Web Search in 2020?

The web, it is a changing.

What will people do in the 2020s?
- Type key words into a search box?
- Use the Semantic Web / Web 2.0
  – (and what will this be?)
- Speak to your computer with natural language search?
- Use social or “human powered” search?
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.”


“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
The future?

- Some scattered ideas from the corporate world
  Courtesy Chris Manning

What’s been happening?

- “Mobilegeddon” (Apr 21, 2015):
  - “Mobile friendliness” as a major ranking signal
- “Pigeon” update (July 2014):
  - More use of distance and location in ranking signals
- “App Indexing” (Android, iOS support May 2015)
  - Search results can take you to an app

Why?
- About half of all searches are now from mobile
- Making/wanting good changes, but obvious self-interest in trying to keep people using mobile web rather than apps
Google: What’s been happening? 2014

• New search index at Google: “Hummingbird”

• Answering long, “natural language” questions better
  – Partly to deal with spoken queries on mobile

• More use of the Google Knowledge Graph
  – Concepts versus words

What’s been happening

• 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

• 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?
Two goals

• Things not strings
• Inference not search

Two paradigms for question answering

• Text-based approaches
  – TREC QA, IBM Watson
• 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)

```
painkillers that don't thin blood
painkillers that don't thin blood
painkillers that don't cause constipation
painkillers that don't cause nausea
painkillers that don't upset stomach
painkillers that don't upset stomach

About 89,300 results (0.14 seconds)
```

```
painkillers that don't upset stomach

About 26,700 results (0.15 seconds)
```

- Symptoms of Crohn's
- Do You Know the Signs of Crohn's? Read About Symptoms and More Now.

- 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. — bonenri ... Perhaps part of the problem is that you don't always have the option to rest your injury if ...
Related concepts may not match question

tylenol $\Rightarrow$ painkiller
Text can identify synonyms

upset stomach ⇒ adverse effect (on stomach)

tylenol ⇒ no adverse effects (on stomach)
⇒ no upset stomach
This took us from Query

- painkiller
- upset
- stomach

To Answer

- painkiller
- upset
- stomach
- tylenol
- adverse effects

Relationships:
- not
- is-a
- a-kind-of
- arg
### “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
**Direct Answer**

45 years (March 27, 1970)

Mariah Carey, Age

Madonna 56 years
Nicky Jam 34 years
Jennifer Lopez 46 years

**Structured Data**

Mariah Carey - Wikipedia, the free encyclopedia

Mariah Carey, Singer-songwriter

Mariah Carey is an American singer, songwriter, record producer, and occasional actress. She rose to prominence after releasing her self-titled debut studio album Mariah Carey in 1990. It went multiplatinum... Wikipedia

Born: March 27, 1970 (age 49), Huntington, NY
Height: 5' 9" (1.75 m)

Albums

**Query Answering**

48 years (September 18, 1965)

Christopher D. Manning, Age

Christopher Manning, Stanford NLP - The Stanford NLP
nlp.stanford.edu/manning/

Mar 1, 2013 - Chris Manning works on systems and formalisms that can intelligently process and produce human languages. His research concentrates on...

Rocchio algorithm - Wikipedia, the free encyclopedia

rocchio.algorithminformationretrieval.page181

Algorithm - Time complexity - Usage - Limitations
But not always…

Patrick Pantel talk
(Then) Current experience
Desired experience: Towards actions

Politician
Actions vs. Intents

User Intents and Goals

Query

- Hilton Orlando reviews
- Sea World location
- How to lose weight

Query Intent

- Informational
- Navigational
- Transactional

[Broder, 2002]

Finer-grained Intents

- Advice
- Locate
- Download
- Obtain
- Interact

[Rose and Levinson, 2004]

Actions on Entities

- Get address (landmark)
- Add to Netflix queue (film)
- Read reviews (hotel)
- Buy (camera)

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

- \(2,164,579\) (query, host) pairs
- \(235,385\) entities
- \(129,088\) contexts
- \(58,123\) hosts
- A hodgepodge of related strings
- Only actionable through search
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

Mentions, Meanings, Mappings

- 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.

Mentions (surface names)

- Sergio
- Ennio
- Eli
- Ecstasy

Entities (meanings)

- Sergio means Sergio_Leone
- Sergio means Serge_Gainsbourg
- Ennio means Ennio_Antonelli
- Ennio means Ennio_Morricone
- Eli means Eli_(bible)
- Eli means ExtremeLightInfrastructure
- Eli means Eli_Wallach
- Ecstasy means Ecstasy_(drug)
- Ecstasy means Ecstasy_of_Gold
- trilogy means Star_Wars_Trilogy
- trilogy means Lord_of_the_Rings
- trilogy means Dollars_Trilogy

KB
Incorporate structured info?

- and linked to a canonical reference
  - Freebase, dbPedia, Yago2, (WordNet)

Facebook Graph Search
The Facebook Graph

• Collection of entities and their relationships
• Entities (users, pages, photos, etc.) are nodes
• Relationships (friendship, checkins, tagging, etc.) are edges
• Nodes and edges have metadata
• Nodes have a unique id – the fbid
Facebook Graph Snippet

fbid: 213708728685
type: PAGE
name: Breville
mission: To design the best ...
...

fbid: 586206840
type: USER
name: Sriram Sankar
...

© 2017 Christopher W. Clifton
Facebook Graph Search

- Uses a weighted context free grammar (WCFG) to represent the Graph Search query language:
  - [start] => [users] $1
  - [users] => my friend friends(me)
  - [users] => friends of [users] friends($1)
  - [users] => {user} $1
  - [start] => [photos] $1
  - [photos] => photos of [users] photos($1)
- A terminal symbol can be an entity, e.g., {user}, {city}, {employer}, {group}; it can also be a word/phrase, e.g., friends, live in, work at, members, etc. A parse tree is produced by starting from [start] and expanding the production rules until it reaches terminal symbols.

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
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
Knowledge: Not just semantics but pragmatics

- Pragmatics = taking account of context in determining meaning

- 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

Inference directly in text: Natural Logic
(van Benthem 2008, MacCartney & Manning 2009)

Q Who is Beyoncé Knowles’s husband?
Q’ Beyoncé Knowles’s husband is X.

Task – Answer Sentence Selection

- Given a factoid question, find the sentence that
  - Contains the answer
  - Can sufficiently support the answer

Q

S1

S2

Scott Wen-tau Yih (ACL 2013) paper

Who won the best actor Oscar in 1973?

Jack Lemmon

Best Supporting Actor Oscar
Mister Roberts

Best Actor Oscar
Save the Tiger

Jack Lemmon - Wikipedia, the free encyclopedia


John Henry "Jack" Lemmon III was an American actor and musician. He starred in more than 60 films including Some Like It Hot and Witness. He was also famous for his role as the lead character in The Apartment and won multiple awards for his performances. He died in 2001.
• What is the fastest car in the world?

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

Full NLP QA: LCC (Harabagiu/Moldovan) [below is the architecture of LCC’s QA system circa 2003]