Introduction to

Information Retrieval

CS276: Information Retrieval and Web Search
Lecture 19: Web Question Answering
“Information retrieval”

The name **information retrieval** is standard, but as traditionally practiced, it’s not really right

All you get is **document retrieval**, and beyond that the job is up to you
Getting information

The common person’s view? [From a novel]

“I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota ... I’m the first guy to get the modem humming. But as a source of information, it sucks. You got a billion pieces of data, struggling to be heard and seen and downloaded, and anything I want to know seems to get trampled underfoot in the crowd.”

Web Search in 2025?

The web, it is a changing.

What will people do in 2025?

- Type key words into a search box?
- Use the Semantic Web?
- Ask questions to their computer in natural language?
- Use social or “human powered” search?
What do we know that’s happening?

- Much of what is going on is in the products of companies, and there isn’t exactly careful research explaining or evaluating it.
- So most of this is my own meandering observations giving voice over to slides from others.
Google
What’s been happening? 2013–2019

- Many updates a year ... and 3rd party sites try to track them
  - e.g., [https://moz.com/google-algorithm-change](https://moz.com/google-algorithm-change) by & aimed at SEOS
- I just mention a few changes here

- New search index at Google: “Hummingbird” (2013)
  - Answering long, “natural language” questions better
  - Partly to deal with spoken queries on mobile

- More use of the Google Knowledge Graph (2014)
  - Concepts versus words

- RankBrain (second half of 2015):
  - A neural net helps in document matching for the long tail
Google
What’s been happening? 2013–2019

- “Pigeon” update (July 2014):
  - More use of distance and location in ranking signals
- “Mobilegeddon” (Apr 21, 2015):
  - “Mobile friendliness” as a major ranking signal
- “App Indexing” (Android, iOS support May 2015):
  - Search results can take you to an app
- Mobile-friendly 2 (May 12, 2016):
  - About half of all searches are now from mobile
- “Fred” (1st quarter 2017):
  - Various changes discounting spammy, clickbaity, fake? sites
Google

What’s been happening? 2013–2019

- Longer snippets in results pages (Nov 2017)
- Mobile-first Index (Mar 2018)
  - Index mobile version of websites in preference to desktop!
- Revert snippet length in results pages (May 2018)
- “Medic” update (Aug 2018)
  - More emphasis on expertise, authoritativeness, trust
  - Big changes for diet, nutrition, medical products sites
- Core Algorithm Update (Mar 2019)
  - Seems kind of like “Medic 2”
- 2019 seems to have been kinda quiet so far …
The role of knowledge bases

- Google Knowledge Graph
- Facebook Graph Search
- Bing’s Satori
- Things like Wolfram Alpha

Common theme: Doing graph search over structured knowledge rather than traditional text search
What’s been happening

- More semi-structured information embedded in web pages
  - schema.org
Mobile

Move to mobile favors a move to *speech* which favors *natural language information search*

- Will we move to a time when over half of searches are spoken?
Mobile

- Mobile proved importance of NLU/QA
- [What is the best time for wildflowers in the bay area]

*Where to See Wildflowers in the Bay Area - Bay Area Parent*
www.bayareaparent.com/Article/Where-to-See-Wildflowers-in-the-Bay-Area/
This is a great time of year to find wildflowers in the Bay Area. Call parks ... Loop Trail, known for being flat, moderate in difficulty and best for viewing the blooms.

*Best Bay Area hikes into wildflower country - Bay Nature*
baynature.org/article/best-bay-area-hikes-into-wildflower-country/
Mar 28, 2013 - It's spring, and all you nature lovers know what that means ... time to go see wildflowers! You've probably already noticed them popping in little ...
Information quality

- There have always been concerns about information **provenance** (the source) and information **reliability**, especially among “information professionals” (reporters, lawyers, spies, ...)
- It wasn’t ignored on the web: ideas like PageRank were meant to find good content, and there has been a decade of work targeting link farms, etc.
- However, a lot of recent events have shown the limited effectiveness of that work, and how “fake” information easily gets upvoted and spreads
Towards intelligent agents

Two goals

- Things not strings
- Inference not search
Two paradigms for question answering

- Text-based approaches
  - TREC QA, IBM Watson, DrQA
- Structured knowledge-based approaches
  - Apple Siri, Wolfram Alpha, Facebook Graph Search

(And, of course, there are hybrids, including some of the above.)

At the moment, structured knowledge is back in fashion, but it may or may not last
Example from Fernando Pereira (GOOG)
Introduction to Information Retrieval

**Search Results**

**Symptoms of Crohn's**

**Best Pain Reliever, Over the Counter Pain Relievers, Prescription ...**
For stomach upset, eating or drinking milk before you take a painkiller can help. ... Don't expect painkillers to relieve all your pain, but the best pain ... www.healthblurbs.com/best-pain-reliever-over-the-counter-pain-relievers-prescription-analgesic-painkillers/ - Cached - Similar

**Why do painkillers make your stomach hurt? - Yahoo! Answers**
Aug 30, 2008 ... stomach upset, reflux/indigestion, or if severe ulcers ... I often take painkillers FOR stomach aches lol. Don't do this again. 2 years ago ...
answers.yahoo.com › Health › Other - Health - Cached - Similar

- How are painkillers bad for you? - Apr 13, 2010
- What effect do painkillers have on you? - Sep 30, 2009
- What's the difference between OTC painkillers? (Advil, Aleve ... - Nov 19, 2008
- Is it bad to take painkillers on an empty stomach? - Aug 20, 2008

More results from answers.yahoo.com »

**Phys Ed: Does Ibuprofen Help or Hurt During Exercise? - NYTimes.com**
Sep 1, 2009 ... When are ibuprofen and other anti-inflammatory painkillers justified? ... Can't take NSAIDS due to upset stomach. — bonemri ... Perhaps part of the problem is that you don't always have the option to rest your injury if ...
Introduction to Information Retrieval

class negation synonyms

painkillers that don't upset stomach

About 29,700 results (0.15 seconds)

Symptoms of Crohn's

Best Pain Reliever, Over the Counter Pain Relievers, Prescription ...
For stomach upset, eating or drinking milk before you take a painkiller can help. ... Don't expect painkillers to relieve all your pain, but the best pain ...
www.healthblurbs.com/best-pain-reliever-over-the-counter-pain-relievers-prescription-analgesic-painkillers/ - Cached - Similar

Why do painkillers make your stomach hurt? - Yahoo! Answers
Aug 30, 2008 ... stomach upset, reflux/indigestion, or if severe ulcers ... I often take painkillers FOR stomach aches lol. Don't do this again. 2 years ago ...
answers.yahoo.com › Health › Other - Health - Cached - Similar

How are painkillers bad for you? - Apr 13, 2010
What effect do painkillers have on you? - Sep 30, 2009
What's the difference between OTC painkillers? (Advil, Aleve ... - Nov 19, 2008
Is it bad to take painkillers on an empty stomach? - Aug 20, 2008
More results from answers.yahoo.com »

Phys Ed: Does Ibuprofen Help or Hurt During Exercise? - NYTimes.com
Sep 1, 2009 ... When are ibuprofen and other anti-inflammatory painkillers justified? ... Can't take NSAIDS due to upset stomach. — bonemri ... Perhaps part of the problem is that you don't always have the option to rest your injury if ...
"painkillers such as * tylenol"

About 19,200 results (0.21 seconds)

TYLENOL® Official Site
Tylenol.com Pain Relief from TYLENOL. Get Information and Coupons Here. Head & Body Pain - Arthritis Pain - Sinus - Cold & Flu

Ovarian hyperstimulation syndrome (OHSS): Treatments and drugs...
To ease abdominal discomfort, take over-the-counter painkillers such as acetaminophen (Tylenol) or ibuprofen (Advil, Motrin, others)....
www.mayoclinic.com/.../DSECTION=treatments-and-drugs - Cached - Similar

tylenol ⇒ painkiller
upset stomach ⇒ adverse effect (on stomach)
tylenol \Rightarrow no adverse effects (on stomach)
\Rightarrow no upset stomach
painkiller    upset    stomach
“Things, not strings”

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>Requires</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term</td>
<td>Concept</td>
<td>Parsing, disambiguation, coreference</td>
</tr>
<tr>
<td>Term identity</td>
<td>Entailment</td>
<td>Concept relations</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>Syntactic relation</td>
<td>Document structure, parsing</td>
</tr>
<tr>
<td>Term index</td>
<td>Semantic index</td>
<td>Concept disambiguation, inference</td>
</tr>
</tbody>
</table>
Slides from Patrick Pantel (MSFT)

Entity Detection

Structured Recommendations

Play trailer

Structured Data
Introduction to Information Retrieval

Direct Answer

Structured Data

45 years (March 27, 1970)
Mariah Carey, Age

Mariah Carey - Wikipedia, the free encyclopedia

Mariah Carey was born on March 27, 1970 in Huntington, New York. Her father, Alfred Roy, was of African American and Venezuelan (including Afro-Venezuelan) descent, while her mother, Patricia (Hickey), is of White Irish descent.

Nick Cannon - Tommy Mottola - Albums - Me. I Am Mariah... The Elusive ...
Patrick Pantel talk
(Then) Current experience
Desired experience: Towards actions

Actions easily accessible

Recognize entity in query

Introduction to Information Retrieval
Politician

Actions easily accessible

Recognize entity in query
Actions vs. Intents

User Intents and Goals

Query

Query Intent

Finer-grained Intents

Actions on Entities

plan vacation

get in shape

hilton orlando reviews

sea world location

how to lose weight

Informational

Navigational

Transactional

get address(landmark)

add to Netflix queue(film)

read reviews(hotel)

buy(camera)

[Broder, 2002]

[Rose and Levinson, 2004]
Learning actions from web usage logs

- Three months of us-en web logs
- Annotate with Freebase entities
- Keep queries with an entity in set of 21 types
- Filter out navigational queries
- Filter out clicked hosts that weren’t clicked at least 100 times

<table>
<thead>
<tr>
<th>2,164,579 (query, host) pairs over 3 months</th>
<th>235,385 entities</th>
<th>21 types</th>
</tr>
</thead>
<tbody>
<tr>
<td>129,088 contexts</td>
<td>58,123 hosts</td>
<td></td>
</tr>
</tbody>
</table>

get reviews  ←  Orlando hotel reviews
read biography
download software
watch shows online
Does Hope Solo have a boyfriend?
Free Winzip download
watch family guy online
- A hodgepodge of related strings
- Only actionable through search
Big Wins

- Typed relations
- User Interface accesses structured data
- Click through experience can now leverage strongly-typed identifier
Big Wins

- Typed relations
- User Interface accesses structured data
- Click through experience can now leverage strongly-typed identifier
- Brokered Actions (one click conversions)
Entity disambiguation and linking

- Key requirement is that entities get identified
  - Named entity recognition (e.g., Stanford NER!)
- and disambiguated
  - Entity linking (or sometimes “Wikification”)
    - e.g., Michael Jordan the basketballer or the ML guy
Sergio talked to Ennio about Eli's role in the Ecstasy scene. This sequence on the graveyard was a highlight in Sergio's trilogy of western films.
and linked to a canonical reference

- Freebase
- dbPedia
- Yago2
- (WordNet)
Understanding questions

Chris Watson
The 3rd Prime Minister of Australia

Chris Watson - Wikipedia, the free encyclopedia

3rd Prime Minister of Australia ... He was the first prime minister from the Australian Labour Party, and the first prime minister from the labour movement in the ... Early life - Union activities and colonial ... - Federation - Prime Minister in 1904

List of Prime Ministers of Australia - Wikipedia, the free ...
2017 ...

Google search for "australia's third prime minister" shows the following results:

Chris Watson - Wikipedia, the free encyclopedia
3rd Prime Minister of Australia. He was the first prime minister from the Australian Labour Party, and the first prime minister from the labour movement in the... Early life - Union activities and colonial... - Federation - Prime Minister in 1904

List of Prime Ministers of Australia by time in office - Wikipe... en.wikipedia.org/...List_of_Prime_Ministers_of_Australia_by...  •  Wikipedia
This is a list of Prime Ministers of Australia by time in office. The basis of... Rank: No: Prime Minister: Portrait: Party: District: Assumed Office.... 3rd, Chris Watson

List of Prime Ministers of Australia - Wikipedia, the free ... en.wikipedia.org/wiki/List_of_Prime_Ministers_of_Australia  •  Wikipedia
Below is a list of Prime Ministers of the Commonwealth of Australia. The parties shown are those to which... 17 February 1917, 9 February 1923, —, 3rd Hughes. Edmund Barton - Protectionist Party - Longest serving Prime Ministers
2017 ...
2019

Chris Watson

People also search for
- Andrew Fisher
- Arthur Fadden
- George Reid
- Frank Forde
- Billy Hughes
- Edmund Barton
- Joseph Cook

More about Chris Watson
3 approaches to question answering: Knowledge-based approaches (Siri)

- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia infoboxes, dbPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases
  - Wolfram Alpha
Text-based (mainly factoid) QA

- **QUESTION PROCESSING**
  - Detect *question type*, *answer type*, focus, relations
  - Formulate queries to send to a search engine

- **PASSAGE RETRIEVAL**
  - Retrieve ranked documents
  - Break into suitable passages and rerank

- **ANSWER PROCESSING**
  - Extract candidate answers (as named entities)
  - Rank candidates
    - using evidence from relations in the text and external sources
Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification
Texts are Knowledge
Knowledge: Jeremy Zawodny says ...

![Jeremy Zawodny](image)
Is the goal to go from language to knowledge bases?

• For humans, going from the largely unstructured language on the web to actionable information is effortlessly easy.
• But for computers, it’s rather difficult!
• This has suggested to many that if we’re going to produce the next generation of intelligent agents, which can make decisions on our behalf
  • Answering our routine email
  • Booking our next trip to Fiji

then we still first need to construct knowledge bases

• To go from languages to information
• But should we rather just have computers work with language?
Knowledge:
Not just semantics but pragmatics

Pragmatics = taking account of context in determining meaning
A natural part of language understanding and use

Search engines are great because they inherently take into account pragmatics (“associations and contexts”)

- [the national] → The National (a band)
- [the national ohio] → The National - Bloodbuzz Ohio – YouTube
- [the national broadband] → www.broadband.gov
Lemmon was awarded the Best Supporting Actor Oscar in 1956 for *Mister Roberts* (1955) and the Best Actor Oscar for *Save the Tiger* (1973), becoming the first actor to achieve this rare double...

Source: Jack Lemmon -- Wikipedia
Assume that there is an underlying alignment

- Describes which words in and can be associated

See if the (syntactic/semantic) relations support the answer

What is the fastest car in the world?

The Jaguar XJ220 is the dearest, fastest and most sought after car on the planet.

[Harabagiu & Moldovan, 2001]
Full NLP QA: LCC (Harabagiu/Moldovan)

[below is the architecture of LCC’s QA system circa 2003]
DrQA: Open-domain Question Answering

WIKIPEDIA

ask me
anything

---

58
Open-domain Question Answering

SQuAD
Q: How many of Warsaw's inhabitants spoke Polish in 1933? A: 833,500

TREC
Q: What U.S. state’s motto is “Live free or Die”? A: New Hampshire

WebQuestions (Berant et al, 2013) ≈ Freebase
Q: What part of the atom did Chadwick discover? A: neutron

WikiMovies (Miller et al, 2016)
Q: Who wrote the film Gigli? A: Martin Brest
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

833,500
**Document Retriever**

- Traditional tf.idf inverted index + efficient bigram hash

---

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Wiki Search</th>
<th>Doc. plain</th>
<th>Retriever +bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>62.7</td>
<td>76.1</td>
<td>77.8</td>
</tr>
<tr>
<td>CuratedTREC</td>
<td>81.0</td>
<td>85.2</td>
<td>86.0</td>
</tr>
<tr>
<td>WebQuestions</td>
<td>73.7</td>
<td>75.5</td>
<td>74.4</td>
</tr>
<tr>
<td>WikiMovies</td>
<td>61.7</td>
<td>54.4</td>
<td>70.3</td>
</tr>
</tbody>
</table>

- 70-86% of questions we have that the answer segment appears in the **top 5 articles**
Stanford Attentive Reader

Input

Passage (P)

Question (Q)

Output

Q Which team won Super Bowl 50?

Answer (A)

Which team won Super Bowl 50?
Who did Genghis Khan unite before he began conquering the rest of Eurasia?

He came to power by uniting many of the nomadic tribes of Northeast Asia. After founding the Mongol Empire and being proclaimed "Genghis Khan", he started the Mongol invasions that resulted in the conquest of most of Eurasia. These included raids or invasions of the Qara Khitai, Caucasus, Khwarezmid Empire, Western Xia and Jin dynasties. These campaigns were often accompanied by wholesale massacres of the civilian populations – especially in the Khwarezmian and Xia controlled lands. By the end of his life, the Mongol Empire occupied a substantial portion of Central Asia and China.
## SQuAD Results (single model)

<table>
<thead>
<tr>
<th>Model Description</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>51.0</td>
</tr>
<tr>
<td>Fine-Grained Gating (Carnegie Mellon U)</td>
<td>73.3</td>
</tr>
<tr>
<td>Match-LSTM (Singapore Management U)</td>
<td>73.7</td>
</tr>
<tr>
<td>DCN (Salesforce)</td>
<td>75.9</td>
</tr>
<tr>
<td>BiDAF (UW &amp; Allen Institute)</td>
<td>77.3</td>
</tr>
<tr>
<td>Multi-Perspective Matching (IBM)</td>
<td>78.7</td>
</tr>
<tr>
<td>ReasoNet (MSR Redmond)</td>
<td>79.4</td>
</tr>
<tr>
<td>DrQA (Chen et al. 2017)</td>
<td>79.4</td>
</tr>
<tr>
<td>r-net (MSR Asia) [Wang et al., ACL 2017]</td>
<td>79.7</td>
</tr>
<tr>
<td>Human performance</td>
<td>91.2</td>
</tr>
</tbody>
</table>
General questions

Combined with **Web search**, we can answer **57.5% of trivia questions** correctly.

**Q:** The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
**A:** The Guns of Navarone

**Q:** American Callan Pinckney’s eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
**A:** Fitness
Demo

Hello! Please ask a question.

What is question answering?

a computer science discipline within the fields of information retrieval and natural language processing

Who was the winning pitcher in the 1956 World Series?

Don Larsen

What is the answer to life, the universe, and everything?

42