Web Search in 2020?

The web, it is a changing.

What will people do in 2020?
• Type key words into a search box?
• Use the Semantic Web?
• Speak to your computer with natural language search?
• Use social or “human powered” search?
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.”


---

“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

What's been happening

• 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?
Towards intelligent agents

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)
class negation synonyms

painkillers that don't upset stomach

About 26,700 results (0.15 seconds)

Related concepts may not match question
Text can identify synonyms

painless "adverse effects include" stomach

About 266,000 results (0.14 seconds)

**Painkiller Comparison Chart**
Oct 31, 2008 ... Chart showing prescription and non-prescription painkillers and what they are used for. ... menstrual pain, fever, Common adverse effects include: nausea, asthma, upset stomach, not for last pregnancy trimester ...

www.drugs.com/pain/chart.html - Cached - Similar

**List of Painkillers and what they are used for: Over the Counter**
File Format: PDF/Adobe Acrobat - Quick View
Common adverse effects include: nausea, dyspepsia, gastrointestinal ulcers/bleeding ... upset stomach, not for last pregnancy trimesters. Individuals ...

healingwithoils.com/.../Painkiller%20side%20effects%20chart%20.pdf

upset stomach => adverse effect (on stomach)

---

tylenol stomach

About 1,390,000 results (0.28 seconds)

**Pain Relievers And Upset Stomach - Tylenol Extra Strength Pain**
Tylenol Extra Strength Pain Reliever Fever Reducer 100 Caplets - User Rating: 5 stars. Review Summary: Taking the wrong pain reliever can cause you stomach ...

www.apinions.com > Wellness & Beauty > Health Aids - Cached - Similar

**Tylenol** has virtually no useful antiinflammatory activity, does not affect the blood's ability to clot, and has almost no adverse effects on the stomach ...

www.merck.com/mmhe/sec06/ch378/ch378d.html - Cached - Similar

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
- is-a
- a-kind-of
- arg
- not
“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
But not always…

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

Politician
Actions vs. Intents

User Intents and Goals
- plan vacation
- get in shape

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 over 3 months
- 235,385 entities
- 129,088 contexts
- 58,123 hosts
- 21 types

→ Orlando hotel reviews
→ Does Hope Solo have a boyfriend?
→ Free Winzip download
→ watch family guy online
Big Wins

- Typed relations
- User Interface accesses structured data
- Click through experience can now leverage strongly-typed identifier
Entity disambiguation and linking

• Key requirement is that entities get identified
  – Named entity recognition (e.g., Stanford NER!)

• and disambiguated
  – Entities get "something like a unique identification”)
  – 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.

Mentions, Meanings, Mappings

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

Eli (bible)
Eli Wallach
Ecstasy (drug)
Ecstasy of Gold
Star Wars Trilogy
Lord of the Rings
Dollars Trilogy

Mentions (surface names)

Entities (meanings)
Incorporate structured info?

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

http://.../wiki/Dollars_Triology
http://.../wiki/The_Good_the_Bad
http://.../wiki/Clint_Eastwood
http://.../wiki/Honorary_Academ

http://.../wiki/The_Good_the_Bad
http://.../wiki/Metallica
http://.../wiki/Bellagio_(casino)
http://.../wiki/Ennio_Morricone

http://.../wiki/Sergio_Leone
http://.../wiki/The_Good_the_Bad
http://.../wiki/For_a_Few_Dollars
http://.../wiki/Ennio_Morricone

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

EVENT
LIKES
FRIEND
TAGGED
PHOTO
Facebook Graph Search

- Uses a weighted context free grammar (WCFG) to represent the Graph Search query language:
  - \([\text{start}] \rightarrow \text{[users]}\) $1$
  - \([\text{users}] \rightarrow \text{my friend}\) friends(me)
  - \([\text{users}] \rightarrow \text{friends of [users]}\) friends($1$)
  - \([\text{users}] \rightarrow \{\text{user}\}\) $1$
  - \([\text{start}] \rightarrow \text{[photos]}\) $1$
  - \([\text{photos}] \rightarrow \text{photos of [users]}\) photos($1$)
- A terminal symbol can be an entity, e.g., \{\text{user}\}, \{\text{city}\}, \{\text{employer}\}, \{\text{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] \(\rightarrow\) The National (a band)
  - [the national ohio] \(\rightarrow\) The National - Bloodbuzz Ohio – YouTube
  - [the national broadband] \(\rightarrow\) www.broadband.gov

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

Q \(\text{Who is Beyoncé Knowles’s husband?}\)
Q’ \(\text{Beyoncé Knowles’s husband is } X.\)

A \(\text{Beyoncé’s marriage to rapper Jay-Z and portrayal of Etta James in Cadillac Records (2008) influenced her third album I Am... Sasha Fierce (2008).}\)
Task – Answer Sentence Selection

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

Q: Who won the best actor Oscar in 1973?
S1: Jack Lemmon was awarded the Best Actor Oscar for Save the Tiger (1973).
S2: Academy award winner Kevin Spacey said that Jack Lemmon is remembered as always making time for others.

Scott Wen-tau Yih (ACL 2013) paper
What is the fastest car in the world?

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