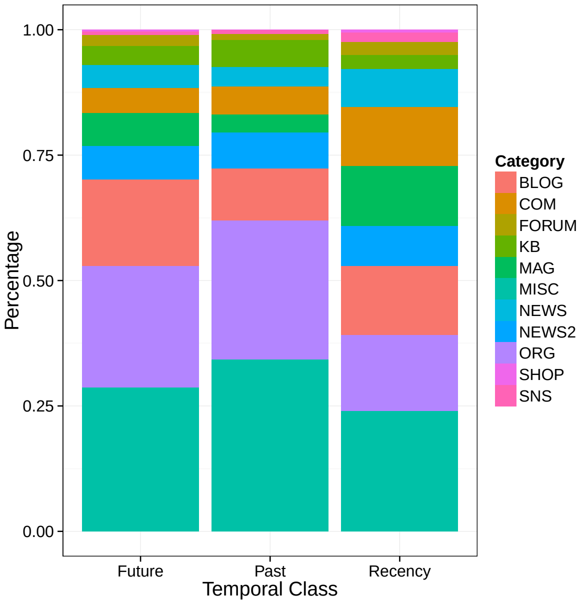


Figure 3: Percentage of visited page categories. Future: $N = 1157$, Past: $N = 1093$, Recency: $N = 1194$.

	Temporal class		
	Future	Past	Recency
Number of clicked pages	1157 (100%)	1033 (100%)	1194 (100%)
.pdf	62 (5.4%)	68 (6.6%)	27 (2.3%)
.html or .htm	414 (35.8%)	304 (29.4%)	409 (34.3%)

Table 11: Frequency and percentage of file extensions of clicked pages. PDF and HTML only.

was that the distribution of Future and Past tasks was more similar to each other than Recency task. For example, for both classes participants visited the ORG category pages more often than for Recency class. Both classes also have fewer percentage of access to the COM and MAG category pages than Recency class. In addition, access to different categories appears to be more evenly distributed in Recency class than the other two classes. In other words, Future and Past classes have more skewed access to a set of categories. However, it is difficult to be conclusive about these trends from our data alone and further analysis on large-scale query logs should be conducted to follow up.

Finally, we looked at the file extensions of visited pages. Here, we were particularly interested in the access to PDF files which was mentioned by participants as one of search strategies (See Section 4.2). The frequency and ratio of PDF files and HTML files are shown in Table 11. We observed that access to PDF files in Future and Past tasks was more frequent than in Recency tasks. Since we do not have an underlying distribution of these file types in search results, it is difficult to interpret the exact numbers. However, it is clear that there was more access to PDF files to complete Future and Past tasks than Recency tasks.

5. Discussions

This study was motivated by lack of understanding regarding temporal search behaviour. Particularly, we were interested in investigating temporal searching behaviour and strategies in a structured way so that effect of temporal classes such as past, recency, and future can be identified. This section first summarises the main findings from our study, then discusses the implications on the design of future temporal IR systems, and finally, clarifies the limitations of our study.

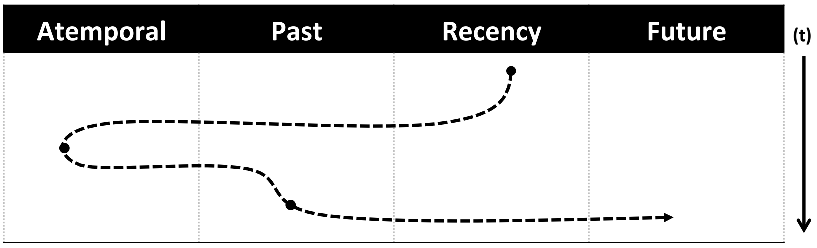


Figure 4: Illustration of time-related information orienteering. It might begin with a recency search result, moving to atemporal information, gaining more knowledge from past information, before finally reaching to an original need of future information.

5.1. Main findings

Our findings on temporal search behaviour and strategies are benefited from a combination of objective measures such as query logs and subjective measures such as questionnaires.

First, participants had a varied degree of search experience across temporal classes. Atemporal search such as looking up a definition of concepts and recency search to find latest information were more common than finding past or future information. This could affect participants' level of confidence and anxiety during search [25]. The overall search performance and experience basically reflected these practice and the Recency class had more positive scores than Past and Future classes.

Second, the analyses on task perceptions showed that participants' perceptions on query formulation and reformulation did not differ significantly across the temporal classes. However, it was the most common strategy to control temporal orientation of search results using temporal controlling expressions (TCEs). Our log analysis identified some of the common expressions employed by participants that are different across temporal classes. We also found that Future tasks had a higher level of TCEs in search vocabulary, suggesting that participants had a difficulty in controlling temporal orientation.

Third, participants found triage on search results and finding relevant information from web pages particularly difficult in Future tasks. They also expressed that search results were often unexpected and their confidence of relevance judgements was lower in Future tasks when compared to Recency or Past tasks. This appears to be closely related to the nature of future-oriented tasks where the credibility of information is crucial. Participants sometimes opted for PDF files in official web pages to increase the level of credibility in their accessed information. These intentions were partially supported by the log analysis. However, more studies are needed.

Fourth, different behavioral patterns might be useful for detecting temporal query intents. For example, query formulation patterns were similar between Future and Past, but different from Recency. On the other hand, click page category pattens were similar between Future and Recency and different from Past. File types might also be used to discriminate temporal intents but a large data is likely to be required. TCEs should be a good starting point for temporal query intent classification, too.

Last, a careful examination is needed to understand the cause of search difficulties across temporal classes. For example, there could simply be a fewer number of future-related documents on the Web than other classes, or, finding future-related documents might be more difficult than other classes even if there is a similar number of documents. A gross estimation using average saved pages and total clicked pages indicates that Future had a lower level of saving ratio than Past and Recency. However, this is far from conclusive. Similarly, further work is needed to understand the distribution of documents across temporal classes on the Web.

5.2. Implications on TIR System Design

Our participants often reported problems with retrieving information related to a particular temporal class, especially, to the Future and Past. To support them, first, search systems should detect temporal intents behind queries or search sessions. Then they can provide a range of supporting services for users searching for time-related information. An important implication from the common strategies was that it is common and intuitive for people to attempt a temporal control of search using broad temporal expressions such as past, history, recent, latest, future, predictions. However, many participants reported that search results containing these temporal expressions did not always reflect their temporal intents. We thus need to enable an IR system to accurately interpret people's temporal intents from queries, and retrieve documents that have information within intended temporal scope, not just documents that contain the temporal expression terms themselves. In other words, temporal control expressions should be seen as a sign of temporal intents of search, rather than as query terms.

Different queries exhibit different difficulty levels when it comes to finding information from a given temporal class. For example, it may be easier to find future-related information about Abenomics than about fashion trends. When the temporal intent of search query is properly detected, search engines can push the results that adhere to the correct temporal class up and thus decrease *temporal information overload* which may result from the abundance of data of different temporal classes. For example, more news articles can be incorporated in search results for queries that clearly relate to current events as suggested in prior research works [11, 26].

Temporal diversification could be another mechanism to satisfy temporal user needs, especially in cases where temporal intents cannot be reliably estimated and there is a high probability that users may search for information pertaining to a different time. Some participants reported that accurately focusing on a particular year in the past was difficult without an overall historical understanding. Some of them directly expressed need to have summary-like overview pages such as biographies in case of person-related queries. Such documents could function as a reference base to overview temporal progress of a topic or an entity and to support users with query formulation. We think that the process of *time-related information orienteering* (see Figure 4) is the first step towards successful time-based search but it however requires considerable effort of collecting, comparing and organizing time-related information from diverse sources. We believe that Wikipedia articles often serve such purpose (evidenced by Wikipedia frequent use for searching past-related information). However, not every topic may be supported with such pages and for some rare topics information about the past or future may be scattered around different documents. In such cases an automatic construction of temporal summaries both for overviewing the past evolution and for displaying a future roadmap could alleviate the time-related information orienteering burden. Recently, there have been several research proposals for generating both past- as well as future-focused summaries and timelines [2, 29, 10]. They could be displayed alongside search results much like maps tend to be shown for queries containing spatial components.

In some cases simple user interface changes could be beneficial. For example, we have found that time constraint was rarely used and even known. Then emphasizing more the temporal search options such as filtering by timestamp could help. However, we note that a simple chronological arrangement or filtering by document timestamp clearly cannot help in the case of future-related information needs. The fact that a document was published at certain time does not necessarily mean it is about that time period, as the document can be related to the future or the past. We believe that organizing documents by their focus time [16], that is the time to which they refer to, would be a strong complementary function to a simple time control based on document creation dates.

Similarly to UI issues, the presentation of results offers potential for search efficiency improvements. A simple way is to increase the understanding of the search results temporality by emphasizing their publication or update times or by the time-aware construction of snippets. Effective temporal snippets should contain explicitly understandable temporal expressions that appear in documents or any hints of their temporal class.

Temporal query suggestion for effective retrieval of data from a particular temporal class could be another useful technique. The suggestions could involve the selection of effective temporal control expressions including explicit dates or topic words to force content from particular temporal class.

The last implication relates to the issue of information credibility. We need a mechanism to allow searchers to judge the credibility of the results, especially, those in future information search. Participants reported the necessity of visiting several pages and comparing their information to synthesise and corroborate the findings. A solution for credibility estimation should incorporate several features such as source type (e.g., company page, newswire site or personal blog), information age, temporal horizon of prediction, popularity of the same prediction in other documents and the occurrence frequency of modality expressions such as maybe, likely, surely, probably, might and so on.

5.3. Limitations

There are limitations in this study. First of all, participants were sampled from university students from Japan, although their academic background varied. A particular caution is needed to interpret the temporal control expressions shown in Table 8, as they are translation from Japanese. Other languages are likely to have their own expressions that are not listed there.

Second, due to time and other constraints only a limited number of topics were used. As topics tend to have varying difficulty in different temporal classes we could not test the whole spectrum of search difficulty. We have also focused on coarse temporal classes such as past, recency and future. Further testing would involve more fine grained divisions (e.g., near/far future/past) or particular decades.

Lastly, one evaluation of search success has been done by considering the numbers of relevant documents saved which may not accurately reflect whether users could indeed find what they were looking for.

6. Conclusions and Future Work

This paper reported the findings of an exploratory user study that investigated people's information seeking behaviour of temporal information searching. The study allowed us to gain insight into the current practice of temporal search, difficulties people tend to encounter, and implications on future temporal search systems. When contrasting search effectiveness across classes, we noticed that search for Future-related information is the most challenging and requires more support for users. On the other hand, searching for Recency-related content was easiest, most common and most successful for users. The suggestions from our study ranged from a prominent presentation of existing search control options in an appropriate context to intelligent interpretation of temporal control expressions in queries, application of expert finding technologies, and smooth linkage across temporal classes. This suggests that there are still many opportunities to improve the search performance and experience of temporal search in the future.

There are several potential future directions for this research. One of them is studying the effect of topics on user search strategies and their behaviour. It is expected that certain topics are well-represented on the Web making it easy to find relevant information (e.g., there may be few documents on the past of genetically modified food but there could be quite much content that relates to future perspectives and predictions regarding this issue). Next, we plan to study the problem of vocabulary mismatch in greater detail. As events and topics evolve, different names can be used to refer to the same or similar objects. It is then interesting to study how users would try to find past information when the concept names and named entities differ greatly. Another future line of work is to investigate the relationship between temporal task familiarity and temporal orienteering behaviour. It is plausible that users which are less familiar with a particular temporal class might require a greater level of temporal orienteering. Therefore, behavioural analysis on temporal orienteering from the viewpoint of task familiarity can be an important aspect of temporal IR. Finally, a more detailed analysis following this exploratory study should be done to measure the probability of transitions between certain strategies, between browsing and searching as well as to quantify the effect of time duration on the search and its success.

7. Acknowledgments

The authors want to thank the participants of the user study. The first author was supported in part by MEXT Grant-in-Aid for Young Scientists B (#24700239). The second author was supported in part by MEXT Grant-in-Aid for Young Scientists B (#22700096) and by the JST research promotion program Sakigake: "Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents". The authors also thank to the NTCIR project at NII for supporting the construction of NTCIR-11 Temporalia test collection. Any opinions, findings, and conclusions described here are the authors and do not necessarily reflect those of the sponsors.

References

[1] Alonso, O., Baeza-yates, R., Strötgen, J., Gertz, M., 2011. Temporal information retrieval: Challenges and opportunities. In: 1st Temporal Web Analytics Workshop at WWW. pp. 1–8.

[2] Alonso, O., Gertz, M., Baeza-Yates, R., 2009. Clustering and exploring search results using timeline constructions. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. CIKM '09. ACM, New York, NY, USA, pp. 97–106.
URL `http://doi.acm.org/10.1145/1645953.1645968`

[3] Anand, A., Bedathur, S., Berberich, K., Schenkel, R., 2011. Temporal index sharding for space-time efficiency in archive search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM, New York, NY, USA, pp. 545–554.
URL `http://doi.acm.org/10.1145/2009916.2009991`

[4] Bates, M. J., Jul. 1979. Information search tactics. Journal of the American Society for Information Science 30 (4), 205–214.
URL `http://dx.doi.org/10.1002/asi.4630300406`

[5] Berberich, K., Bedathur, S., Alonso, O., Weikum, G., 2010. A language modeling approach for temporal information needs. In: Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010). pp. 13–25.

[6] Borlund, P., 2000. Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56, 71–90.

[7] Campos, R., Dias, G., Jorge, A. M., 2011. What is the temporal value of web snippets? In: Proceedings of TWAW '11. TWAW '11. pp. 9–16.

[8] Campos, R., Dias, G., Jorge, A. M., Jatowt, A., Aug. 2014. Survey of temporal information retrieval and related applications. ACM Comput. Surv. 47 (2), 15:1–15:41.
URL `http://doi.acm.org/10.1145/2619088`

[9] Campos, R., Jorge, A. M., Dias, G., Nunes, C., 2012. Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01. WI-IAT '12. IEEE Computer Society, Washington, DC, USA, pp. 1–8.
URL `http://dl.acm.org/citation.cfm?id=2457524.2457656`

[10] Demartini, G., Missen, M. M. S., Blanco, R., Zaragoza, H., 2010. Taer: Time aware entity retrieval. In: Proc. of ACM Conference on Information and Knowledge Management (CIKM), Toronto, Canada, 2010.

[11] Diaz, F., 2009. Integration of news content into web results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. WSDM '09. ACM, New York, NY, USA, pp. 182–191.
URL `http://doi.acm.org/10.1145/1498759.1498825`

[12] Fetterly, D., Manasse, M., Najork, M., Wiener, J., 2003. A large-scale study of the evolution of web pages. In: Proceedings of the 12th International Conference on World Wide Web. WWW '03. ACM, New York, NY, USA, pp. 669–678.
URL `http://doi.acm.org/10.1145/775152.775246`

[13] Fidel, R., 1985. Movesin online searching. Online Review 9 (1), 61–74.

[14] Harter, S. P., 1986. Online Information Retrieval: Concepts, Principles, and Techniques. Academic Press.

[15] Hearst, M. A., 1999. User interfaces and visualization. In: Baeza-Yates, R., Ribeiro-Neto, B. (Eds.), Modern information retrieval. Harlow: Addison-Wesley, New York, NY, USA, pp. 257–323.

[16] Jatowt, A., Au Yeung, C.-M., Tanaka, K., 2013. Estimating document focus time. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. CIKM '13. ACM, New York, NY, USA, pp. 2273–2278.
URL `http://doi.acm.org/10.1145/2505515.2505655`

[17] Joho, H., Jatowt, A., Blanco, R., 2014. NTCIR Temporalia: a test collection for temporal information access research. In: Proceedings of the companion publication of the 23rd international conference on World wide web companion. pp. 845–850.

[18] Joho, H., Jatowt, A., Blanco, R., Naka, H., Yamamoto, S., 2014. Overview of ntcir-11 temporal information access (temporalia) task. In: Proceedings of the NTCIR-11 Conference.

[19] Joho, H., Jatowt, A., Roi, B., 2013. A survey of temporal web search experience. In: Proceedings of the 22nd International Conference on World Wide Web Companion. WWW '13 Companion. pp. 1101–1108.
URL `http://dl.acm.org/citation.cfm?id=2487788.2488126`

[20] Jones, R., Diaz, F., Jul. 2007. Temporal profiles of queries. ACM Trans. Inf. Syst. 25 (3).
URL `http://doi.acm.org/10.1145/1247715.1247720`

[21] Kanhabua, N., Blanco, R., Matthews, M., 2011. Ranking related news predictions. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM, New York, NY, USA, pp. 755–764.
URL `http://doi.acm.org/10.1145/2009916.2010018`

[22] Kanhabua, N., Nørvåg, K., 2010. Determining time of queries for re-ranking search results. In: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries. ECDL'10. Springer-Verlag, Berlin, Heidelberg, pp. 261–272.
URL `http://dl.acm.org/citation.cfm?id=1887759.1887796`

[23] Kanhabua, N., Nørvåg, K., 2012. Learning to rank search results for time-sensitive queries. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. CIKM '12. ACM, New York, NY, USA, pp. 2463–2466.
URL `http://doi.acm.org/10.1145/2396761.2398667`

[24] Kato, M. P., Yamamoto, T., Ohshima, H., Tanaka, K., 2014. Cognitive search intents hidden behind queries: A user study on query formulations. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. WWW Companion '14. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 313–314.
URL `http://dx.doi.org/10.1145/2567948.2577279`

[25] Kuhlthau, C., 1991. Inside the search process: Information seeking from the users perspective. Journal of the American Society for Information Science 42 (5), 361–371.

[26] Kulkarni, A., Teevan, J., Svore, K. M., Dumais, S. T., 2011. Understanding temporal query dynamics. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. WSDM '11. ACM, New York, NY, USA, pp. 167–176.
URL `http://doi.acm.org/10.1145/1935826.1935862`

[27] Li, X., Croft, W. B., 2003. Time-based language models. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management. CIKM '03. ACM, New York, NY, USA, pp. 469–475.
URL `http://doi.acm.org/10.1145/956863.956951`

[28] Marchionini, G., 1995. Information Seeking in Electronic Environments. Cambridge University Press, New York, NY, USA.

[29] Matthews, M., Tolchinsky, P., Blanco, R., Atserias, J., Mika, P., Zaragoza, H., 2010. Searching through time in the new york times. In: Bridging Human-Computer Interaction and Information Retrieval.

[30] Mazur, P., 2012. Broad-coverage rule-based processing of temporal expressions. In: PhD thesis, Australia Macquarie University. pp. 1–245.

[31] Metzler, D., Jones, R., Peng, F., Zhang, R., 2009. Improving search relevance for implicitly temporal queries. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09. ACM, New York, NY, USA, pp. 700–701.
URL `http://doi.acm.org/10.1145/1571941.1572085`

[32] Ntoulas, A., Cho, J., Olston, C., 2004. What's new on the web?: The evolution of the web from a search engine perspective. In: Proceedings

of the 13th International Conference on World Wide Web. WWW ’04. ACM, New York, NY, USA, pp. 1–12.
URL `http://doi.acm.org/10.1145/988672.988674`

[33] Nunes, S., Ribeiro, C., David, G., 2008. Use of temporal expressions in web search. In: Proceedings of the 30th European Conference on Advances in Information Retrieval. ECIR’08. Springer-Verlag, Berlin, Heidelberg, pp. 580–584.
URL `http://dl.acm.org/citation.cfm?id=1793274.1793347`

[34] Strötgen, J., Alonso, O., Gertz, M., 2012. Identification of top relevant temporal expressions in documents. In: Proceedings of the 2nd Temporal Web Analytics Workshop. TempWeb ’12. ACM, New York, NY, USA, pp. 33–40.
URL `http://doi.acm.org/10.1145/2169095.2169102`

[35] Strötgen, J., Gertz, M., Popov, P., 2010. Extraction and exploration of spatio-temporal information in documents. In: Proceedings of the 6th Workshop on Geographic Information Retrieval. GIR ’10. ACM, New York, NY, USA, pp. 16:1–16:8.
URL `http://doi.acm.org/10.1145/1722080.1722101`

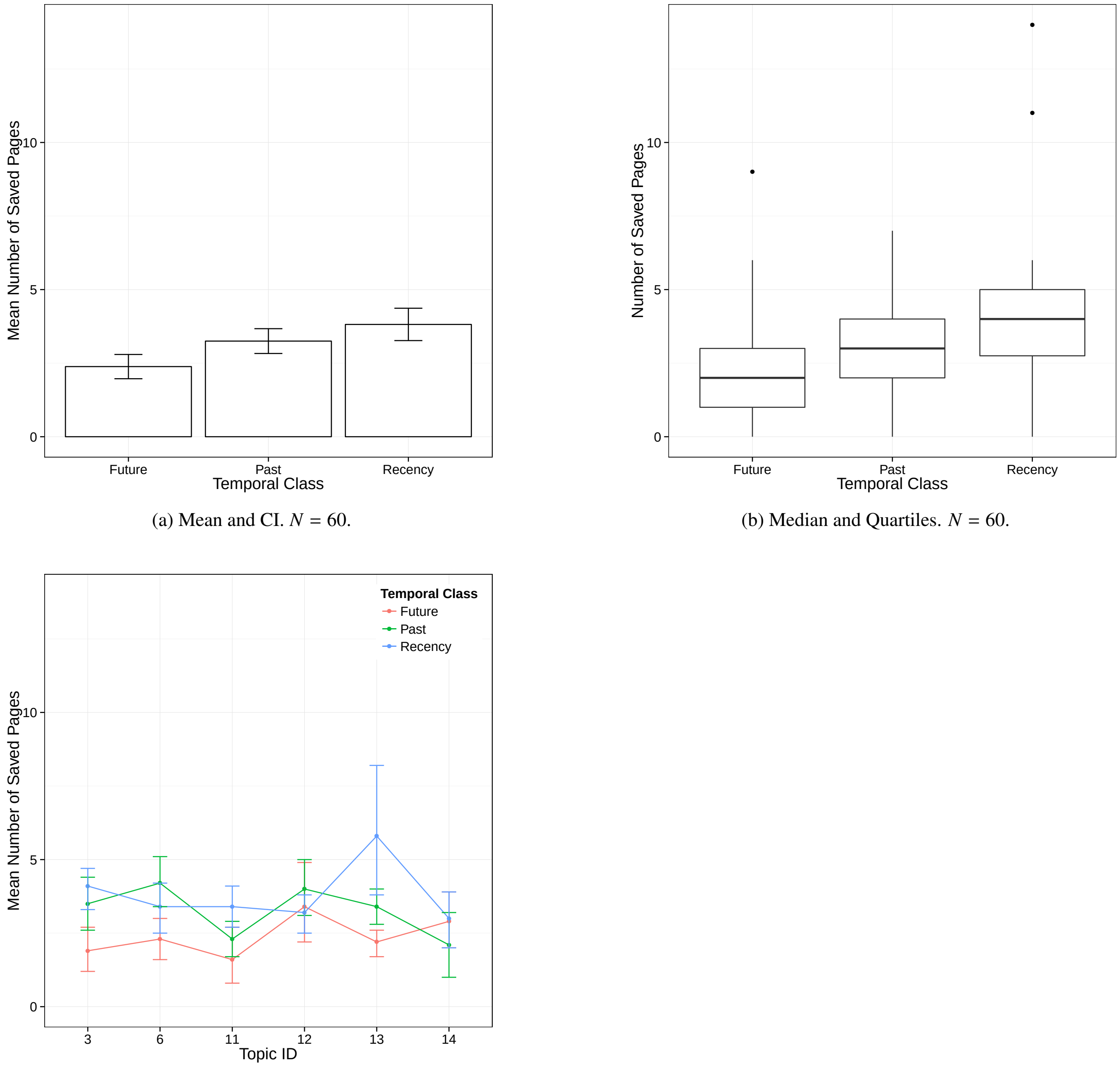
[36] Wang, P., 2011. Information behavior and seeking. In: Ruthven, I., Kelly, D. (Eds.), Interactive Information Seeking, Behaviour, and Retrieval. Facet Publishing, pp. 15–42.

[37] Wilson, M. L., schraefel, M. C., White, R. W., Jul. 2009. Evaluating advanced search interfaces using established information-seeking models. J. Am. Soc. Inf. Sci. Technol. 60 (7), 1407–1422.
URL `http://dx.doi.org/10.1002/asi.v60:7`

[38] Xie, I., Joo, S., 2010. Transitions in search tactics during the web-based search process. Journal of the American Society for Information Science and Technology 61 (11), 2188–2205.
URL `http://dx.doi.org/10.1002/asi.21391`

A. Supporting figures.

The following figures should be referred to gain detail differences among temporal classes. The bar plots are used to present a mean value and its confidence interval. The box plots are used to show the median and quartiles values. The topic breakdown charts are used to show how the statistics are distributed across the six topics used in our experiment.

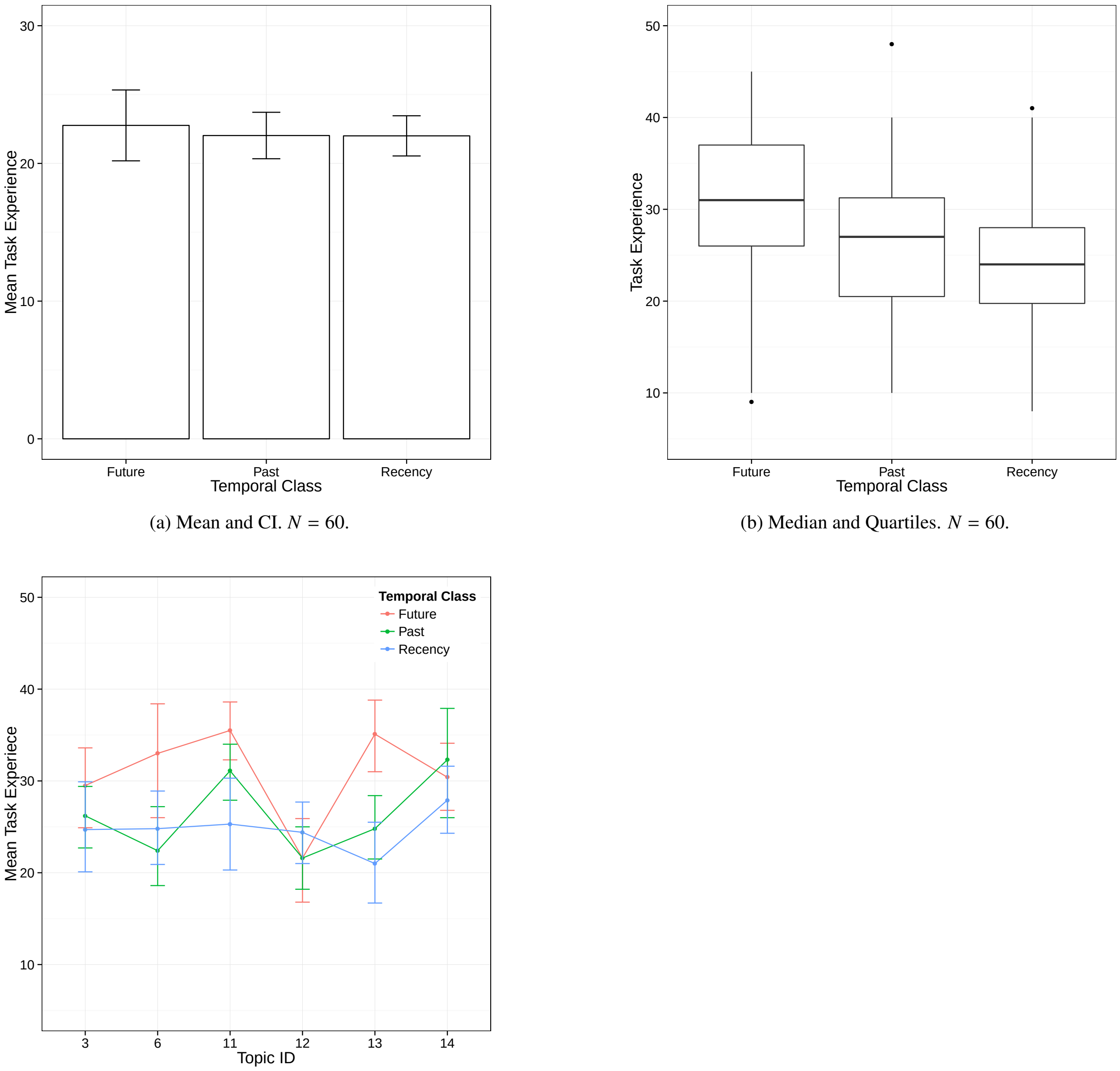


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 5: Number of saved pages.

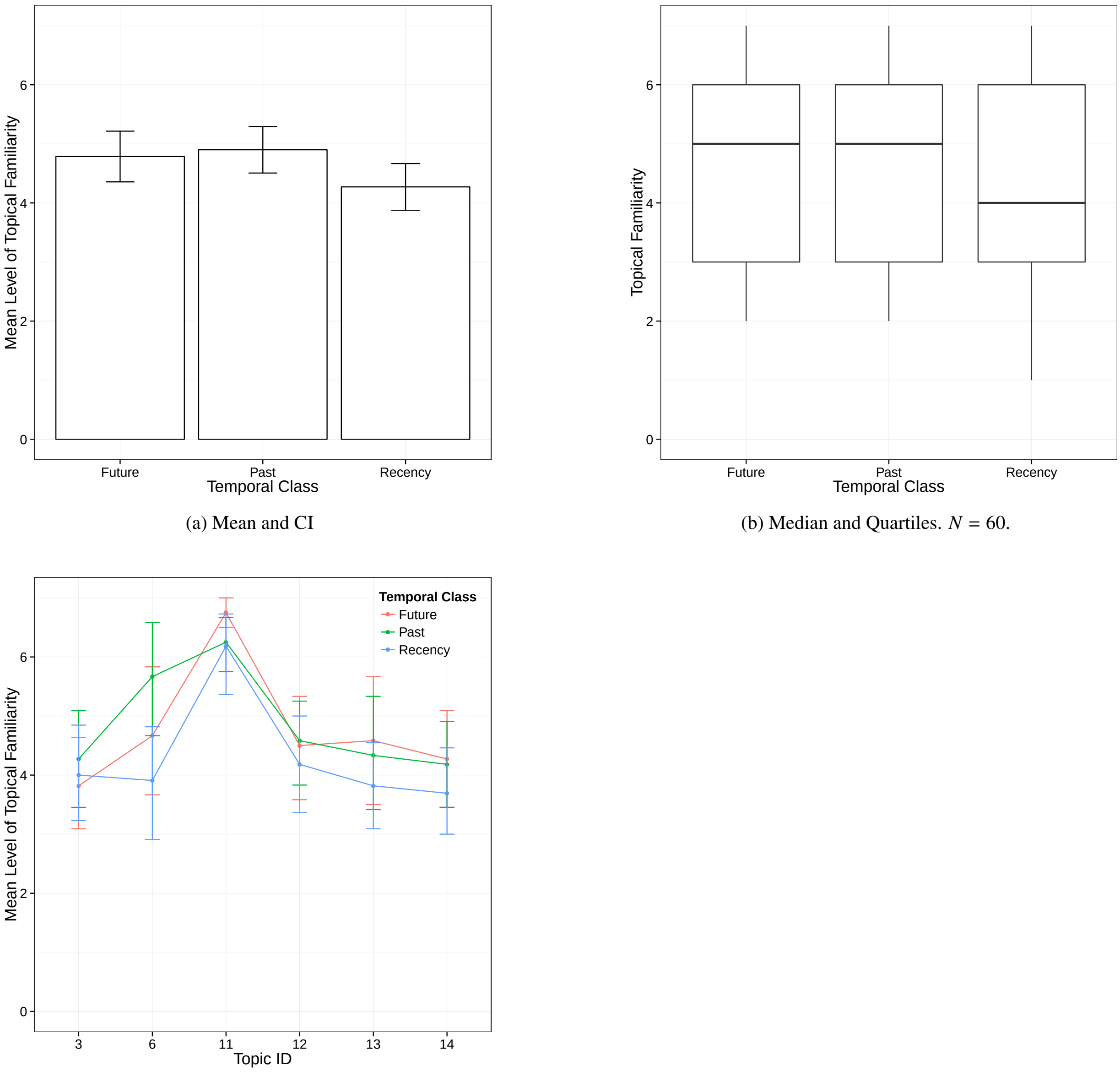


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 6: Total Experience Score (High: Negative, Low: Positive).

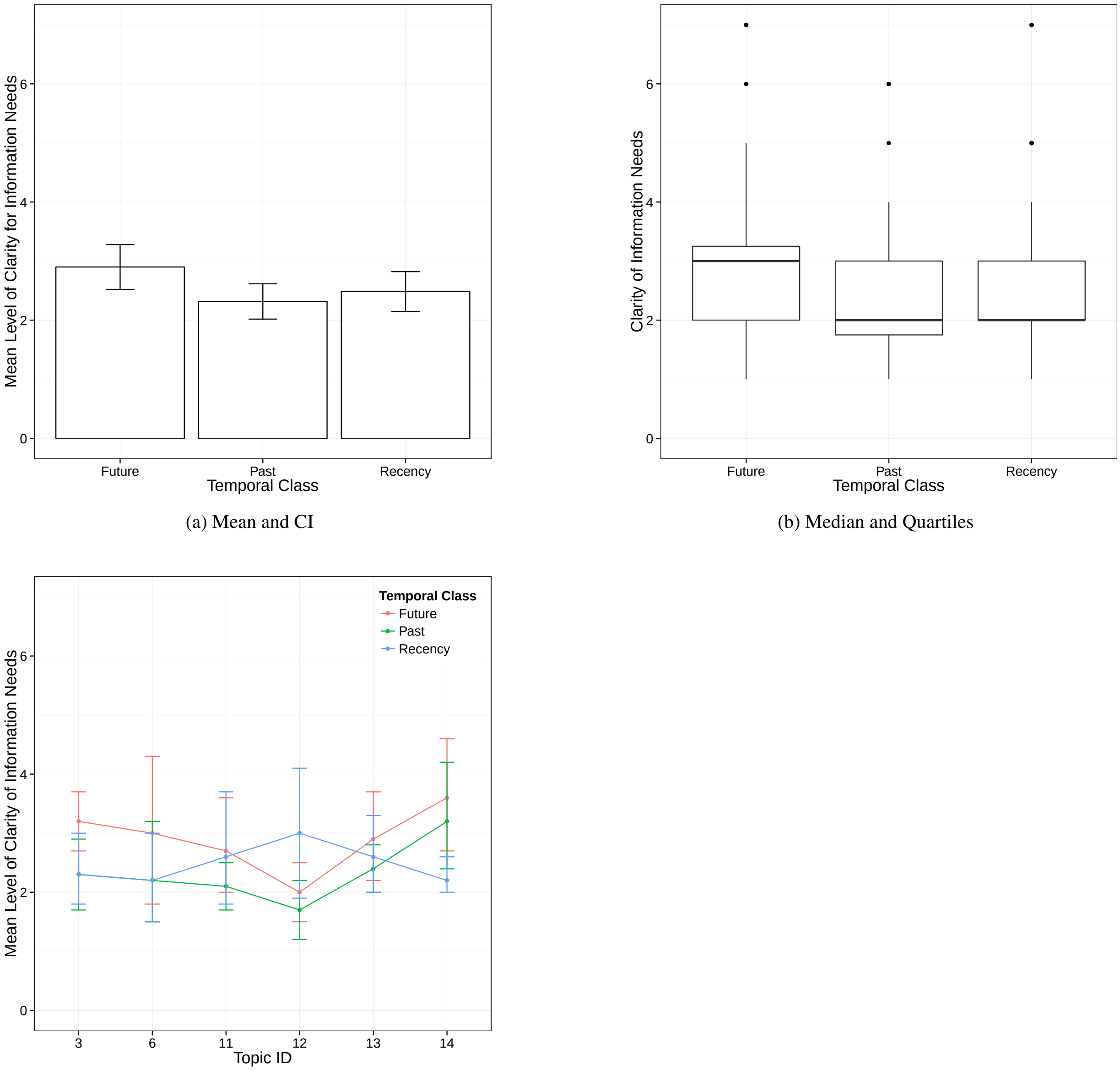


(a) Mean and CI

(b) Median and Quartiles. N = 60.

(c) Interaction with topics. N = 10.

Figure 7: Pre-Task Topical Familiarity. (1: High, 7: Low). N = 60.

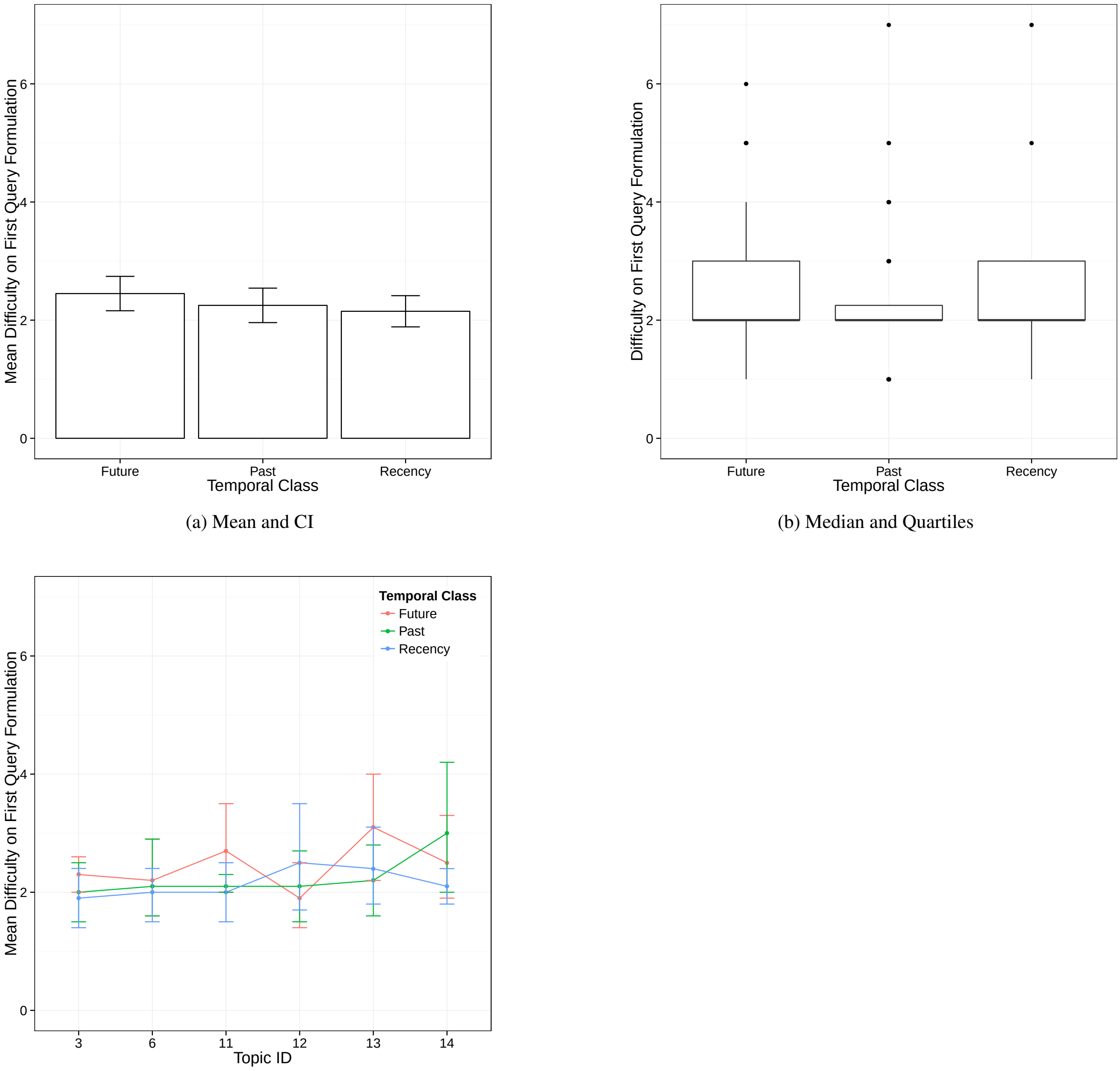


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 8: Clarity of information needs. (1: High, 7: Low). $N = 60$.

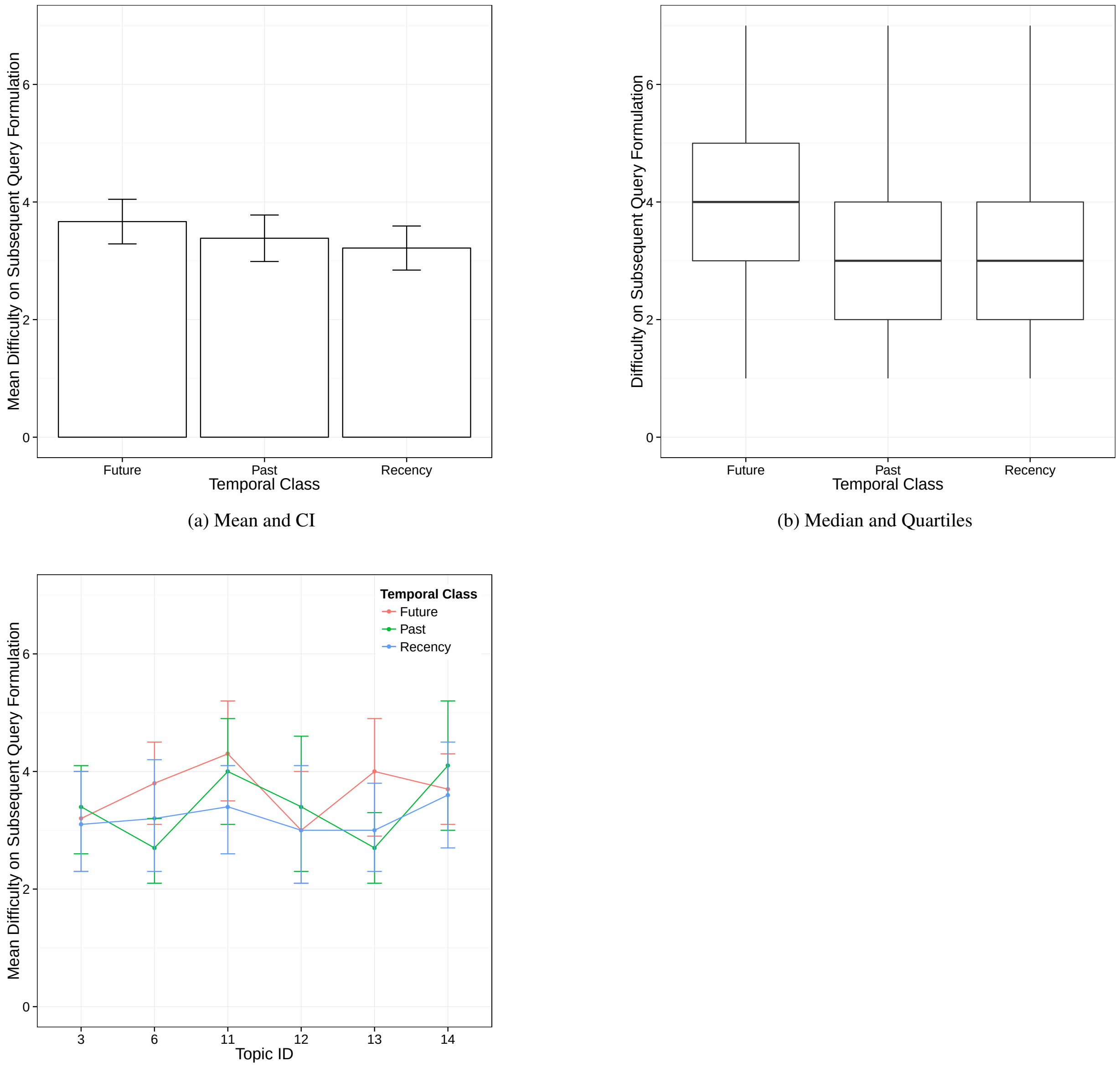


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 9: Ease of first query formulation. (1: High, 7: Low). $N = 60$.

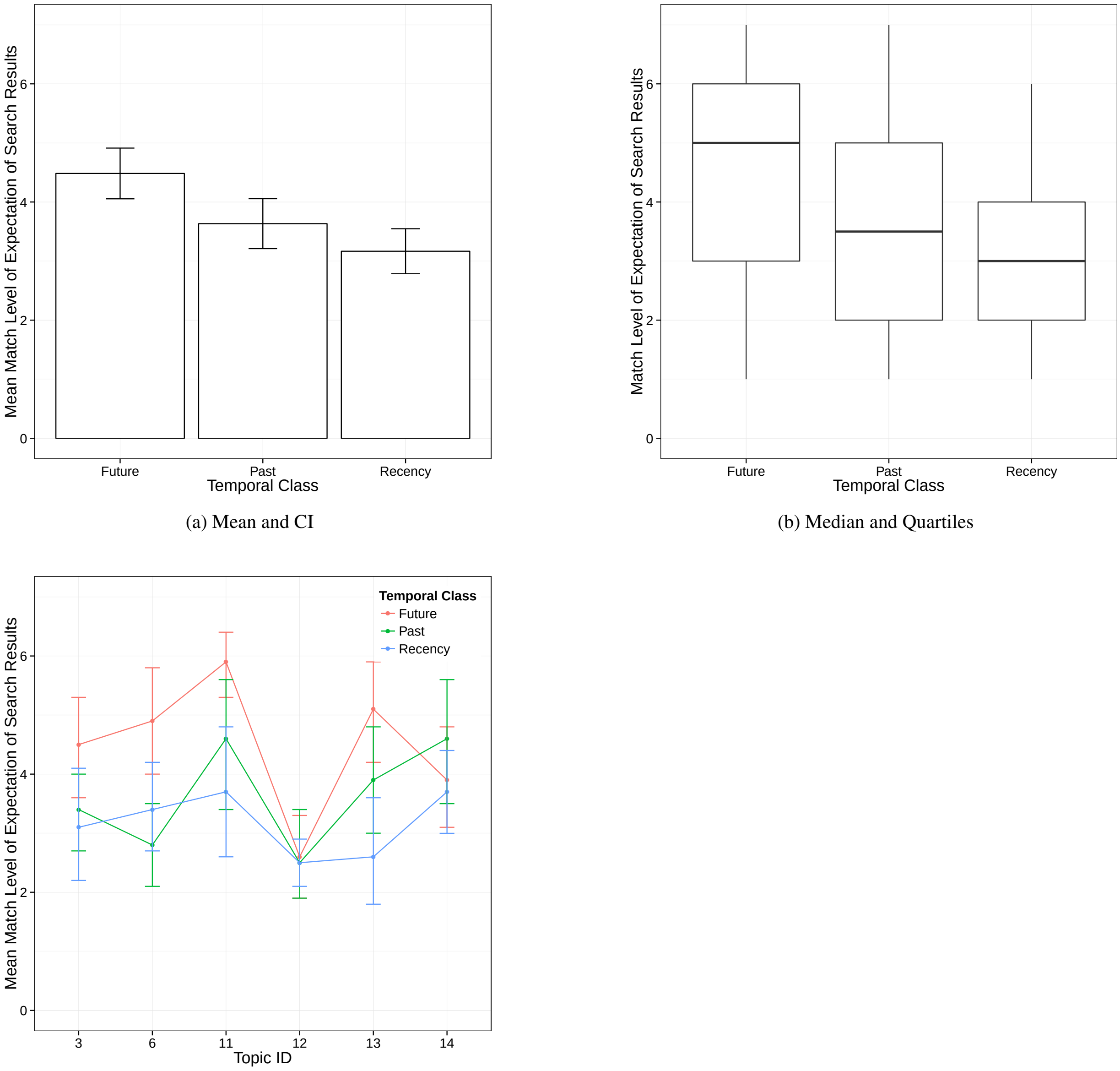


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 10: Ease of subsequent query formulation. (1: High, 7: Low). N = 60.

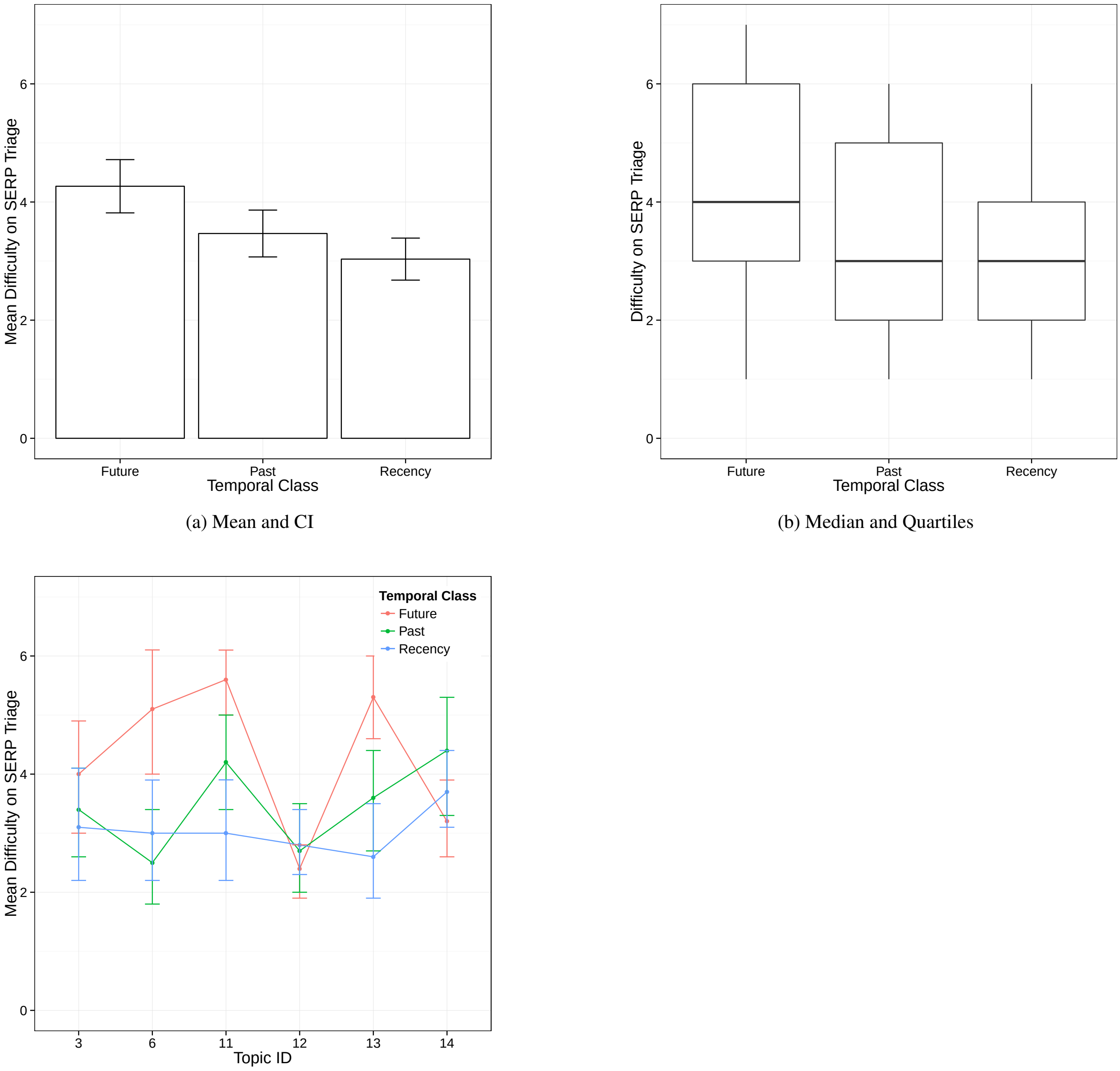


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 11: Match to expected search results. (1: High, 7: Low). $N = 60$.

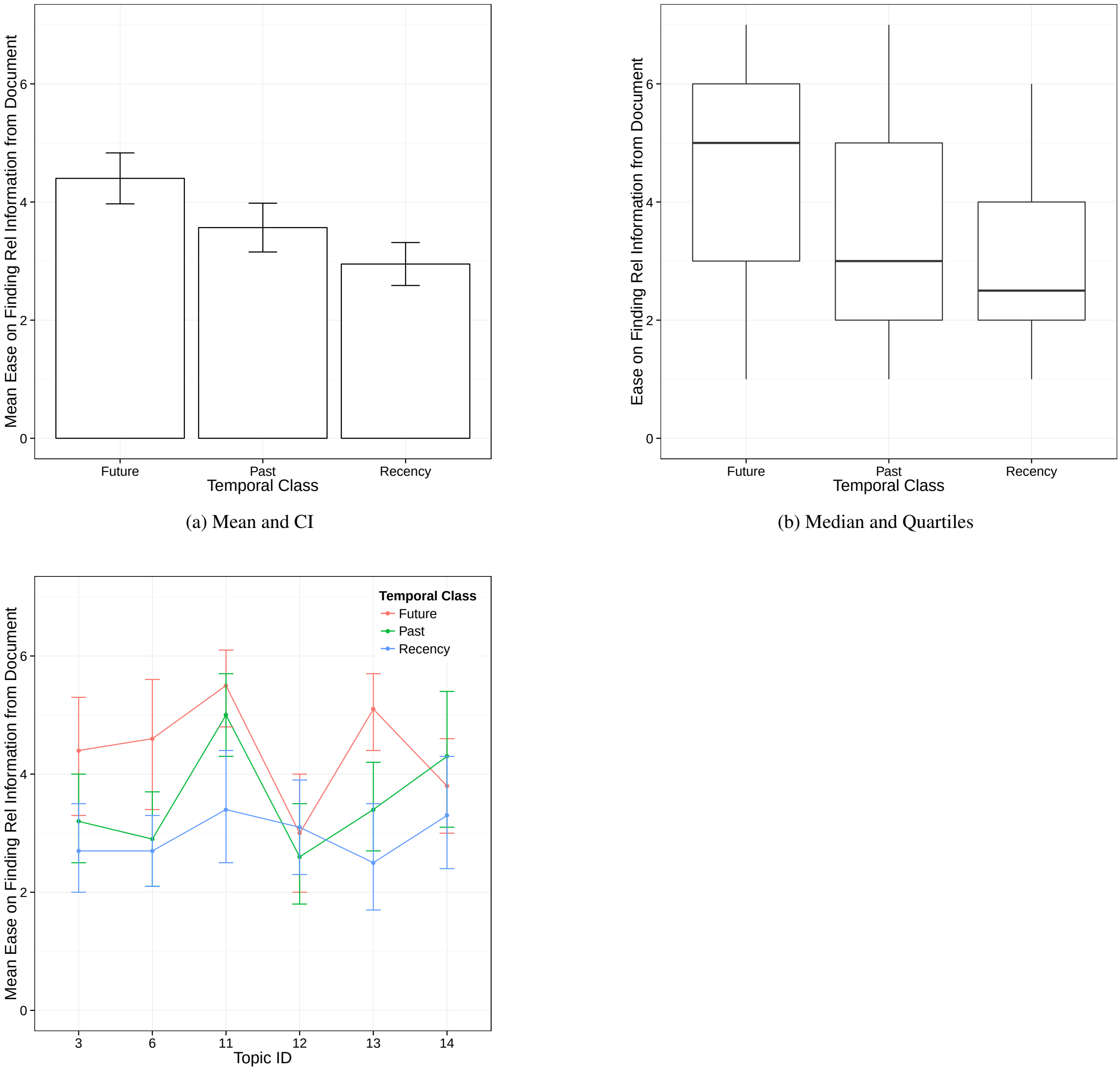


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 12: Ease of SERP triage. (1: High, 7: Low). $N = 60$.

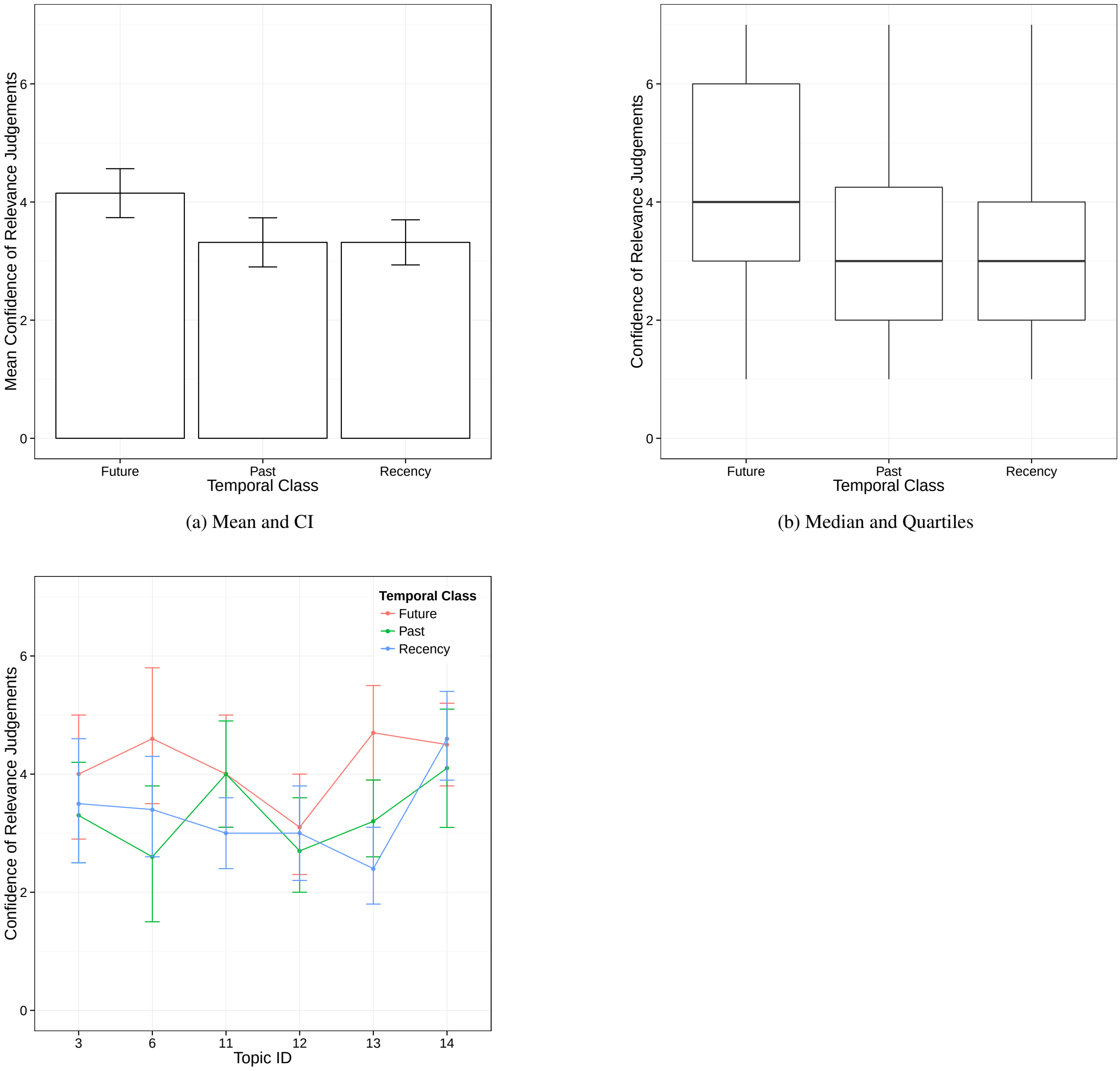


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 13: Ease of finding relevant information from documents. (1: High, 7: Low). $N = 60$.

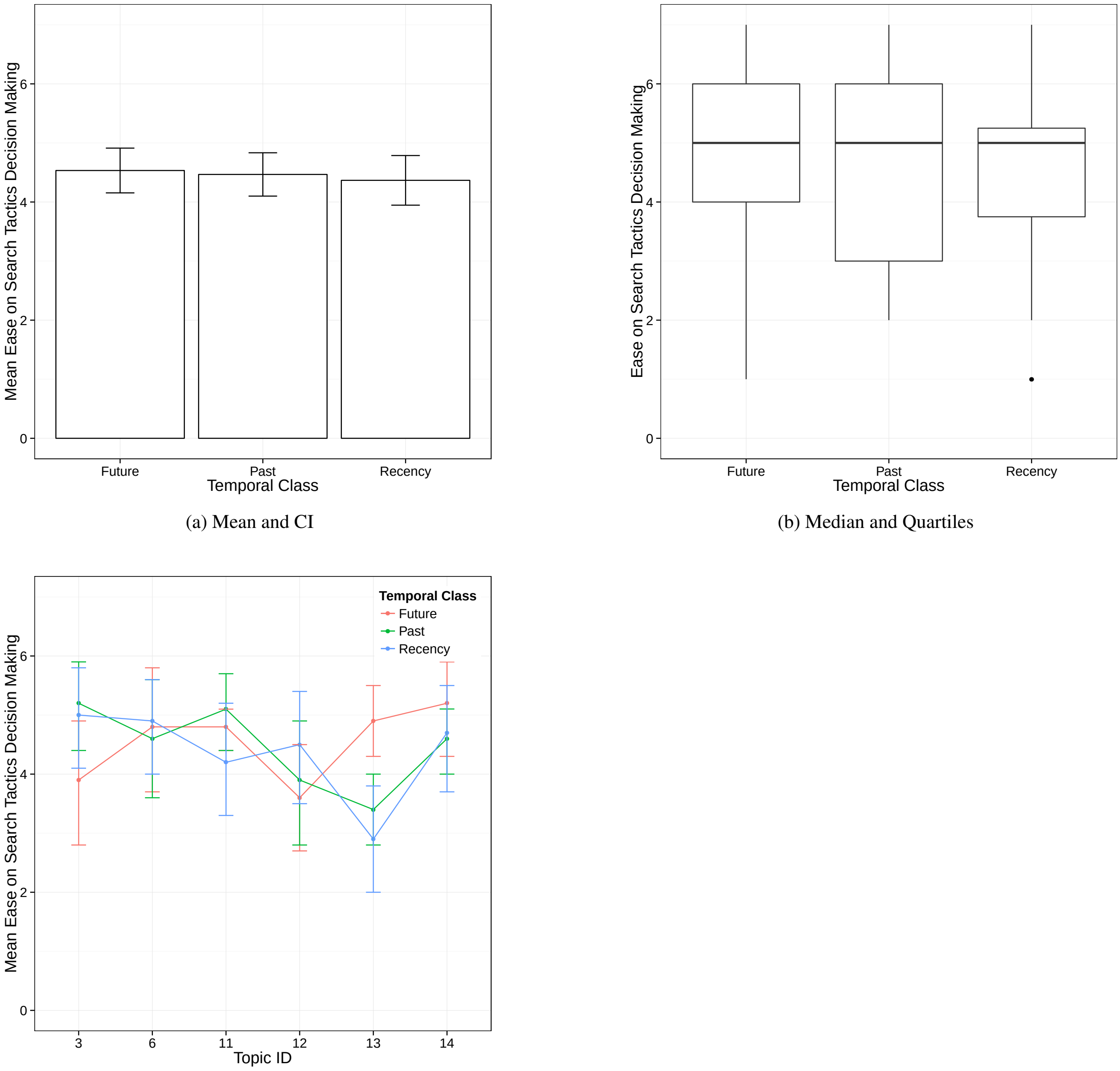


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 14: Confidence of relevance judgements. (1: High, 7: Low). $N = 60$.

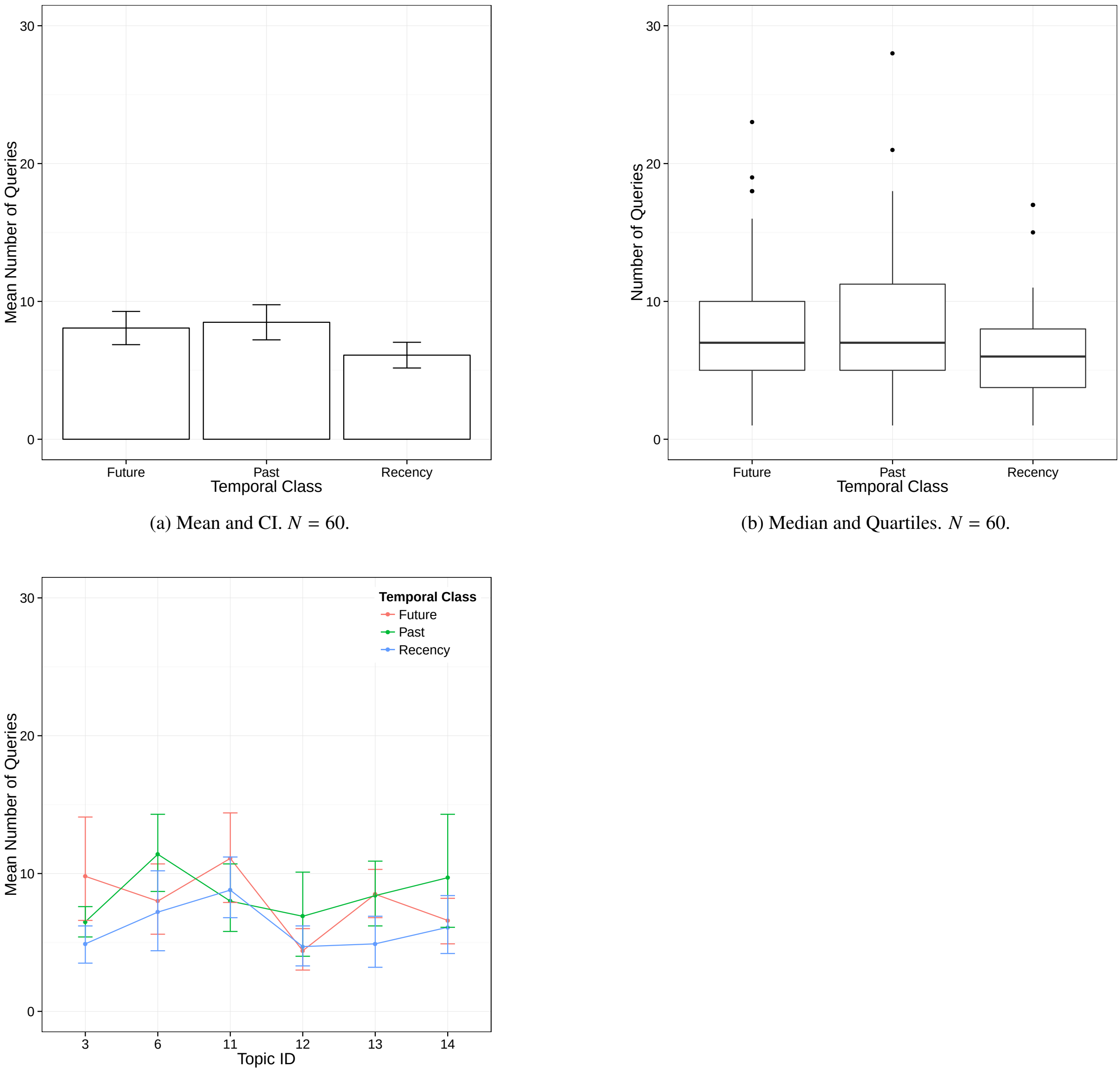


(a) Mean and CI

(b) Median and Quartiles

(c) Interaction with topics

Figure 15: Ease of search strategies decision making. (1: High, 7: Low). $N = 60$.

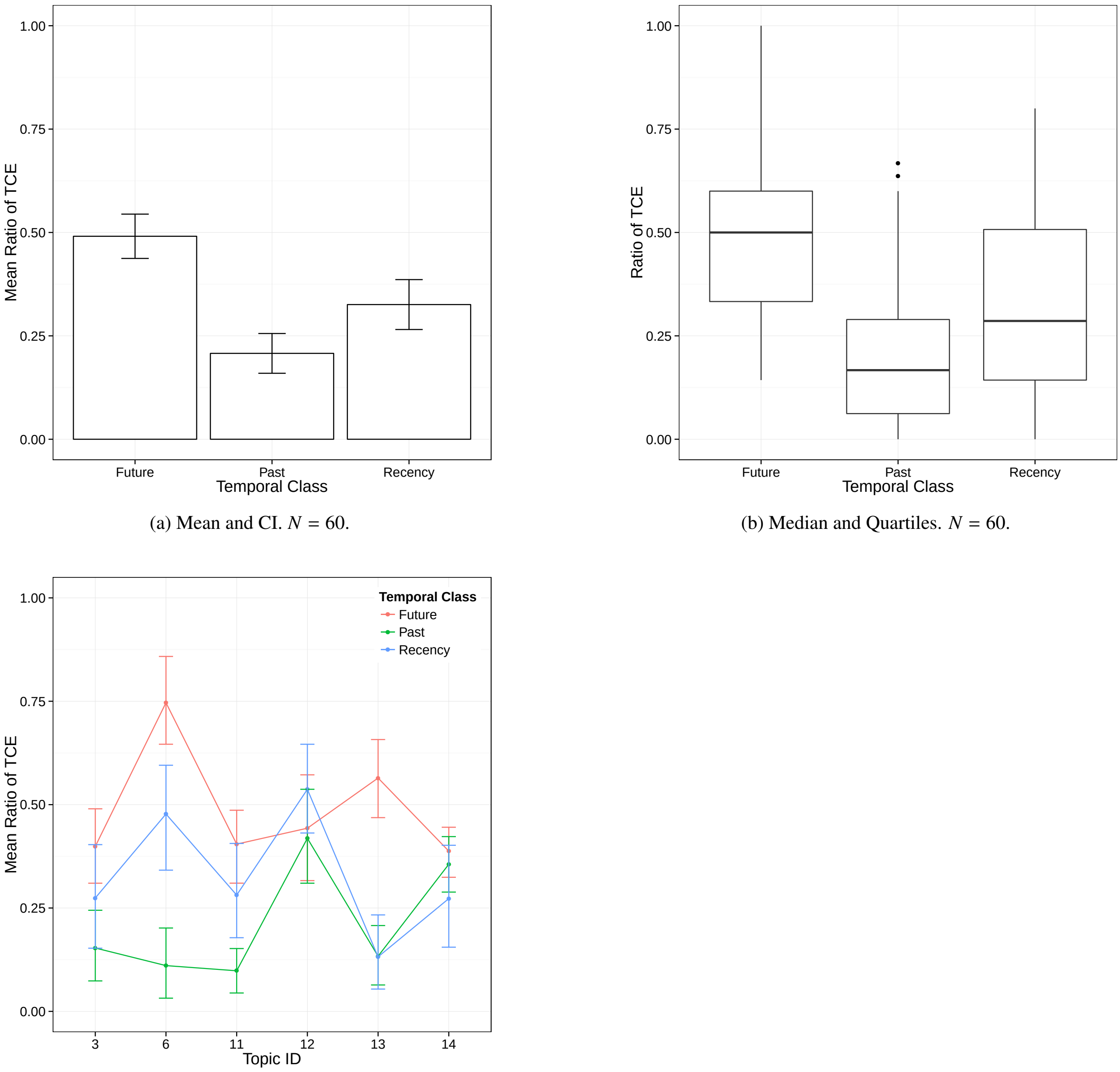


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 16: Number of Queries.

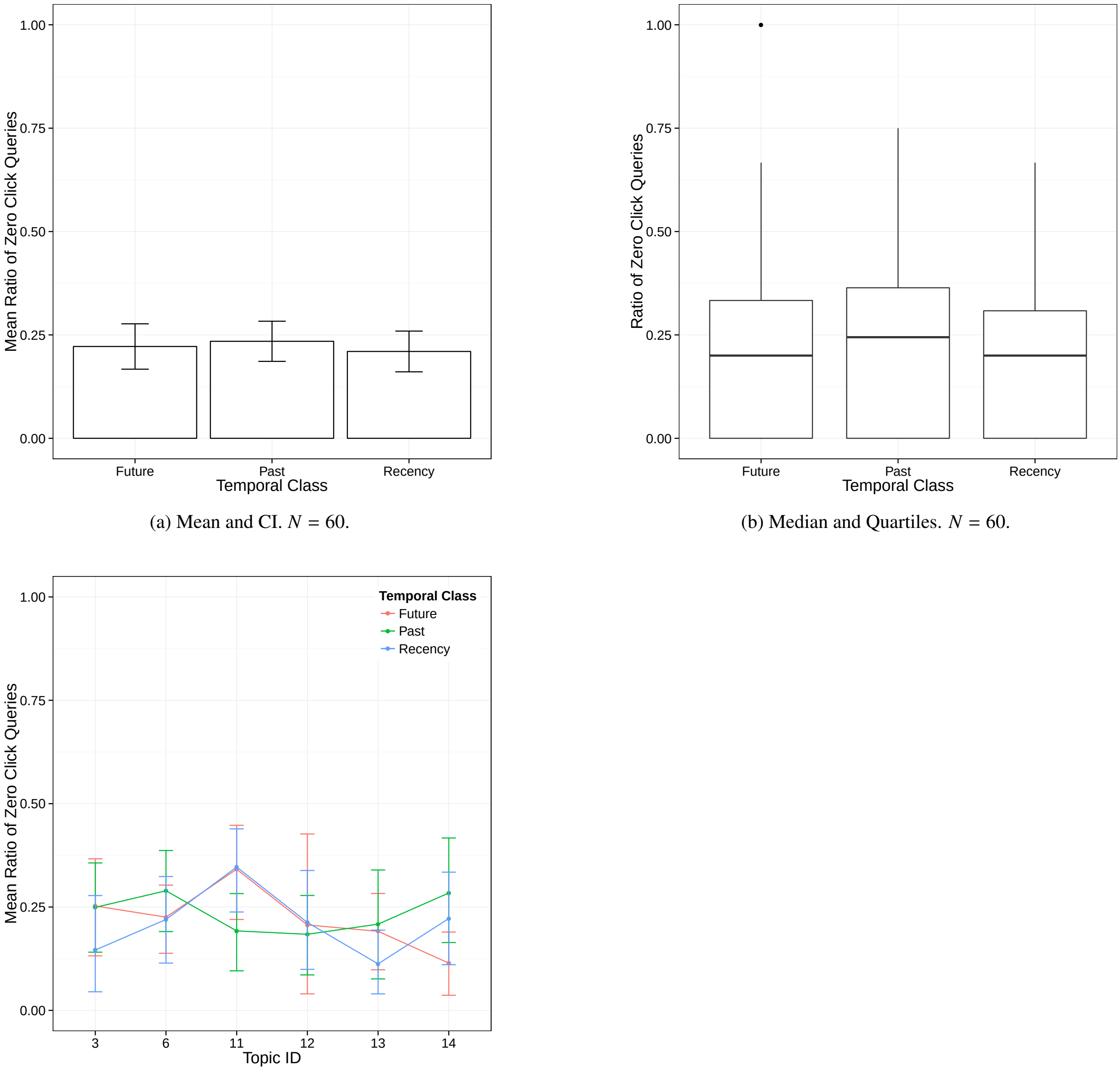


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 17: Ratio of Temporal Control Expressions.

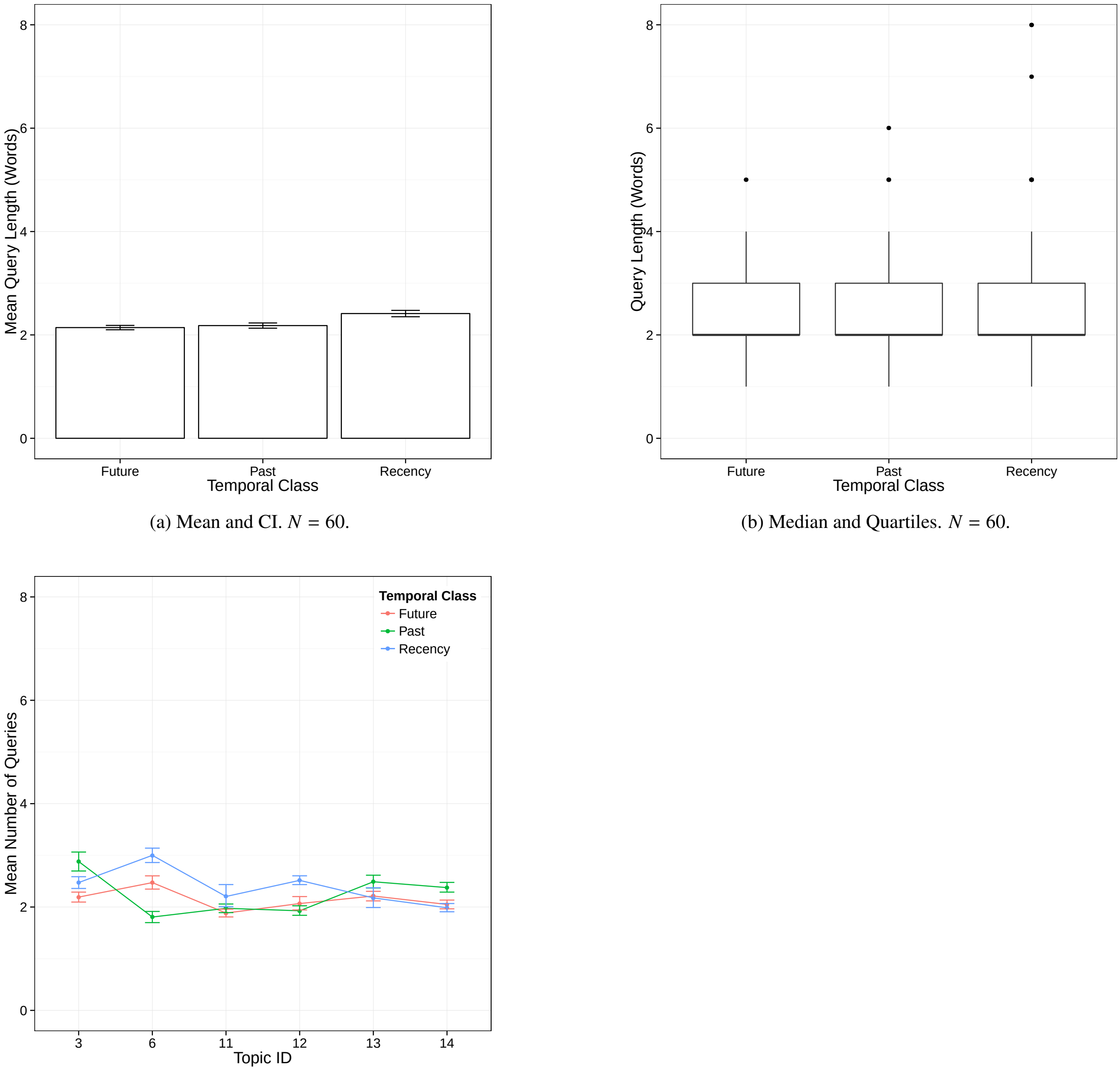


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 18: Ratio of Zero-Click Queries.

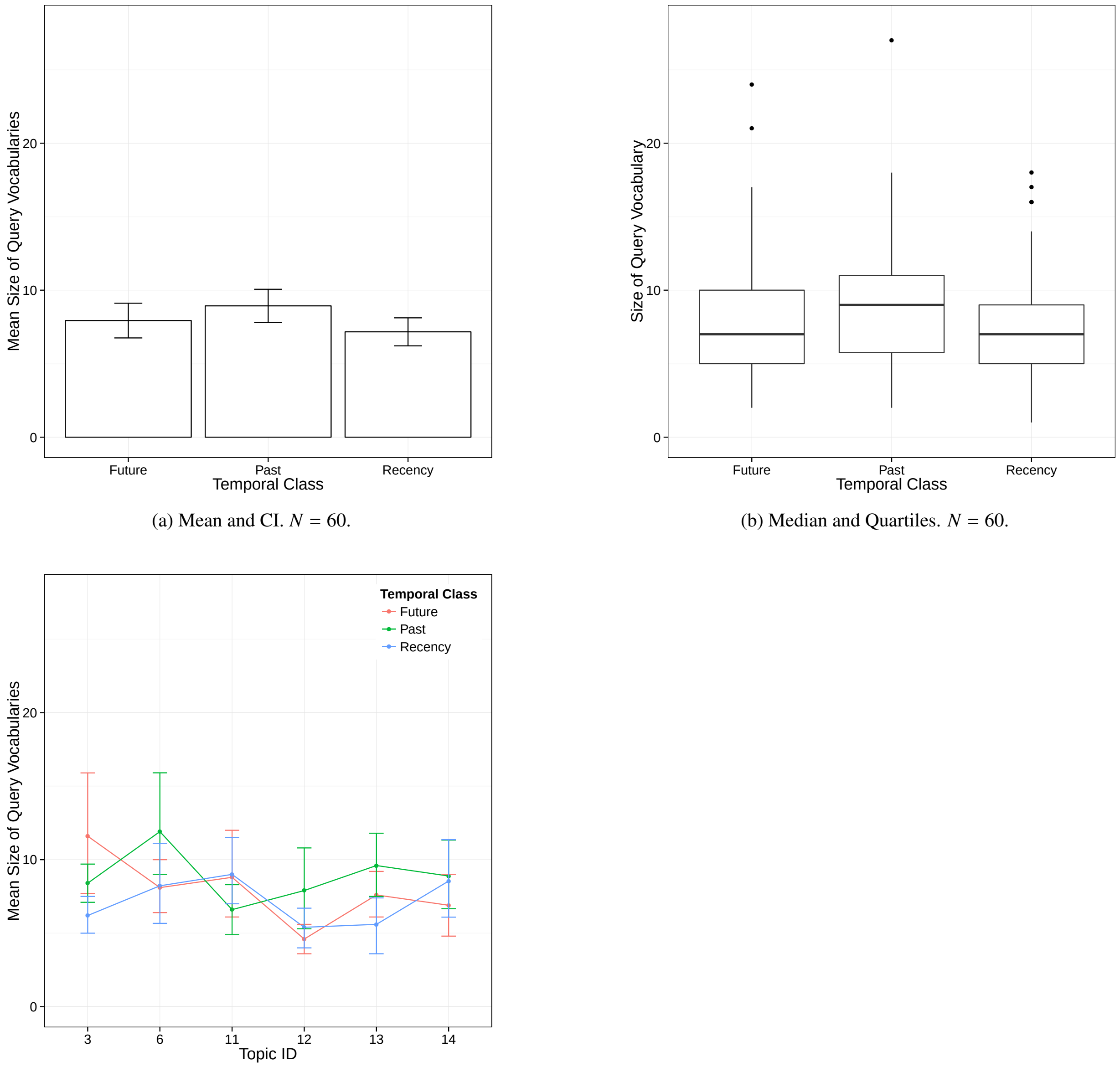


(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 19: Query Length (Terms).



(a) Mean and CI. $N = 60$.

(b) Median and Quartiles. $N = 60$.

(c) Interaction with topics. $N = 10$.

Figure 20: Search Vocabulary Size.