Know Your Onions: Understanding the User Experience with the Knowledge Module in Web Search

Ioannis Arapakis
Yahoo Labs
Barcelona, Spain
arapakis@yahoo-inc.com

Luis A. Leiva
PRHLT, Universitat Politècnica
de València, Spain
llt@acm.org

B. Barla Cambazoglu
Yahoo Labs
Barcelona, Spain
barla@yahoo-inc.com

ABSTRACT

The increasing availability of large volumes of human-curated content is shifting web search towards a paradigm that introduces seamlessly more semantic information to search engine result pages. This trend has resulted in the design of a new element known as the knowledge module (KM), where certain facts about named entities, obtained from various knowledge bases, are shown to users. So far, little has been done to uncover the role that this module plays on user experience in web search and whether it is perceived by users as a useful aid for their search tasks. Our work is an early attempt to bridge this gap. To this end, we conducted a crowdsourcing study aimed at understanding the effect of the KM on users' search experience and its overall utility. In particular, our study is the first to provide insights about the noticeability and usefulness of the KM in web search, together with comprehensive analyses of usability and workload.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.1.2 [**User/Machine Systems**]: Human factors

Keywords

web search engine; knowledge module; user experience

1. INTRODUCTION

In recent years, the knowledge module (KM) has become a standard component on search engine result pages (SERPs) of all major web search engines (Fig. 1). This module provides users with information about the named entities they are searching for as part of their search tasks. The content presented in the KM is typically obtained in a semi-structured format from curated entity databases, such as Freebase or Wikipedia, and often includes both quantitative and qualitative information about the queried entity. This raw information can be further enriched by the search engine;

CIKM'15, October 19–23, 2015, Melbourne, Australia.

ACM 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806591.

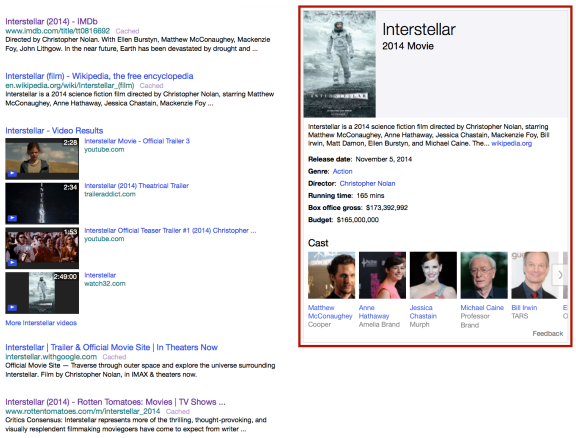

Figure 1: The KM (in red border) displayed on the Yahoo SERP for the query "interstellar".

e.g., by showing a ranking of related entities, accompanied with explanations of their relationship. Moreover, the KM is often complemented with additional content, such as multimedia or social media content associated with the entity, typically obtained from third-party data sources.

In this context, most research has focused on general backend system tasks, the most important being knowledge base construction [1, 4, 6, 7, 11], or more specific backend tasks, such as related entity recommendation [2, 3]. With the exception of a recent query log analysis on exploratory search [10], so far little has been done to understand the way web search users interact with the frontend system, i.e., the knowledge component presented as part of the SERP. Our work makes an early attempt to understand the impact of the KM on users' overall search experience in entity-centric search tasks. In particular, we try to answer questions of the following kind: Do users notice the presence of the KM on SERPs? If they notice it, do they find it useful? Does the KM really ease web search? Is it cognitively or physically demanding?

Our contribution can be summarized as follows. To understand the user experience with the KM, we conducted a crowdsourcing study. The study involved questionnaires and self-reported feedback from 533 users about noticeability, usefulness, usability, and workload toward the KM shown on the Yahoo web search engine. We observed that the majority of the participants (about three-fourths) who performed the search tasks noticed the presence of the KM on the SERP, and they felt that, overall, it provided a useful aid to accomplish their search tasks better and faster.

2. CROWDSOURCING STUDY

To understand the impact of the KM in web search, we conducted a crowdsourcing study and collected feedback from users who performed short, entity-centric search tasks using the Yahoo web search engine. With this study, we aimed to determine: (i) what percentage of users notice the KM on the SERP, (ii) to what extent they perceive it as a useful aid to their search tasks, and whether the presence of the KM can affect (iii) the perceived usability and (iv) experienced workload due to web search engine usage.

Crowdsourcing offers several advantages not available in other experimental settings [9], such as access to a large and diverse pool of participants with stable availability, as well as collection and analysis of real usage data at a large scale. Another advantage of crowdsourcing is the low cost of the tasks, which makes it a preferable solution over the more expensive laboratory-based experiments. On the downside, a limited range of parameters can be explored in a controlled manner and experimenters have to account for potential threats to ecological validity, distractions in the physical environment of the user, and privacy issues, to name a few.

In our study, we used the Amazon Mechanical Turk service. All of the aforementioned limitations were taken into consideration and preventive measures were put into practice to discount low-quality responses. Also, strict selection criteria were applied to exclude unsuitable participants (e.g., HIT approval rate $\geq$ 98%, number of HITs approved $\geq$ 1,000).

2.1 Experimental Design

The experiment had a repeated measures design with one independent variable: KM (with two levels: "visible" or "hidden"). The KM visibility was controlled with client-side scripting, removing the KM from the SERP in the "hidden" condition. The dependent variables (Section 2.3) were: (i) KM noticeability, (ii) KM usefulness (ease of use and speed of use), (iii) perceived usability of the search engine (Tables 1 and 2), and (iv) overall workload (Table 3). The experiment consisted of two short search tasks that were completed using the Yahoo search engine, one task displaying the KM on the SERP and one without it. To control for order effects, we counterbalanced task assignments using Latin square design.

Participants accessed the search engine through a custom proxy which did not alter the original look and feel of the SERPs. This allowed us to instrument the browsed pages on the fly and capture user interactions with the SERP without interfering with the actual web search engine interface in production. The proxy had a common entry page for all participants. For each search task, participants were presented with a question and were suggested a search query to begin with. Finally, the suggested queries were all picked from a pool of queries that triggered the KM on the SERP, independent of the KM visibility (Section 2.2).

2.2 Search Query Sample

Our query set consisted of 32 unique query patterns that were selected after a large-scale query log analysis All queries would trigger the KM on the Yahoo SERP, so we could ensure that in all tasks the KM would be displayed on the SERP, thus allowing us to choose between leaving it visible or hiding it, depending on the experimental condition.

The selected query patterns belonged to four different themes (famous people, movies, athletes, sport teams) and required either single or multiple answers. An example of a single-answer query pattern is "Who is the head coach of the team X?" while an example of a multi-answer query is "Who are X's children?". To diversify our search query pool, we produced three questions per query pattern while we introduced some additional multi-answer questions to increase the difficulty of the search tasks. In total, our query set included 144 different queries.[1] In the study, the query set was repeated as many times as needed to accommodate all participants. Each query was answered under each condition by at least two participants and at most six participants.

2.3 Self-Reported Measures

We used three different post-task questionnaires to elicit participants' subjective experience about the search engine and search tasks. More specifically, participants were asked to complete the Computer System Usability Questionnaire (CSUQ), the Perceived Usefulness and Ease of Use questionnaire (PUEU), and the NASA Task Load Index (NASA-TLX), together with custom statements described later.

The CSUQ [8] is a multi-dimensional user satisfaction questionnaire designed for use in scenario-based usability evaluations. Out of the four scales it contains, we considered only the scores from the system usefulness (SYSUSE) subscale (Table 1). The PUEU questionnaire [5] is a psychometric scale with significant empirical relationships with self-reported measures of usage behavior. It focuses on two theoretical constructs, perceived usefulness and ease of use, which are fundamental determinants of system usage. In our study, we considered only the perceived usefulness scale, which consists of the six statements shown in Table 2. The NASA-TLX is a multi-item assessment tool that allows participants to perform subjective workload assessments of human-computer interaction systems. NASA-TLX employs a rating procedure based on the six questions shown in Table 3. Combined together, CSUQ, PUEU, and NASA-TLX gauged key aspects of participants' experience with the search tasks and the search engine. The questions were all forced-choice type and appeared at random to mitigate order effects. A 7-point Likert scale was used in all questionnaires, with high scores representing a stronger agreement with the given statement.

In addition to the above psychometric scales, we also collected demographic information as well as information about participants' agreement to the following statements: (i) "This search engine helped me accomplish my task in a reasonable amount of time", (ii) "I feel satisfied with the retrieved results". Finally, we inquired about the KM through a mini-questionnaire, which only appeared on the SERPs that displayed the KM. The mini-questionnaire was initially hidden, in order not to interfere with regular browsing, and was shown to the user just before unloading the SERP on closing the browser tab. The mini-questionnaire contained three questions: (i) "Did you notice the KM? (yes/no)", (ii) "To what extent did you find the KM useful in answering the question? (1: not useful at all, ..., 5: completely useful)", (iii) "To what extent did the KM help you answer the question faster? (1: not faster at all, ..., 5: extremely faster)".

2.4 Participants

We recruited 612 participants through Amazon Mechanical Turk. From this original sample, we approved assignments

[1] `http://personales.upv.es/luileito/kme/queries.tsv`

Table 1: CSUQ-SYSUSE subscale items

1. Overall, I am satisfied with how easy it is to use this search site.
2. I feel this search site is simple to use.
3. I can effectively complete my work using this search site.
4. I am able to complete my work quickly using this search site.
5. I am able to efficiently complete my work using this search site.
6. I feel comfortable using this search site.
7. It was easy learning to use this search site.
8. I believe I became productive quickly using this search site.

Table 2: PUEU subscale items

1. Using this search site would allow me to accomplish my search tasks more quickly.
2. Using this search site would improve my performance.
3. Using this search site would increase my productivity.
4. Using this search site would enhance my effectiveness.
5. Using this search site would make it easier to do my search tasks.
6. I would find this search site useful in my search tasks.

Table 3: NASA-TLX factor definitions

Factor	Question
Mental Demand	How mentally demanding was the task?
Physical Demand	How physically demanding was the task?
Temporal Demand	How hurried or rushed was the pace of the task?
Performance	How successful were you in accomplishing what you were asked to do?
Effort	How hard did you have to work to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, and annoyed were you?

for 533 participants (female = 226, male = 307), aged from 18 to 66. Participants were of mixed nationality (e.g., Belgian, Finnish, British, American) and had varying educational backgrounds: 29.98% had a high school diploma, 18.98% had a college diploma, 41.56% had a BSc degree, 7.97% had an MSc, and 1.52% had a PhD. All participants were proficient in English, 98.31% being native speakers. When asked about their search engine at home or work, participants reported using primarily Google, followed by Yahoo and Bing.

2.5 Procedure

At the beginning, participants were informed about the terms and conditions of the study, followed by a short description of the SERP. The study had to be done in a single session. The participants could opt out at any moment, in which case they would not be compensated. Participants were asked to "evaluate two different backend systems of Yahoo Search, by performing two search tasks". Therefore, participants were not informed of the actual intent of the study ("understand the impact of KM in web search"), in order to avoid a potential bias. For each task, participants had to answer a question by searching for relevant information on the proxified search engine. They were also presented with a suggested query to begin their search, although participants were free to submit additional queries (e.g., if the suggested query did not lead to the answer) and examine as many results as necessary to complete the search task.

We used informational, entity-centric queries to introduce a common starting point across all participants who tested each particular combination of query and backend system. Upon finishing each task, participants were instructed to submit their answer and complete the post-task questionnaire. The study concluded with a demographics questionnaire. The payment for participation was $1.20 and each participant could take the study only once.

3. RESULTS

In the following, we discuss our findings based on the 1,066 search tasks performed by 533 participants. The results are presented in three sections. The first section discusses the noticeability and the usefulness of the KM. The second section shows the effect of the KM on the web search engine's perceived usability. The third section presents findings of the workload analysis. To quantify the statistical significance of our results, we used the Wilcoxon signed-rank test at an α level of 0.05.

3.1 Noticeability and Usefulness

The first research questions we answered are whether the KM is being noticed by web search users, and to what extent it is considered a useful aid to their search activities. According to the responses from the mini-questionnaire (Section 2.3), out of the 533 participants who were involved in our study, the majority (78.86%) reported noticing the KM on the SERP. Considering that the KM is a relatively new element introduced in SERPs, the high percentage of participants who engaged with it is a first positive indication of its noticeability, even if this was demonstrated for only one of the available commercial search engines. The KM was also found to be very useful in answering the search task questions ($M = 4.03, SD = 1.48$). Moreover, the KM helped the participants who noticed it to answer the search task questions fairly faster ($M = 3.84, SD = 1.60$). These findings suggest that the KM is both noticeable and useful to web search users.

Furthermore, we performed a correlation analysis and computed the point-biserial correlation coefficient (r_{pb}) for the above variables. In the case of r_{pb}, the sign of the correlation depends on the way the coding of the variables was made, therefore we ignore all information about direction. Our findings indicated that noticeability is significantly correlated with both ease of use ($r_{pb} = 0.60, p < .0001$) and speed of use ($r_{pb} = 0.54, p < .0001$). In short, users who noticed the KM felt that they could accomplish their tasks better and faster.

3.2 Perceived Usability

Next, we examined the impact of the KM on perceived system usability. To this end, we looked at the participants' responses to our two custom statements as well as the 8-item CSUQ-SYSUSE and 6-item PUEU scales shown in Table 4. We averaged the responses to obtain the final scores and then contrasted and compared what the participants reported in the experimental conditions (visible or hidden KM).

The Wilcoxon signed-rank test showed that participants found the search engine to be significantly more helpful in accomplishing their search tasks in a reasonable amount of time ($z = 8.13, p < .001, r = 0.35$) when the KM was visible ($Mdn = 7$) compared to when it was hidden ($Mdn = 6$).

Table 4: Usability results for custom statements (CS-1 and CS-2), CSUQ-SYSUSE, and PUEU

	KM status	
Scale	**Visible**	**Hidden**
CS-1: This search site helped me accomplish my task in a reasonable amount of time	6.38 ± 1.10	5.80 ± 1.57
CS-2: I feel satisfied with the retrieved results	6.40 ± 1.04	5.85 ± 1.58
CSUQ-SYSUSE	6.22 ± 1.09	5.66 ± 1.52
PUEU	5.25 ± 1.51	4.73 ± 1.72

Table 5: Workload results: NASA-TLX factors

	KM status	
Factor	**Visible**	**Hidden**
Mental	1.70 ± 1.18	1.93 ± 1.32
Physical	1.33 ± 0.91	1.30 ± 0.82
Temporal	1.78 ± 1.25	1.88 ± 1.29
Performance	6.37 ± 1.42	6.19 ± 1.48
Effort	1.96 ± 1.48	2.29 ± 1.58
Frustration	1.55 ± 1.12	1.86 ± 1.41
Sum	14.70 ± 4.40	15.44 ± 4.73

Moreover, participants felt significantly more satisfied with the retrieved results ($z = 7.36, p < .001, r = 0.32$) when having seen the KM ($Mdn = 7$) rather than not ($Mdn = 6$). Participants also perceived the search engine to be significantly more usable ($z = 9.06, p < .001, r = 0.39$) when the SERP displayed the KM ($Mdn = 6.66$) than when it did not ($Mdn = 6$). Indeed, the CSUQ-SYSUSE scores were higher for the "visible" condition, as observed in Table 4. Moreover, the reported PUEU scores were significantly higher ($z = 8.58, p < .001, r = 0.37$) for the "visible" condition ($Mdn = 5.5$) than the "hidden" condition ($Mdn = 5$).

3.3 Overall Workload

Finally, we looked at the perceived workload experienced by our participants as they performed the search tasks. Table 5 shows the NASA-TLX scores reported for each factor (lower is better). The individual factor scores were summed up to obtain the overall workload scores. Participants who interacted with the SERP that did not display the KM ($Mdn = 14$) experienced a significantly higher workload ($z = 4.40, p < .001, r = 0.19$) than the participants who were shown the KM ($Mdn = 13$). Table 5 also presents the relative contribution of each factor to the overall workload score for both experimental conditions. More specifically, the participants in the "hidden" condition ($Mdn = 1$) reported a significantly higher mental demand ($z = 4.81, p < .001, r = 0.20$) than those in the "visible" condition ($Mdn = 1$). Participants in the "hidden" condition also reported lower physical and temporal demand scores than those in the "visible" condition, although these differences were not statistically significant.

When examining how successful they were in accomplishing the search tasks, participants in the "visible" condition ($Mdn = 7$) reported significantly higher performance ($z = 3.39, p < .001, r = 0.14$) than those in the "hidden" condition ($Mdn = 7$). Furthermore, the search task demanded significantly more effort ($z = 5.25, p < .001, r = 0.22$) in the "hidden" condition ($Mdn = 2$) compared to the "visible" condition ($Mdn = 1$). Lastly, participants in the "hidden" condition ($Mdn = 1$) reported significantly higher levels of frustration ($z = 5.56, p < .001, r = 0.10$) than those in the "visible" condition ($Mdn = 1$).

4. CONCLUSION AND FUTURE WORK

This work entails an early attempt to understand the impact of the KM on users' search experience and provides empirical evidence of its overall utility. To this end, we conducted a crowdsourcing study which revealed the potential benefits of the KM, when dealing with entity-centric search tasks. In particular, we showed that the KM was noticed by most participants and was perceived to be a valuable help in web search. Moreover, the KM was perceived to ease the search process for the users. Our ongoing work is on correlating mouse cursor tracking data and user engagement with the KM. We believe that this is a research avenue worth pursuing given the lack of explicit user feedback about engagement (e.g., clicks or dwell time) in the context of the KM. Finally, we anticipate that further research on the topic may have an impact on future web search interfaces.

Acknowledgments

This work is part of the Valorization and I+D+i Resources program of VLC/CAMPUS and has been funded by the Spanish MECD as part of the International Excellence Campus program.

5. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proc. ISWC*, 722–735, 2007.

[2] B. Bi, H. Ma, B.-J. P. Hsu, W. Chu, K. Wang, and J. Cho. Learning to recommend related entities to search users. In *Proc. WSDM*, 139–148, 2015.

[3] R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *Proc. ISWC*, 33–48. 2013.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proc. SIGMOD*, 1247–1250, 2008.

[5] F. D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340, 1989.

[6] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proc. SIGKDD*, 601–610, 2014.

[7] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artif. Intell.*, 194:28–61, 2013.

[8] J. R. Lewis. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *Intl. J. Hum.-Comput. Interact.*, 7(1):57–78, 1995.

[9] W. Mason and S. Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 1–23, 2010.

[10] I. Miliaraki, R. Blanco, and M. Lalmas. From "Selena Gomez" to "Marlon Brando": Understanding explorative entity search. In *Proc. WWW*, 765–775, 2015.

[11] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proc. WWW*, 697–706, 2007.