An Investigation of the Accuracy of Knowledge Graph-base Search Engines: Google knowledge Graph, Bing Satori and Wolfram Alpha

* Farouk Musa Aliyu and Yusuf Isah Yahaya

Abstract— In this paper, we carried out an investigation on the accuracy of two knowledge graph driven search engines (Google knowledge Graph and Bing's Satori) and a computational knowledge system (Wolfram Alpha). We used a dataset consisting of list of books and their correct authors and constructed queries that will retrieve the author(s) of a book given the book's name. We evaluate the result from each search engine and measure their precision, recall and F1 score. We also compared the result of these two search engines to the result from the computation knowledge engine (Wolfram Alpha). Our result shows that Google performs better than Bing. While both Google and Bing performs better than Wolfram Alpha.

Keywords — Knowledge Graphs, Evaluation, Information Retrieval, Semantic Search Engines..

1 INTRODUCTION

Search engines have played a significant role in helping web users find their search needs. Most traditional search engines answer their users by presenting a list of ranked documents which they believe are the most relevant to their user's query. Sometimes, these documents may or may not tally with the user's need. Consequently, some traditional search engines try to understand the exact information users are seeking by building a graph of entities and their relationships which help them to disambiguate user's query and answer user's query by giving a direct answer to a query at one spot and therefore taking search engines a step ahead, just like some inference engines or computational engines like Wolfram Alpha does. Two of these kinds of search engines are Google's Knowledge Graph (GKG) and Bing's Satori. They achieved this by building an entity knowledge graph that store information about real world entities and their relationships. The question is: are the results from these search engines always correct? or how reliable are the result from these search engines? In this paper, we investigated the accuracy of these search engines. We used a dataset consisting of list of books and their authors. We constructed queries to retrieve the author(s) of a book given the book's name. We evaluate the result from each search engine and measure their precision and recall. We also compare the result of these two search engines to the result from a computation knowledge engine (Wolfram Alpha). Our result shows that Google performs better than Bing. While both Google and Bing performs better than Wolfram Alpha. The rest of the paper is organized as follows: In section 2, we discuss the works in the literature that relates to ours, in section 3 we explained the methodology we used to carry out the experiment. In section 4 we discuss our results and findings and then conclude in section 5.

1.1 Research Questions

The questions we intend to answer in this research include:

1. What are the accuracy of semantic search engines leveraging knowledge graphs or how reliable are the result outputted by the semantic search engines?
2. How are the accuracies of semantic search engines as compare with computational knowledge engines?

.

2 RELATED WORKS

In any new information retrieval system or method, one of the most common components needed is regular and standardized evaluation in order to determine the progress or weakness of the system [1]. This is increasingly important as major search engines continue to move from their traditional text base to semantic base by incorporating large amount of semantic data (entities and their relationship) into their system.

There are a lot of works that evaluate and compare search engines, however, there is little effort by scholars to evaluate search engines incorporating knowledge graph into their search system. This may be due to the infancy of the technique. Our work [2] was the first to investigate the semantic aspect of Google knowledge graph and Bing's Satori after their introduction [3],[4] in 2012. In the work, we investigated the coverage of entity types, the capabilities of their natural language query interfaces and the extent of their support for list search services. Our findings shows that only common entity types were covered by the two search engines as of then and they only support queries of simple or moderate complexities. Also list search service is provided for a small percentage of entity types.

Some works on semantic search has focus on building a standardize evaluation mechanism for entity/object retrieval system. As a first step towards a standardized methodology, Pound et al. [5] defined the task of Ad-hoc Object Retrieval, where semantic search is considered to be the retrieval of objects represented as Semantic Web data, using keyword queries for retrieval. In [6], the first evaluation campaign that specifically targets the task of ad-hoc object retrieval was proposed. Blanco et al [7] developed an evaluation framework

for semantic search and show that the framework is repeatable and reliable.

Zhao et al [8] developed evaluation framework for the evaluation of Google as a Question Answering (QA) system rather than a typical search engine. The question answering feature was introduced by Google in 2012 [8] as a feature snippet [9] displayed in their search engine result page (SERP). The content in the answer box might be an exact answer generated from Google Knowledge Graph or a featured snippet extracted automatically from a webpage [8]. Their study focused on the evaluation of Google QA quality across four target types and six question types. Their finding shows that Google provides significantly higher-quality answers to person related questions than to thing-related, event-related and organization-related questions. Google also provided significantly higher-quality answers to where- questions than to who, what and how questions. The more specific a question is, the higher the QA quality would be. In contrast, our work evaluates not just Google but Bing's Satori and wolfram Alpha. Moreover, our work focuses on the evaluation of the accuracy of result from the search engines knowledge graph and not the question type or target types they support.

Strzelecki and Rutecka [9] examined the direct answer feature capability of Google search engine that is provided as a feature snippet in their SERP displayed from their Google Knowledge Graph (GKG). Their study focused on three issues viz: What is the expected length of keywords for triggering direct answers?, What grammar forms are significant in direct answers? and How important are keywords in URL and answer content in choosing websites as reliable sources for direct answers? Their finding shows that Keywords should be built in the form of short, two-to-four-word sentences comprising the subject and its attribute. Using relative pronouns, articles, and prepositions, as well as using questions as queries, can help to properly define a query and display the best direct answer. Their main aim is to help search users and web master to understand and utilize more about Google direct answer. Our work focused on the accuracy of the results by Google and other search engines that is coming from Knowledge Graph.

3 RESEARCH METHODOLOGY

The method we adopted to investigate the accuracy of these semantic search engines is to manually formulate and test queries with known result in each of the semantic search engine. We use human judgment to evaluate each result from the search engines using three metrics. In order to be consistent and unbiased in our judgment, one human judge was selected for the evaluation. Also the searches and evaluations were carried out in minimal non-distant time frame.

3.1 Data Set

We use the dataset from [10] consisting of 149 book names and there corresponding authors. For each book-author pair, we constructed a query that will retrieve the author(s) of the book by appending the phrase "who is the author of" and <*book name*>. Eg. For the book "A Tale of Two Cities" the query will be "who is the author of a Tale of Two Cities". **Figure 1** shows a section of the dataset. In total, we run 149 queries in each of these search engines and note down the result from each.

	S/	Books	Authors	Query	Result In GKG	GKG Eva	Result in Satori	Satori	Result in Walfram	Walfr	n Alpha Eval.
2	1	My experiments with Truth	Mahatma M.K.Gandhi	Author of My experiments with Truth	Mahatma Gandhi	2	Mahatma Gandhi	2	No result from Walfram A	1	
3	2	Far from the Madding Crowd	Thomas Hardy	Author of Far from the Madding Crowd	Thomas Hardy	2	Thomas Hardy	2	Thomas Hardy	2	
4	3	Geetanjali	Rabindra Nath Tagore	Author of Geetanjali	Rabindranath Tagore	2	Rabindranath Tagore	2	43390 people (2009)	0	
5	4	One Day in the Life of Ivan Denisovitch	Alexander Solzhenitsyn	Author of One Day in the Life of Ivan Denisovi	Aleksandr Solzhenitsyn	2	One Day in the Life of Ivan Denis	0	Alexander Solzhenitsyn	2	
6	5	The Merchant of venice	William shakespeare	Author of The Merchant of venice	William Shakespeare	2	William Shakespeare	2	William Shakespeare	2	
7	6	The Moon and Six pense	Somerset Maughan	Author of The Moon and Six pense	W. Somerset Maugham	2	The Moon and Sixpence Entity w	0	43390 people (2009)	0	
8	7	Pilgrim's Progress from this world to that v	John Bunyan	Author of Pilgrim's Progress from this world to t	John Bunyan	2	John Bunyan	2	No result from Walfram A	1	
9	8	A Tale of Two Cities	Charles Dickens	Author of A Tale of Two Cities	Charles Dickens	2	Charles Dickens	2	Charles Dickens	2	
10	9	Utopia	Sir Thomas Moor	Author of Utopia	Thomas More	2	Entity...Utopia: UK TV Series	0	Sir Thomas More	2	
11	10	Origin of species	charles Darwin	Author of Origin of species	Charles Darwin	2	Charles Darwin	2	Charles Darwin	2	
12	11	David Copperfield	Charles Dickens	Author of David Copperfield	Charles Dickens	2	Charles Dickens	2	Charles Dickens	2	
13	12	A passage to India	E.M.Forster	Author of A passage to India	E. M. Forster	2	E. M. Forster	2	E. M. Forster	2	
14	13	Gulliver's Travels	Jonathan Swift	Author of A Gulliver's Travels	Jonathan Swift	2	Jonathan Swift	2	Jonathan Swift	2	
15	14	Discovery of India	Pandit Jawaharlal Nehru	Author of Discovery of India	Jawaharlal Nehru	2	Jawaharlal Nehru	2	Space Shuttle Discovery missic	0	
16	15	The Vicar of Wakefield	Oliver Goldsmith	Author of The Vicar of Wakefield	Oliver Goldsmith	2	The Entity....The Vicar of wa	0	Oliver Goldsmith	2	
17	16	The Decline and Fall of the Roman Empir	Edward Gibbon	Author of The Decline and Fall of the Roman E	Edward Gibbon	2	Edward Gibbon	2	Western Roman Empire decline	0	
18	17	The Lady of the Last Minstrel	Sir Walter Scott	Author of The Lady of the Last Minstrel	Walter Scott	2	The Entity...The Lay of the Last M	1	minstrel (English word)	0	
19	18	Pride and Prejudice	Jane Austen	Author of Pride and Prejudice	Jane Austen	2	Jane Austen	2	Jane Austen	2	
20	19	Time Machine	H.G. Wells	Author of Time Machine	H. G. Wells	2	A list of authors: Time Machine (r	2	H. G. Wells	2	
21	20	Arthashastra	Kautilya	Author of Arthashastra	Chanakya However, a search	2	Chanakya	2	Chanakya	2	
22	21	Le Contract Social	Jean Jacques Rousseau	Author of Le Contract Social	Jean-Jacques Rousseau	2	Jean-Jacques Rousseau	2	No result from Wolfram Alpha	1	
23	22	Avigyan Sakuntalam	Kalidas	Author of Avigyan Sakuntalam	Kālidāsa	2	No result from Satori	1	43390 people (2009)	0	
24	23	Anand Math	Bankimchandra Chattopadhy	Author of Anand Math	Bankim Chandra Chattopadh	2	No result from Satori	1	Bimal Dutta | Hrishikesh Mukher	0	
25	24	Mein Kampf	Adolf Hitler	Author of Mein Kampf	Adolf Hitler	2	Adolf Hitler	2	Adolf Hitler	2	
26	25	Ain-i-Akbari	Abul Fazal	Author of Ain-i-Akbari	Abu'l-Fazl ibn Mubarak	2	Abu'l-Fazl ibn Mubarak	2	43390 people (2009)	0	
27	26	Akbar-Nama	Abul Fazal	Author of Akbar-Nama	Abul Fazl	2	Entity...Akbarnama No author at	1	George Lucas (1944-)	0	

Sheet1 Sheet2 Sheet3

Figure 1 A cross section of the queries and the result from the search engines.

3.2 Evaluation Scale

We use three evaluation scale i.e. (relevant, no-result and irrelevant) to evaluate each result after comparing it with the known answer from the dataset. For simplicity, we used the number '2' for a relevant result, '1' for queries without result and '0' for irrelevant results. While evaluating the retrieved documents the following criteria were used to rank the result:

Relevant (2): if the result from the search engine contains the expected result and in the right context i.e. the correct author of the book searched for, either as a direct answer shown in the regular search engine result page (SERP) i.e. as attribute result or on the knowledge graph result panel (entity result). We mark the result from the search engines as relevant if it tallies with the result from our dataset. However, we are not strict at enforcing naming convention for authors. For example we take the following result to be the same: "G.B.Shaw", "George Bernard Shaw" or "Shaw". Sometime the search engines may display the result both on the SERP and on the knowledge graph result panel. For Example, a search for "Author of The age of Reason" in Bing display Thomas Pain as one of the attributes of the book on top of the SERP and again Thomas Pain as an entity on the knowledge graph result panel as shown in Figure 2.

No result (1): If the knowledge engine doesn't show any result for the query, either as a direct answer in the SERP or in the Knowledge graph display area.

Irrelevant (0): if the result displayed from the knowledge engine contains the wrong author or any other information contrary to the expected answer such as displaying the book in the knowledge graph panel instead of the author for a query. i.e if the search engine return as an entity the book that was searched. This clearly means the search engine does not understand the query.

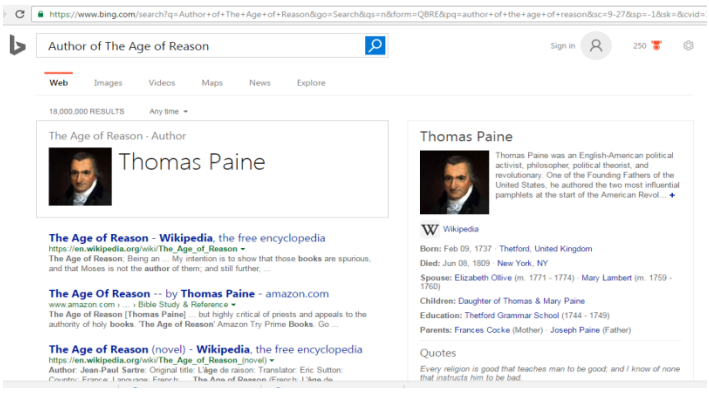

Figure 2 Result of the search "Author of The Age of Reason" in Bing which displays the result on both SERP and the knowledge graph result panel

3.3 Evaluation Metrics:

We use set retrieval evaluation metric [11] (precision, recall and F1 score) to evaluate the systems rather than ranked retrieval evaluation metric (P@k, NDCG, MAP) since most of the results of our query set are expected to produce a single result from all the semantic search engines and not a rank list of items.

4 RESULTS AND DISCUSSION

In this section, we report the findings of our investigation.

From the result in Table 1, Google produces a total of 124 relevant results, 13 irrelevant or wrong result and 12 out of the 149 queries failed to produce any output. Satori on the other hand produces a total of 97 relevant, 29 irrelevant and 23 queries has no outcome. Lastly, Wolfram Alpha has a total of 74 relevant, 45 irrelevant and 30 queries without result.

TABLE 1 The Result Of Testing Our 149 Book Queries The Search Engines.

Search Engine	Book Queries		
	Relevant	Irrelevant	No Result
GKG	124	13	12
Satori	97	29	23
Wolfram Alpha	74	45	30

It can be observed clearly that the number of irrelevant results in all the three systems surpasses that without output in the systems. This indicates that there is significant amount of noise in all the search systems. For example, a search for "Author of The Age of Reason" in Bing gives 'Thomas Paine' as the result (Figure 2). Google gives similar outcome for the same query. However, their result is wrong as the author of The Age of Reason is Jean Paul Sartre from our data. A search for " Jean Paul Sartre books" gives a list of books including The Age of Reason in both Google and Bing. Figure 3 shows the result for Google.

Table 2 gives the performance of the search systems base on precision, recall and F1 scores. The result indicated that the semantic search engines performs better that the computational engine. This may be due to the larger amount of entities indexed by the semantic search engines and also the algorithms they use. However, we expect wolfram Alpha to perform better for queries that requires some computation or aggregation of data since it was originally design to perform some computations. For example, the query "How many rivers are found in Colorado?" require the search engines to perform

some computations. First a list of rivers in Colorado need to be obtained and then counted to give the final answer of the query. This type of query is not well supported by the semantic search engines.

TABLE 2 Result showing the precision, recall and F1 score of the engines.

Search Engine	Book Queries		
	Precision	Recall	F1 Score
GKG	0.9051	0.8322	0.8672
Satori	0.7698	0.6510	0.7054
Wolfram Alpha	0.6218	0.4966	0.5522

Overall Google performs better with an F1 measure of 0.8672 follow by Satori with F1 measure of 0.7054 and lastly Wolfram Alpha with F1 score of 0.5522 as shown in Table 2. The result shows that Wolfram Alpha trails behind the two semantic search engines, as it has the highest number of incorrect result and queries without any result. Figure 4 gives a picture of the results obtained from each search engine including number of queries not having any result in the search engines, the number of incorrect result and the number of queries with accurate result. The reason behind the low performance of wolfram Alpha may be due to lower number of entities indexed by the knowledge engine compared to the semantic search engines while the low accuracy may be as a result of the inability of the knowledge engine to parse some natural language queries:

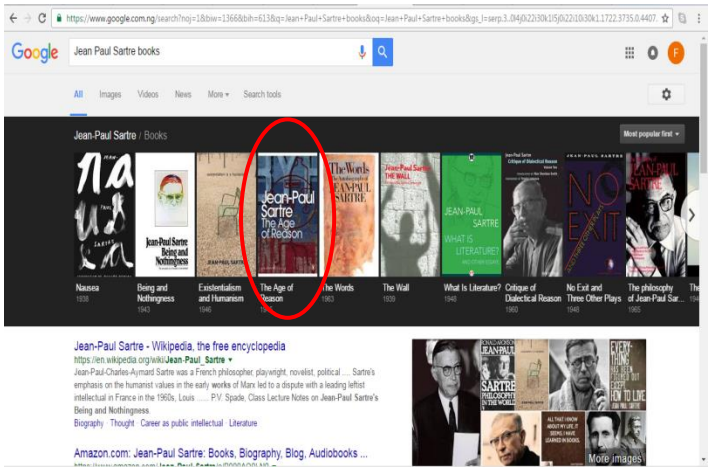

Figure 3: The result for searching Jean Paul Sartre books in Google

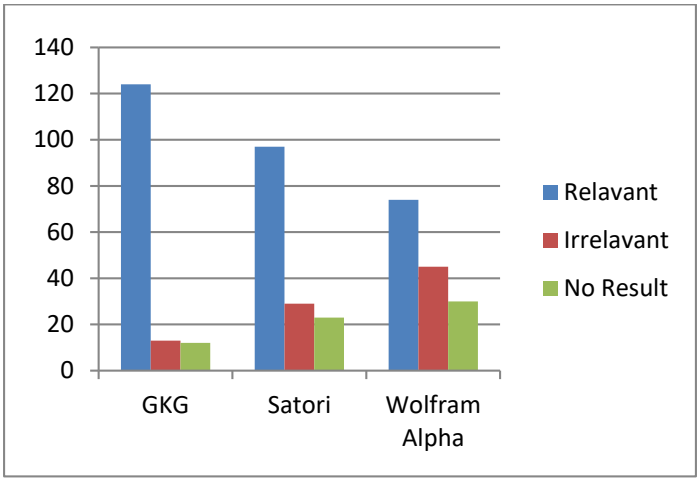

Figure 4: A chart displaying the result.

Our result indicated that the search engines are doing great in giving direct answers to user's query. However, there is need for the search engines to make improvements in their system for better and more accurate result. First, the search engines can improve in their natural-language processing technology. This will allow the search engines to parse user's query accurately and hence a good understanding of the query. Secondly, the search engines can improve their systems by indexing new entities and update information for the existing ones. Thirdly, the search engines can improve user's satisfaction by supporting more query types e.g. aggregate query and other complex queries.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the accuracy of results from Google Knowledge Graph (GKG), Bing's Satori and Wolfram Alpha. We use a dataset from Gutenberg consisting of 149 books and their authors and manually constructed queries to retrieve the author of a book given its name. The result of testing the queries shows that the search engines are doing great in directly answering user's query. However, a good number of results from the search engines indicated that there is still noise in their data which gives misleading result for user's query. Our investigation also shows that semantic search engines like Google Knowledge Graph and Satori performs better than computation engines like Wolfram Alpha in terms of giving direct answer with regards to searches about entity. Overall, Google performs better than Bing. While both Google and Bing performs better than Wolfram Alpha.

Suggestions were given to information professionals on how to improve their search engines to better increase their user's satisfaction. The search engines need to improve their natural-language processing ability, expand their systems by indexing new entities and updating information for the existing ones. Lastly, the search engines can improve user's satisfaction by supporting more query types e.g. aggregate query and other complex queries.

This study has several limitations. First the queries used are of the same type, as such the result could not be used to generalize the overall performance of the systems. Other query type can be used that retrieve different type of entity data such as attributes, relation and entity list. Also, the performance of the systems can be tested with more complex queries. It will also be good to know how the search engines perform on real world queries.

In the future, we wish to investigate the accuracies of the search engines with real world queries and for queries that retrieve multiple entities. We hope to also study the accuracies of the search engines for complex queries. In addition, the quality of information provided by these search engines as results for a given set of queries can be investigated. For example, while GKG and Satori often gives some set of attributes of entity and their values which they think users are mostly interested about, Wolfram Alpha often gives statistical information about entity in addition.

REFERENCES

[1] Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., & Duc, T. T. (2011, July). Entity search evaluation over structured web data. In Proceedings of the 1st international workshop on entity-oriented search workshop (SIGIR 2011), ACM, New York (Vol. 14, pp. 2181-2187).

[2] Uyar, A., & Aliyu, F. M. (2015). Evaluating search features of google knowledge graph and bing satori. Online Information Review.

[3] Qian, R. (2013), "Understand your world with Bing", 21 March, available at: www.bing.com/blogs/site_blogs/b/search/archive/2013/03/21/satorii.aspx

[4] Singhal, A. (2012), "Introducing the knowledge graph: things, not strings", 16 May, available at: http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.

[5] J. Pound, P. Mika, and H. Zaragoza. Ad-hoc Object Ranking in the Web of Data. In WWW, pages 771–780, Raleigh, USA, 2010

[6] Halpin, H., Herzig, D. M., Mika, P., Blanco, R., Pound, J., Thompon, H., & Tran, D. T. (2010, November). Evaluating ad-hoc object retrieval. In IWEST@ ISWC.

[7] Blanco, R., Halpin, H., Herzig, D. M., Mika, P., Pound, J., Thompson, H. S., & Tran, T. (2013). Repeatable and reliable semantic search evaluation. Journal of web semantics, 21, 14-29.

[8] Zhao, Y., Zhang, J., Xia, X., & Le, T. (2019). Evaluation of Google question-answering quality. Library Hi Tech.

[9] Strzelecki, A., & Rutecka, P. (2020). Direct Answers in Google Search Results. IEEE Access, 8, 103642-103654.

[10] Lahiri, Shibamouli (2014). Gutenberg Dataset. http://web.eecs.umich.edu/~lahiri/gutenberg_dataset.html accessed June, 2018.

[11] Jaime Arguello (2013). Evaluation Metrics. https://www.coursehero.com/file/33942586/10-EvaluationMetricspdf/ accessed March, 2020