Evaluation of Search Methods on Community Documents

Kushagra Singh Bisen, Sara Assefa Alemayehu, Pierre Maret, Alexandra Creighton, Rachel Gorman, Bushra Kundi, Thumeka Mgwgwi, Fabrice Muhlenbach, Serban Dinca-Panaitescu, Christo El Morr

Evaluation of Search Methods on Community Documents

Kushagra Singh Bisen[1,5][0000−0003−0950−6043], Sara Assefa Alemayehu[1], Pierre Maret[1,4], Alexandra Creighton[2], Rachel Gorman[2], Bushra Kundi[2], Thumeka Mgwgwi[3], Fabrice Muhlenbach[1], Serban Dinca-Panaitescu[2], and Christo El Morr[2]

[1] Université Jean Monnet Saint Etienne, France
[2] York University, School of Health Policy and Management, Canada
[3] School of Gender, Sexuality and Women's Studies, York University, Canada
[4] The QA Company, France
[5] IDLab, Ghent University - imec, Belgium

Abstract. Searching for domain-specific information on the web is tough. Community documents are therefore made searchable with a dedicated search platform. Search Methods employed on a document corpora are often evaluated over the aspect of efficiency and not focusing on the often-overlooked user experience. In the paper, we present an evaluation of search methods over domain-specific document corpora over search methods. The document corpora are represented in RDF as well as free-text. We describe the search methods as well as present the evaluation environment prepared. Moreover, we present the result of the user study to understand the experience of a user with the search methods.

Keywords: Meta-Data · Domain-Specific Documents · Question Answering · Information Retrieval · UEQ · Human-Computer Interaction

1 Introduction

Search for domain-specific information is tough on the web. Search engine usually returns a response from pages which are popular and indexed above. Communities with an interest in a particular domain often find results which are not relevant to their query searching on the web. For instance, this is the case for activists in disability and human rights advocacy groups. This community, as well as many other distributed communities, need domain-specific repositories where they will have more chances to find the information they are searching for. These data and document repositories may implement a different mechanism for accessing the information. Moreover, the questions asked by stakeholders in a project can utilise the meta-data, the document's content or both. Therefore, there is a need for an evaluation of methods that can be employed for searching over community data and documents. In this paper, we present an evaluation of search techniques available over domain-specific document corpus. The paper is organised as follows, In section 2, we describe the existing work done in this

direction. In section 3, we explain the experiment we conducted and the evaluation methods. We further show the results in section 4.2. We conclude and point to future work in the section 5

2 Related Work

The evaluation of search methods generally concentrates on the performance of the search mechanism: calculation of precision, recall, and F-Measure. Benchmarks have been proposed for this kind of task. Apart from the quality of the search method, it is also important to provide the user with an efficient search experience to retrieve the information in the corpus. Therefore, there is a need for evaluating the quality of the user search experience when searching for information. And this is especially necessary because of the emergence and popularity of new search techniques such as Elastic search[16] [26], Question Answering over free text[3] [6] [4] [2], and Question Answering over knowledge graphs[8] [9] [27] [30]. The search techniques are *one-field one-shot search* i.e users retrieve information by building a question/query through only a text field and receive the answer in response. There have been end user evaluations on semantic web to improve the human-semantic web interaction [21] [10]. There are various methods for evaluating a method such as Concept Testing [22], Heuristic Evaluation [23] and User Experience Evaluation [29]. We need to adapt the evaluation of user experience to the search methods. To the best of our knowledge, there has been no user experience evaluation on search methods over community documents.

3 Description of an Evaluation Environment for Search Methods

In this section, we describe the environment prepared to experiment with and evaluate the search methods on community documents. In the subsection 3.1 we describe the search methods we compare and in the subsection 3.2 we describe the experiment.

3.1 Search Methods

QAnswer over Knowledge Graphs. QAnswer KG [8] is a search engine over RDF datasets with which users can search information with both questions as well as keywords. The data stored in a knowledge graph can be exported as a triple pattern for QAnswer KG. The input query is expanded and n-grams from the query are mapped with the properties and resources. The properties and resources are used to generate possible SPARQL queries by combining the triples which share a variable. The queries are ranked with a machine-learning model and the query with the highest confidence is chosen as the response to the user's input.

Elastic Search over Documents. Elastic Search over Documents (ESDoc) Elastic Search [12] is a JSON based full-text search engine capable to search over big data in a real-time fashion. Elastic Search can be used out of the box, although some users prefer to tweak parameters. Elastic Search is built over Apache Lucene, a java library. The elastic search uses algorithms such as okapi bm25 and NMSLIB to return a relevant document in response to the query.

An elastic search system's architecture is composed of [17]

- *Document* for storing an entity, it has an identifier and is part of an index where it is stored. They can be separated over multiple nodes.
- *Node* is an instance of elastic search active for a query. Connected active nodes are referred to as a *cluster*.
- *Shards* are used to improve efficiency through processing in parallel by further dividing the index into shards. Shards are often stored as *Replicas* to provide throughput of the data for efficiency in search.

QAnswer Search over Documents. QAnswer Search over Documents (QADoc) employs RoBERTa [19] for question answering over the document corpus. The documents are uploaded, split into paragraphs and pre-processed. The questions to the documents are answered with the content of the document.

3.2 Experiment

Search Instruction Questionnaire. An experiment was conducted where the candidates were requested to search for domain-specific information as illustrated in the table 2. The instructions provided are *search-method agnostic*. We did not propose questions but rather proposed search instructions to which a user can formulate the question by himself in a search technique. The reason was to prevent biases that could arise by providing questions to search the method, as the methods are different in their implementation. For example, telling the users to search for "*keyword1, keyword2, keyword3*" will work in keyword-based elastic search but could not be good in other search methods. The candidates used each search method listed in the section 3.1 to search for the domain-specific information. The candidates have presented a questionnaire with the search instructions with a 7-point Likert Scale [13] from -3 to 3 (the higher the better) to record the relevancy of the information retrieved, as per the user. We present 6 search instructions to the candidate of which 5 are True i.e there is any information related to the instruction in the community document corpus and 1 is False i.e there is no information related to the instruction in the community document corpus (see Table 1). The candidates use a stopwatch to record the time they spent searching for the information. The candidates are instructed to stop searching if more than 2 minutes are spent and further record that they didn't find an answer. The candidates are also instructed to record if they found an answer with three scales which are yes, no and maybe. We chose two minutes as a threshold for search as the search method should be able to provide an answer in that time, if not the search method is not considered efficient.

User Experience Questionnaire The second questionnaire presented to the candidate after each search method was a user experience questionnaire (UEQ) [18]. We choose UEQ as it provides a benchmark to classify the values obtained from the result.The objective of UEQ is to allow a quick assessment done by end users covering a preferably comprehensive impression of user experience. It should allow users to express feelings, impressions and attitudes that arise when experiencing the search method under investigation simply and immediately [24]. We employ the standard version of UEQ which contains 26 items. The items are divided into 6 scales. The 6 scales focus on different experience aspects of the search method:

- Attractiveness: Signifying the overall impression of the search method.
- Perspicuity: Describing if it is easy to get familiar with the search method and if it is easy to understand.
- Efficiency: Describing if the search method is fast to provide them with information.
- Dependability: Describing if the user feels confident while using the search method.
- Stimulation: Describing if the user finds the search method exciting and motivating.
- Novelty: Describing if the user finds the search method innovative and creative.

We use the UEQ scales to find how the candidates feel about the attractiveness of the search method (with the scale *Attractiveness*), the usefulness of the search method i.e Pragmatic Value (employing the scales *Perspicuity, Efficiency and Dependability*), the ease of use of the search method i.e Hedonic Value (employing the scales *Stimulation and Novelty*)

4 Experimental Results

4.1 Domain of Experiment

The communities working for disability rights advocacy require better information access [20] [1]. In the domain of healthcare, the stakeholders perform a variety of search tasks like literature reviews, scoping reviews, rapid evidence reviews and systematic reviews [5]. The domain of the documents employed in the experiment is disability studies. The motivation to choose disability studies documents was to provide a better search experience for the disability documents in the WikiDisability Project [11]. We search over the disability documents which are annotated with the Disability Wiki Website[6] in the Disability Knowledge Graph[7] as well as in free-text in PDF to be searchable by Elastic Search3.1 and QAnswer Search over Documents 3.1. The table 2 describes the disability domain-specific search instructions given to the user. The search instructions, as

[6] https://disabilityrightsweb.univ-st-etienne.fr/
[7] https://disabilitywiki.univ-st-etienne.fr/wiki/The_Disability_Wikibase

well as the UEQ, were provided to the user in form of a response form[8] to be filled.

Documents	Candidates	Search Methods	True Questions	False Questions
24	17	3	5	1

Table 1. The table shows the details related to the search instruction questionnaire experiment

Instruction to the User	Is answer available?
Find text about the racism faced by black feminists	Yes
Find text about elitism in american womens movement	Yes
Find text about human rights of minors	No
Find text about racism in United States	Yes
Find text about ableism in prison	Yes
Find text about police violence for disabled people	Yes

Table 2. The table shows the search instructions given to the user for search with each method

4.2 Results of the Experiment

The value for UEQ scales are described in the table 4. We further compare the search methods with an ANOVA test [7] on each 6 sub-scale of UEQ for the 17 candidates. However, there was a statistically significant difference between the groups as determined by One-Way ANOVA of the scales Perspicuity, Efficiency and Novelty. An ANOVA test signifies if there is an overall difference between the groups. We, therefore, perform a Tukey-Kramer Test [14] to find out the differences between the groups. The results from the experiment are summarised as,

- ESDoc provided the most relevant answers as per the candidates of the experiment, although it provided them with a false sense of information as the users found information for the instruction with no information available in the community corpus (see figure 1).
- For search instruction with information available in the document corpus, users found most information with ESDoc followed by QAnswer KG and QADoc. However, for the instruction with no information available, users *found* most information with ESDoc as well (see figure 2).
- The values obtained from UEQ (see table 4) falls in the category of *bad* according to the UEQ benchmark [25]

[8] https://forms.gle/bjKqpdRGCuFSQFFG9

- There were no statistically significant differences between the group means as determined by One-Way ANOVA of the scales Attractive, Dependability and Stimulation (see table 5)
- There was a statistically significant difference between the groups as determined by One-Way ANOVA of the scales Perspicuity, Efficiency and Novelty. (see table 5)
- From Tukey-Kramer Test, we find that there is a significant difference between ESDoc vs QAnswer KG and ESDoc vs QADoc in the Perspicuity Scale.
- There is a significant difference between ESDoc vs QADoc for both Novelty and Efficiency scale (see 6)
- QADoc had the highest Hedonic Value i.e users found QADoc to be the most pleasant while interacting with it. It is followed by ESDoc and QAnswer KG. (see figure 3)
- QAnswer KG had the highest Pragmatic Value i.e users found QAnswer KG to be the most efficient and useful. It is followed by QADoc and ESDoc. (see figure 3)

	Att.	**Eff.**	**Per.**	**Dep.**	**Sti.**	**Nov.**
Excellent	>= 1.75	>= 1.78	>= 1.9	>= 1.65	>= 1.55	>= 1.4
Good	>= 1.52 < 1.75	>= 1.47 < 1.78	>= 1.56 < 1.9	>= 1.48 < 1.65	>= 1.31 < 1.55	>= 1.05 < 1.4
Above Avg	>= 1.17 < 1.52	>= 0.98 < 1.47	>= 1.08 < 1.56	>= 1.14 < 1.48	>= 0.99 < 1.31	>= 0.71 < 1.05
Below Avg	>= 0.7 < 1.17	>= 0.54 < 0.98	>= 0.64 < 1.08	>= 0.78 < 1.14	>= 0.5 < 0.99	>= 0.3 < 0.71
Bad	**< 0.7**	**< 0.54**	**< 0.64**	**< 0.78**	**< 0.5**	**< 0.3**

Table 3. Benchmark scores for classifying the experiences of users to the scales of UEQ [25]. The scores obtained in the experiment belong to the scale **bad**.

Scales	QAnswer KG	ESDoc	QADoc
Attractive	-0.272	**-0.114**	-0.433
Perspicuity	**-0.014**	-1.205	-0.05
Efficiency	-0.22	**0.014**	-0.583
Dependability	**-0.132**	-0.014	-0.266
Stimulation	-0.161	**0.0588**	-0.1
Novelty	-0.088	-0.191	**0.266**

Table 4. The scores obtained from UEQ on different scales

Fig. 1. Likert scale scores for the search methods in the relevancy of the information found from the search methods on instructions with information was available and for the instruction, information was not available

Scale	$df_{between}$	df_{within}	F-Ratio	P-Value
Attractive	2	46	1.269	0.29
Perspicuity	2	46	36.20	0
Efficiency	2	46	5.284	0.008
Dependability	2	46	0.861	0.429
Stimulation	2	46	1.78	0.179
Novelty	2	46	3.2	0.049

Table 5. Values of F-Ratio and P-Value for One-Way Anova on the UEQ data for Search Methods

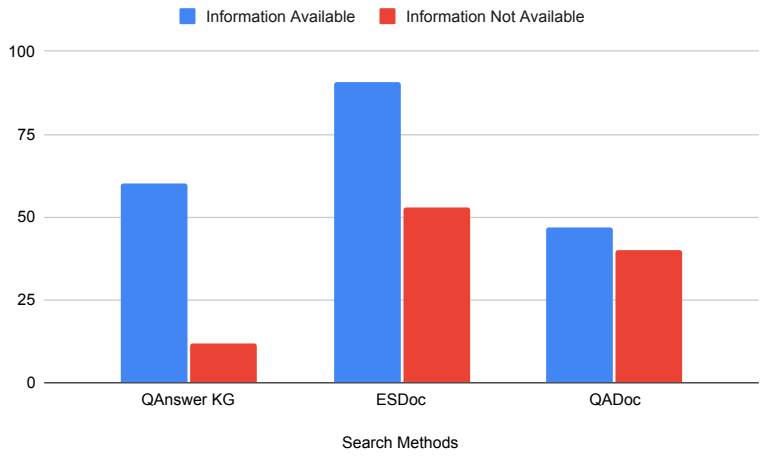

Fig. 2. Percentage of users who found an answer within 2 minutes with search methods on the instruction for which information was available and for the instruction information was not available

Scale	Comparison Groups	q_{tukey}
Perspicuity	QAnswerKG vs ESDoc	**10.6742**
	ESDoc vs QADoc	**10.029**
	QAnswerKG vs QADoc	0.306
Efficiency	QAnswerKG vs ESDoc	1.86
	ESDoc vs QADoc	**4.579**
	QAnswerKG vs QADoc	2.777
Novelty	QAnswerKG vs ESDoc	0.798
	ESDoc vs QADoc	**3.439**
	QAnswerKG vs QADoc	2.665

Table 6. q_{tukey} values for the scales showing a statistically significant difference. As there are three groups with the degree of freedom within groups being 46, the critical value for 3 groups, df = 46 for 5% significance level is **3.425** [15]

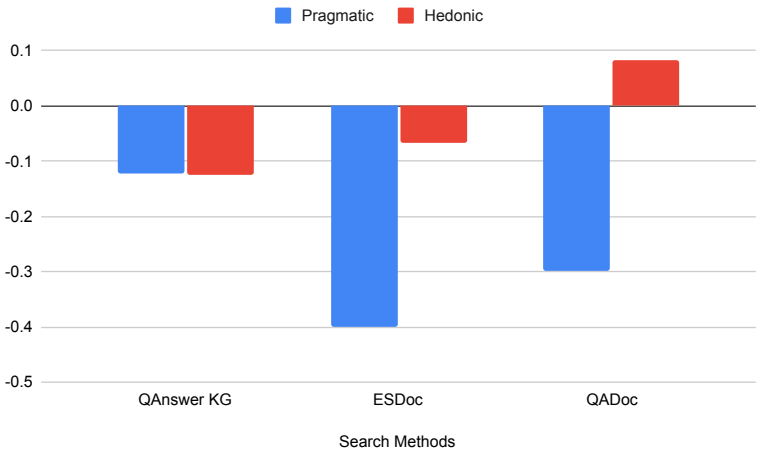

Fig. 3. Pragmatic (i.e perceived use-fullness, efficiency) and Hedonic (i.e perceived innovation) values for the search methods

5 Conclusion and future work

In the paper, we have presented the evaluation of search methods on domain-specific community document corpus based on the user experience. We found out that using Elastic search over the documents can provide relevant answers. Although, it also provided a false sense of relevancy of the information to the user for the search instruction with no information available in the document corpora. Thus, for non-exploratory question answering with an exact answer, we need more than the Elastic search technique. We find out that QADoc was the search method with perceived innovation by the users, but it did not perform as well as ESDoc or QAnswer KG for information retrieval. QAnswer KG was perceived as the most useful search method by the users. We conclude that we need to combine various search methods over community documents to provide a better search experience to the user. After conclusion, we developed a demo[9] on the same document corpus where we combine the three search methods [28] (manuscript under evaluation). We used wikibase as a knowledge graph to store data around the docment (the meta-data) and QADoc for data inside the document (the actual content). We introduced a fallback to other search methods in case the confidence of the response to the query is below a predefined threshold. In future, we have planned for a heuristic evaluation of the user interface for each search method to improve the search experience. We have also planned to introduce a new set of documents to the corpus and evaluate the search with the application where the search methods are combined.

References

1. Ola Abualghaib, Nora Groce, Natalie Simeu, Mark T. Carew, and Daniel Mont. Making visible the invisible: Why disability-disaggregated data is vital to "leave no-one behind". *Sustainability*, 11(11), 2019.
2. Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries. *arXiv preprint arXiv:2205.01290*, 2022.
3. Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W. Cohen. Open question answering over tables and text, 2020.
4. Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*, 2020.
5. Alexandra Collins, Deborah Coughlin, James Miller, and Stuart Kirk. The production of quick scoping reviews and rapid evidence assessments: a how to guide. Technical report, December 2015. Freely available via Official URL link.
6. Silviu Cucerzan and Eugene Agichtein. Factoid question answering over unstructured and structured web content. In *TREC*, volume 72, page 90, 2005.
7. Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman. An anova test for functional data. *Computational statistics & data analysis*, 47(1):111–122, 2004.

[9] http://demo-disabilityrightsweb.univ-st-etienne.fr/

8. Dennis Diefenbach, José Giménez-García, Andreas Both, Kamal Singh, and Pierre Maret. Qanswer kg: Designing a portable question answering system over rdf data. In Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez, editors, *The Semantic Web*, pages 429–445, Cham, 2020. Springer International Publishing.
9. Daniel Diomedi and Aidan Hogan. Question answering over knowledge graphs with neural machine translation and entity linking, 2021.
10. Roberto García and Rosa Gil. Improving human-semantic web interaction: The rhizomer experience. In *SWAP*, 2006.
11. Rachel Gorman, Pierre Maret, Alexandra Creighton, Bushra Kundi, Fabrice Muhlenbach, Alexis Buettgen, Enakshi Dua, Geoffrey Reaume, Thumeka Mgwigwi, Serban Dinca-Panaitescu, et al. The potential of an artificial intelligence for disability advocacy: The wikidisability project. In *Public Health and Informatics*, pages 1025–1026. IOS Press, 2021.
12. Clinton Gormley and Zachary Tong. *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine.* " O'Reilly Media, Inc.", 2015.
13. Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015.
14. HJ Keselman and Joanne C Rogan. The tukey multiple comparison test: 1953–1976. *Psychological Bulletin*, 84(5):1050, 1977.
15. Stephen Kokoska and Christopher Nevison. Critical values for the studentized range distribution. In *Statistical tables and formulae*, pages 64–66. Springer, 1989.
16. Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W. Godfrey. Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, page 328–331, New York, NY, USA, 2014. Association for Computing Machinery.
17. Rafal Kuc and Marek Rogozinski. *Elasticsearch server*. Packt Publishing Ltd, 2013.
18. Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*, pages 63–76. Springer, 2008.
19. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.
20. Mitchell Loeb. Disability statistics: an integral but missing (and misunderstood) component of development work. *Nordic journal of human rights*, 31(3):306–324, 2013.
21. Rob McCool, Andrew J Cowell, and David A Thurman. End-user evaluations of semantic web technologies. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.
22. William L. Moore. Concept testing. *Journal of Business Research*, 10(3):279–294, 1982.
23. Jakob Nielsen. How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1(1):8, 1995.
24. Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and evaluation of a short version of the user experience questionnaire (ueq-s). *International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108.*, 2017.
25. Martin Schrepp, Jorg Thomaschewski, and Andreas Hinderks. Construction of a benchmark for the user experience questionnaire (ueq). 2017.
26. Dikshant Shahi. *Apache solr*. Springer, 2016.

27. Sangjin Shin, Xiongnan Jin, Jooik Jung, and Kyong-Ho Lee. Predicate constraints based question answering over knowledge graph. *Information Processing Management*, 56(3):445–462, 2019.
28. Kushagra Singh Bisen, Sara Assefa Alemayehu, Pierre Maret, Alexandra Creighton, Rachel Gorman, Bushra Kundi, Thumeka Mgwgwi, Fabrice Muhlenbach, Serban Dinca-Panaitescu, Dennis Diefenbach, Kunpeng Guo, and Christo El Morr. Wikibase as an Infrastructure for Community Documents: The example of the Disability Wiki Platform. In *Semantics 2022 - 18th International Conference on Semantics Systems*, Vienna, Austria, September 2022.
29. Arnold POS Vermeeren, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. User experience evaluation methods: current state and development needs. In *Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries*, pages 521–530, 2010.
30. Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. Question answering over knowledge graphs: Question understanding via template decomposition. *Proc. VLDB Endow.*, 11(11):1373–1386, jul 2018.