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.

Basic example

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

The system returns a (ranked) list of results.

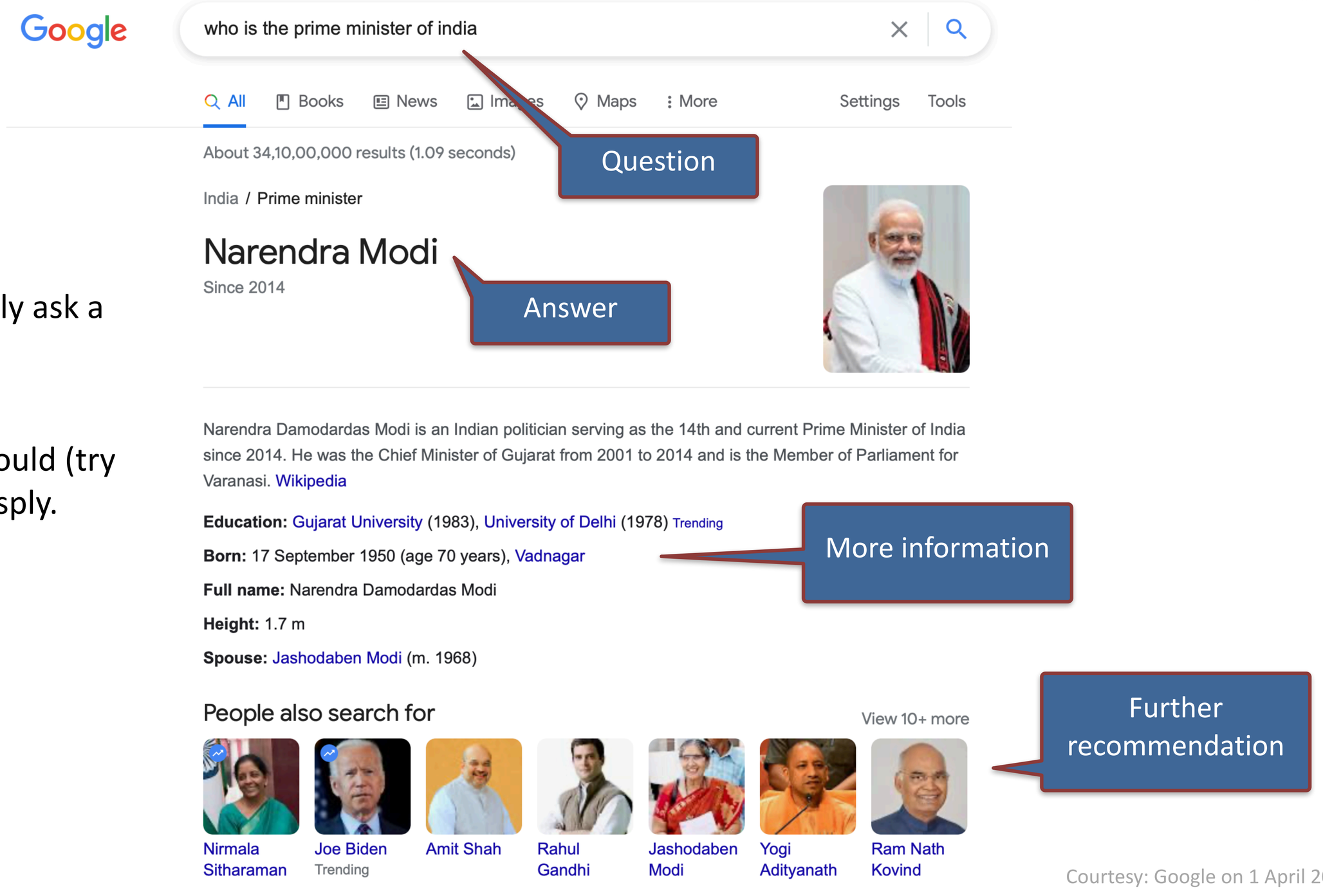
Google in 2000
Courtesy: http://googlesystem.blogspot.com/2007/12/google-in-2000.html

Yahoo search in 2006
Courtesy: https://www.searchenginewatch.com/2006/11/08/in-the-election-results-race-yahoos-the-winner/

Question Answering

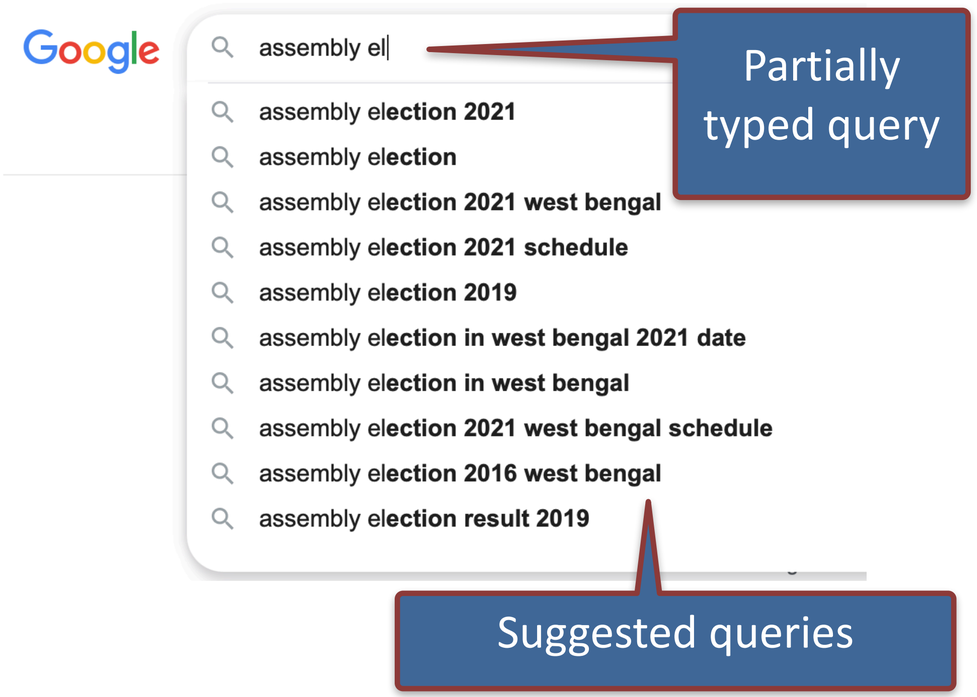
User can simply ask a question.

The system would (try to) answer crisply.



Courtesy: Google on 1 April 2021

Query suggestions



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

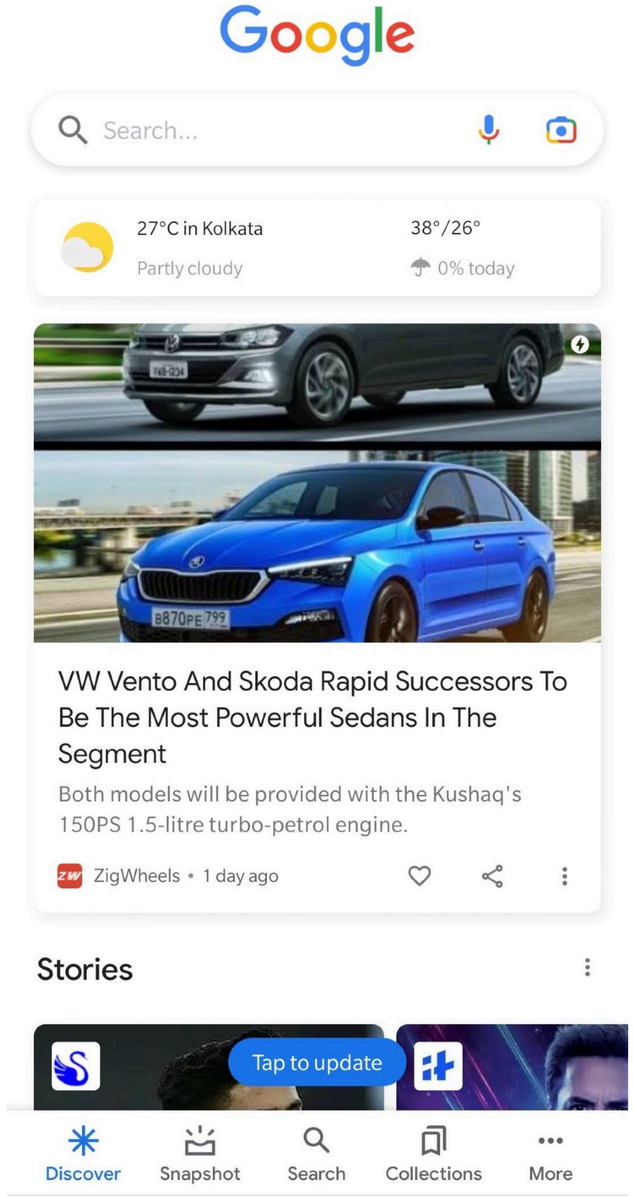
The system helps the user to formulate a query as the user types on.

Interactive IR.

Recommendation

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

Proactive IR.



Courtesy: Google on 1 April 2021

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.

Tentative Course Outline: Part 2

- 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.

Sanderson, Mark, and W. Bruce Croft. "The history of information retrieval research."
Proceedings of the IEEE 100, no. Special Centennial Issue (2012): 1444-1451.

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.
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. "Introduction to Information Retrieval", Cambridge University Press. 2008.