



Google Squared

*Web scale, open domain information
extraction and presentation*

Dan Crow
Google



Project aims

- *Web scale*: extract from tens of billions of pages
- *Open domain*: answer questions on any topic
- Automatic extraction, no manual intervention
- Solve real user problems
- Learn from user feedback
- Not limited by traditional search UI
- No technology religion: solve problems using any methodology available

Google Squared



currencies

Square it

Add to this Square







Unsaved

Share

Export

Save

currencies 20 item

Item Name	Image	Description	Symbol	Central Bank	Banknotes	Add columns	Add
<input checked="" type="checkbox"/> Euro		The euro is the second largest reserve currency and the second most traded currency in the world after the U.S. dollar. As of October 2009, with more than ...	EC	European Central Bank	€5 · €10 · €20 · €50 · €100 · €200 · €500		
<input checked="" type="checkbox"/> Japanese yen		The yen declined during the Japanese asset price bubble and continued to do so The exchange rate for the Japanese yen is en.wikipedia.org	JPY	Bank of Japan	¥1000, ¥2000, ¥5000, ¥10000		
<input checked="" type="checkbox"/> Australian dollar		American Dollar, Argentine Peso, Australian Dollar , Brazilian Real, British Pound, Bulgarian Lev, Canadian Dollar, Chilean Peso, Chinese Yuan ...	AUD	Reserve Bank of Australia	\$5, \$10, \$20, \$50, \$100		
<input checked="" type="checkbox"/> Canadian dollar		Argentine Peso . Australian Dollar . Bahraini Dinar . Botswana Pula . Brazilian Real . British Pound . Brunei dollar . Bulgarian Lev . Canadian Dollar ...	CAD	Bank of Canada	1 possible value		
<input checked="" type="checkbox"/> Swiss franc		The franc (German: Franken, French and Romansh: franc , Italian: franco; code: CHF) is the currency and legal tender of Switzerland and Liechtenstein; ...	CHF	Swiss National Bank	10, 20, 50, 100, 200 & 1000 francs		
<input checked="" type="checkbox"/> Malaysian ringgit		The Malaysian ringgit (plural: ringgit; currency code MYR; formerly the Malaysian dollar) is the currency of Malaysia. It is divided into 100 sen (cents)	RM	Bank Negara Malaysia	RM1, RM5, RM10, RM50, RM100		



Comparison: an interesting search problem

Many users want to compare items in a topic:

- I'm going on safari in South Africa
- Write a school paper about the US presidents
- Research digital cameras
- Choose a restaurant near the British Museum
- Who were the conspirators in the Gunpowder Plot?
- Compare sedimentary rocks

Need to gather data from many sources and the same data about multiple objects

Tedious, time consuming, but high value



How users compare today

Users in "comparison mode" look for information, not pages

Two main phases:

- **Research** - learn about the domain
- **Acquire** - find specific answers

People use: spreadsheets, email, post-its, memory to record and organize searches

They are frustrated by the inability to find information, by the effort involved, give up before the task is complete

Oh, and, users love tables



What Google Squared does

Query to list of names:

[us presidents] -> Ford, Nixon



What Google Squared does

Extend list of names:

Ford, Nixon -> Obama, Carter, Reagan

Ford, Chrysler -> BMW, Honda, Audi



What Google Squared does

Find attributes:

Ford, Nixon, Obama, Carter, Reagan:

Date of birth

Preceded by

Party

Vice President

Religion



What Google Squared does

Find values:

	<i>Date of birth</i>	<i>Vice president</i>	<i>Party</i>	<i>Religion</i>
<i>Ford</i>	14 July 1913	Nelson Rockefeller	Republican	Episcopalian
<i>Nixon</i>	9 January 1913	Gerald Ford	Republican	Quaker
<i>Obama</i>	4 August 1961	Joe Biden	Democrat	United Church of Christ
<i>Carter</i>	1 October 1924	Walter Mondale	Democrat	Baptist

How it works: query analysis

Is the query about an item or a category?

[Obama] or [US Presidents]?

Is this a product or a local query?

[mp3 players] or [cambridge restaurants]

If not, its a web search query:

[active baseball players named in the mitchell report]



Extraction: Query to list of names

Offline:

- Find web pages that contain lists and tables
- Look for likely entity names
- Look for likely subject names (headers, page titles)
- Aggregate over the entire web
- Find synonyms and alternatives

Query time:

- Run searches, e.g. [List of <query>], Wikipedia category pages
- Find extracted lists from search results



Extraction: find attributes

Offline table extractor:

- Ignore layout tables
- Extract row and column headers
- Aggregate tables

Hundreds of millions of tables extracted

Query time:

- Search for tables containing list of items
- Look for attribute candidates in headers

Large scale synonym data to find canonical attribute:

born, birthdate, birth date,
birthday, date of birth -> date of birth



Extraction: find values

Offline:

- Table extractors
- NLP extractors (verb and possessive fact extraction)
- Type-specific extractors (dimensions, price, date, location...)
- Page structure analysis
- Score extractors using Rifle classifier

Web scale: tens of billions of extracted facts





Query time:

- Run: [context, item, attribute]
- Search snippets to find similar values



Learn from user feedback

Look for consistent value corrections, increase confidence

Item Name	Image	Description	Preceded By	Vice President
Gerald Ford		Gerald Rudolph Ford, Jr. (born Leslie Lynch King, Jr.; July 14, 1913 – December 26, 2006) was the 38th President of the United States, serving from 1974 to ...	Spiro Agnew	Nelson Rockefeller
James Monroe		On New Year's Day, 1825, at the last of his annual White House receptions, President James Monroe made a pleasing impression upon a Virginia lady who shook ...	<input checked="" type="radio"/> Spiro Agnew Preceded by for Gerald Ford simple.wikipedia.org - all 3 sources »	
John Tyler		Dubbed "His Accidency" by his detractors, John Tyler was the first Vice President to be elevated to the office of President by the death of his predecessor. ...	<input type="radio"/> Richard Nixon <small>Low confidence</small> Preceded by for Gerald Ford simple.wikipedia.org - all 2 sources »	
Dwight Eisenhower		Dwight David "Ike" Eisenhower (pronounced /'aɪzənhaʊər/ EYE-zən-how; October 14, 1890 – March 28, 1969) was a five-star general in the United States Army	<input type="radio"/> Richard M. Nixon <small>Low confidence</small> Preceded by for Gerald Ford en.wikipedia.org Search for more values »	

Enhancing search results

Bias the result snippet to show and highlight facts, where we have high confidence:

From:

[How tall is the Eiffel tower? - Yahoo! Answers](#)




According to The Oxford Dictionary of Phrase and Fable: "**Eiffel Tower** a wrought-iron structure erected in Paris for the World Exhibition of 1889, designed and built ...

[answers.yahoo.com/question/index?qid...](#) - [Cached](#) - [Similar](#) -   

To:

[How tall is the Eiffel tower? - Yahoo! Answers](#)

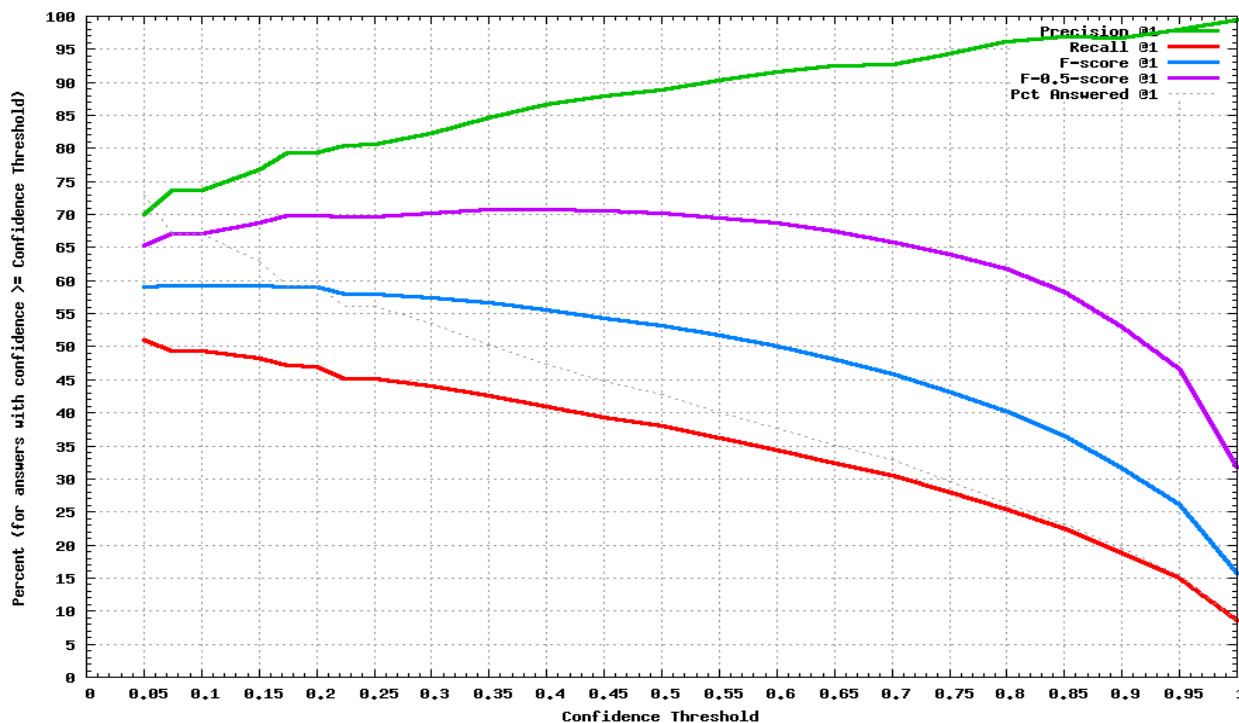
How tall is the Eiffel tower? ... Including the 24 m (79 ft) antenna, the structure is **324 m (1063 ft)** high. 2 years ago. 14% 1 Vote ...

[answers.yahoo.com/question/index?qid...](#) - [Cached](#) - [Similar](#) -   



Evaluation

Continually evaluate precision and recall of individual components and overall system across sets of thousands of hand-evaluated questions



Improving quality

Significant quality improvement:

- Search for more items and attributes than required
- Find values for all items and attributes then prune
- Remove items/attributes that are:
 - **Wrong** - radically different value types
 - **Duplicates** - likely synonyms
 - **Not useful** - no values available

Improves precision and recall around 20%



What we've learned: part I

- Precision is key:
 - Precision from 50%→60%; user satisfaction 50%→ 60%
- Recall also critical:
 - anesthetic solubility, titanium rings, design software, novels of kurt vonnegut, artificial tears, boutiques in san antonio texas, swiss cantons, japanese instruments, green rating systems...
- Deep semantics are hard to extract in the general case
 - We don't support computation across values
 - No-one seems to mind
- Small blacklists can greatly improve quality (uncyclopedia)
- Combine large-scale offline and query-time extraction
 - Search engine ranking is very effective



What we've learned: part II

- Context is important
 - helps disambiguate (Ford vs Ford)
 - Improves precision
- Scale allows you to aggregate the wisdom of the web:
 - Occurrence count
 - Web rank (~=authority)
- If you fail, you can always ask the user
- Users understand tables

- Open domain extraction is a hard and satisfying problem to work on

