Information Retrieval Introduction and a Brief History

Debapriyo Majumdar Indian Statistical Institute debapriyo@isical.ac.in

Information Retrieval



User needs some information.



An information retrieval system tries to bridge this gap.



Assumption: the required information is present somewhere.

The goal of an information retrieval system is to satisfy user's information need.

Query

Ranked

results

Google results 1-10 of about 23,000 for april fool's. Search took 0.07 seconds

April Fool's Day

Quick response

Basic example

User expresses the information need in the form of a query.

The system returns a (ranked) list of results. Society > Honor

aprilfools.infospace.com/ - Cached - 10k - GoogleScout

AprilFools.com

...BUILD-A-DAD Foolishness · April Fool's History ·... Description: AprilFools.com is Gag Central for April Fools 1999, the cool Category: Society > Holidays > April Fool's Day

April Fool's Day 1999 - Practical Jokes, Pranks, Tips and

...Bangkok.com April Fool's Day Page April... ... Fool's Day is here! Do you have a favorite April... bangkok.com/fools/ - Cached - 42k - GoogleScout

Ranked results

Google in 2000

Courtesy: http://googlesystem.blogspot.com/

2007/12/google-in-2000.html

Many matches

TAHOO! SEARCH election results

Also try: 2004 election results, election results 2006, 2000 election results More

- Looking for Election Result Products? www.ebay.com - Find exactly what you want today.
- www.nytimes.com Continuous coverage of the 2006 elections and results from The New York
- Y News Results for election results Election results from the South - AP via Yahoo! News - 45 minutes ago Election results from the Midwest - AP via Yahoo! News - 1 hour, 37 minutes ago Election results from the Northeast - AP via Yahoo! News - 2 hours, 0 minute ago
 - Elections 2006 CNN Ongoing analysis and coverage of the 2006 midterm elections. Includes an issue spotlight and Category: 2006 Midterm Elections
- You Decide 2006 Fox News Features coverage and news analysis of the 2006 midterm elections. Category: 2006 Midterm Elections www.foxnews.com/politics/youdecide2006 - 68k - Cached - More from this site
- Midterm Elections 2006 Full Coverage on Yahoo! News Ongoing coverage of the key races for the 2006 midterm elections. Category: 2006 Midterm Elections
 news.yahoo.com/fc/US/Midterm_Elections_2006 - 50k - Cached - More from this site

Courtesy: https://www.searchenginewatch.com/

2006/11/08/in-the-election-results-race-yahoos-the-winner/

www.cnn.com/ELECTION/2006 - 57k - Cached - More from this site

Yahoo search in 2006



who is the prime minister of india

X

Tools

Question Answering

User can simply ask a question.

The system would (try to) answer crisply.

About 34,10,00,000 results (1.09 seconds)

News

India / Prime minister

Narendra Modi

Books

Since 2014

Answer

Maps

: More

Question



Settings

Narendra Damodardas Modi is an Indian politician serving as the 14th and current Prime Minister of India since 2014. He was the Chief Minister of Gujarat from 2001 to 2014 and is the Member of Parliament for Varanasi. Wikipedia

Education: Gujarat University (1983), University of Delhi (1978) Trending

-

Born: 17 September 1950 (age 70 years), Vadnagar

Full name: Narendra Damodardas Modi

Height: 1.7 m

Spouse: Jashodaben Modi (m. 1968)

More information

People also search for



Nirmala Sitharaman



Trending



Gandhi





Modi





Adityanath

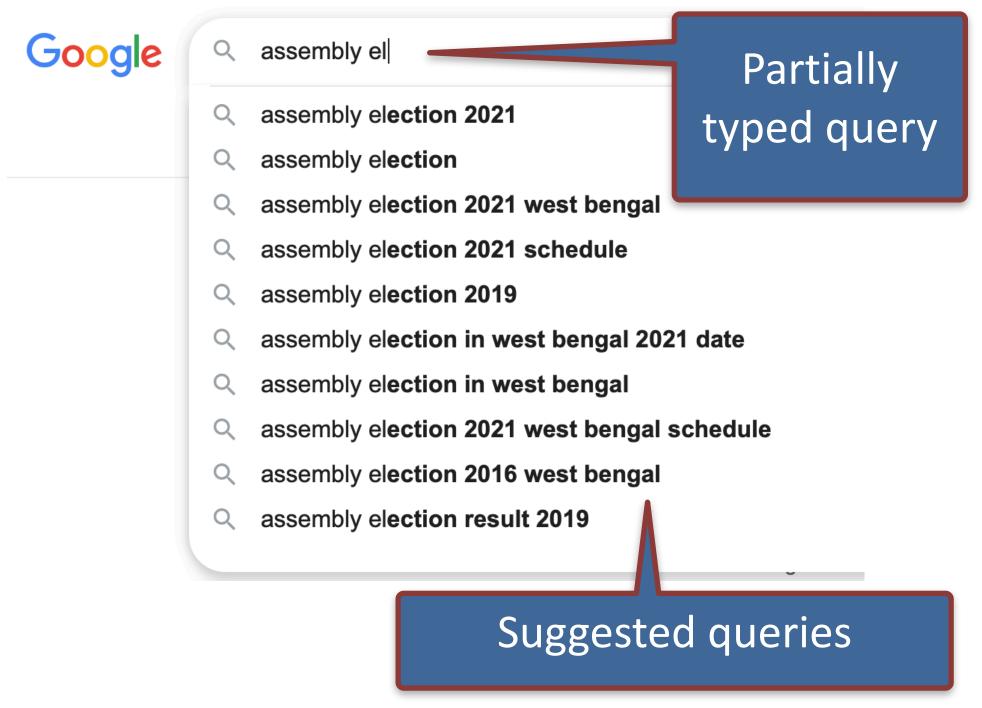
Ram Nath

View 10+ more

Kovind

Further recommendation

Query suggestions



User may not be sure what would be a good query.

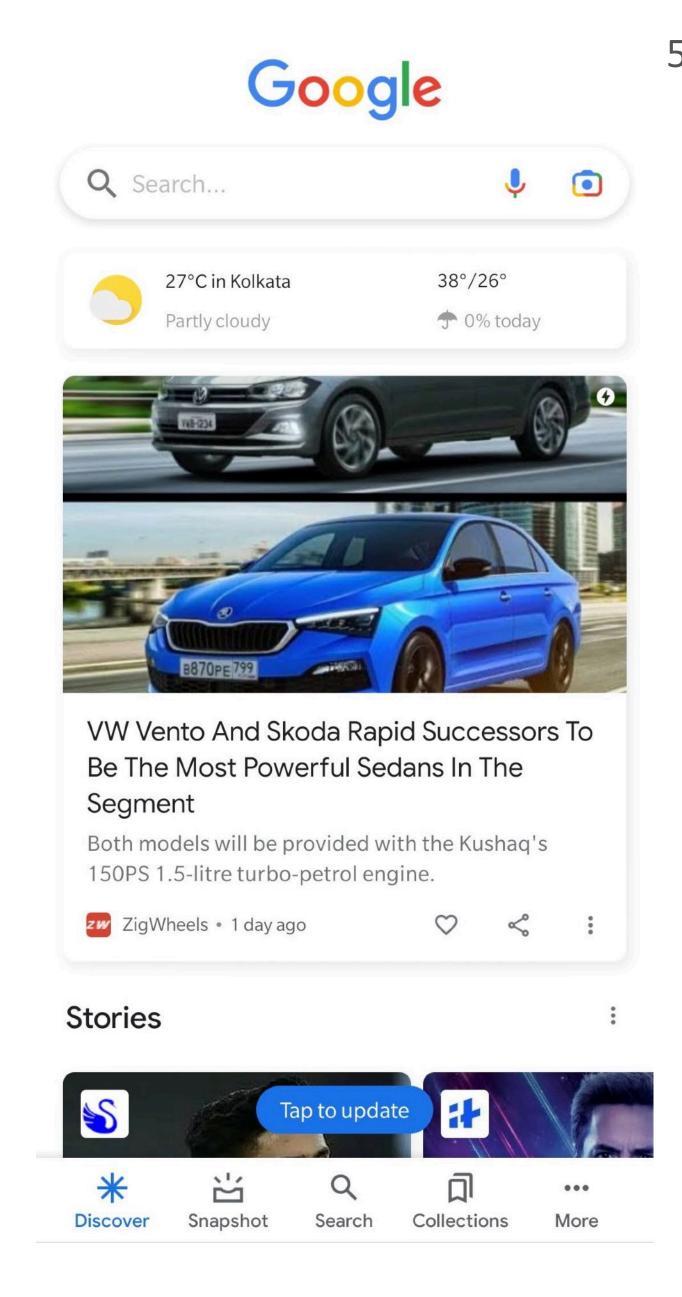
The system helps the user to formulate a query as the user types on.

Interactive IR.

Recommendation

User may not even have to ask, the system tries to "feed" information which may be "helpful".

Proactive IR.



Tentative Course Outline: Part 1

- Introduction to IR. Applications. Brief history.
- MapReduce and Hadoop.
- Apache Spark and tutorial.
- Terms and documents, term-document matrix, inverted index. Ranked retrieval, Vector space model, basic term weighting.
 Index structure, index creation.
- Query processing using inverted index, skip lists, champion lists, gd-ordering. Index compression.
- Tokenisation, stemming, lemmatization, stopword removal. Dictionary structure, tolerant retrieval.
- Apache lucene tutorial. Discussion on Solr and ElasticSearch.
- Evaluation: precision, recall, average precision, NDCG, other metrics, test collections, evaluation forums, sound experimental methods.
- Relevance feedback, pseudo relevance feedback, query expansion. Latent semantic indexing.
- Probabilistic models for IR. Okapi BM25. Language models for IR.
- Web search: crawling and indexing.
- Advertising on the web: Adwords.
- Deduplication: min-hashing, locality sensitive hashing.
- Link analysis: term-spam, Markov chain, PageRank, link spam, hubs and authorities.

- Supervised learning basics. Classification methods: Naive Bayes, SVM. (Assumed to be covered in ML1)
- Clustering basics: hierarchical clustering, point-assignment clustering, k-means. (ML1). Search result clustering.
- Logistic regression, Deep neural network, gradient descent.
- Overview of deep learning optimization techniques.
- Recurrent neural networks, LSTM, GRU.
- Convolutional neural networks and its applications on text data.
- Word embeddings: neural LM, word2vec, GloVe, FastText.
- Attention, Transformer.
- BERT, fine tuning BERT for transfer learning. Text classification and applications in IR.
- Learning to Rank.
- Query suggestion.
- Recommender systems.
- Personalization of search results.
- Text summarization.
- Question answering.
- Summary and overview of recent trends in IR.

Brief history

- 1800s and till 1930s: only librarians or paralegals had to retrieve information by searching (manually).
- 1950s: use of computer for information retrieval started.
 - IR started as a discipline: how to index documents, and how to retrieve them.
 - From Boolean retrieval to ranked retrieval: not all matches are equally good.
- 1978: ACM SIGIR conference started.
- Standard test collections became important to measure success of IR methods.
 - 1992: Text Retrieval Conference (TREC) started.
- 1990s: The world wide web started growing and web search engines came up: Altavista, Yahoo.
- 1998: Google originated from Stanford University with the invention of PageRank algorithm.
- 2000s: Search advertising made Google a tech giant.
 - Also, internet reached a huge portion of the world's population, search became a necessity.
 - Machine learning (later deep learning and reinforcement learning) became integral parts of IR.

References

- Sanderson, Mark, and W. Bruce Croft. "The history of information retrieval research."
 Proceedings of the IEEE 100, no. Special Centennial Issue (2012): 1444-1451.
- <u>Christopher D. Manning</u>, <u>Prabhakar Raghavan</u> and <u>Hinrich Schütze</u>. "<u>Introduction to Information</u>
 <u>Retrieval</u>", Cambridge University Press. 2008.