

Practical Data Science – COSC2670

Textual Data: Query Suggestions for Search

Dr. Yongli Ren

(yongli.ren@rmit.edu.au)

Computer Science & IT
School of Science

Outline

- Query suggestions and related searches
- Search engine query logs
- Approaches for search suggestions
 - Session-based
 - Click data
 - Statistical (relevance feedback)

Query Suggestions

Game of ?



game of

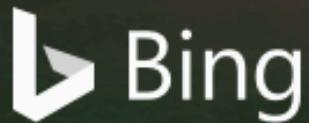


- game of **thrones**
- game of **thrones season 8**
- game of **thrones cast**
- game of **thrones season 7**
- game of **thrones season 8 release date**
- game of **thrones season 1**
- game of **thrones season 8 trailer**
- game of **thrones episodes**
- game of **thrones characters**
- game of **thrones map**

Google Search

I'm Feeling Lucky

Report inappropriate predictions



game of



game of thrones

game of thrones season 8

game of thrones season 7

game of thrones trailer

game of thrones season 1

game of thrones google

game of thrones cast

game of life



DuckDuckGo

game of

X



game of thrones

game of thrones season 8

game of thrones cast

game of thrones trailer

game of thrones characters

game of thrones map

game of thrones wiki

game of thrones books

Google 2010

game of death

X

Search

game of **death**

game of **thrones**

game of **the year**

game of **life**

game of **the year 2010**

About 30,300,000 results (0.06 seconds)

[Advanced search](#)

Bing 2010

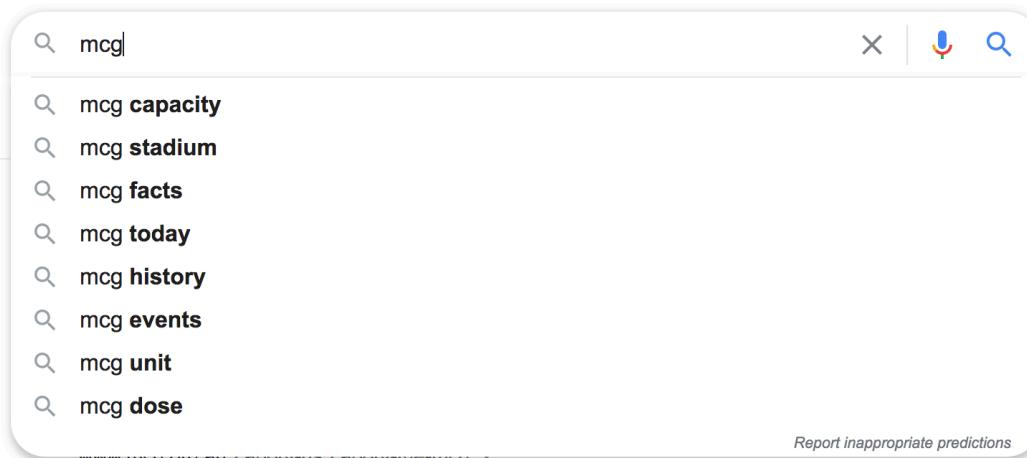
The screenshot shows the Bing search interface. At the top, there is a search bar containing the text "game of". To the right of the search bar is an orange square button with a white magnifying glass icon. Below the search bar, a list of suggested search terms is displayed in a white box with a dark blue border. The suggestions are:

- game of **life**
- game of **death**
- game of **3 halves**
- game of **thrones**
- game of **life board** game
- game of **the year**
- game of **life download**
- game of **love**

At the bottom of the search bar area, there is a link labeled "Manage search history". The background of the entire interface is a dark blue color with some abstract white patterns.

Location-based Differences

Google



mcg

- mcg capacity
- mcg stadium
- mcg facts
- mcg today
- mcg history
- mcg events
- mcg unit
- mcg dose

[www.mcg.org.au / about-us / about-the-mcg](http://www.mcg.org.au/about-us/about-the-mcg) · Report inappropriate predictions

About the MCG - Melbourne Cricket Ground

The **MCG** is more than just a sports venue. It's a place where memories are made and childhood dreams come alive. Ask any Victorian and they'll be aware of ...

www.mcg.org.au › Tours ▾

MCG Tours - Melbourne Cricket Ground

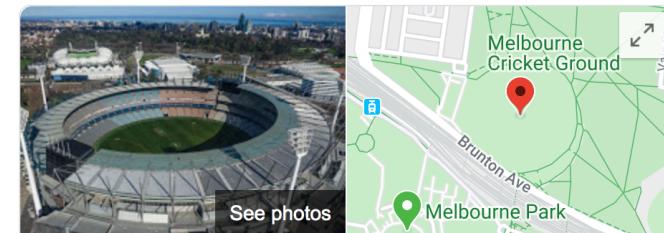
Go behind the scenes of one of the world's most iconic stadiums, the **MCG**. Located just minutes from the CBD, discover the secrets of the **Melbourne Cricket** ...

People also ask

What is the meaning of MCG?

What is the MCG made of?

What is the size of the MCG?



Melbourne Cricket Ground

[Website](#)

[Directions](#)

[Save](#)

4.7 ★★★★★ 17,635 Google reviews

Ground in East Melbourne, Victoria

The Melbourne Cricket Ground, also known simply as "The G", is an Australian sports stadium located in Yarra Park, Melbourne, Victoria.

[Wikipedia](#)

Address: Brunton Ave, Richmond VIC 3002

Capacity: 100,024

Opened: 1853

Owner: MCG Trust

Phone: (03) 9657 8888

[Suggest an edit](#)

Location-based Differences

The Google logo, featuring the word "Google" in its signature multi-colored font.

mcg|



- mcgraw hill connect**
- mcgregor**
- mcgraw hill**
- mcg to mg**
- mcgill university**
- mcg**
- mcgriddle**
- mcguffin**
- mcgee and co**
- mcgregor retired**

Google Search

I'm Feeling Lucky

Report inappropriate predictions

Related Searches

Game of Thrones

5 Apr 2011 ... A fan site dedicated to HBO's **Game of Thrones**, with News, Information and Comment.

www.game-of-thrones.co.uk/ - Cached - Similar

Variety Reviews - Game of Thrones - TV Reviews -- Review by Brian ...

10 Apr 2011 ... Westeros, a mythical land of seven kingdoms where dragons once lived is full of plotting, intrigue and fractious families on all sides, ...

www.variety.com/review/VE1117944992

Searches related to game of thrones

[george rr martin](#)

[game of thrones ccg](#)

[game of thrones hbo](#)

[game of thrones board game](#)

[game of thrones review](#)

[game of thrones rpg](#)

[game of thrones movie](#)

[game of thrones map](#)

Goooooooooooooogle ►

1 2 3 4 5 6 7 8 9 10

[Next](#)

Query Suggestions and Related Searches

- Provided by all major web search engines
- Currently the implementations look very similar
- Function as a kind of “*auto-complete*” feature
- Suggestions are *listed in a selection box*; related searches are embedded as *clickable links* in the results page
 - Both can be selected easily
- They are usually whole queries

Query Suggestion

- Suggested queries provided by search engines may differ based on various features
 - Location of user
 - Search history of user (if available)
 - Connection speed
 - Device being used (smartphone, tablet, computer, ...)
 - Browsing or searching preferences (e.g. language settings)
- The suggestions that you and I see when we type in a query might be different

Why Provide Search Suggestions?

- Make life easier for the user
 - Less typing
 - Help to specify their information need
 - Disambiguate queries
 - Minimise spelling errors
- Make life easier for search engines
 - Reduces overall variability of queries entered
 - Efficiency benefits
 - Effectiveness benefits

Implementing Search Suggestions: Sources of Evidence

- There are three main sources of evidence that are used to generate query suggestions and related searches
 - Search sessions
 - Click data
 - Term statistics

Approach 1: Search Sessions

A Search Log

[16:34:02] Q: godzilla

[16:34:05] C (1): www.imdb.com/title/tt0831387/

[16:34:10] C (5): www.godzillamovie.com/

[16:36:39] Q: godzilla resurgence

[16:36:45] Q: shin gojira

[16:36:47] C (1): www.imdb.com/title/tt4262980/

[16:41:22] Q: python for data science

.....

Search Logs

- Key items
 - Timestamps
 - Queries
 - Clicks (on answer items)
- Advantages
 - Provide direct evidence about searcher behaviour, in a natural setting
 - Once instrumentation is in place, gathering log data is relatively cheap
 - Can quickly accumulate large amounts of evidence



Search Logs: Complications

- Can be difficult to determine **session boundaries**
 - When does **one search “task”** end and another begin?
 - Non-linear search processes / tabbed browsing
- **Amount of data** can become a problem
 - $1\text{k bytes/record} \times 10 \text{ records/query} \times 10 \text{ mil queries/day} = 100 \text{ Gb/day}$
- **What data can be collected**
 - Privacy issues / Ethics considerations / Terms of service
- **How to identify users**
 - Cookies / IP address – but people might share a machine
 - Login / plugin – identifies users more reliably, but not all users will login/use plugin, so data will be limited

Search Logs: Complications...

- Key limitation: logs don't tell us **why** the user did something
- Instead, we need to make assumptions about what those observed actions mean
 - Did the user actually like the item that they clicked on?
 - Did they finish the search session because their information need was *resolved*, or because they were *frustrated* and gave up?

Session Data for Search Suggestions

- Given a query log L , pre-process the log
 - Split into *sessions*
 - For example, based on an inactivity threshold (5 minutes) and a maximum session length (30 minutes)
 - For each query q_i in the log, count how often it has occurred with each other query q_j in the same session
- For a new user query q_{new}
 - Look up the query in the processed log data
 - Return the N most frequently co-occurring past user queries

Session Data for Search Suggestions...

- Query pairs and counts in 10,000,000 sessions

—....

- <jaguar, jaguar wild cat> 3,251
- <jaguar, jaguar car> 7,795
- <jaguar, OSX jaguar> 1,406
- <jaguar, movie jaguar > 341
- <jaguar, jaguar drink> 2,599

—....

- New query: *jaguar* → return:

- jaguar car
- jaguar wild cat

—....

Session Data for Search Suggestions

- Query log pre-processing can be extremely resource intensive
 - But fortunately, can be done offline
 - Needs to be repeated periodically since search behaviour (and therefore trends) change over time
- After processing, the *suggestion* part can be implemented efficiently using a fast lookup data structure (such as a hash table, or a suffix array)

Approach 2: Click Data

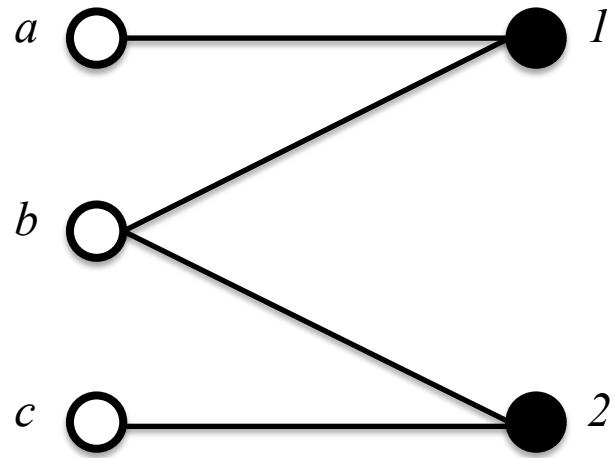
Click Data for Search Suggestions

- In addition to finding **direct relationships** from query co-occurrences in a *search session*, can also use *click data* to try and identify new trends
- Intuition: if different queries lead to clicks on the same page, then those queries are related
- This approach uses a technique called *agglomerative clustering*

Agglomerative Clustering of Click Data: Setup

- Input: Search log L , in the form of (*query*, *URL*) pairs
 - Output: A bipartite graph G
1. Collect Q , the set of unique queries in L
 2. Collect U , the set of unique URLs in L
 3. For each of the queries in Q , create a “white” vertex in G
 4. For each of the URLs in U , create a “black” vertex in G
 5. For each URL (in U) that was clicked in response to a query q (in Q), draw an edge

Agglomerative Clustering of Click Data...

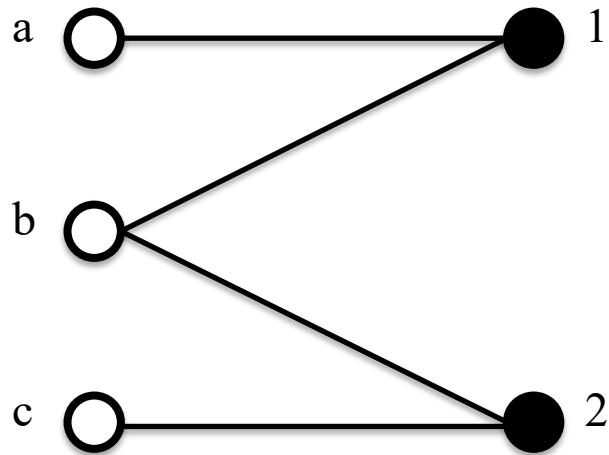


Measuring Similarity

- Let $N(x)$ be the set of vertices that neighbour a vertex x in a graph
 - For a black (white) vertex, the neighbours are the white (black) vertices that it is linked to
- Then the *similarity* of a vertex x with another vertex y can be measured based on the *Jaccard similarity* (or overlap) of their neighbours:

$$sim(x, y) = \begin{cases} \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}, & \text{if } |N(x) \cup N(y)| > 0 \\ 0, & \text{otherwise} \end{cases}$$

Agglomerative Clustering of Click Data...

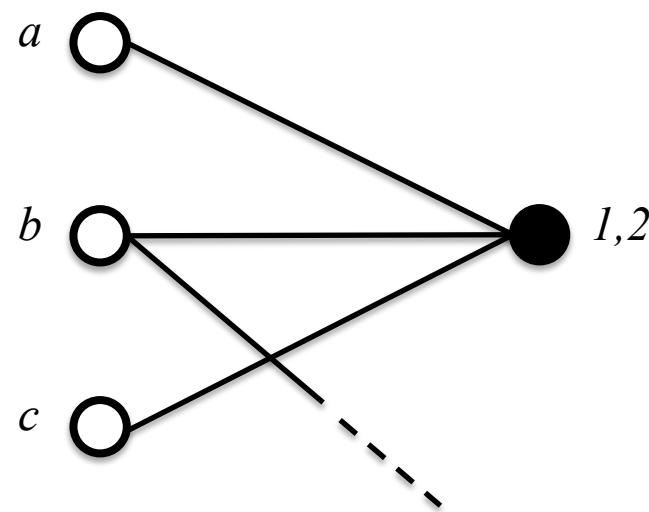
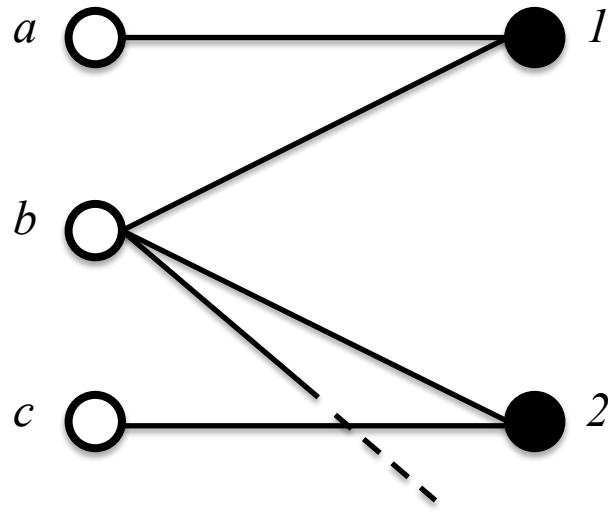


- $\text{sim}(a,b) = 1/2$
- $\text{sim}(a,c) = 0$
- $\text{sim}(b,c) = 1/2$

Agglomerative Clustering of Click Data: Processing

- Input: Bipartite graph G
 - Output: A new bipartite graph G' where each white (black) vertex in G' corresponds to one or more white (black) vertices of G
1. Score all pairs of white vertices in G by $\text{sim}(x,y)$
 2. Merge the two white vertices for which $\text{sim}(x,y)$ is largest
 3. Score all pairs of black vertices in G by $\text{sim}(x,y)$
 4. Merge the two black vertices for which $\text{sim}(x,y)$ is the largest
 5. If (termination condition not met) return to 1.

Agglomerative Clustering of Click Data...



- Iterative clustering is able to uncover latent relationships
- Initially, $\text{sim}(a,c) = 0$
- After merging black nodes 1 and 2 (which have a similarity of 1/3), vertices a and c are actually shown to have something in common

Using Click Data for Query Suggestions

- For example, suppose some of the following queries initially didn't have any clicks directly in common, but shared some URLs through neighbours:

- godzilla
 - shin gojira
 - kaiju movie 2016
 - kamata kun



- The queries appear unrelated from a string-similarity perspective, and didn't need to occur in the same search session, but can now be clustered
- Using the created clusters for query suggestions
 - For a new user query q_{new} that is submitted
 - Find the cluster that contains q_{new}
 - Return up to N members of the cluster

Approach 3: Term Statistics

Statistical Query Suggestions

- When a query is submitted to a search engine, a *ranked list* of search results is returned
 - The ranking is based on the *expected relevance* of each resource to the user's query
 - Items that are higher in the ranking can be expected to be a better match than items that are lower in the list
- This observation can be exploited to identify related terms that might help to improve a query

Automatic Relevance Feedback

- Input: a user query q
 - Output: a set of term suggestions s
1. Submit q to a search engine and obtain a *ranked list* of results
 2. Assume that the top N results are relevant
 - Let the list of all *terms* occurring in these documents be a candidate set, C
 3. Process set C , to *count how often each unique term occurs*
 4. Select the R most frequently occurring *terms*, and return them as query suggestions

Step 1: Run Initial Query

Google godzilla

All Images Videos News Shopping More Settings Tools

About 36,400,000 results (0.63 seconds)

Godzilla - Wikipedia
<https://en.wikipedia.org/wiki/Godzilla> ▾
Godzilla (Japanese: ゴジラ, Hepburn: Gojira) is a monster originating from a series of tokusatsu films of the same name from Japan. The character first appeared ...
Created by: Tomoyuki Tanaka; Ishirō Honda; Eiji ... First appearance: Godzilla (1954)
Aliases: : King of the Monsters; Gigantis; Monst... Family: Minilla and Godzilla Junior (adopted s...
Godzilla (2014 film) · Son of Godzilla · Godzilla · Godzilla (franchise)

Godzilla (2014 film) - Wikipedia
[https://en.wikipedia.org/wiki/Godzilla_\(2014_film\)](https://en.wikipedia.org/wiki/Godzilla_(2014_film)) ▾
Godzilla is a 2014 American monster film directed by Gareth Edwards and written by Max Borenstein, from a story by David Callaham. The film is a reboot of ...
MUTO · High-altitude military parachuting · Akira Takarada

Godzilla (2014) - IMDb
<https://www.imdb.com/title/tt0831387/> ▾
★★★☆☆ Rating: 6.4/10 - 331,626 votes
Action Ken Watanabe and Gareth Edwards in **Godzilla (2014)** David Strathairn and Ken ... **Godzilla**. **Pacific Rim**. **Transformers: Dark of the Moon**. **The Mummy**.

Godzilla (1998) - IMDb
<https://www.imdb.com/title/tt0120685/> ▾
★★★☆☆ Rating: 5.3/10 - 163,859 votes
Action ... **Godzilla (1998)** Matthew Broderick in **Godzilla (1998)** Jean Reno and ... People who liked this also liked... **Godzilla**. **King Kong**. **The Mummy**. **Twister**.



More images

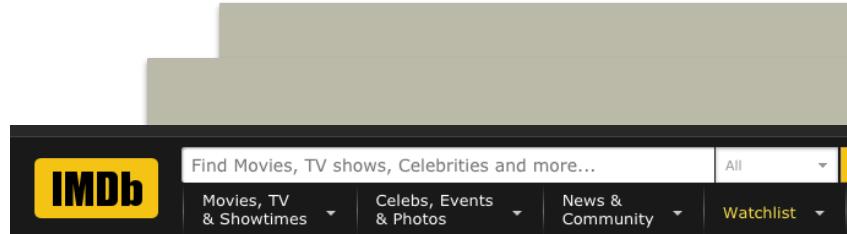
Godzilla

Film character

Godzilla is a monster originating from a series of tokusatsu films of the same name from Japan. The character first appeared in Ishirō Honda's 1954 film **Godzilla** and become a worldwide pop culture icon, ...
[Wikipedia](#)

Species: Mutated dinosaur
First appearance: **Godzilla (1954)**
Designed by: Akira Watanabe, Teizō Toshimitsu, Eiji Tsuburaya
Movies: **King Kong vs. Godzilla**, **Godzilla**, **Godzilla: Final Wars**, **MORE**
Family: **Minilla**, **Godzilla Junior** (adopted sons)

Step 2: Extract Terms From Top Documents



Find Movies, TV shows, Celebrities and more... All

Movies, TV & Showtimes Celebs, Events & Photos News & Community Watchlist



Not logged in Talk Contributions Create account Log in

Article Talk Read View source View history Search Wikipedia

Wiki Loves Earth: An international photographic contest where you can showcase Australia's unique natural environment and potentially win a prize.

Godzilla

From Wikipedia, the free encyclopedia

This article is about the monster. For the film franchise, see [Godzilla \(franchise\)](#). For other uses, see [Godzilla \(disambiguation\)](#).

"ゴジラ" redirects here. For other uses of "Gojira", see [Gojira \(disambiguation\)](#).

Godzilla (Japanese: ゴジラ Hepburn: *Gojira* /*go.dz̥ɪ.ra*/) is a monster originating from a series of tokusatsu films of the same name from Japan. The character first appeared in Ishirō Honda's 1954 film *Godzilla* and became a worldwide pop culture icon, appearing in media including 29 films produced by Toho, three Hollywood films, and numerous video games, novels, comic books, television shows. It is often dubbed the "King of the Monsters", a phrase first used in *Godzilla, King of the Monsters!*, the Americanized version of the original film.

Godzilla is depicted as an enormous, destructive, prehistoric sea monster awakened and empowered by nuclear radiation. With the nuclear bombings of Hiroshima and Nagasaki and the *Lucky Dragon* 5 incident still fresh in the Japanese consciousness, Godzilla was conceived as a metaphor for nuclear weapons.^[20] As the film series expanded, some stories took on less serious undertones, portraying Godzilla as an *anthero*, or a lesser threat who defends humanity. With the end of the Cold War, several post-1984 Godzilla films shifted the character's portrayal to themes including Japan's forgetfulness over its imperial past,^[21] natural disasters, and the human condition.^[22]

First appearance *Godzilla* (1954)

Created by Tomoyuki Tanaka
Ishirō Honda
Eiji Tsuburaya

Portrayed by Shōwa era:
Haruo Nakajima^[1]
Katsumi Tezuka^[2]
Hiroshi Sekida^[3]
Seiji Onaka^[3]
Shinji Takagi^[4]



godzilla, dinosaur,
kaiju,
japan,
toho, monster,
movie, 1964,
ghidora,
...

Step 3: Count Term Occurrences And Sort

59: godzilla
23: dinosaur
41: kaiju
31: japan
11: toho
5: monster
74: movie
2: 1964
6: ghidora
...



74: movie
59: godzilla
41: kaiju
31: japan
23: dinosaur

Step 4: Run Expanded Query

Google godzilla movie kaiju japan dinosaur

All Images Videos Shopping News More Settings Tools

About 407,000 results (0.47 seconds)

Images for godzilla movie kaiju japan dinosaur



→ More images for godzilla movie kaiju japan dinosaur Report images

[Kaiju - Wikipedia](#)
<https://en.wikipedia.org/wiki/Kaiju> ▾
Kaijū (怪獣, kaijū) (from Japanese "strange beast") is a Japanese film genre that features giant ...
Gojira (transliterated to Godzilla) is regarded as the first kaiju film and was released in 1954. 1975 – present); Dinosaur War Izenborg (Tsuburaya Productions; October 17, 1977 – June 30, 1978); Spider-Man (Toei Company; ...

[List of films featuring giant monsters - Wikipedia](#)
https://en.wikipedia.org/wiki/List_of_films_featuring_giant_monsters ▾
This is an alphabetical list of films featuring giant monsters, known in Japan as kaiju. One of the ...
Later in 1954, the Japanese film Godzilla was released. This was See also[edit]. List of films featuring dinosaurs · List of monster movies ...

[Godzilla: Monster Planet Final Trailer \(2018\) 2017 Godzilla Anime ...](#)
<https://www.youtube.com/watch?v=24BDBiUwyZ0>
Jan 8, 2018 - Uploaded by New Trailer Buzz
Godzilla Monster Planet Trailer 3 - 2017 Japanese Anime Movie Subscribe for more: ... Also, Godzilla looks ...

Summary

- Query suggestions can be useful for both users and search systems
- Query logs are a key source of information about searcher behaviour
- Query suggestions can be made based on
 - Query co-occurrence data in a *search session*
 - Agglomerative clustering of *click data*
 - Automatic relevance feedback based on *term statistics*

References

- Ryen White. *Interactions with Search Systems*. Cambridge University Press, 2016.
- D. Beeferman and A. Berger. *Agglomerative clustering of a search engine query log*. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 407--416, 2000.
- D. Kelly, K. Gyllstrom, and E. Bailey. *A comparison of query and term suggestion features for interactive searching*. ACM SIGIR Conference on Research and Development in Information Retrieval, pp 371—378, 2009.
- J.M. Yang, R. Cai, F. Jing, S. Wang, L, Zhang, and W, Ma. *Search-based query suggestion*. ACM Conference on Information and Knowledge Management, pp 1439—1440, 2008.
- P. Nayak. *Search with fewer keystrokes and better spelling*. The Official Google Blog, <http://googleblog.blogspot.com/2010/04/search-with-fewer-keystrokes-and-better.html>



Data Science

Thanks!