

A Corpus of Realistic Known-Item Topics with Associated Web Pages in the ClueWeb09

Matthias Hagen Daniel Wägner Benno Stein

Bauhaus-Universität Weimar
matthias.hagen@uni-weimar.de
[@matthias_hagen](https://twitter.com/matthias_hagen)

ECIR 2015
Vienna, Austria
April 1, 2015

The scenario



This is not just a problem of philosoraptor!



Known-item search

Re-finding previously seen/heard items like

- Documents
- Websites
- Emails
- Tweets
- Movies
- Music
- Books
- TV



Known-item search

Re-finding previously seen/heard items like

- Documents
- Websites
- Emails
- Tweets
- Movies
- Music
- Books
- TV



Remarks: Users have some knowledge about their need.
Only very few relevant documents out there.

Problem

How do users search for known items?

Studies on re-finding known items

Web search	[Sadeghi et al., ECIR 2015] [Tyler and Teevan, WSDM 2010] [Edar et al., CHI 2008] [Azzopardi et al., SIGIR 2007] [Teevan, TOIS 2008, UIST 2007] [Beitzel et al., SIGIR 2003]
Twitter search	[Meier and Elsweiler, IUI 2014]
Email search	[Elsweiler et al., SIGIR 2011, ECIR 2011, TOIS 2008]
PIM	[Kim and Croft, SIGIR 2010, CIKM 2009] [Kelly et al., IUI 2008] [Blanc-Brude and Scapin, IUI 2007] [Boardman and Sasse, CHI 2004] [Dumais et al., SIGIR 2003] [Barreau and Nardi, SIGCHI Bulletin 1995]

Studies on re-finding known items

Web search	[Sadeghi et al., ECIR 2015] [Tyler and Teevan, WSDM 2010] [Edar et al., CHI 2008] [Azzopardi et al., SIGIR 2007] [Teevan, TOIS 2008, UIST 2007] [Beitzel et al., SIGIR 2003]
Twitter search	[Meier and Elsweiler, IIiX 2014]
Email search	[Elsweiler et al., SIGIR 2011, ECIR 2011, TOIS 2008]
PIM	[Kim and Croft, SIGIR 2010, CIKM 2009] [Kelly et al., IIiX 2008] [Blanc-Brude and Scapin, IUI 2007] [Boardman and Sasse, CHI 2004] [Dumais et al., SIGIR 2003] [Barreau and Nardi, SIGCHI Bulletin 1995]

Problem: Most corpora and queries not freely available.

Exceptions: Known-item query generation

Automatic extraction

- ① Select some document
- ② Draw most discriminative terms
- ③ Add random noise

Web [Azzopardi et al., SIGIR 2007]

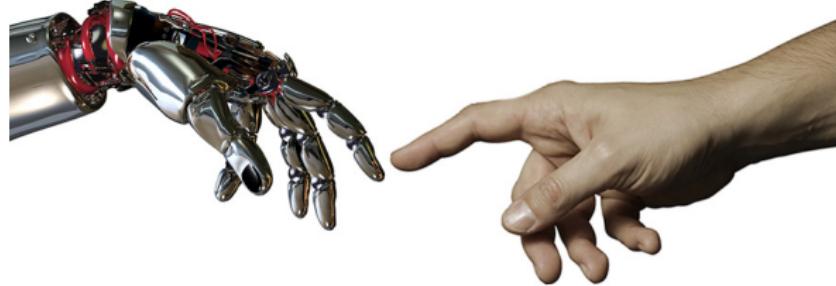
PIM [Kim and Croft, CIKM 2009]

Email [Elsweiler et al., SIGIR 2011]

Human computation game

- ① Select some document
- ② Show it to a user for some time
- ③ Ask for a query retrieving it top-ranked

PIM [Kim and Croft, SIGIR 2010]



Exceptions: Known-item query generation

Automatic extraction

- ① Select some document
- ② Draw most discriminative terms
- ③ Add random noise

Web [Azzopardi et al., SIGIR 2007]

PIM [Kim and Croft, CIKM 2009]

Email [Elsweiler et al., SIGIR 2011]

Human computation game

- ① Select some document
- ② Show it to a user for some time
- ③ Ask for a query retrieving it top-ranked

PIM [Kim and Croft, SIGIR 2010]



Problem: Not really “natural” settings.

Human memory: Not perfect but also not random



Reasons for memory failure?

Reasons for memory failure?

Psychology, man!



Our goal

A large corpus of **difficult and realistic** known-item needs.

Our goal

A large corpus of **difficult and realistic** known-item needs.

Remark: Will be freely available!

- ① Fetch known-item questions from Yahoo! Answers
 - To ensure realistic human information needs
 - Websites, movies, music, books, TV series
- ② Link questions to ClueWeb09 documents
 - Environment for repeatable research
 - ClueWeb12 has no Wikipedia in it
- ③ Construct queries from questions
 - Maybe via crowdsourcing
 - Not part of this paper

Question acquisition

Querying Yahoo! Answers API:

- forgot AND name AND film
- forgot AND title AND song
- remember AND title AND movie
- forgot AND url AND (website OR (web site))
- (remember OR forgot) AND (name OR title) AND book
- 37 such queries in total

24,765 answered questions returned on January 21, 2013

Question acquisition

Querying Yahoo! Answers API:

- forgot AND name AND film
- forgot AND title AND song
- remember AND title AND movie
- forgot AND url AND (website OR (web site))
- (remember OR forgot) AND (name OR title) AND book
- 37 such queries in total

24,765 answered questions returned on January 21, 2013

Problems: Not all questions are really “answered.”

Not all questions are known-item intents.

Not all questions are linkable to the ClueWeb09.

Corpus cleansing

Answered status

- Keep when best answer selected by asker
- 8,825 questions remain (only about 36% of original crawl)

Known-item status and ClueWeb linkage need manual assessment

- Two independent annotators
- About 400 hours of work
- 3,406 questions with known-item information need
- 2,755 can be linked to ClueWeb09 documents
- Only these form the Webis-KIQC-13

Corpus cleansing

Answered status

- Keep when best answer selected by asker
- 8,825 questions remain (only about 36% of original crawl)

Known-item status and ClueWeb linkage need manual assessment

- Two independent annotators
- About 400 hours of work
- 3,406 questions with known-item information need
- 2,755 can be linked to ClueWeb09 documents
- Only these form the Webis-KIQC-13

Problem: Hardly any website questions remained.

ClueWeb09 coverage

Over the years

Question from	2006	2007	2008	2009	2010	2011	2012
Webis-KIQC-13	68	176	369	701	578	477	364
Coverage	89.5%	92.2%	86.0%	86.2%	79.6%	77.3%	71.9%

Type of associated URL

- 95% Wikipedia
- 5% other

Corpus analysis

An initial observation related to a famous IR movie

False memories hinder total recall



False memories in questions

Entertainment & Music > Movies

Next >



What's that cartoon movie where it's about dogs and it starts off with a box full of free puppies?

only one puppy is left of at the end, and its raining and such. then later he meets another dog in the park or something and theres a big singing thing. and later he finds a young girl, and she like takes the puppy in or something. what's that movie called?



Follow



6 answers

Answers

Relevance ▾



💡 **Best Answer:** Yeah, that's Oliver and Company and it's a box of free cats, Oliver being the last one. Meets the Billy Joel dog for the big singing thing

Blake B · 7 years ago



Comment

Movie "... starts off with a box full of free puppies..."



Question

Movie "... starts off with a box full of free puppies..."



Question



Actual known item

Note a difference?!

False memories in questions

Entertainment & Music › Movies

Next ›



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... [show more](#)

Update: It's not the first three answers sorry.

Update 2: Not wanted either it nevermind.



1 following



6 answers

Answers

Rating



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

Source(s):

<http://www.imdb.com/title/tt0822854/>

Jennifer · 3 years ago



Comment

Asker's rating

Movie "... Morgan Freeman offers him a job to kill ..."



Question

Movie "... Morgan Freeman offers him a job to kill ..."



WANTED

IN THEATERS JUNE 27TH



Question

Actual known item

Note a difference?!

Yeah, funny! But these are just a few outliers?!

False memories statistics

- At least 240 questions (9% of corpus) contain false memories
- Most frequent false memories: **Person names!**

False memories statistics

- At least 240 questions (9% of corpus) contain false memories
- Most frequent false memories: **Person names!**

Remark: Makes me think ...

Does my mail search take this into account?

Potential usage of the corpus

Observation: False memories hinder good results.

Might even yield zero-result lists!

IR systems should

- Detect false memory situations
- “Repair” the query
 - Leave out the false memory or
 - Replace it with correction

Potential usage of the corpus

Observation: False memories hinder good results.

Might even yield zero-result lists!

IR systems should

- Detect false memory situations
- “Repair” the query
 - Leave out the false memory or
 - Replace it with correction

Our corpus might be a starting point in that direction.

Potential usage of the corpus

Observation: False memories hinder good results.

Might even yield zero-result lists!

IR systems should

- Detect false memory situations
- “Repair” the query
 - Leave out the false memory or
 - Replace it with correction

Our corpus might be a starting point in that direction.

Did I mention that it is freely available?!

Other fields: False memory implantation

Scientists Implant False Memories Into Sleeping Mice

March 10, 2015 | by Justine Alford



Remark: We are not working on that!

A little scary, isn't it?

Let's finish the talk in a better mood!

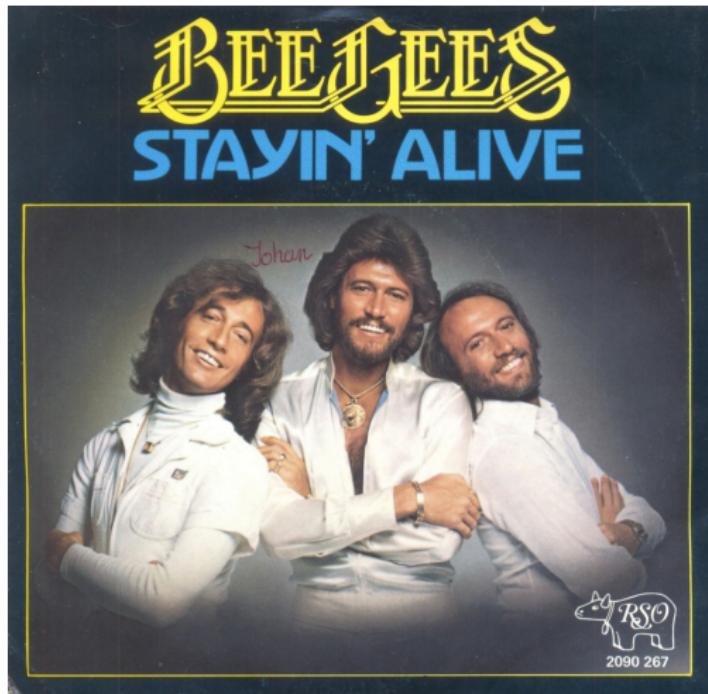
You know this song?!



One more hint needed?!



Yes, the Bee Gees!



Ah, ha, ha, ha, **steak and a knife, steak and a knife**

Some funny false memories really are Mondegreens.

Some funny false memories really are Mondegreens.

. . . that are misheard lyrics.

Almost the end: The take-home messages!

What we have done

Results

- Webis-KIQC-13 available
- 2,755 known-item questions
- Posted by real human users
- Linked to the ClueWeb09
- False memories annotated
- Often refer to persons
- Or song lyrics

Future Work

- Enlarge the corpus
- Website known-items esp.
- Web queries for the questions
- False memory handling in IR
- False memory detection

What we have (not) done

Results

- Webis-KIQC-13 available
- 2,755 known-item questions
- Posted by real human users
- Linked to the ClueWeb09
- False memories annotated
- Often refer to persons
- Or song lyrics

Future Work

- Enlarge the corpus
- Website known-items esp.
- Web queries for the questions
- False memory handling in IR
- False memory detection

What we have (not) done

Results

- Webis-KIQC-13 available
- 2,755 known-item questions
- Posted by real human users
- Linked to the ClueWeb09
- False memories annotated
- Often refer to persons
- Or song lyrics

Future Work

- Enlarge the corpus
- Website known-items esp.
- Web queries for the questions
- False memory handling in IR
- False memory detection

Thank you
😊