The Factoid Queries Collection

Ido Guy
Ben Gurion University of the Negev, Israel
Yahoo Research, Israel
idoguy@acm.org

Dan Pelleg
Yahoo Research, Israel
pellegd@acm.org

ABSTRACT

We present a collection of over 15,000 queries, issued to commercial web search engines, whose answer is a single fact. The collection was produced based on queries landing on questions within a large community question answering website, each with a best answer no longer than 3 words and an explicit reference to a Wikipedia page. We describe the collection generation process and provide a variety of descriptive characteristics, demonstrating the collection's uniqueness compared to existing datasets and its potential use for research of factoid question answering and retrieval.

Keywords: dataset; fact retrieval; factoid question answering

1 INTRODUCTION

In recent years, commercial search engines have started to provide inline results to factoid questions, which present a short answer (e.g., a number, a unit, or a few words) directly on the search results page (SERP) [5]. This is part of a broader trend of presenting search results, such as weather, sport scores, or dictionary definitions, directly on the SERP, sparing the user the need to click. Commercial search engines put considerable effort to satisfy this kind of need, usually under the "card" paradigm, where a dedicated module is built to respond to a particular type of query [7].

Despite the increasing popularity of such fact retrieval, little has been published about the technology behind it, and, in particular, little is known about how web search users formulate factoid queries and express "factual" information needs. Previous research collections synthesized queries by making assumptions on the actual query distribution; most noticeably, these queries were generally thought of as starting with one of the WH-question words[1] [3, 4, 6]. Yet, as we will later show, this common perception is inaccurate, as many factual queries "in the wild" do not match this pattern. In addition, synthesis of the queries, however accurate, gives no information about their distribution. Without click logs, there is no way to determine which facts are more popular, and what phrasing is more commonly used to query for them.

In this paper, we present the *FactQueries* collection[2]: a collection of 15,557 queries that landed on factoid question pages within the Yahoo Answers Community Questions Answering (CQA) website. Factoid questions were identified by considering questions with best answers that include no more than 3 words and an explicit link within the "source" field to English Wikipedia. The factuality of the questions and the correctness of the corresponding answers were validated by manual judgment. We then fetched, for each factoid question, up to 5 queries from commercial web search engines that landed on the question's page. Each entry in the collection includes a query, the corresponding question's title and best answer, an integer indicating the query's landing frequency on the question's page, the answer's source field, and a link to the question's page on Yahoo Answers.

The *FactQueries* collection reflects queries by web search users that are formulated to retrieve factual answers. To the best of our knowledge, it is the first to include web search queries that express factual intent and are not restricted to predefined patterns. The collection offers the following contributions:

- It includes queries from commercial web search engines, exactly as they were generated by users. Over 70% of the questions have multiple landing queries, demonstrating the diverse ways users may query for the same fact.
- It includes an indication of the frequency of each query.
- Nearly 60% of the queries are not phrased as WH-questions, reflecting the difference between query language and question language.
- As opposed to other datasets, the generation process of *FactQueries* did not apply any restrictions on the question or query, but rather on the answer. It therefore reflects a broader set of topics and questions for facts web users are interested in, which span the diverse categories of Yahoo Answers, from Science and Mathematics to Music and Entertainment.
- While the collection includes a link to the knowledge source (Wikipedia), it is not restricted to facts that map to a knowledge graph structure [3, 4]. On the other hand, the collection's facts are not as complex as trivia questions that deliberately combine multiple facts [8], as these are not typically pursued by web search users.
- The correctness of the answer is validated both by the Yahoo Answers "best answer" mechanism and in-house annotators, reducing the likelihood of wrong answers, such as recently reported for existing factoid datasets [2].

[1] Who, where, why, when, how, what, and which.

[2] http://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=76

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914676

2 COLLECTION GENERATION

The collection was generated based on questions from the Yahoo Answers (YA) CQA website, which satisfy the following two preconditions:

1. The text of the question's "best answer" contains between 1 to 3 words (based on white-space tokenization). Note that this precondition implies that the question has a best answer, chosen either by the asker or by the community [1].
2. The "source" field of the best answer includes a link to English Wikipedia. Note that this precondition implies that the question has a non-empty source field (Yahoo Answers allows answerers to add their source(s) in an optional field below the answer's text).

We conjectured that these two stipulations would yield a subset of answers likely to be factual. Applying these two preconditions on the entire set of Yahoo Answers non-deleted English questions, posted between 2006 and 2014, produced a set of over 31,000 different questions. We applied additional rule-based filtering to detect answer text that does not contain the actual answer, including reference pointing (e.g., answers that contain "try", "check", "go to", "wikipedia", "http://", "visit", "hope this helps", or "there you go"), multiple choice selection (e.g., "yes", "nope", or starting with "2.", "c ", or "3)"), and expression of doubt (e.g., "think", "believe", "maybe", "probably", "i"). This filtering left us with nearly 19,000 questions. For 7,500 of these questions, we had a non-empty set of "landing queries", i.e., queries from commercial web search engines that resulted in a click on their YA page. The landing queries we inspected were issued between 2012 and 2014.

We sent this set of 7,500 questions to manual judgment by 25 professional in-house annotators ("editors"). Each question was evaluated by a single editor. The editors were given the question's title, its (up-to-3-words) best answer, the link to the question's page, and the link(s) provided in the source field of the best answer. They were asked the following questions:

- Q1: Does the entry describe a question that has exactly one factual answer, which can be expressed in a few words?
- Q2: If Q1 is affirmative, is the provided answer correct?

In addition, the editors were given the following notes with regards to Q1:

- We seek facts, not opinions. So for a question like "What is the best book series by J.K. Rowling?", Q1 should be negative.
- If the answer requires a longer explanation than a few words, Q1 should be negative.
- For classic Yes/No questions, Q1 should be negative (e.g., "Is Sidney the capital of Australia?").
- We are not looking for lists. If there is more than one answer, Q1 should be negative (e.g., "What is the cast of Star Wars?" or "Who are Barack Obama's Children?").

For determining the answer's correctness (Q2), editors were instructed to use Wikipedia and, if needed, a web search.

Overall, for 79.7% of the 7,500 questions given to editors, Q1 was affirmative, i.e., they were indicated to contain a factual question (19.8% negative, 0.5% "don't know"). This high portion indicates that our methodology for extracting factual questions is effective, to the degree of achieving almost 80% precision without human interference. Out of the questions indicated to be factual, for 86.4% the answer was indicated to be correct (Q2 affirmative). Table 1 lists a few examples for question titles marked non-factual (Q1 negative), while table 2 lists a few examples for question titles marked as factual, but with incorrect answers (Q1 affirmative and Q2 negative). For the collection, we only considered questions for which both Q1 and Q2 were affirmative (68.9% of the original set of questions).

question	editor's comment
Who was the african american that discovered penicillin?	question is wrong
How much do 3ds cost on black friday?	no definitive answer
What was Alexander Graham Bell's Most unknown invention?	opinion-based
What is it called when a person thinks they are always right?	no definitive answer
Who Played Michael Jackson in The Jacksons: American Dream?	different actors for each age
Lands discovered between 17-19 century?	list
What is an RDF file, and how do I use it?	detailed answer required
How old is Ke Huy Quan?	time-dependent
Where was ezra J. warner born?	two persons with that name
Did 50 cent real get shot 9 times?	yes/no question

Table 1: Example questions marked non-factual.

question	best answer	editor's comment
How many seasons does the cartoon winx have?	3 seasons.	answer is outdated
Who invented the muffin?	Fannie Merritt Farmer	no definitive answer
World War 2 Total Deaths?	Approx. 72 million	estimated number
What is the legal smoking age in italy?	it is 14.	should be 18

Table 2: Example questions marked factual but incorrect.

Figure 1 shows the portion of questions marked as factual (Q1 affirmative) and answers marked as correct (Q1 and Q2 affirmative) according to the answer length in words. The portion of factual answers decreases from 82.4% for 1-word and 81.1% for 2-word answers, to 74% for 3-word answers, indicating that as answers become longer, they are less likely to be factual.

After selecting our subset of questions based on the feedback from annotators, we sampled a list of up to 5 landing queries per question. We filtered out a small subset of the queries, such as queries that included "yahoo answers" or very long queries. This left us with a set of 15,557 queries, referring to 4,691 distinct YA questions. Each query is represented as a tab-separated line in the collection, with the following fields:

1. **Query**. The query's text as logged by the search engine.
2. **Occurrence Score (*OcSc*)**. An integer between 1 and 10 representing the occurrence frequency of the query's landing on the YA page (a larger integer indicates a higher landing frequency)[3].
3. **Question Title**. The title of the question in YA.
4. **Answer**. The best answer for the question (1-3 words).
5. **Source**. The source field for the answer, which includes at least one URL of an English Wikipedia page.
6. **URL**. The URL of the Yahoo Answers page. In addition to the information directly included in this collection, the Yahoo Answers page may provide additional information, such as the question's description, the question's

[3]Generated by manual bucketing, to obscure sensitive data.

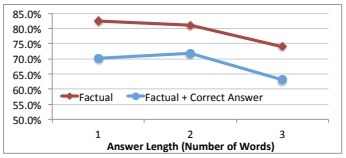

Figure 1: Percentage of questions marked factual and answers marked correct by answer length.

Question's Title	Best Answer	Query	*OcSc*
What is altitude of Ouray, Colorado?	7,792 ft	ouray colorado elevation	8
		ouray altitude	7
		altitude of ouray	2
		altitude ouray	1
		ouray altitude co	1
What is it called when you dance on a ribbon in the air?	Aerial silk	what is it called when you dance on a ribbon	8
		What is ribbon dancing called	7
		ribbon dancing in air	2
		what is dancing with ribbons in the air called	1
		air ribbon dancing	1
What episode of the simpsons does homer smoke weed?	Weekend at Burnsie's	simpsons weed episode	10
		homer smokes weed	10
		simpsons weed episodes	6
		the simpsons episode homer smokes weed	6
		homer simpson smoke weed	5

Table 3: Collection examples.

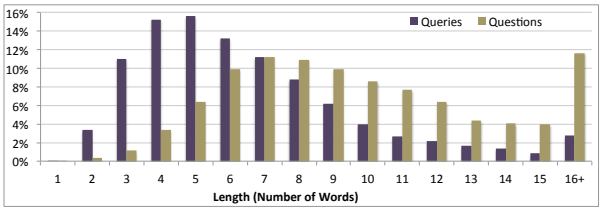

Figure 2: Query and Question Length Distribution.

category, other answers, votes on the answers, links to the profile page of the asker and answerers, and more.

It should be noted that the matching of the landing queries to their corresponding questions was not manually verified. We generally observed a good match across the collection, but the *OcSc* can be used for filtering to queries with more "landing evidence".

Table 3 includes a few examples from the collection. The examples demonstrate differently-phrased queries that landed on the same YA page for three different questions.

3 COLLECTION CHARACTERISTICS

In this section, we describe the characteristics of the collection. We first focus on queries and then discuss questions.

3.1 Queries

The average length of a query in the collection is 6.6 words (stdev: 3.5, median: 6, max: 53). Figure 2 shows the detailed distribution. The portions of 1-word queries (0.02%) and 2-word queries (3.33%) are very low, whereas 11.4% of the queries contain more than 10 words. This distribution of query length is in accordance with previous findings reported for queries that landed on CQA websites [9].

Table 4 shows the distribution of queries per question in the collection. As can be seen, 28.4% of the questions have only one landing query, while almost half have 5 queries, which is the maximum number allowed for this collection.

The next analysis focuses on the portion of question queries [11]. The "queries" row of Table 5 shows the distribution of queries starting with a WH-word. The most popular WH-word, by a large margin, is "what", opening over 20% of all queries in the collection, followed by "who" and "how" (half of the latter start with "how many"). On the other hand, not even one query starts with "why", which typically represents broader, open-ended types of questions [10].

1	2	3	4	5
28.4%	11.5%	7.1%	6.0%	47.0%

Table 4: Queries per Question Distribution.

	What	Who	How	When	Which	Where	Why	Total
Queries	20.9%	7.0%	5.4%	3.1%	2.9%	1.6%	0	40.9%
Questions	42.9%	12.8%	5.7%	4.2%	4.2%	1.7%	0.1%	71.6%

Table 5: Distribution of queries and questions starting with WH-words.

	1	2	3	4	5	6	7	8	9	10
All	62.9%	11.4%	4.8%	4.8%	3.6%	3.1%	2.7%	2.7%	2.4%	1.6%
1 query	83.4%	8.6%	2.8%	1.4%	0.8%	0.7%	0.6%	0.5%	0.8%	0.5%
5 queries	55.4%	11.9%	5.5%	6.0%	4.6%	4.0%	3.7%	3.6%	3.3%	2.1%
5 queries, 1st	7.9%	12.1%	8.3%	11.5%	10.3%	9.9%	10.5%	10.6%	10.7%	8.3%
5 queries, 5th	97.5%	1.7%	0.5%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 6: Distribution of query occurrence score.

As a sanity check to the collection generation process, we also inspected yes/no queries. Overall, very few queries follow a yes/no pattern [11]: no queries start with "was", "did", "does", "are", "do", "will", "would", "should", "can", or "could". Very small portions of the queries start with "is" (4 in total) and "were" (10 queries), with many of the latter resulting from misspelling of the more common "where". In total, it can be seen that only slightly over 40% of the queries are question queries, leaving nearly 60% of the queries not phrased as classic questions. Considering queries that contain a WH-word anywhere in the text, although these are often not WH-questions (e.g., "person who studies music"), the portion grows to 47.6%. Other common words to open a query in the collection are "the" (4.4% of the queries), "a" (1.6%), "name" (0.8%), "in" (0.7%), "movie" (0.7%), and "first" (0.4%).

As mentioned in Section 2, the occurrence score, marked *OcSc*, represents the frequency of the query in our inspected log, and ranges from 1 to 10. Table 6 shows the distribution of the *OcSc* across all queries ("All" row). The majority of the queries (62.9%) have an *OcSc* of 1, while lower portions have higher *OcSc*, down to 1.6% for *OcSc* of 10. For questions with 1 query, as can be seen in the respective row of Table 6, the *OcSc* is typically lower, while for questions with 5 queries, the *OcSc* is higher across all queries. The two lowest rows of Table 6 show the *OcSc* distribution of the 1st query (the one with highest *OcSc*) and the 5th query (lowest *OcSc*) for questions with 5 landing queries. It can be seen that the *OcSc* for the 1st query is distributed roughly evenly across the 10 buckets, while the *OcSc* score of the 5th query is all but minimal.

3.2 Questions

The average question title length is 10 words (stdev: 4.2, median: 9, max: 24). The full distribution is shown in Figure 2, as compared to the query length distribution. The correlation between the question length and the average length of its landing queries is positive, but not extremely high ($r=0.36$, $p<0.001$). As can be seen in Table 5 (bottom row), a substantially higher portion of the questions (over 70%) start with a WH-word, with "what" alone opening 42.9% of the questions, and "why" opening as few as 4 questions.

As previously mentioned, the source field of the best answer is provided as part of the collection. While this field usually contains a single URL, the user interface does not strictly impose it and in some cases the field may include multiple URLs and even some free-form text. In rare cases (2.9% of the questions), the field is null due to various processing issues. For the rest of the questions (4,555 in total), 95.5% have exactly one URL in their source field (linking to English Wikipedia), while 4.5% have multiple URLs (up

	Zero	Avg	Std	Max
# other answers	32.1%	2.4	3.5	55
description length (words)	42.9%	16.2	31.6	665
# best answer's upvotes	65.8%	0.56	1.11	26
# best answer's downvotes	91.6%	0.11	0.44	6

Table 7: Additional question characteristics

Category Name	% Questions	Freq-Ratio
Entertainment & Music	31.7%	**2.23**
Science & Mathematics	22.0%	**3.98**
Education & Reference	9.0%	1.96
Arts & Humanities	8.6%	**2.44**
Society & Culture	4.3%	0.56
Sports	3.8%	0.91
Politics & Government	3.5%	0.73
Health	2.3%	0.29
Travel	2.2%	1.03
Computers & Internet	2.1%	0.36

Table 8: Most common top-level categories.

to 4, except one case with 7), to a total of 455 URLs. Of these, 67% link to English Wikipedia, while the rest link to a variety of websites (114 in total), with `imdb.com` (14 URLs) and `youtube.com` (12 URLs) the most popular. Inspecting all the URLs to English Wikipedia within the source field, including the cases where the field contains multiple URLs (a total of 4,651 URLs), we see that they cover 4,391 different Wikipedia values. The vast majority of these values (95%) appear exactly once in *FactQueries*, 4.2% appear twice, 0.7% appear 3 times, and 0.1% (5 values) appear 4 times in the collection: Adenosine triphosphate, Cellular respiration, DNA, List of Bleach episodes, and Generation Y.

Table 7 shows several statistics of question metadata not directly included in the collection, including the number of answers aside from the best answer, the length of the question's description (body), and the number of upvotes and downvotes the best answer received.

Questions on Yahoo Answers are assigned to categories, in a 3-level taxonomy with over 1500 nodes [1]. The collection's questions span all 26 top-level categories of Yahoo Answers. Table 8 shows the 10 most common top-level categories, alongside the corresponding portion of questions in the collection. The two most common categories, jointly covering over 50% of the questions, are Entertainment & Music and Science & Mathematics. The top-level categories with the smallest portions in the collection are Pregnancy & Parenting (8 questions), Dining Out (4), and Local Businesses (3).

Since some categories are more frequent than others on YA, it is also interesting to inspect the frequency of the category in the collection relative to its general YA frequency. The "Freq-Ratio" column of Table 8 indicates the ratio between the portion of category questions in *FactQueries* and on YA in general (considering over 150M non-deleted English questions with a best answer). A ratio greater than 1 indicates higher frequency of the category across *FactQueries*' questions than across all YA's questions. The categories with highest ratio are Science & Mathematics (3.98), Arts & Humanities (2.44), and Entertainment & Music (2.23). The three categories with the lowest ratio (not shown in the table) are Beauty & Style (0.12), Pregnancy & Parenting (0.05), and Family & Relationships (0.02).

Thus far, we inspected the top-level categories in the YA taxonomy. To gain a more fine-grained sense of the collection's topics, we also examine the specific categories assigned to questions (any category in the taxonomy, at any level, can be assigned to a question). Table 9 shows the 15 most common categories within *FactQueries*. The most popular categories are Movies, Biology, Celebrities, History, Chemistry, and Comics, reflecting the mix of science and entertainment. The categories with the highest freq-ratio are Geography, Trivia, Comedy, and Biology. The categories with the lowest freq-ratio among the 30 most common YA categories are Polls & Surveys, Hair, Diet & Fitness, Pregnancy, Singles & Dating, Friends, and Marriage & Divorce (the last two without any associated questions in the collection).

Category Name	Top-level Category	% Qstns	Freq-Ratio
Movies	Entertainment & Music	6.5%	5.20
Biology	Science & Mathematics	4.9%	**10.07**
Celebrities	Entertainment & Music	4.8%	5.60
History	Arts & Humanities	4.7%	7.63
Chemistry	Science & Mathematics	4.5%	5.78
Comics & Animation	Entertainment & Music	4.0%	4.98
Words & Wordplay	Education & Reference	3.6%	3.97
Geography	Science & Mathematics	3.0%	**19.70**
Homework Help	Education & Reference	2.3%	1.84
Other - Music	Entertainment & Music	2.3%	2.81
Comedy	Entertainment & Music	2.0%	**10.92**
Books & Authors	Arts & Humanities	1.9%	1.77
Astronomy & Space	Science & Mathematics	1.8%	5.72
Drama	Entertainment & Music	1.7%	6.15
Trivia	Education & Reference	1.7%	**15.16**

Table 9: Most common categories.

4 CONCLUSION

The *FactQueries* collection provides researchers and practitioners with over 15,000 queries used by web search users to find answers to a variety of factual questions. The factual nature of the questions and the correctness of the answers were validated in a large editorial effort. The collection strives to help the community better understand the domain of factoid queries, which plays an interesting role in today's web search.

5 REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proc. WWW*, pages 665–674, 2008.

[2] H. Bast and E. Haussmann. More accurate question answering on freebase. In *Proc. CIKM*, pages 1431–1440, 2015.

[3] J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on freebase from question-answer pairs. In *Proc. EMNLP*, pages 1533âĂŞ–1544, 2013.

[4] Q. Cai and A. Yates. Large-scale semantic parsing via schema matching and lexicon extension. In *Proc. ACL*, pages 423–433, 2013.

[5] A. Chuklin and P. Serdyukov. Good abandonments in factoid queries. In *Proc. WWW Companion*, pages 483–484, 2012.

[6] H. T. Dang, D. Kelly, and J. J. Lin. Overview of the trec 2007 question answering track. In *Proc. TREC*, 2007.

[7] I. Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proc. SIGIR*, 2016.

[8] A. Kalyanpur, S. Patwardhan, B. Boguraev, A. Lally, and J. Chu-Carroll. Fact-based question decomposition in deepqa. *IBM Journal of Research and Development*, 56(3.4):13:1–13:11, 2012.

[9] G. Tsur, D. Carmel, Y. Pinter, and I. Szpektor. Identifying queries with question intent. In *Proc. WWW*, pages 783–793, 2016.

[10] S. Verberne. Paragraph retrieval for why-question answering. In *Proc. SIGIR*, pages 922–922, 2007.

[11] R. W. White, M. Richardson, and W. Yih. Questions vs. queries in informational search tasks. In *Proc. WWW*, pages 135–136, 2015.