The Impact of Entity Cards on Learning-Oriented Search Tasks

Salimzadeh, S.; Maxwell, D.M.; Hauff, C.

DOI
10.1145/3471158.3472255

Publication date
2021

Document Version
Final published version

Published in
ICTIR 2021 - Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval

The Impact of Entity Cards on Learning-Oriented Search Tasks

Sara Salimzadeh, David Maxwell, Claudia Hauff
Delft University of Technology
Delft, The Netherlands
{s.salimzadeh,d.m.maxwell,c.hauff}@tudelft.nl

ABSTRACT

Entity cards are a common occurrence in today's web *Search Engine Results Pages (SERPs)*. SERPs provide information on a complex information object in a structured manner. Typically, they combine data from several search verticals. They have been shown to: *(i)* increase users' engagement with the SERP; and *(ii)* improve decision making for certain types of searches (such as health searches). In this paper, we investigate whether the benefits of showing entity cards also extend to the *Search as Learning (SAL)* domain. *Do learners learn more when entity cards are present on the SERP?* To answer this question, we designed a series of learning-oriented search tasks (with a minimum search time of 15 minutes), and conducted a crowdsourced *Interactive Information Retrieval (IIR)* user study ($N = 144$) with four interface conditions: *(i)* a control with no entity cards; *(ii)* displaying relevant entity cards; *(iii)* displaying somewhat relevant entity cards; and *(iv)* displaying non-relevant entity cards. Our results show that *(i)* entity cards do not have an effect on participants' learning, but *(ii)* they do significantly impact participants' search behaviours across a range of dimensions (such as the dwell time and search session duration).

CCS CONCEPTS

• **Information systems → Search interfaces**; • **Human-centered computing → User studies**.

KEYWORDS

Entity cards, information cards, search as learning, user study

ACM Reference Format:
Sara Salimzadeh, David Maxwell, Claudia Hauff. 2021. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21), July 11, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3471158.3472255

1 INTRODUCTION

Learning is an important aspect of our lives. Thanks to the prevalence of the *World Wide Web (WWW)*, learning is today often achieved in an informal way, with *web search engines* acting as the information source. Marchionini [30] defined these search episodes as a part of *exploratory search*. Known as *Search as Learning (SAL)* [9] today, SAL is an iterative process where the goal of the learner is to gain knowledge about their specific *information need*, or *learning objective*. A large body of research now encompasses the SAL domain [8, 10, 15, 18, 19, 22, 28, 29, 35, 45, 52, 53, 55], with most of these works concerning the analysis of query logs to attain insights into how those subjected to learning-oriented search tasks behave. Another prominent research direction is how to measure learning in a scalable manner, as only with cheap to compute metrics derived from observable search behaviours at scale will we be able to fulfil the vision of a search interface that adapts to a user's learning needs. In fact, the adaptation of the search system itself—either at the front-end or the back-end—have largely been left unexplored in the SAL domain. Exceptions to this are a small number of works that propose retrieval functions that surface documents suitable for learning [44–46], and works that designed and evaluated search engine result page widgets for learning purposes [5, 7, 41].

Modern web search engines do not provide interfaces that are explicitly designed for learning-oriented searches, though they have changed remarkably over their lifespan. Until a few years ago, the *ten blue links* paradigm dominated the look and feel of SERPs. Contrasted to contemporary SERPs, results are now shown from multiple modalities and search verticals. One prominent result type is the *entity card*. Each entity card (or *information card*) contains a summary of the *entity* (e.g., the name, description, associated images, and related entities)—and thus often helps users find information without the need to interact with other search results.

Although research into the usability and usefulness of entity cards is somewhat limited, several studies [3, 21, 34] have shown that entity cards can enhance the search experience in several ways. Entity cards provide concise content corresponding to the user query by merging information from various information sources, such as images, maps, Wikipedia, or social media [3]. They assist users in accomplishing their task [24, 34], and increase users' engagement with organic search results [3].

Despite these advantages however, entity cards have not been *evaluated* in the SAL context. To this end, *we investigate in this paper whether entity cards are beneficial to users that undertake learning-oriented search tasks* in terms of the achieved learning outcomes. We conduct an *Interactive Information Retrieval (IIR)* study, and design four SERP variants: *(i)* the control condition which provides a standard SERP without an entity card (**No-EC**); *(ii)* a SERP with an entity card relevant to the query (**Good-EC**); *(iii)* a SERP with an entity card that is somewhat relevant to the query (**Fair-EC**); and *(iv)* a SERP with a non-relevant entity card (**Bad-EC**).[1] We implemented these variants on top of the `SearchX` framework [37], and conducted a between-group study with $N = 144$ participants.

This research has been supported by NWO project *SearchX* (639.022.722), NWO project *Aspasia* (015.013.027) and *ICAI AI for Fintech Research.*

ICTIR '21, July 11, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8611-1/21/07.
https://doi.org/10.1145/3471158.3472255

[1] As a concrete example from our query log, for the query `radioactivity`, a good entity card is *radioactive decay*, a fair one is *radionuclide* and a poor one is *time*.

Each participant was assigned to one of the four conditions to assess how different variants of entity cards impact human learning while searching. Concretely, our research questions are as follows.

RQ1 Does the inclusion of entity cards of various quality impact the amount of learning taking place during a learning-oriented search task?

RQ2 Does the inclusion of entity cards of various quality impact users' search behaviours during a learning-oriented search task?

Our main findings can be summarised as follows. *(i)* The inclusion (or not) of an entity card has no discernible impact on participants' learning gains. *(ii)* In contrast, the quality of the entity card with respect to the query has a significant effect on participants' search behaviour across a range of dimensions (such as the dwell time and search session duration).

This paper is the first work to begin to shed light on the influence of entity cards on users' learning gain. Despite observed changes in search behaviour led to no positive changes in learning gain, these findings point to many open issues in terms of entity card design optimised for human learning.

2 RELATED WORK

2.1 Entity Cards

Despite the fact that entity cards are ubiquitous in web search engines today, there is a limited amount of research published about them. Most research focuses on exploring the impact of entity cards on users' search behaviour. Navalpakkam et al. [34] undertook a user study to determine the impact a non-linear SERP layout has on eye and mouse movement behaviours. They were able to show that users spend more time on relevant entity cards than their non-relevant counterparts. When entity cards are relevant, they are beneficial to reduce the task completion time (at least sometimes). This is because the information need can be directly answered by the card's content. Lagun et al. [24] interleaved entity cards within organic search results and carried out a user study in a mobile setting. In line with [34], they found that in the presence of non-relevant entity cards, users gloss over them. Upon not finding an answer, they continue to examine results below, leading to an increased amount of time spent further down the SERP. Bota et al. [3] explored how entity cards affect users' search behaviours and perceived workload. While they went in-depth into generating different types of entity cards (i.e., on-topic and off-topic), results generally showed that participants were more likely to interact (in terms of clicks and mouse hovers) with cards that are relevant to their information need. Furthermore, the presence of entity cards on search result pages increases the users' engagement with organic search result pages. Relevant entity cards also do not significantly increase users' workloads.

Apart from behavioural aspects, prior works have considered how to generate and present entity cards on SERPs. Hasibi et al. [20] examined the content of entity cards, introduced the task of *dynamic entity summarisation*, and proposed an approach to generate query-dependent entity summaries. Their user study found participants to favour dynamic summaries over static ones. Recently, Jimmy et al. [21] have shown in the *health IR* setting that when searching for information about a particular condition, users typically consider the entity card presented first—and then continue to the remainder of the SERP. In addition, they proposed an entity-focused SERP with *multiple* entity cards, and showed that the presence of relevant entity cards regardless of the interface type (i.e., one or multiple cards) leads to a higher probability of making correct decisions. Lastly, the SERP variant with multiple entity cards shown at once allowed participants to make health decisions with significantly less effort as measured by the number of clicks.

2.2 Keyphrase Extraction

In order to determine which entity card(s) to show for a given query (and without access to a large query log for training), we rely on the top retrieved documents for that query. As a first step, we need to extract the *keyphrases* from each of those documents. The task of keyphrase extraction can be defined as *"automatically selecting a small set of phrases that best describe a given free text document"* [2]. Here, we only focus on unsupervised methods, as they are most suitable for our user study due to their domain independence and no required training data. Unsupervised algorithms are divided into two primary groups: *(i) corpus-dependent approaches* [13, 16, 31, 38, 48, 49, 54] which rely on the entire corpus that the current document may be linked to; and *(ii) corpus-independent approaches* [2, 4, 6, 17, 23, 26, 27, 32, 36, 42, 48, 50], which rely on the current document only. Within the corpus-independent category, approaches follow different strategies such as: *(i) graph-based methods* [4, 17, 32, 42, 48, 50] which exploit graph-based language representations to detect keywords; *(ii) embedding-based approaches* [2, 23, 26, 27, 36]; *(iii) statistical-based methods* [6] which rely on statistical features of the text. For our work, we picked one graph-based [42], one embedding-based [2], and one statistical-based approach [6]—each reporting state-of-the-art effectiveness within their category. We picked the best model for our use case in a validation study, as described in the following section.

2.3 Search As Learning (SAL)

SAL is concerned with exploring how search engines can aid users in learning, in both the formal and informal setting. Prior studies [18, 40] explored the impact of domain expertise on learning. Gadiraju et al. [18] observed that participants who are less familiar with a particular topic achieve slightly greater knowledge than users already familiar with the topic (though it is not yet clear whether this finding is mainly an artefact of the topics and the manner of how learning is measured). Roy et al. [40] noticed the difference between experts and non-experts in terms of their learning toward the end of the search task. Previous research also suggests that domain experts employ different search strategies (in terms of queries posed, documents viewed, etc.) to find what they are looking for compared to non-experts [40, 52, 52].

An important aspect of SAL are cheap and easy to measure user behaviours that allow us to estimate the amount of learning taking place—this in turn would allow us to adapt search algorithms and interfaces on-the-fly. Eickhoff et al. [15] studied the flow of evolving expertise within search sessions purely based on users' search behaviours. It was shown that SERP snippets and documents viewed inspire users' queries and reveal information about users' domain knowledge. Other proxies for learning explored include:

eye movement patterns [8]; documents saved and opened [1, 18, 56]; as well as SERP clicks [1, 10]. While most studies focus on lower cognitive levels, Kalyani and Gadiraju [22] studied how search behaviours correlate with information needs at different cognitive levels. They found that users' search interactions with the SERP increased as participants moved towards tasks with higher cognitive levels of complexity.

To *explicitly* measure the learning gain (instead of inferring it from search behaviours), many lab-based user studies assess the knowledge of users before and after the search sessions via vocabulary tests, mind maps and the writing of summaries [10, 28, 29, 35, 45, 53]. Following this setup, we investigate in this work the impact that entity cards have on users' vocabulary learning and search behaviour during a learning-oriented search task.

3 ENTITY CARD IMPLEMENTATION

The present study was undertaken using `SearchX` [37], an open-source, modular retrieval framework that allows one to undertake crowdsourced IIR experiments. Out-of-the-box, `SearchX` provides quality assurances and basic logging functionalities, ensuring that only high-quality participants complete a study—and that the necessary interactions are logged. As entity cards are not yet supported by `SearchX`, we implemented a novel entity card component for it.

Figure 1 demonstrates the user interface that was used by the participants of our study. Users can issue their queries in the query box which also offers query auto-completion provided by the *Bing Autosuggest API*[2]. We present ten search result snippets per page, drawn from the *Bing Search API*. Pagination is provided at the bottom of the SERP. Participants can easily bookmark documents and access them in the *Saved Documents* box. In addition, *Recent Queries* a user issued are also shown in a separate box. The description of the search task appears in the top right corner of the SERP. The timer above the task box helps users gauge the elapsed time. Our entity card is always presented at the position shown in Figure 1 and presents concise information regarding one significant entity within the query and search results. The remainder of this section discusses the structure of our entity cards, and the three variants of entity cards we evaluate—good (**`Good-EC`**), fair (**`Fair-EC`**) and poor/bad (**`Bad-EC`**) quality cards.

3.1 Entity Card Structure

Figure 2 illustrates the structure of our entity cards. Each entity card consists of *up to* four components: *(i)* a set of *images*, which were obtained from the *Bing Image Search API*; *(ii)* the entity's *title*; *(iii)* its' *Wikipedia-based summary*; and *(iv)* multiple *attributes* whose existence and number are dependent on *DBPedia's* open knowledge graph, which contains structured content of various *Wikipedia* projects. In Figure 2, only attributes for the entity *Barack Obama* are shown. This is *not* the consequence of the experimental condition, but instead due to our decision of filtering out rare attributes. More concretely, we processed DBpedia version 2016-10[3], and remove all attributes that occurred in fewer than 20% of attributes of a particular type to avoid distracting users by the presence of unusual attributes (such as *eye colour* for entities of type *Person*).

3.2 Entity Card Rankers

The most important question in the setup of our study is how to determine the ranking of entities: for a given query, once a ranking of entities has been established, we are able to determine the good, fair and bad entities for a query by considering the ranks at which entities are retrieved. For each query a user submits, we concatenate the user query and the top 10 search results snippets. We opted to not include the actual document content in this step as this would require an additional ten HTTP requests, slowing down our SERP's responsiveness significantly (and a slow responsiveness is known to decrease user engagement [25]).

After setting up the context as the concatenation of the query and top ten retrieved document snippets, we then need to retrieve the ranking of the entities through keyword extraction methods. As described in Section 2, several unsupervised approaches for keyphrase extraction exist. Besides the already noted advantages of unsupervised approaches, we also aim to detect keyphrases on-the-fly, thus requiring a fast algorithm (and inference of a large neural network for instance has significant speed constraints). Based on prior works, we selected three keyphrase extraction approaches that are all corpus-independent: `Yake` [6], `RaKUn` [42] and `EmbedRank` [2]. We select these algorithms as: *(i)* they have functioning open-source implementations; *(ii)* they are lightweight, unsupervised algorithms that produce output in a timely manner; and *(iii)* they are robust (i.e., they do not degrade in effectiveness significantly) to changes in collections and domains. For each of these algorithms, we provide our query and document snippets as input, and consider the resulting top 20 ranking of keyphrases. Highly-ranked keyphrases have the highest relevance score with respect to query and document snippets. In order to convert the ranking of these 20 keyphrases into a ranking of 20 entities, we employ the `TagMe` API[4]. This API links each keyphrase to at least one pertinent Wikipedia page. In any cases, `TagMe` returns at least one output. We chose the output of `TagMe` with the highest probability score and fixed this as the entity corresponding to the keyphrase.

For simplicity, we refer to the keyphrase extraction algorithms now as our *entity rankers*, as the procedure to convert the extracted keyphrases to entity rankings (via the `TagMe` API) is the same for all three. Next, we describe the user study we conducted to determine which of the three entity rankers provides us with the best ranking.

3.3 Comparison of Entity Card Rankers

First, we fixed a list of ten topics[5] randomly drawn from the *TREC 2019 Decision (Health Misinformation) Track*[6]. We asked ten volunteers of a computer science lab to provide up to five queries for each of the topics, whose TREC topic description we provided to them. This resulted in between 10 and 27 unique queries per topic, with a median number of 12 unique queries. Each of these queries was submitted to the Bing Search API, from which the top ten result

[2] All Bing APIs used can be found at https://www.microsoft.com/en-us/bing/—all URLs listed in this paper were last accessed on April 27th, 2021.
[3] https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10
[4] https://sobigdata.d4science.org/web/tagme/tagme-help
[5] The ten TREC topic titles are *acupuncture insomnia*, *ear drops remove ear wax*, *honey wound*, *melatonin jet lag*, *magnesium muscle cramps*, *insulin gestational diabetes*, *vaccine common cold*, *antibiotics children pneumonia*, *caffeine asthma*, and *surgery obesity*.
[6] https://trec.nist.gov/data/misinfo2019.html

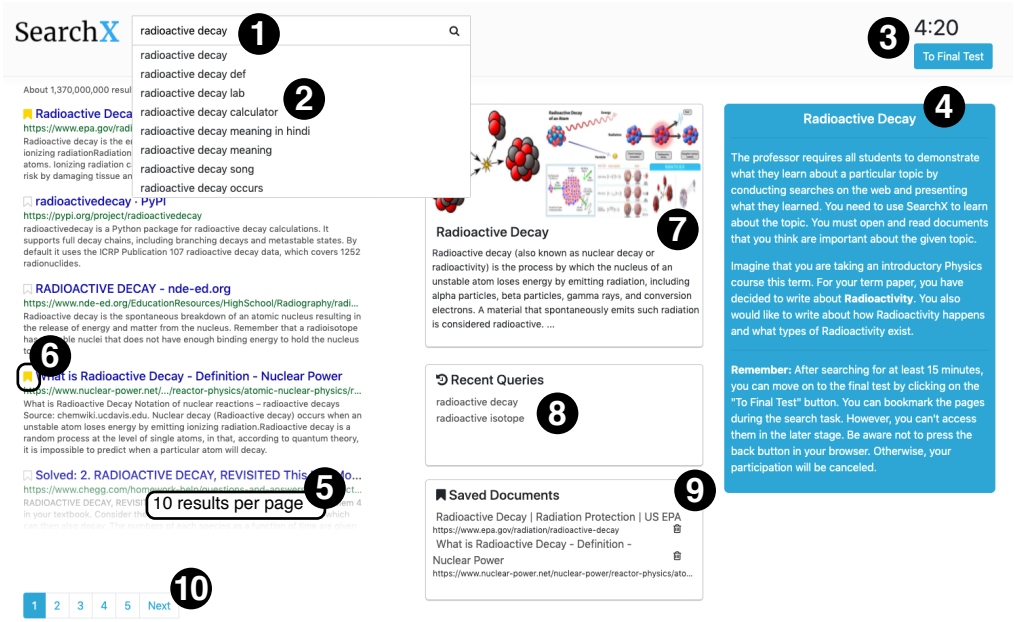

Figure 1: The SearchX interface as used for this study. Included in this screenshot at the 10 superimposed annotation marks: (1) the query box with (2) autocompletion; (3) the timer that indicates the time spent in the search session so far; (4) the task description; (5) the ten search results per page which can be (6) saved to the (9) *Saved Documents* box; (7) the entity card; (8) the list of *Recent Queries*; and finally (10) pagination. Note that this figure shows an entity card from the Good-EC condition.

Figure 2: Demonstration of the three types of entity cards (for the query barack obama). From left to right: *(a)* Good-EC, a high-quality, on-topic entity card; *(b)* Fair-EC, another somewhat relevant entity card, but not the first choice; and *(c)* Bad-EC, an entity card not relevant to the query.

snippets were extracted. Based on this input, we retrieved the entity rankings from Yake, RaKUn and EmbedRank respectively.

As in our actual experiment, we only show one entity card per query as we are most interested in the top-ranked entity retrieved by each algorithm. For this reason, we now focus on the top ranked entity. We discard queries for which at least two of the three algorithms produced the same entity, leaving us with 70 queries—and the three respective top-retrieved entities.

We then randomly selected three topics from our initial list of 10 topics. For each topic, we randomly drew four queries from the collected queries and assigned them to 32 volunteers to judge which of the three top-ranked entity cards from Yake, RaKUn, and EmbedRank respectively are *useful* given the information need (i.e., the TREC topic). They could select multiple options or select *None* to signify that they do not consider any of the presented entity cards to be useful. Overall, EmbedRank's top retrieved entity was selected as the useful entity for 35% of the queries, in contrast to RaKUn's and Yake's 17% and 15% respectively. For 21% of queries, none of the algorithms returned a useful top-ranked entity.

Based on these results, we opted to take EmbedRank forward as our entity ranker throughout the remainder of the experiment discussed in this paper.

3.4 Entity Card Types

Given a query and an entity ranking produced by EmbedRank, we create three types of entity cards.

- **Good-EC** The top ranked entity is selected.
- **Fair-EC** The entity at the rank five is selected.
- **Bad-EC** The entity at rank 20 is selected.

To provide the reader with an impression of the entities retrieved for the three entity card conditions, we refer to Table 1. For example, for query *cellular respiration* (the third query), the entities in descending order of quality are as follows: *cellular respiration* (**Good-EC**); *adenosine triphosphate* (**Fair-EC**); and *art* (**Bad-EC**). For completeness, the **No-EC** condition is our control condition: here, no entity card is presented on the SERP.

Note that besides retrieving different entities for the different conditions, we do not alter the way the entity card looks for each entity type. In particular, what type of attributes and imagery is shown depends on the information available on Wikipedia/DBPedia,

Table 1: An example query chain, as drawn from a search session under the *Glycolysis* topic. Corresponding entity cards that were presented for each query across conditions `Good-EC`, `Fair-EC`, and `Bad-EC` are shown in their respective columns.

Query	`Good-EC`	`Fair-EC`	`Bad-EC`
1. `glycolysis`	glycolysis	nicotinamide adenine dinucleotide	river source
2. `pyruvate pyruvic acid`	lysosomal acid lipase	pyruvic acid	ration
3. `cellular respiration`	cellular respiration	adenosine triphosphate	art
4. `major phases of glycolysis`	glycolysis	sugar	steps

and not the entity card type. As stated previously, Figure 2 provides an example query and the three entity cards generated.

Evaluation of Entity Cards. In order to determine whether our intended quality levels of the entity cards were in fact correct, we manually evaluated the quality of the entity cards the participants received in response to their queries across all conditions.

In our manual labelling effort, we labelled entity cards as *good* when the entity card title exists in the query. The label *fair* is given to entity cards when their title aligns with any concepts related to the query. Lastly, we mark entirely off-topic entity cards as *bad*. The manual annotation of 743 participants' queries (and the corresponding entity card) led to the following results: 87.2% of entity cards shown in the **`Good-EC`** condition were annotated as *good*, 82.4% of entity cards shown in the **`Fair-EC`** condition were labelled as *fair*, and 99.2% of entity cards shown in the **`Bad-EC`** condition were labelled as *bad*. This gives us confidence that the entity cards presented to participants during the experiment fell into line with our expectations.

Additionally, we asked the participants post-test to evaluate the relevance of the entity cards to their queries. 82.8% of the participants in the **`Good-EC`** condition asserted that entity cards are *Mostly/Always Relevant* to the queries, while the proportion is 76.1% for the participants of the **`Fair-EC`** condition, and 7.5% for participants in the **`Bad-EC`** condition.

4 USER STUDY SETUP

We now describe our user study in more detail. We go over the search topics, how to measure learning, and the workflow the participants of the study followed.

4.1 Topics

We employ three of the topics introduced by Moraes et al. [33]'s search as learning study: *Glycolysis*, *Radioactive Decay*, and *Qubits*. Each of these topics comes with a list of 10 vocabulary terms that have been manually curated by the authors.

For example, for the *Glycolysis* topic, vocabulary terms include *krebs cycle*, *electron transport chain*, and *cellular respiration*. These vocabulary terms are terms that: *(i)* were mentioned in a specific video lecture about the topic at least once; and that *(ii)* do not frequently occur outside of this domain-specific context. In their work, Moraes et al. [33] proposed a list of in total 10 topics. We chose the three listed above based on the availability of entity cards: concretely, we received the query log of Moraes et al. [33], submitted all the queries for each topic to the Bing Search API, and ran `EmbedRank` to retrieve the respective entity rankings. We then selected the three topics with the largest number of relevant entities. Specifically, the topic *Radioactive Decay* (with a rate of 6.02 entities per query) has the greatest number of entities per query, followed by topics *Glycolysis* and *Qubits* with 5.4 and 4.7 entities per query, respectively.

4.2 Learning Gain

We measure our participants' learning gain by measuring their difference in knowledge in a pre-test (conducted right before the search session) and a post-test (conducted right after the search session) in line with [18, 33, 40, 44, 45, 55]. As in [33, 40], we employ the slightly modified *Vocabulary Knowledge Scale (VKS)* test [43, 45, 51], which demonstrate the incremental stage of stages of word learning [12]. For every vocabulary term, our participants are asked about their knowledge across four levels:

(1) I don't remember having seen this term/phrase before.
(2) I have seen this term/phrase before, but I don't think I know what it means.
(3) I have seen this term/phrase before, and I think it means ___ .
(4) I know this term/phrase. It means ___.

Note that for levels *(3)* and *(4)*, we require participants to write their definition of the term. The difference between the two is in the certainty of the participants' knowledge: in level *(3)* the uncertainty is high; with level *(4)*, participants are certain about their knowledge.

Again, in line with prior works [11, 33, 45, 46], we employ *Realised Potential Learning (RPL)* to measure the learning gain which normalises *Absolute Learning Gain (ALG)* by the *Maximum possible Learning Gain (MLG)*. ALG is an aggregated difference in knowledge level before and after the search session across all vocabulary terms—with the added proviso that *knowledge cannot degrade over the time of the search session (between the pre- and post-tests)*.

Here, $vks^{pre}(v_i)$ and $vks^{post}(v_i)$ indicate the scores assigned to vocabulary term v_i in the pre- and post-test, respectively. We set the *vks* score to 0 knowledge levels *(1)* or *(2)*. We also assign the score of 1 for both knowledge levels *(3)* and *(4)*, which is in line with the binary setup employed in [33]. RPL is computed as follows.

$$ALG = \frac{1}{n}\sum_{i=1}^{n} \mathbf{max}(0, vks^{post}(v_i) - vks^{pre}(v_i))$$

$$MLG = \frac{1}{n}\sum_{i=1}^{n} maxScore - vks^{pre}(v_i)$$

$$RPL = \frac{ALG}{MLG}$$

4.3 Workflow

When a participant enters the study, the online learning experiences questionnaire is presented, consisting of seven questions. These

Table 2: Example annotations of participants' definitions of vocabulary terms for the topic *Glycolysis*.

Vocabulary term **pyruvate**	
Correct	*A compound that is produced via glycolysis and is related to pyruvic acid.*
Partially correct	*It is a product of glycol is it can help with fat burning.*
Incorrect	*A molecular unit of sugar.*
Vocabulary term **krebs cycle**	
Correct	*The Krebs cycle is also called the citric acid cycle. It's a series of chemical reactions which require oxygen and get energy from food. It can only be aerobic. It produces ATP and also other compounds used by the electron transport chain.*
Partially correct	*Also known as they citric acid cycle.*
Incorrect	*A cellular process that helps an organism live.*

questions are inspired by Rovira et al. [39], and focus on online learning experiences with the goal to prime participants for the upcoming task. Then, we present the pre-test for the three topics to each participant. For each topic (in addition to the 10 vocabulary knowledge questions), we include three more general questions to probe the participants.

- *How much do you know about this topic?*
- *How interested are you to learn more about this topic?*
- *How difficult do you think it will be to search for information about this topic?*

Thus, in the pre-test, each participant answers a total of $7 + 3 \times 13 = 46$ questions. Subsequently, the participants move on to the search phase where they are randomly assigned to one of our four experimental conditions. For the topic, the one with the least amount of prior knowledge (computed from the answers to their pre-test questions) is selected. Before starting the search task, a tutorial is shown to the participant providing information about how to interact with different interface components. The search task presented to the participants is the following (the underlined phrases are specific to each search topic).

> Imagine that you are taking an introductory Physics course this term. For your term paper, you have decided to write about Radioactivity. You also would like to write about how Radioactivity happens and what types of Radioactivity exist.

The minimum search time was fixed to fifteen minutes to provide sufficient time to search and learn while alleviating fatigue. We relied on the Bing Search API as our search backend, and filtered out any search results originating from Wikipedia or any of its mirrored pages. As we aim for our participants to search in order to learn, we removed this source of information to avoid participants spending their search time reading a single Wikipedia document.

During the search session, participants can search, view, and bookmark documents. We disable copy and paste options and limit the tab changes to a maximum of two to avoid participants searching the web to answer our questions. At three browser tab changes, a participant is disqualified from the study.

The experiment ends with a post-test, which contains the same vocabulary knowledge test as the pre-test this time though only focused on the one topic assigned to the participant. Additionally, participants are tasked with writing a summary with a minimum of 100 words, and the term paper's outline as indicated in the search task description. Lastly, we include 10 questions regarding the entity cards, their experience working with our search system, their perceived learning, and perceived search success.

4.4 Study Participants

We conducted our user study on the *Prolific* Academic Platform[7]. We required our $N = 144$ participants to: *(i)* have at least 15 accepted Prolific task submissions; *(ii)* be native English speakers (limiting participants to be from only the United Kingdom); and *(iii)* have a minimum approval rate of 85%. The study took approximately 40 minutes to complete. We paid our participants GBP£6.43 per hour for the experiment. Among our participants, 64.5% were female, and 35.5% were male. We report a mean age of 32.4 (minimum 18 years, maximum 74 years). Due to the nature of crowdsourced studies, we continued to add more participants to our Prolific task until we reached 36 participants for each condition.

4.5 Vocabulary Knowledge Assessment

In total, participants provided us with 394 concept definitions (across both the pre-test definitions written for the topic that was eventually selected for the respective participant, and the post-test definitions) when self-assessing their knowledge as level *(3)* or *(4)* (see §4.2). We manually evaluated all provided concept definitions and labelled each one as either *correct*, *partially correct*, or *incorrect*. Examples of definitions and the labels we assigned to them are provided in Table 2. More formally, we employed the following criteria to judge each definition provided by a participant.

- ***(2)*** **Correct** If a participant explains one related concept without any errors, their definition was assigned the highest score. Furthermore, the highest score was given to the participant's definition which explains multiple related concepts, while leeway was given if an error was in *one* of the concepts.
- ***(1)*** **Partially Correct** The participant's definition describing one related concept with any errors was given a score of 1. This score also applied to participants whose definition provided a correct synonym for the term. For example, the *Krebs cycle* is also known as the *citric acid cycle*.
- ***(0)*** **Incorrect** Definitions that are either entirely incorrect or trivial (e.g., *'beta-minus decay is a kind of decay'*).

As a first step in our annotation, we randomly sampled 50 of the vocabulary term definitions (13% of the total available terms). The authors then annotated them independently according to the above correctness criteria. *Inter-annotator agreement*, computed as *Cohen's kappa*, is 0.83. With this high rate of agreement, we

[7] https://www.prolific.co/

then split the remaining definitions and annotated them independently. In contrast to prior works [5, 33, 41], we did *not* rely on self-assessments of knowledge. Instead, we instead manually verified to what extent these self-assessments were correct. We found that for knowledge level *(3)* (see §4.2): 31% of the provided term definitions were identified as being correct; 38% were partially correct; with the remaining 31% incorrect. From the vocabulary terms self-assessed as knowledge level *(4)*: 48% were correct; 25% were partially correct; with the remaining 27% incorrect.

5 RESULTS

To address our two research questions as outlined in Section 1, measures were analysed by using a two-way ANOVA. These were conducted considering both the conditions and topics as factors; main effects were examined where $\alpha = 0.05$. For post-hoc analysis, the TukeyHSD pairwise test was used. For results in all tables, ± values denote the standard deviation.

5.1 Entity Cards and Learning (RQ1)

RQ1 asks to what extent entity cards of varying quality impact the amount of learning taking place. Table 3, row **X**, presents the RPL across the four experimental conditions. To complement this, rows **XI**-**XII** also report the RPL achieved over each of the three topics. As these measures only provide a high-level overview of the learning gain, Figure 3 plots the *distribution* of learning gain across participants for each of the four conditions and each of the three topics respectively.

We first focus on the learning gain across four experimental conditions. For our control condition (**No-EC**), the average *RPL* is 0.18, which means that participants gain on average 18% of the knowledge they could have gained at best. When comparing **No-EC** with the other conditions, we do not observe significant differences in learning gain. We also observed that the learning gain for the **Bad-EC** is the lowest compared to other conditions; lower than even **No-EC**. Additionally, Figure 3(a) shows that for both **Good-EC** and **Fair-EC**, the variability in RPL scores across participants is larger than for the other two conditions. Although there is no significant difference across conditions, these findings suggest that (at least partially) relevant entity cards may improve learning gain, but only marginally. In contrast, poor entity cards could negatively impact on learning—with the suggestion that a bad entity card may distract participants from learning within complex topics. Rows **XI**-**XIII** of Table 3 also report the RPL across each condition, splitting it up by each of the three topics trialled.

Table 4 presents a summary of the RPL (amongst behavioural measures) from a per-topic perspective. We can see on row **XI** a large variation in the mean RPL attained over the three topics: 0.12±0.16 for *Radioactive Decay*; 0.16±0.16 for *Qubits*; and 0.30±0.25 for *Glycolysis*. Indeed, *Glycolysis* was found to have a significantly higher level of RPL than either *Radioactive Decay* or *Qubits*. This meant that *Glycolysis* was considered the easier topic on average, with *Radioactive Decay* appearing to be the more complex. The differences between topics are also visible in Figure 3(b): *Glycolysis* has the highest median with the greatest variability in learning gain. What had an impact on knowledge gain is the distribution of topics among participants.

Given these observations, we find the presence of entity cards (no matter their quality with respect to the issued queries) to not lead to higher learning gains (thus addressing **RQ1**). However, comparing the RPL across our conditions, we can see that bad/poor entity cards (**Bad-EC**) have detrimental impact on an individual's learning. Results show that topic difficulty does play a major role, with significant differences found between the mean performance of participants when the three topics are considered separately.

5.2 Entity Cards and Search Behaviours (RQ2)

We return to Table 3 for insights into the search behaviours exhibited by participants over each of the four conditions trialled, as shown on rows **I**-**IX**.

We first examine the recorded search session duration reported on row **II** of Table 3. With results presented in minutes and seconds, we observe that for both **Good-EC** (16:43±4:22) and **Fair-EC** (16:15±2:04), the mean session time is approximately one minute longer than for **No-EC** (15:44±0:45). We also note that participants spent significantly longer using interface **Bad-EC**(16:58±3:06) (with a higher variance) than on **No-EC**. Together, these findings suggest a slightly higher engagement with the task and interface when entity cards were present on our search interface, regardless of the quality of the cards provided. Looking deeper, we find that this pattern was repeated when considering average document dwell times, with the same patterns once again being observed (see row **VIII**). Examining the interactions with the entity cards themselves, we note that the mean number of hovers over the entity cards was found to be approximately 15 for all three conditions containing them (Table 3, row **VI**). No significant differences were observed. A similar number of documents were examined across all four conditions, once again without any observed significant differences (Table 3, row **IX**). Here, **No-EC** has the highest number of viewed documents on average, at 9.75 ± 3.98. This intuitively makes sense: no entity cards means the only source that participants could gain information was to go and read the linked documents. To complement this, we observe a trend: a decrease in the number of unique documents viewed as the quality of the presented entity cards increases. Here, we hypothesise that as entity card quality increased, participants had a greater likelihood of being able to satisfy their information need on the SERP without having to resort to clicking links.

In terms of the number of queries issued, participants in the **No-EC** condition on average issued the greatest number of queries on average (6.03±2.89), though this was not significantly so (Table 3, row **III**). We observe a consistent increasing trend in the number of queries issued as the entity card quality drops (or the entity card is absent), starting from **Good-EC** (4.89 ± 2.23) and ending at **No-EC** (6.03 ± 2.89). When receiving (partially) relevant information from the entity cards, we speculate that participants were able to obtain important information for their information need from them. Correspondingly, in terms of the average time between queries, participants in the **Good-EC** condition recorded the highest time (221.08 seconds) with that time dropping as we move along the conditions towards poor entity cards (Table 3, row **V**).

Within the post-test, we also asked participants how much they paid attention to entity cards to ensure the impact of entity cards on their behaviour. A total of 90% of participants of **Good-EC** stated

Table 3: Mean (± standard deviations) of RPL and search behaviour measures across all participants over each of the four experimental conditions. A † indicates two-way Anova significance, while ^{G,F,B,N} reveals post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) increases vs. Good-EC, Fair-EC, Bad-EC, and No-EC conditions, respectively.

			Good-EC	Fair-EC	Bad-EC	No-EC
Behavioural	I	Number of participants	36	36	36	36
	II	Search session duration (mm:ss)†	16:43 (± 4:22)	16:15 (± 2:04)	16:58 (± 3:06)^{N}	15:44 (± 0:45)^{B}
	III	Number of queries	4.89(±2.23)	5.18(±2.41)	5.62(±2.69)	6.03(±2.89)
	IV	Fraction of entity card terms within the subsequent query†	0.69(±0.30)^{FB}	0.34(±0.29)^{GB}	0.02(±0.05)^{GF}	-
	V	Average time between queries (secs)	221.08(±187.32)	197.56(±162.22)	187.99(±113.41)	164.5(±80.87)
	VI	Number of hovers over entity cards	14.94(±9.01)	16.62(±9.73)	13.94(±6.35)	-
	VII	Average time between documents (secs)	79.41(±82.75)	63.96(±43.36)	73.57(±60.10)	67.43(±49.68)
	VIII	Average document dwell time (secs)†	126(±69.6)	117(±65.4)	151.2(±93.6)^{N}	113.4(±48.6)^{B}
	IX	Number of unique documents viewed	8.25(±4.22)	8.56(±3.8)	8.68(±3.24)	9.75(±3.98)
Learning	X	RPL (over all topics)	0.19(±0.21)	0.22(±0.22)	0.17(±0.22)	0.18(±0.17)
	XI	RPL for topic *Radioactive Decay*	0.16(±0.19)	0.12(±0.17)	0.09(±0.14)	0.13(±0.16)
	XII	RPL for topic *Qubits*	0.15(±0.19)	0.22(±0.23)	0.11(±0.12)	0.15(±0.11)
	XIII	RPL for topic *Glycolysis*	0.26(±0.26)	0.35(±0.20)	0.34(±0.31)	0.25(±0.22)

Table 4: Summary statistics for the three topics used in our study (± standard deviations). A † indicates two-way Anova significance, while ^{R,Q,G} indicate post-hoc significance (TukeyHSD pairwise test, $p < 0.05$) vs. *Radioactive Decay*, *Qubits*, and *Glycolysis*, respectively.

			Radioactive Decay	*Qubits*	*Glycolysis*
#	I	Total participants	48	48	48
	II	# in conditions **Good-EC**, **Fair-EC**, **Bad-EC**, and **No-EC**	12	12	12
Behavioural	III	Average number of queries	5.0(±2.40)	5.90(±2.44)	5.39(±2.93)
	IV	Median number of queries	4	5.5	4
	V	Average time between queries (sec)†	210.56(±102.47)^{Q}	145.48(±55.35)^{RG}	224.99(±215.84)^{Q}
	VI	Median time between queries (sec)	183.53	142.07	148.54
	VII	Average number of bookmarks	2.88(±2.83)	3.15(±3.67)	3.25(±2.97)
	VIII	Median number of bookmarks	2.5	2	3
	IX	Average number of unique documents viewed†	7.90(±3.66)^{G}	8.83(±3.72)	9.80(±3.99)^{R}
	X	Median number of unique documents viewed	7	8.5	9
RPL	XI	RPL†	0.12(±0.16)^{G}	0.16(±0.16)^{G}	0.30(±0.25)^{RQ}
	XII	Median RPL	0.10	0.11	0.30

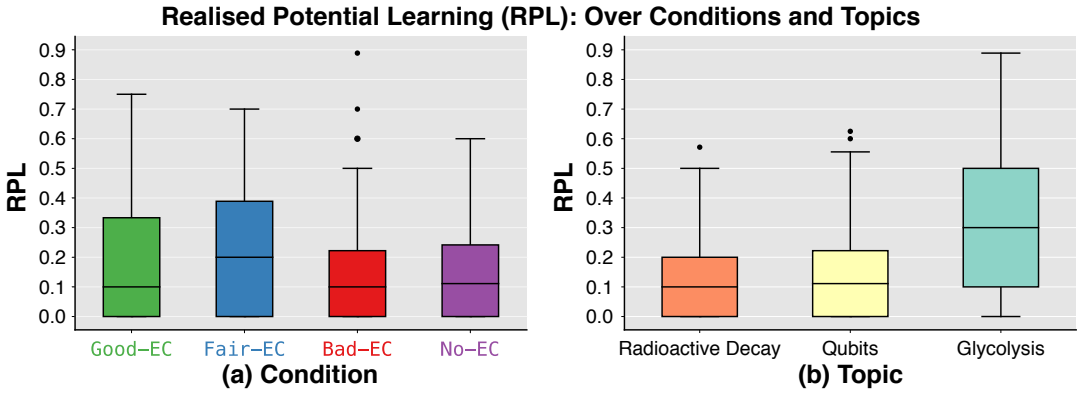

Figure 3: RPL, considered over both the four experimental conditions (a), and (b) the three topics trialled.

Table 5: Source of terms for query reformulations. A † indicates two-way Anova significance, while ^{G,F,B,N} indicate post-hoc significance (TukeyHSD pairwise test, p < 0.05) vs. Good-EC, Fair-EC, Bad-EC, and No-EC conditions, respectively.

		Good-EC	Fair-EC	Bad-EC	No-EC
I	Fraction of query terms from prior snippets†	0.29(±0.25)	0.18(±0.19)^{N}	0.25(±0.23)	0.34(±0.17)^{F}
II	Fraction of query terms from prior documents†	0.50(±0.33)^{N}	0.35(±0.33)^{N}	0.41(±0.36)^{N}	0.58(±0.21)^{GFB}
III	Fraction of query terms from prior entity card titles†	0.24(±0.17)^{FB}	0.12(±0.14)^{GB}	0.01(±0.03)^{GF}	-
IV	Fraction of query terms from prior entity card summaries†	0.53(±0.24)^{FB}	0.39(±0.31)^{GB}	0^{GF}	-

they examined entity cards regularly. 90.3% and 82.5% of **Fair-EC** and **Bad-EC** participants respectively also self-reported paying attention to them regularly.

In order to examine whether entity cards influenced the terms that appeared in subsequent queries in the search sessions, we examined the fraction of entity card terms occurring within the queries issued by the participants (Table 3, row **IV**). We found significant differences between the three conditions that presented entity cards, with a clear, increasing trend from **Bad-EC** (0.02 ± 0.05), through to **Fair-EC** (0.34 ± 0.20), up to **Good-EC** (0.69 ± 0.30). Significant differences existed between all conditions, suggesting that participants were able to judge the quality of the entity cards and employed them when formulating their queries (e.g., through the learning of terms to then issue to the search engine). Spurred by this finding, we examined this phenomenon in more detail.

In terms of query reformulations, we observed entity cards to have a considerable impact. In Table 5, we examined the *source* of participant's query terms. We report the following statistics.

- **Fraction of query terms from prior snippets** Here, we consider *previously observed* snippets as a potential source for query terms.
- **Fraction of query terms from prior documents** For each query, we consider all *previously viewed* documents, and compute the fraction of query terms that appeared in at least one of them.
- **Fraction of query terms from prior entity card titles** Here, instead of considering previously viewed documents, we consider only entity card titles.
- **Fraction of query terms from prior entity card summaries**. Finally, we consider the entity card summary text, instead of the title.

We acknowledge that this can only be considered an approximation, as we do not know whether for instance a term present in a viewed document was even read by a participant (this likely requires eye-tracking hardware and analysis, as per [14]). However, significant differences were found across all four additional measures. If we first consider the measures corresponding to the entity cards, it is unsurprising to note that the fraction of query terms from both entity card titles and summaries were significantly higher for **Good-EC** than **Bad-EC**, with **Fair-EC** once again, on average, landing in between the two extremes. From the **Good-EC** summaries, for example, the fraction of terms in participant queries jumped from 0.53 ± 0.24 down to a flat 0 for **Bad-EC**—this acts as a sanity check, confirming that **Bad-EC** entity cards always yielded entity cards that did not correspond to the given query.

Taking this analysis further, we also extracted *query chains* from our gathered interaction logs to examine what terms were actually used. Table 1 presents an example query chain drawn from a participant's interaction log over the *Radioactive Decay* topic. Along the first column are the queries issued by the participant, with the associated entity card titles shown for each of the three conditions. We can see that the terms that appear in the issued queries correspond closely to those in **Good-EC**, with the third query's terms matching those of the suggested **Good-EC** exactly.

These results show that there is at least some interaction effect in the search and learning process, where entity cards are priming and providing participants with query terms to assist in their query formulation patterns. Further work is required to investigate this.

6 CONCLUSIONS

In this paper, we examined to what extent entity cards impact users' learning gains (**RQ1**) and search behaviours (**RQ2**) for learning-oriented search tasks.

To answer our two research questions, we conducted a crowdsourced user study, where $N = 144$ participants were assigned to one of four conditions. The conditions controlled whether entity cards were present on the SERP, and if present, dictated whether they were *good* (relevant to the query), *fair* (contained a degree of relevant information), or *bad* (not relevant to the query). We evaluated participants' knowledge with a vocabulary learning test.

Our results show that entity cards—as used in our experimental setup—do not significantly affect human learning, with RPL scores consistently low and without significant differences between conditions. On the other hand, significant differences were found when examining topic effects.

When considering the search behaviours of participants, we did observe a number of significant differences across the four conditions. For example, varying the entity cards presented significantly impacted on the dwell time spent over documents, and overall session duration. We also observed a consistent trend that with lower quality entity cards the number of queries increase, although this was not significant. Similarly, as the entity card quality decreases, the number of unique documents viewed was shown to increase consistently across conditions (though again, not significantly so). When examining query terms issued by our participants, we began to see evidence that demonstrated that participants may indeed be examining the entity cards and using them to reformulate their queries, assisting in the learning process. Significant differences were observed when considering the fraction of query terms appearing in entity card title and summaries.

Our study has several limitations related to the task (artificial in nature), evaluation regime (we only consider vocabulary learning) and study setup (we are limited to a single search session).

Our study did not regard *concept difficulty*, and instead focused purely on providing entity cards based on the entity rankings derived from `EmbedRank`. We also opted to show a single entity card, as this is the common web search setup. However, some evidence [21] suggests that multiple entity cards may also be suitable for a learning environment. Introducing different entity card styles (depending on a participant's prior knowledge levels or their search strategies) would also be an interesting direction for future work. Instead of simply taking a Wikipedia summary and some basic attributes for the entity in question, richer content could be included based upon prior search history. In order to gain insights into the impact of entity cards on higher-level learning, we also need to explore more complex learning tasks and move beyond a single search session setup. As continuation of work by Urgo et al. [47], we may also want to study the effect of entity cards in different domains along various cognitive processes (apply, evaluate, create) and knowledge types (factual, conceptual, procedural).

REFERENCES

[1] J. Arguello and R. Capra. 2016. The Effects of Aggregated Search Coherence on Search Behavior. *ACM Trans. Inf. Syst.* 35, 1, Article 2 (Sept. 2016), 30 pages.

[2] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi. 2018. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proc. 22nd CoNLL*. Association for Computational Linguistics, 221–229.

[3] H. Bota, K. Zhou, and J.M. Jose. 2016. Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload. In *Proc. 1st ACM CHIIR*. 131–140.

[4] A. Bougouin, F. Boudin, and B. Daille. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proc. 6th AACL-IJCNLP*. 543–551.

[5] A. Câmara, N. Roy, D. Maxwell, and C. Hauff. 2021. Searching to Learn with Instructional Scaffolding. In *Proc. 6th ACM CHIIR*. 209–218.

[6] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* 509 (2020), 257–289.

[7] B. Choi, J. Arguello, R. Capra, and A.R. Ward. 2021. OrgBox: A Knowledge Representation Tool to Support Complex Search Tasks. In *Proc. 6th ACM CHIIR*. 219–228.

[8] M.J. Cole, J. Gwizdka, C.L., N.J. Belkin, and X. Zhang. 2013. Inferring user knowledge level from eye movement patterns. *IP&M* 49, 5 (2013), 1075 – 1091.

[9] K. Collins-Thompson, P. Hansen, and C. Hauff. 2017. Search as learning (dagstuhl seminar 17092). In *Dagstuhl reports*, Vol. 7.

[10] K. Collins-Thompson, S.Y. Rieh, C.C. Haynes, and R. Syed. 2016. Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies. In *Proc. 1st ACM CHIIR*. 163–172.

[11] H. Colt, M. Davoudi, S. Murgu, and N. Rohani. 2011. Measuring learning gain during a one-day introductory bronchoscopy course. *Surgical endoscopy* 25 (01 2011), 207–16.

[12] E. Dale. 1965. Vocabulary Measurement: Techniques and Major Findings. *Elementary English* 42, 8 (1965), 895–948.

[13] K. Eichler and G. Neumann. 2010. DFKI KeyWE: Ranking Keyphrases Extracted from Scientific Articles. In *Proc. 5th SemEval*. 150–153.

[14] C. Eickhoff, S. Dungs, and V. Tran. 2015. An eye-tracking study of query reformulation. In *Proc. 38th ACM SIGIR*. 13–22.

[15] C. Eickhoff, J. Teevan, R. White, and S. Dumais. 2014. Lessons from the Journey: A Query Log Analysis of within-Session Learning *(WSDM '14)*. 223–232.

[16] G. Ercan and I. Cicekli. 2007. Using lexical chains for keyword extraction. *Inf. Process. Manag.* 43 (2007), 1705–1714.

[17] C. Florescu and C. Caragea. 2017. A Position-Biased PageRank Algorithm for Keyphrase Extraction. In *AAAI*.

[18] U. Gadiraju, R. Yu, S. Dietze, and P. Holtz. 2018. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web *(Proc. 3rd ACM CHIIR)*. 2–11.

[19] S. Ghosh, M. Rath, and C. Shah. 2018. Searching as Learning: Exploring Search Behavior and Learning Outcomes in Learning-related Tasks. *Proc. 3rd ACM CHIIR* (2018), 22–31.

[20] F. Hasibi, K. Balog, and S.E. Bratsberg. 2017. Dynamic Factual Summaries for Entity Cards. In *Proc. 40th ACM SIGIR*. 773–782.

[21] G. Jimmy, Zuccon, G. Demartini, and B. Koopman. 2020. Health Cards to Assist Decision Making in Consumer Health Search. *AMIA* 2019 (03 2020), 1091–1100.

[22] R. Kalyani and U. Gadiraju. 2019. Understanding User Search Behavior Across Varying Cognitive Levels. In *Proc. 30th ACM HT*. 123–132.

[23] R. Kiros, Y. Zhu, R. Salakhutdinov, R.S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. 2015. Skip-Thought Vectors.

[24] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *Proc 37th ACM SIGIR*. 113–122.

[25] M. Lalmas and L. Hong. 2018. Tutorial on Metrics of User Engagement: Applications to News, Search and E-Commerce. In *Proc. 11th ACM WSDM*. 781–782.

[26] J.H. Lau and T. Baldwin. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proc. 1st RepL4NLP*. 78–86.

[27] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. arXiv:cs.CL/1405.4053

[28] C. Liu and X. Song. 2018. How Do Information Source Selection Strategies Influence Users' Learning Outcomes'. In *Proc. 3rd ACM CHIIR*. 257–260.

[29] H. Liu, C. Liu, and N.J. Belkin. 2019. Investigation of users' knowledge change process in learning-related search tasks. *Proc. ASIS&T* 56, 1 (2019), 166–175.

[30] Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[31] O. Medelyan, E. Frank, and I.H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proc. EMNLP*. 1318–1327.

[32] R. Mihalcea and P. Tarau. 2004. TextRank: Bringing Order into Text. In *Proc. EMNLP*. 404–411.

[33] F. Moraes, S.R. Putra, and C. Hauff. 2018. Contrasting Search as a Learning Activity with Instructor-Designed Learning. In *Proc. 27th ACM CIKM*. 167–176.

[34] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. 2013. Measurement and Modeling of Eye-Mouse Behavior in the Presence of Nonlinear Page Layouts. In *Proc. 22nd WWW*. 953–964.

[35] H.L. O'Brien, A. Kampen, A.W. Cole, and K. Brennan. 2020. The Role of Domain Knowledge in Search as Learning *(CHIIR '20)*. 313–317.

[36] M. Pagliardini, P. Gupta, and M. Jaggi. 2018. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proc. NAACL HLT*. 528–540.

[37] S.R. Putra, F. Moraes, and C. Hauff. 2018. SearchX: Empowering Collaborative Search Research. In *Proc. 41st ACM SIGIR (SIGIR '18)*. 1265–1268.

[38] Minghui Qiu, Y. Li, and Jing Jiang. 2012. Query-Oriented Keyphrase Extraction. In *AIRS*.

[39] J. Rovira, Joan María Senent, and Miquel Àngel Essomba Gelabert. 2016. Educational leadership and teacher involvement as success factors in schools in disadvantaged areas of Spain. *RELIEVE* 22 (2016), 4.

[40] N. Roy, F. Moraes, and C. Hauff. 2020. Exploring Users' Learning Gains within Search Sessions. In *Proc. 5th ACM CHIIR*. 432–436.

[41] N. Roy, M.V. Torre, U. Gadiraju, D. Maxwell, and C. Hauff. 2021. Note the Highlight: Incorporating Active Reading Tools in a Search as Learning Environment. In *Proc 6th ACM CHIIR*. 229–238.

[42] B. Škrlj, A. Repar, and S. Pollak. 2019. RaKUn: Rank-based Keyword Extraction via Unsupervised Learning and Meta Vertex Aggregation. In *Stat. Lang. & Speech Proc.* 311–323.

[43] K. Stahl and M. Bravo. 2010. Contemporary Classroom Vocabulary Assessment for Content Areas. *READ TEACH* 63 (04 2010), 566–578.

[44] R. Syed and K. Collins-Thompson. 2017. Optimizing search results for human learning goals. *IRJ* 20 (2017), 506–523.

[45] R. Syed and K. Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In *Proc. 40th ACM SIGIR*. 555–564.

[46] R. Syed and K. Collins-Thompson. 2018. Exploring Document Retrieval Features Associated with Improved Short- and Long-Term Vocabulary Learning Outcomes. In *Proc. 3rd ACM CHIIR*. 191–200.

[47] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2020. The Effects of Learning Objectives on Searchers' Perceptions and Behaviors. In *Proc 6th ACM ICTIR*. 77–84.

[48] X. Wan and J. Xiao. 2008. Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proc. 23rd AAAI*. 855–860.

[49] J. Wang, J. Liu, and C. Wang. 2007. Keyword Extraction Based on Pagerank. In *PAKDD*. 857–864.

[50] R. Wang. 2015. Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors.

[51] M. B. Wesche and T. Paribakht. 1996. Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Lang. Review* 53 (1996), 13–40.

[52] R.W. White, S.T. Dumais, and J. Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proc. 2nd WSDM*. 132–141.

[53] M.J. Wilson and M.L. Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *JASIST* 64, 2 (2013), 291–306.

[54] W.-T. Yih, J. Goodman, and V.R. Carvalho. 2006. Finding Advertising Keywords on Web Pages. In *Proc. WWW*. 213–222.

[55] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, and S. Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. 75–84.

[56] X. Zhang, M. Cole, and N. Belkin. 2011. Predicting Users' Domain Knowledge from Search Behaviors. In *Proc. 34th ACM SIGIR*. 1225–1226.