Bauhaus-Universität Weimar
Faculty of Media
Computer Science and Media

Construction and Analysis of a Known-Item Question Corpus for the ClueWeb09

Bachelor Thesis

Daniel Wägner

1. Supervisor: Prof. Dr. Benno Stein
Advisor: Dr. Matthias Hagen

Date of submission: June 26, 2013

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, June 26, 2013

. . . .
Daniel Wägner

Abstract

Known-item finding is a strongly personalized task, which involves the recall and retrieval of a previously accessed item based on the searcher's memory of it. Despite being very dependent on the searcher and their re-access strategies, most scientific work in the field relies on methodologies that make unrealistic assumptions about user behavior or don't allow for repeatable research.

As an alternative to existing approaches, this thesis proposes a methodology to generate known-item topic sets from information needs posed by users of a question-answering platform. To ensure that we match the correct items to a query, we only consider answer that have been rated as correct by the asker.

Based on the methodology outlined in this thesis, we have developed an annotated corpus of 2755 known-item questions sourced from the Yahoo! Answers service. Additionally, the corpus contains 240 false memories along with their corrections. To allow for re-use in a scientific context, the known items have been mapped to the publicly available ClueWeb09 corpus.

Contents

Chapter 1

Introduction

In the field of information retrieval, *known-item search* is the common task of re-finding a previously accessed item. Types of known items include personal documents, emails, visited web sites, books in a library or songs heard on the radio.

In contrast with informational or transactional searches, which can have a multitude of viable results, the goal of a known-item search is usually to retrieve a single, specific item (or syntactic/semantic aliases of it) [Bro02]. In some cases a hub that is "one step away from the target [item]" can also be a less desirable, but still acceptable result [Bro02]. An example for such a hub could be the track listing of a music album, with one of the songs being the desired known item.

Consequently, the number of potentially useful results tends to be smaller for known-item queries than for other query types. On the other hand, the user often has a larger amount of information which can be used to narrow down the results of a known-item query. These two points, the number of acceptable results and the available knowledge, are the main factors that separate known-item searches from other search tasks.

While a larger amount of available information can make it easier to re-find a known item, particular attention needs to be paid to incomplete or false memories. Studies have shown that humans remember some kinds of details better than others [BS07; ERJ07; KCF+08]. For example, a user looking for a movie might misremember details about the setting (by thinking that it took place in Ireland, rather than Scotland), the cast (by confusing Danny Glover with Morgan Freeman) or misquote a specific line (Darth Vader never says the exact phrase "Luke, I am your father" during *The Empire Strikes Back*). False memories are problematic in that they can lead to the desired item being excluded from the results of a query.

Current research on the topic of known-item retrieval relies heavily on corpora of known-item queries and their respective known items [HHB+12]. Unfortunately, many of those corpora either

- are proprietary and not publicly available,
- consist of automatically generated queries [ARB07; KC09] or
- consist of queries generated manually from a known item itself, in a in a human computation game [KC10].

Proprietary corpora are unsuitable for further research, as they do not allow for repeatable experiments. Likewise, queries generated from the known item itself, whether automatically or manually, are rather artificial and not representative of real-world user queries. Additionally, they make unrealistic assumptions about human memory, namely

- randomly failing memory in automatic query generation or
- perfect memory in the human computation game, due to the item being seen as is.

To provide an alternative to existing corpora, Hauff et al. proposed the creation of a known-item topic set built from questions posted by users of the Yahoo! Answers platform[1], with the aim to address the lack of public data and the unrealistic approaches to query generation they identified in prior work [HHB+12]. As a proof of concept, Hauff et al. crawled 103 questions by Yahoo! Answers users. Among those, they identified and manually assessed 64 information needs, consisting of 32 website and 32 movie known items.

The goal of this thesis is to expand on the ideas of Hauff et al., to build a larger corpus with a broader coverage of different information needs, suitable for use in further research. Furthermore, we want to analyze the effect of *false memories* in known-item retrieval tasks.

To ensure the usability of our experiments in a broad context, we only examine known items whose URL is included in the ClueWeb09 corpus [CC09; CHY+09]. For non-website items, like movies or books, this is usually their corresponding entry in the English Wikipedia. The ChatNoir search engine [PHS+12] is utilized to map URLs to their respective entry in the ClueWeb09. The end result will be a public, reusable topic set of information needs dealing with known item tasks that can be used in conjunction with the ClueWeb09.

The thesis is structured as follows. Chapter 2 documents the construction the Webis Known-Item Question Corpus 2013, while Chapter 3 provides a thorough analysis of it. Finally, in Chapter 4 we present the conclusion of our work and give an outlook for possible future research.

[1] `http://answers.yahoo.com`

Chapter 2

Construction of the Webis-KIQ-13 Corpus

As discussed in the previous chapter, the existing approaches to constructing known-item corpora tend to yield rather artificial results. The Webis Known-Item Question Corpus 2013 (Webis-KIQ-13) is proposed as an alternative to those corpora, with the goal of being a known-item corpus based on real information needs expressed by real humans. This chapter will detail the process of its construction.

Internet-based community question-answering (cQA) services allow users to pose questions to a group, rate answers by others and receive rewards for providing good answers to open questions. Depending on the service, these rewards can range from monetary payments to a mere increase in reputation within the community. cQA websites can cover specialized fields, such as questions related to computer programming on Stack Overflow[1] and refinding books on Whatsthatbook.com[2], or a broader range of subjects, like Answers.com[3] and Yahoo! Answers[4]. For this thesis, the Yahoo! Answers platform and its public API were chosen as a source of information needs.

2.1 The Yahoo! Answers Platform

Yahoo! Answers was launched on December 13, 2005 as a replacement for Yahoo!'s former Q&A platform, Ask Yahoo!. As of July 3, 2012, Yahoo! Answers is available in 12 different languages with 26 international variants and caters to 250 million unique users worldwide, 54 million of which are from the United States [Yah12]. The service covers information needs for a range of 26 top-level

[1] http://stackoverflow.com/

[2] http://www.whatsthatbook.com/

[3] http://www.answers.com/

[4] http://answers.yahoo.com

Action	Points
Begin participating on Yahoo! Answers	One time: 100
Log in to Yahoo! Answers	Once daily: 1
Ask a question	-5
Choose the best answer for one's question	Points returned: 3
Vote for no best answer, best answer was selected by voters	Points returned: 0
No best answer was selected by voters on one's question	Points returned: 5
Answer a question	2
Have an answer selected as the best answer	10
Vote for an answer	1
Receive a "thumbs-up" rating on a best answer	1 per "thumbs-up" (up to 50)
Receive a violation	-10

Table 2.1: User actions and their associated point values

categories, such as *Entertainment & Music* or *Computers & Internet*, which are further divided into over 300 sub-categories.

Users are able to submit questions expressed in a natural language. These are then opened for other users to answers for a period of four days, with the possibility to be extended to a maximum of eight days. Users can also vote for the best answer to a question, both on questions they asked themselves and those entered by others. If no best answer gets selected by the asker during the open period, the community votes given by other users potentially determine the chosen answer. In both cases, the question is marked as *resolved*. If no best answer can be chosen through either method, the question is labeled as *undecided*.

The Yahoo! Answers platform operates under a point system designed to "encourage participation and reward great answers" among its users [Yah13b]. They can spend points to ask questions and gain points for answering questions and voting on answers, as well receiving votes on their own answers. An overview of user actions and their associated point values is given in Table 2.1. By accumulating points, users can raise their member level. Increasing levels allow them to post a larger number of questions, answers and comments and cast more votes per day.

This reward system has been criticized on grounds of encouraging quantity over quality, as every answerer and voter receives credits, whether the provided answers were useful or not [Lei07]. Likewise, Liu et al. and have observed a

Attribute	Description
`Id`	The Yahoo! Answers question ID
`Type`	The state of the question (either `"Answered"`, `"Open"`, or `"Voting"`)
`Subject`	The subject/headline of the question
`Content`	The body text of the question
`Timestamp`	The time the question was submitted
`Link`	Link to the question
`CategoryName`	The category that the question is listed in
`CategoryId`	The Yahoo! Answers category ID
`UserId`	The Yahoo! Answers user ID of the asker
`NumAnswers`	The number of answers to the question
`NumComments`	The number of comments on the question
`ChosenAnswer`	The answer chosen by the user(s) as the best
`ChosenAnswererId`	The Yahoo! Answers user ID of the answerer
`ChosenAnswer-Timestamp`	The time the best answer was submitted
`ChosenAnswer-AwardTimestamp`	The time the best answer was awarded

Table 2.2: Attributes returned by the `questionSearch` API query

trend of increasingly passive participation—more voting, less answers—as well as a decrease quality as early as 2007 [LA08].

Hauff et al., did not discard answers chosen by the community, sometimes leaving unclear whether the selected answer actually contained the desired item. To avoid this effect, we only kept resolved questions whose best answer had been chosen by the asker. Additionally, questions and answers were assessed to ensure that they represent satisfied known-item information needs.

2.2 Querying the Yahoo! Answers API

Yahoo! Answers exposes questions posted on the Q&A platform to a public API [Yah13a], which allows querying details about a question entry (`getQuestion`) as well as retrieving a list of questions by their category (`getByCategory`), by the user who asked or answered them (`getByUser`), or by using a search term (`questionSearch`).

To generate a set of information needs, we used the `questionSearch` API function to retrieve resolved questions matching a given search query. Results can be requested either in order of relevance, or ascending/descending by

`(remember) AND (title) AND (movie)`
`(forgot) AND (name) AND (film)`
`(forgot) AND (title) AND (song)`
`(forgot) AND (url) AND (website OR (web site))`
`(remember OR forgot) AND (name OR title) AND (book)`

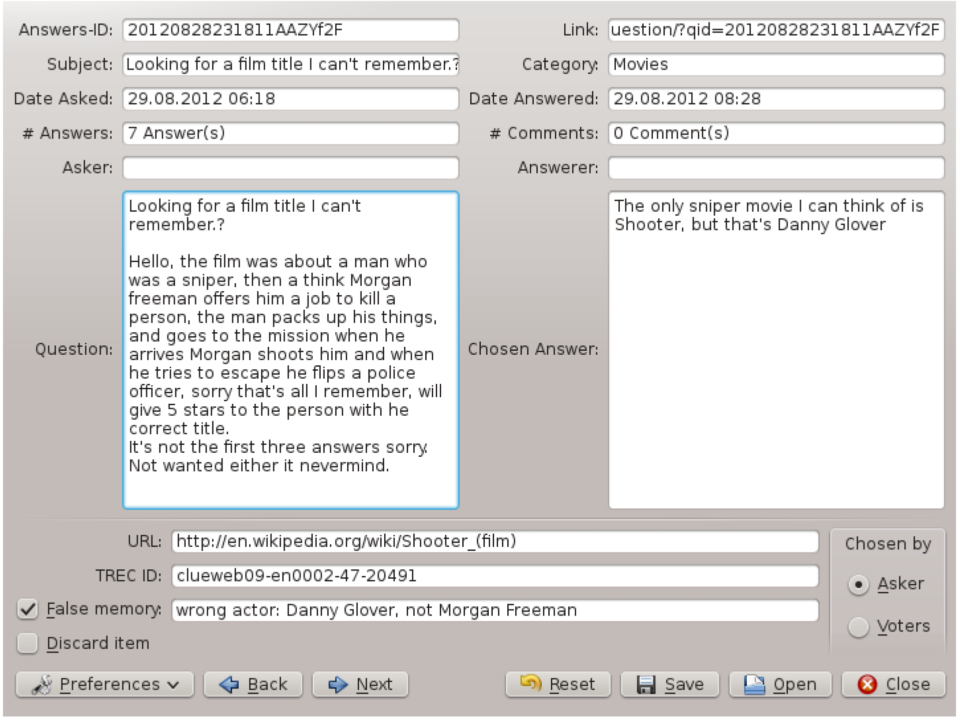
Table 2.3: Examples of search terms used to retrieve suitable information needs from Yahoo! Answers

date. The returned JSON output consists of question objects, up to a limit of 1050 questions per search query. Each question object maps to attributes of a question and its chosen answer. A selection of returned attributes used for the Webis-KIQ-13 corpus is presented in Table 2.2. These query results have been dumped into one JSON file per query and were stored locally for further assessment.

In building a topic set for the Webis-KIQ-13 corpus, the primary focus was placed on three types of known items that are often searched for: websites, movies and musical works (songs and music albums). To account for synonyms, different ways of posing a known-item question, and to retrieve more results than the API limit of a maximum of 1050 returned questions, nine separate sub-queries were formulated for each of those types. To provide a broader range of topics, ten additional types of known-item information needs were identified, such as re-finding a book or TV series, with one search query used for each of them. Examples of the used search terms are shown in Table 2.3. On January 21, 2013, these 37 distinct search queries were submitted to the Yahoo! Answers API, which resulted in a combined set of 24,765 unique questions.

Unfortunately, the `questionSearch` function does not provide information on whether a best answer was selected by the asker, or chosen by community vote. According to its documentation, this information is supposed to be provided by the `getQuestion` API call [Yah13a]. At the time of the crawl, however, this functionality did not work as described. Additionally, comments that the asker added to an answer are not exposed to the API. These comments can sometimes be a valuable indication of whether an answer actually contained the searched item. As a workaround, this information has been scraped from each question's HTML version on the Yahoo! Answers website. Six questions returned by the API were no longer accessible on the website, which usually means they have been deleted after they were answered. As it could no longer be determined who selected the best answer, they were discarded.

Among the 24,759 retrieved questions, the 15,934 that had been decided by a community vote were discarded, while the 8825 questions that had their best answer chosen by the original asker were kept for manual assessment.



Answers-ID: 20120828231811AAZYf2F

Link: uestion/?qid=20120828231811AAZYf2F

Subject: Looking for a film title I can't remember.?

Category: Movies

Date Asked: 29.08.2012 06:18

Date Answered: 29.08.2012 08:28

Answers: 7 Answer(s)

Comments: 0 Comment(s)

Asker:

Answerer:

Question: Looking for a film title I can't remember.?

Hello, the film was about a man who was a sniper, then a think Morgan freeman offers him a job to kill a person, the man packs up his things, and goes to the mission when he arrives Morgan shoots him and when he tries to escape he flips a police officer, sorry that's all I remember, will give 5 stars to the person with he correct title.
It's not the first three answers sorry.
Not wanted either it nevermind.

Chosen Answer: The only sniper movie I can think of is Shooter, but that's Danny Glover

URL: http://en.wikipedia.org/wiki/Shooter_(film)

TREC ID: clueweb09-en0002-47-20491

☑ False memory: wrong actor: Danny Glover, not Morgan Freeman

□ Discard item

Chosen by
● Asker
○ Voters

Figure 2.1: GUI form used for the manual assessment of the corpus

2.3 Assessment of the Retrieved Questions

To browse the Yahoo! Answers API dumps and facilitate the assessment of the therein contained information needs, a simple Qt-based GUI was developed. Assessors are presented with a form that collects the data fields retrieved by the API query and HTML scraper, as well as additional fields that are to be filled out manually (Figure 2.1). An external window provides a web view, which allows the assessors to view questions as they are presented to Yahoo! Answers users, to follow hyperlinks and to perform web searches. The data filled into form fields by the assessors is stored in a JSON dictionary, independently from the API dumps.

2.3.1 Assessment of Question Intent

Each of the 8825 questions with a best answer chosen by the asker was first judged whether the intent was to re-find a previously known item, and whether the answer contained the desired known item. For example, questions like “What is the weirdest movie you remember from your childhood?” or “What songs are similar to ’Remember The Name’ by Fort Minor?” are posed with the intention to generate a discussion or to receive a recommendation, rather than to satisfy a known-item information need.

For some known-item questions, the asker commented that an answer did not contain the known item, but still chose it as the best answer. This would happen if the answer was still useful to the asker (e.g. as a recommendation for a similar item), or merely so the asker would regain some of their spent points (see Table 2.1). Either led to the questions being omitted from the corpus, as the desired known item could not be determined.

All in all, 5419 questions were discarded in this step, further narrowing down the topic set to 3406 known-item information needs. Although similar search terms were chosen for all types of items, the proportion of discarded questions varied widely. While only about 35% of movie questions had to be discarded, the ratio was over 95% for websites. Possible explanations for this occurrence could be:

- The search terms used with the `questionSearch` function are ill-suited for finding known-item questions for website items. Askers may use other cue phrases more often for different types of known-item queries.

- The default behavior of the `questionSearch` function, to search in both the question and the answer, led to a large number of unwanted results. For instance, one of the website API queries returned almost one-hundred site support questions answered by the same user, with the same or similar stock answers containing every part of the search term. All of these had to be discarded.

- Askers are less interested in re-finding a specific website than they are for other item types. Frequently, users are also content with an alternative website offering the same functionality, even if it is not the one they originally accessed.

- Websites re-finding questions are less often posed on Yahoo! Answers, compared to those for movies or music.

Website re-finding information needs were originally intended to form a major part of the Webis-KIQ-13 corpus. However, due to the small number of remaining website items after this step, we had to dismiss this idea.

2.3.2 Mapping of Known Items to their ClueWeb09 ID

In the next step, the assessors checked whether a known item's URL is included in the ClueWeb09 corpus. For website queries, this would be the website's URL itself. For most other types of items, the most appropriate URL would

Known item	False memory	Correction
Shooter (film)	[...] Morgan freeman offers him a job to kill a person [...]	wrong actor: Danny Glover, not Morgan Freeman
Tokio Hotel	What's the english emo rock band [...] They are american [...]	origin: German band, not English or American
An American Tail	[...] a Disney cartoon about a little mouse [...]	company: Amblin Entertainment, not Disney
theforgottenlair.net	[...] it went somethin like the underground lair [...]	URL: "forgotten", not "underground"

Table 2.4: Examples of false memories in Yahoo! Answers questions

be the corresponding article in the English Wikipedia, if there is one. It should be noted that a known item may have multiple semantically or syntactically equivalent aliases [Bro02]. For example, a movie can have both a Wikipedia article and a corresponding IMDb entry, or a notable website may in turn have a Wikipedia article. In these cases, the more appropriate known-item URL was preferred.

As noted by Broder, a so-called *hub*-type result, which is one step away from the target, can also be an acceptable, although less desirable result. Examples where hub-type results were deemed acceptable include songs not represented through a Wikipedia article of their own, but through the music album they were released on, or specific pages on a website where only the main page is included in the ClueWeb09.

The ChatNoir [PHS+12] search engine was used as an interface to the ClueWeb09 corpus, and to map an item's URL to the corresponding TREC ID in it. 651 known items not included in the ClueWeb09 were marked as `"not found"` in the TREC ID field. To allow for further analysis and re-use with other web corpora the corresponding questions were kept separately, but they are not part of our final corpus. Only the 2755 known-item questions with matching ClueWeb09 entries form the base of the Webis-KIQ-13 corpus.

2.3.3 Annotation of False Memories

Lastly, it was determined whether a known-item question contained false memories. In these cases, the assessors tagged the question as such and added a short annotation documenting the type of error, a correction and the misremembered property. For example, in the case of an asker confusing actors in the 2007 action movie *Shooter*, the annotation is: "wrong actor: Danny Glover, not Morgan Freeman". More examples of false memories in Yahoo! Answers questions are shown in Table 2.4. Of the 2755 known-item questions in the Webis-KIQ-13 corpus, 240 (8.7%) contained at least one false memory.

	Movies	Music	Websites	Total
Retrieved questions	5896	6481	5343	24759
Best answer chosen by voters	-3718	-4112	-3637	-15934
Best answer chosen by asker	2178	2369	1706	8825
Not known-item questions	-768	-1451	-1624	-5419
Known-item questions	1410	918	82	3406
Not in ClueWeb09	-250	-219	-20	-651
In ClueWeb09	1160	699	62	2755
Containing false memories	81	74	4	240

Table 2.5: Summary of assessment steps and the respectively removed items

2.3.4 Summary

Although we started from a base of 24,759 unique questions retrieved from the Yahoo! Answers API, the final topic set consists of only 2755 suitable known-item information needs (11.1% of the original crawl). This is mostly due to the decision to exclude questions decided by community vote, which account for about two in three questions across all crawled categories. A summary of the items removed in the further assessment steps is given in Table 2.5.

We were surprised by the large amount of non-known-item questions that we had to discard for some topics. Possible explanations for the large amount of unsuitable website information needs have already been hypothesized in Section 2.3.1. These explanations might to a lesser degree be applicable to other categories as well.

Finally, the mapping step required us to discard 651 information needs, most of them for known items more recent than 2009. Given the age of the ClueWeb09 web corpus, we expected such an outcome. The differences in coverage over time will be further analyzed in Chapter 3.

The amount of false memory effects identified in the corpus met our initial expectations to be in the range of 5–10%. The actual number of false memories may be even higher. As the annotators mostly had to rely on the answer text and the corresponding document for a known item, it is likely that we missed memory errors that were not explicitly mentioned by the answerer.

As argued by Azzopardi et al. in [ARB07], the manual construction of the known-item corpus on the scope of the Webis-KIQ-13 proved to be a laborious and time-consuming process. The assessors spent approximately 200 hours on the evaluation of the 8825 questions that had an answer chosen by the asker, or an average of about 80 seconds per information need.

Attribute	Description
`ChosenBy`	By whom the best answer was selected (either `"Asker"` or `"Voters"`)
`ContainedIn`	The Yahoo! Answers API dump(s) containing the question
`FalseMemory`	`True` if the question contains at least one annotated memory error
`FalseMemoryComment`	Annotation describing the memory error(s)
`KnownItemUrl`	The URL of the known item
`KnownItemId`	The TREC ID of the known item's corresponding ClueWeb09 entry
`TimeDeltaAnswer`	The difference between the submission time of the question and the time the best answer was posted
`TimeDeltaAward`	The difference between the submission time of the and the time the best answer was awarded

Table 2.6: Additional attributes of question objects in the Webis-KIQ-13

2.4 Structure of the Webis-KIQ-13 Corpus

After the annotators had finished their assessment of the retrieved questions, their annotations were merged with the question objects of the remaining 2755 known-item information needs to generate the Webis Known-Item Question Corpus 2013 (Webis-KIQ-13).

The corpus is structured as a JSON array containing question objects. Each question object has the same attributes as the ones returned by the Yahoo! Answers API, most of which were outlined in Table 2.2, as well as additional attributes defined for the Webis-KIQ-13 corpus. Table 2.6 describes the new attributes in the corpus, which include the annotations made by the assessors as well as some simple, automatically generated features that were used to support the assessment.

Question objects are sorted by their Yahoo! Answers question ID, which is increasing over time. Consequently, they are also sorted in chronological order by their date of submission.

Listing 2.1 shows the structure of an example question object in the finished Webis-KIQ-13 corpus.

```
[
  [...]
  {
    "CategoryId": 396545138,
    "CategoryName": "Movies",
    "ChosenAnswer": "Maybe you're thinking of \"More Than a Feeling\" by Boston.",
    "ChosenAnswerAwardTimestamp": 1228205648,
    "ChosenAnswerTimestamp": 1228190651,
    "ChosenAnswererId": "5eaad8a838b5906ad84659f496c62cd8aa",
    "ChosenAnswererNick": "MaryAn",
    "ChosenBy": "Asker",
    "ContainedIn": [
      "music-rel-r-n-s.json",
      "music-rel-rf-nt-sa.json"
    ],
    "Content": "I con't remember the name of the song and it has been bugging me.\n",
    "Date": "2008-12-01 19:04:09",
    "FalseMemory": true,
    "FalseMemoryComment": "wrong artist: by Boston, not Journey",
    "Id": "20081201190409AA3JkOK",
    "KnownItemId": "clueweb09-enwp02-06-02945",
    "KnownItemUrl": "http://en.wikipedia.org/wiki/More_Than_A_Feeling",
    "Link": "http://answers.yahoo.com/question/?qid=20081201190409AA3JkOK",
    "NumAnswers": 2,
    "NumComments": 0,
    "Subject": "What is the name of the song by Journey in the movie Madagascar 2?",
    "TimeDeltaAnswer": 3602,
    "TimeDeltaAward": 18599,
    "Timestamp": 1228187049,
    "Type": "Answered",
    "UserId": "o5oBeiwraa",
    "UserNick": "Olivia",
    "UserPhotoURL": "http://l.yimg.com/q/users/1ZM_mwwWpAAED1YSlpJYPCg==.medium.jpg"
    },
    [...]
]
```

Listing 2.1: Structure of an example question object in the Webis-KIQ-13

Chapter 3

Analysis of the Webis-KIQ-13 Corpus

After outlining the creation process of the Webis Known-Item Question Corpus 2013 (Webis-KIQ-13), we move on to provide a thorough analysis of the retrieved information needs and their associated properties. First, we will examine the content of known-item questions and their chosen answers. Second, we are going to discuss the coverage of the ClueWeb09 corpus over the course of time. Finally, we will analyze the types of false memories exhibited in 8.7% of the known-item questions.

3.1 Text and Readability Measures

As the categorization and annotation of information needs posted on Yahoo! Answers, whether done manually or in an automated way, involves processing large amounts of text, questions about the complexity of the content get raised.

We use the Phantom Readability Library [Ott13], which utilizes regex-based language processing for sentence counting and tokenization. To estimate the count of syllables, it employs a rule-based algorithm ported from an improved version of the Lingua::EN::Syllable Perl module. First, we compute the following simple text features.

- Character count: The number of characters in the text, excluding whitespace and punctuation.
- Sentence count: The number of sentences in the text, delimited by one of the following punctuation marks: . ! ? : ; ...
- Syllable count: The estimated number of syllables in the text.

- Word count: The number of words in the text.

Based on these features, the following well-known readability formulas can be calculated. As all of them have been designed to estimate the U.S. grade level equivalent to the education required for understanding a text, we expect them to yield comparable results.

- Automated Readability Index (ARI): Readability formula developed by Smith et al. Designed for being easily automatable. Instead of being based on syllable count, which the authors argue is unreliable and dependent on the assessor, their formula uses the average character count per word as one of its components. [SS67].

$$\text{ARI} = 4.71 \cdot \frac{\text{Character count}}{\text{Word count}} + 0.5 \cdot \frac{\text{Word count}}{\text{Sentence count}} - 21.43$$

- Gunning fog index: Readability formula developed by Gunning. Suggested to count the number of so-called *hard words*, which he defined as words containing more than three syllables (polysyllables), rather than the entire number of syllables. Some additional exceptions are made, such as proper names always counting as *easy words.* [Gun52]. As we cannot handle all of these exceptional cases, we expect our computed estimates to be somewhat higher than those that would be generated by human calculation of the fog index.

$$\text{Gunning fog} = 0.4 \cdot \left(\frac{\text{Word count}}{\text{Sentence count}} + 100 \cdot \frac{\text{Polysyllable count}}{\text{Word count}}\right)$$

- Flesch-Kincaid grade level: Readability formula derived by Kincaid et al. from the original Flesch Reading Ease index. Weights from the original formula have been modified to directly estimate U.S. grade level [KFR+75].

$$\text{Flesch-Kincaid} = 11.8 \cdot \frac{\text{Syllable count}}{\text{Word count}} + 0.39 \cdot \frac{\text{Word count}}{\text{Sentence count}} - 15.59$$

- Simple Measure of Gobbledygook (SMOG): Readability formula developed by McLaughlin, which only uses the number of polysyllabic words, as first suggested by Gunning, and sentence count as properties. Originally defined for text samples with a length of 30 sentences [McL69]. The generalized form is as follows:

$$\text{SMOG} = 1.0430 \cdot \sqrt{30 \cdot \frac{\text{Polysyllable count}}{\text{Sentence count}}} + 3.1291$$

	Movies (1160 questions)		Music (699 questions)		Webis-KIQ-13 (2755 questions)	
	Mean	Std	Mean	Std	Mean	Std
# Characters	306.98	132.75	251.59	149.05	287.09	132.37
# Sentences	4.68	2.33	4.00	2.33	4.47	2.25
# Syllables	87.76	37.48	69.21	41.44	81.79	37.94
# Words	76.13	32.23	63.95	38.43	71.42	32.73
ARI	7.22	7.69	7.42	11.29	6.97	8.00
Gunning fog	7.73	6.16	8.22	9.07	7.57	6.41
Flesch-Kincaid	2.83	6.66	1.37	9.94	2.22	7.10
SMOG	6.87	3.67	6.51	3.66	6.95	3.64

Table 3.1: Text and readability measures for known-item questions

These indexes will be used to estimate the difficulty of question and answer texts for known-item information needs posted on the Yahoo! Answers platform. It should however be noted many messages submitted by Yahoo! Answers users have a shorter length and different structure from what the readability indexes have originally been defined for, which can lead to worse performance than would normally be expected. Consequently, it would be invalid to make definite statistical judgments based on them. Instead, the readability indexes are suggested mainly as orientation.

3.1.1 Readability of known-item questions

In Table 3.1, we compare the text and readability measures calculated for the known-item questions in the Webis-KIQ-13 corpus to its subsets of movie- and music-related questions. The mean text counts suggest that the average questions are of similar length across the movie and music subsets as well as the entire corpus, with movie-related questions being slightly longer and music-related ones being slightly shorter than the average. Likewise, the standard deviation for each text measure shows similar degrees of variation across the corpus and its subsets.

Looking at the readability formulas, we see that the Flesch-Kincaid grade level, compared to the other measures, makes very low estimates for the required level of education. This can be linked to the much larger weight that is given to *syllables per word* compared its *words per sentence* component. As the mean ratio of syllables per word ranges from only 1.08 for music-related to 1.15 for movie-related questions, the Flesch-Kincaid estimate places most questions on elementary-school level. In contrast, the ARI, Gunning fog and

	Movies (1160 answers)		Music (699 answers)		Webis-KIQ-13 (2755 answers)	
	Mean	Std	Mean	Std	Mean	Std
# Characters	212.66	328.81	335.09	629.51	246.63	430.06
# Sentences	3.05	4.00	3.69	8.05	3.21	5.22
# Syllables	59.11	93.54	91.66	175.63	68.51	121.70
# Words	46.97	71.04	81.57	154.02	56.32	99.81
ARI	7.11	6.79	10.80	22.77	8.19	13.05
Gunning fog	6.05	6.03	9.58	18.51	7.03	10.80
Flesch-Kincaid	−0.04	7.63	2.96	18.55	1.12	11.41
SMOG	7.70	4.03	7.42	4.40	7.87	4.23

Table 3.2: Text and readability measures for known-item answers

SMOG indexes provide very similar estimates, placing the average reading level between sixth and eight grade, or middle school in the U.S. education system. This leads us to the assumption that the Flesch-Kincaid grade level is not a very suitable metric for the questions posted on Yahoo! Answers. Additionally, the third quartiles of the ARI, Gunning fog and SMOG indexes for known-item questions in the Webis-KIQ-13 corpus are respectively at 7.9, 8.0 and 8.8, while it is only 3.6 for Flesch-Kincaid. Based on this, we assume that at least three in four questions would be easily understood by readers with the equivalent of entry-level high school education.

The SMOG index is the only formula with virtually the same standard deviation across the corpus and its movie and music subsets. This can probably attributed to SMOG being independent from the *words per sentence* property, which is used as a component by the other indexes. We expect a large variation of this property for the music subset, which frequently includes song lyrics without any standard punctuation.

3.1.2 Readability of answers

In Table 3.2, we compare the text and readability features for the best answers of all known-item questions in the Webis-KIQ-13 corpus, as well as the subsets of movies and music information needs. The sentence count is similar across the corpus and its subsets. However, the number of words (and consequently, characters and syllables) is much higher for music-related answers. Again, this can be attributed to song lyrics being included in the answer text. All of our text measures exhibit more variation than for known-item questions. This can be expected, as known-item information needs can be satisfied by

	Webis-KIQ-13					
	False Memory (81 questions)		No False Memory (1079 questions)		Most recent movie queries (8560 questions)	
	Mean	Std	Mean	Std	Mean	Std
# Characters	329.32	149.05	305.30	131.30	234.96	255.13
# Sentences	4.85	2.54	4.66	2.31	3.54	3.60
# Syllables	94.57	43.37	87.25	36.95	67.83	73.16
# Words	80.48	36.30	75.81	31.88	56.54	60.51
ARI	8.30	8.48	7.14	7.62	7.61	8.07
Gunning fog	8.35	6.72	7.68	6.11	7.31	6.47
Flesch-Kincaid	3.33	7.52	2.79	6.59	2.34	7.09
SMOG	7.49	4.15	6.83	3.62	7.38	4.07

Table 3.3: Text and readability measures for movie questions

merely identifying the title or URL of the desired known item. However, some answerers choose to include additional information, such as song lyrics or movie synopses to their answer. Consequently, the answer length varies widely.

For the entire corpus and the movie subset, our readability indexes return similar results for both the mean grade level required to understand an answer. For music-related answer, we surmise that the aforementioned effect of unpunctuated song lyrics comes into play. As a result, metrics including the *words per sentence* as a component have more variation and, on average, judge music answers as harder to understand than other types of answers. Again, the SMOG formula is unaffected by this, as it ignores sentence length. For the entire Webis-KIQ-13 corpus, the third quartiles for the ARI, Gunning fog and SMOG indexes respectively are at 10.2, 7.6 and 11.2. However, identifying the known item in an answer does not require the full understanding of the answer text, so the required education level for this task can be assumed to be lower.

3.1.3 Effect of known-item information needs and the presence of false memories

After examining the differences in text and readability measures of known-item questions and their respective answers for different subsets of the Webis-KIQ-13 corpus, we want to evaluate how much, if at all, the presence of false memories affects the content of a question, and how known-item questions differ from other information needs posted on Yahoo! Answers.

For this purpose, we used the `getByCategory` function exposed by the

	Webis-KIQ-13					
	False memory (74 questions)		No false memory (625 questions)		Most recent music queries (7673 questions)	
	Mean	Std	Mean	Std	Mean	Std
# Characters	257.61	109.07	250.88	153.08	220.39	227.81
# Sentences	4.46	2.82	3.95	2.26	3.44	3.33
# Syllables	70.64	29.87	69.04	42.60	63.77	66.91
# Words	66.01	27.52	63.70	39.52	53.86	55.70
ARI	6.54	6.35	7.53	11.73	7.29	9.79
Gunning fog	7.65	4.97	8.29	9.44	7.33	7.67
Flesch-Kincaid	0.64	7.09	1.46	10.22	1.62	8.91
SMOG	6.12	3.62	6.55	3.67	6.99	4.09

Table 3.4: Text and readability measures for known-item music questions

Yahoo! Answers API to crawl the most recently resolved questions in the categories *Movies* and *Music* and computed the same text and readability measures as for the known-item questions in the Webis-KIQ-13 corpus. The calculated measures are shown in Tables 3.3 for movie and 3.4 for music.

We recieved the 8560 most recently answered questions in the *Movies*, covering a range from April 19, 2013 to May 21, 2013 (the day the request was sent). Questions containing false memories are both longer by each of our text metrics and are judged as more demanding by our readability formulas. However, the variation increases likewise for all metrics, so the differences might not be representative. Known-item questions for movies are consistently longer and have much less variation than the unfiltered set of recent queries. Clearly, this is caused by different types of questions being present in the set. Despite the strong variation, the readability metrics are again similar both in terms of arithmetic mean as well as standard deviation. Apparently, the average difficulty of questions is more dependent on categories than on the question's intent.

For the most recently resolved music queries, the Yahoo! Answers API returned 7673 unique questions covering a range from May 3, 2013 to May 21, 2013. As with the movies, known-item questions are slightly longer and have less variance than the set of all queries. Somewhat unexpectedly, known-item questions containing false memories are judged as simpler to read. However, the difference is too small to draw any conclusions.

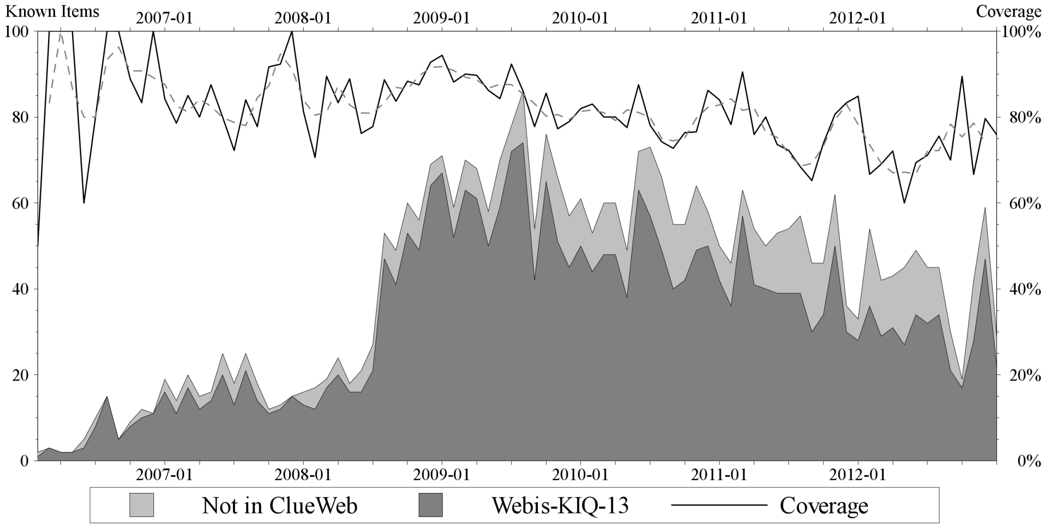


Figure 3.1: Monthly ClueWeb coverage over time

	2006	2007	2008	2009	2010	2011	2012
Webis-KIQ-13	68	176	369	701	578	477	364
Not in ClueWeb	8	15	60	112	148	140	142
Total	76	191	429	813	726	617	506
Coverage	89.5%	92.2%	86.0%	86.2%	79.6%	77.3%	71.9%

Table 3.5: ClueWeb coverage by year

3.2 ClueWeb09 Coverage

At the start of this writing, the ClueWeb09 was still most recent publicly available, static web corpus. As most of its content had been crawled from the live web in January and February 2009, the corpus increasingly showed its age when the assessors moved to more recent questions.

While its successor corpus, the ClueWeb12, had already been crawled in the time between February 10, 2012 and March 10, 2012, unfortunately it did not become available in time for this thesis. On the other hand, this allows us to show the effect of a static corpus becoming gradually outdated for a subset of the assessed information needs.

Figure 3.1 presents the ClueWeb09 coverage of the retrieved known item queries per month. The stacked graph shows the number of retrieved known items that have a corresponding ClueWeb09 entry (dark gray) as well as the number of those that have no entry (light gray). The line graph shows the

	Wikipedia	IMDb	Others	No link
Webis-KIQ-13	2618	3	134	–
Not in ClueWeb	405	66	94	86
Total	3023	69	228	86

Table 3.6: ClueWeb coverage by target domain of known-item URLs

relative coverage per month (solid line) and a sliding three-month average (dashed line). In the second half of 2008, there is a steep incline in the number of retrieved known items that can probably be related to an increase in Yahoo! Answers usage. Beginning from 2009, ClueWeb09 coverage predictably decreases due to the occurrence of known items that did not exist at the time of the crawl.

The decrease in relative coverage becomes even more obvious when shown by year, as in Table 3.5. While in 2007 a record high of 92.2% could be achieved, the known-item coverage fell to only 71.9% by 2012. Interestingly, this decrease did not set in immediately, but was delayed until 2010.

There are two possible reasons for this. First, several movies and, to a lesser degree, music albums that were recent in 2009 already had a Wikipedia article or other appropriate document in the ClueWeb09 in 2008, while they were still in production.because their originally planned release was moved back. Second, users did not have to turn to a question-answering platform like Yahoo! Answers yet, as information about known items from 2009 was still more readily available.

We noticed that there were two major groups of re-finding needs: Queries for items that have not been accessed for a long time (e.g. users searching for the favorite movie of their childhood), and for items that have only been incompletely accessed more recently (e.g. by hearing a song on the radio or watching the trailer of a movie). Obviously, the latter type is affected more easily by a corpus becoming outdated.

Finally, we examine the domains of the ClueWeb09 documents used to represent known items. Table 3.6 shows the frequency with which websites were chosen by the assessors. As can be seen from the table, Wikipedia was usually the first source assessors checked when searching for a known item's URL, and the majority of known items were matched their article there.

This decision was made because the ClueWeb09 corpus contains a nearly complete dump of the English at the time of its crawl [CC09]. At the time of assessment, 3023 known items either had an article of their own or, as per Broder's definition in [Bro02], a *hub*-type result. Of these, 405 did not have a ClueWeb09 entry, usually because the article did not yet exist at the time

Category	False memories relating to...	#
character	attributes of a character in a work of fiction	34
lyrics	the lyrics of a song or poem	29
title	the title of a work	27
format	the way a work was released	21
wrong artist	wrong attribution of an artist to a musical work	22
time	the time a work has been produced or released	18
origin	the geographical background of a work or artist	15
wrong actor	wrong attribution of an actor in a movie or TV series	11
plot	key elements of a work's plot	9
setting	the time or place a work is set in	9
company	the company involved in the production of the item	6
scene	a single scene in a movie	5
prop	an object in a movie or theater play	5
mix-up	confusing or mixing attributes of one item with another	5
URL	the URL of a website	4

Table 3.7: Common types of false memories in the Webis-KIQ-13

of the crawl. We expect that most of these documents could have been linked to an entry in the ClueWeb12, which like its predecessor included a current Wikipedia dump as of the time of its crawl. In a smaller number of cases, technical restrictions prevented an article from being crawled.

Coverage was much less reliable for domains outside of Wikipedia. For instance, IMDb was usually used as a second resort for movies or series not listed on Wikipedia. However, only three out of 69 IMDb entries were actually part of the ClueWeb09.

In some cases, the assessors could not find a suitable document representing the known item on the live web. These were usually rather obscure and included poems or songs not released on an album with a Wikipedia entry. It is unlikely that these known items would be covered by the ClueWeb12.

3.3 False Memories

At least of 240 of the 2755 known items in the Webis-KIQ-13 corpus contain some kind of false memory. Categories were defined ad-hoc by the assessors and were unified in a second pass over the information needs with memory errors. Given the search terms used to retrieve our topic set, most of them relate to works of art and entertainment. The most common types of memory errors are shown in Table 3.7, with an explanation and their number of occurrences.

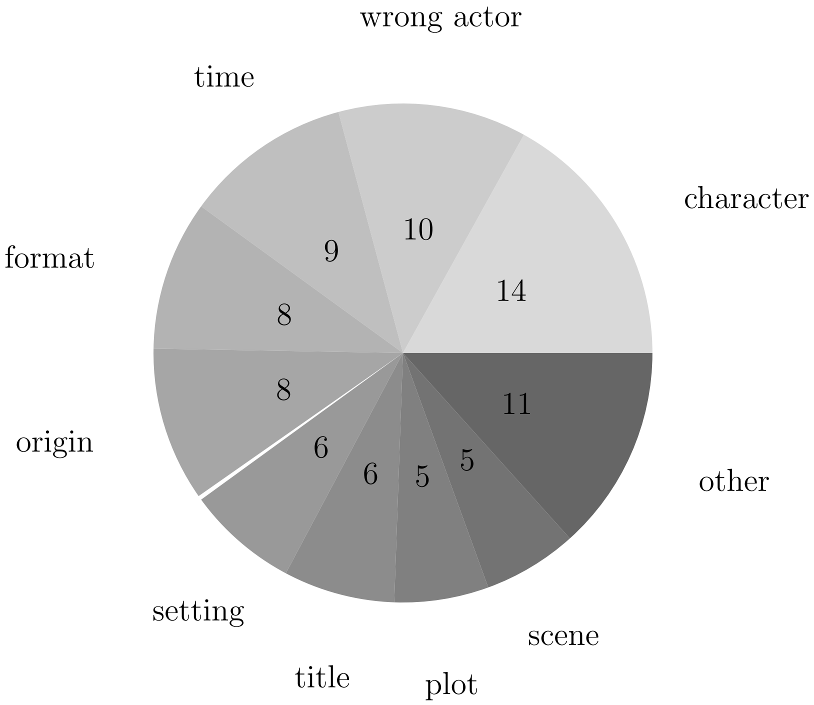


Figure 3.2: Common sources of false memories for movies

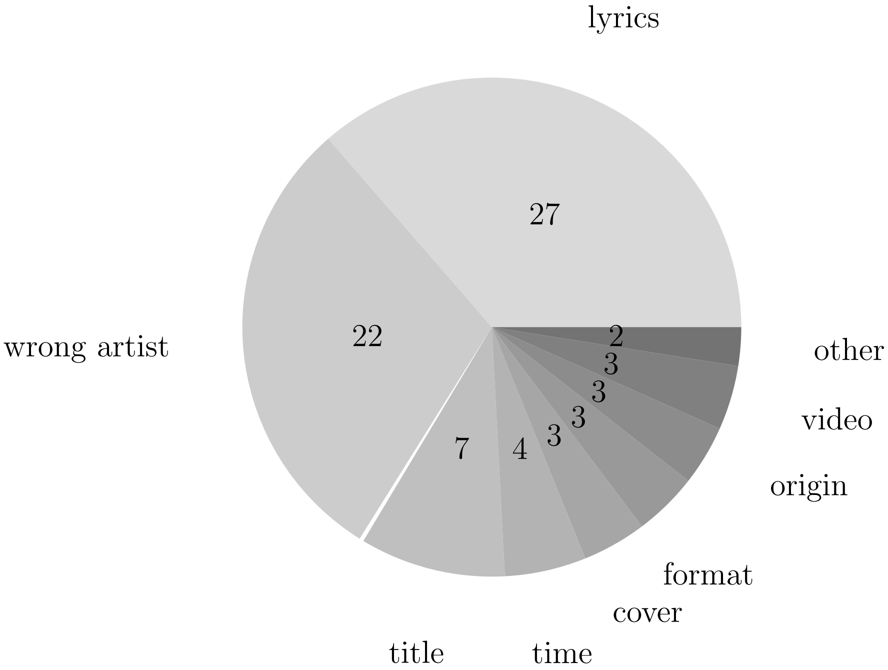


Figure 3.3: Common sources of false memories for music

Chapter 4

Conclusions and Future Work

The Webis Known-Item Question Corpus 2013 generated during the course of this thesis enables a new approach to the evaluation of known-item retrieval tasks, based on the idea of using real information needs with a clearly stated intent of known-item re-finding. We believe that by constraining the topic set to answers selected as correct by their asker, we could minimize the error in our known-item mappings. We hope that in conjunction with the ClueWeb corpus, this topic set allows for repeatable and realistic testing of known-item queries.

Although the corpus was originally developed as a testbed for known-item search queries, other uses could be considered as well.

Most of the search terms we used acquired known-item questions from the Yahoo! Answers categories *Arts & Humanities* as well as *Entertainment & Music*. This places a large number of information needs close to the field of media or video retrieval, although from a different vantage point.

However, this also means that other types of known items that could be searched for, such as geographical landmarks or electronic devices have mostly been neglected by us. It should be noted that while each of the known-item questions corresponds to a real information need, we cannot be sure that the Webis-KIQ-13 corpus provides a representative sample of the known-item questions posed on Yahoo! Answers. We experienced this especially with website items, where the search terms that yielded acceptable results on other categories hardly returned usable known-item information needs. Future research could try to extend the scope of the corpus to other types of information needs.

Finally, the false memories we identified could be used to research the recall of different kinds of information in audiovisual media.

Bibliography

[ARB07] L. Azzopardi, M. de Rijke, and K. Balog. “Building simulated queries for known-item topics: an analysis using six European languages”. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* SIGIR ’07. Amsterdam, The Netherlands: ACM, 2007, pp. 455–462. ISBN: 978-1-59593-597-7.

[Bro02] A. Broder. “A taxonomy of web search”. In: *SIGIR Forum* 36.2 (September 2002), pp. 3–10. ISSN: 0163-5840.

[BS07] T. Blanc-Brude and D. L. Scapin. “What do people recall about their documents?: implications for desktop search tools”. In: *Proceedings of the 12th international conference on Intelligent user interfaces.* IUI ’07. Honolulu, Hawaii, USA: ACM, 2007, pp. 102–111. ISBN: 1-59593-481-2.

[CC09] W. B. Croft and J. Callan. *The ClueWeb09 Dataset.* 2009. URL: `http://lemurproject.org/clueweb09/` (visited on January 15, 2013).

[CHY+09] J. Callan, M. Hoy, C. Yoo, and L. Zhao. *The web09-bst Dataset.* February 4, 2009. URL: `http://boston.lti.cs.cmu.edu/Data/web08-bst/planning.html` (visited on January 15, 2013).

[ERJ07] D. Elsweiler, I. Ruthven, and C. Jones. “Towards memory supporting personal information management tools”. In: *J. Am. Soc. Inf. Sci. Technol.* 58.7 (May 2007), pp. 924–946. ISSN: 1532-2882.

[Gun52] R. Gunning. *The technique of clear writing.* New York: McGraw-Hill, 1952.

[HHB+12] C. Hauff, M. Hagen, A. Beyer, and B. Stein. “Towards realistic known-item topics for the ClueWeb”. In: *Proceedings of the 4th Information Interaction in Context Symposium.* IIiX ’12. Nijmegen, The Netherlands: ACM, 2012, pp. 274–277. ISBN: 978-1-4503-1282-0.

[KC09] J. Kim and W. B. Croft. “Retrieval experiments using pseudo-desktop collections”. In: *Proceedings of the 18th ACM conference on Information and knowledge management.* CIKM ’09. Hong Kong, China: ACM, 2009, pp. 1297–1306. ISBN: 978-1-60558-512-3.

[KC10] J. Kim and W. B. Croft. “Ranking using multiple document types in desktop search”. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* SIGIR ’10. Geneva, Switzerland: ACM, 2010, pp. 50–57. ISBN: 978-1-4503-0153-4.

[KCF+08] L. Kelly, Y. Chen, M. Fuller, and G. J. F. Jones. “A study of remembered context for information access from personal digital archives”. In: *Proceedings of the second international symposium on Information interaction in context.* IIiX ’08. London, United Kingdom: ACM, 2008, pp. 44–50. ISBN: 978-1-60558-310-5.

[KFR+75] J. P. Kincaid, R. Fishburne, R. L. Rogers, and B. S. Chissom. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Tech. rep. 1975.

[LA08] Y. Liu and E. Agichtein. “On the evolution of the yahoo! answers QA community”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval.* SIGIR ’08. Singapore, Singapore: ACM, 2008, pp. 737–738. ISBN: 978-1-60558-164-4.

[Lei07] J. Leibenluft. “A Librarian’s Worst Nightmare”. In: *Slate* (December 2007). ISSN: 1091-2339. URL: `http://www.slate.com/articles/technology/technology/2007/12/a_librarians_worst_nightmare.html` (visited on February 10, 2013).

[McL69] G. H. McLaughlin. “SMOG grading: A new readability formula”. In: *Journal of reading* 12.8 (1969), pp. 639–646.

[Ott13] N. Ott. *Phantom Readability Library.* 2013. URL: `http://niels.drni.de/s9y/pages/phantom.html` (visited on May 12, 2013).

[PHS+12] M. Potthast, M. Hagen, B. Stein, J. Graßegger, M. Michel, M. Tippmann, and C. Welsch. “ChatNoir: A search engine for the ClueWeb09 Corpus”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval.* SIGIR ’12. Portland, Oregon, USA: ACM, 2012, p. 1004. ISBN: 978-1-4503-1472-5.

[SS67] E. Smith and R. Senter. *Automated readability index.* Tech. rep. 1967.

[Yah12] Yahoo! Inc. *Yahoo! Answers ≫ Blog Archive ≫ Yahoo! Answers Ranked 5th Largest Social Media Network.* July 3, 2012. URL: `http://yanswersblog.com/index.php/archives/2012/07/03/yahoo-answers-ranked-5th-largest-social-media-network/` (visited on March 18, 2013).

[Yah13a] Yahoo! Inc. *Yahoo! Answers API - YDN.* 2013. URL: `http://developer.yahoo.com/answers/` (visited on February 17, 2013).

[Yah13b] Yahoo! Inc. *Yahoo! Answers - Point System.* 2013. URL: `http://answers.yahoo.com/info/scoring_system` (visited on February 10, 2013).