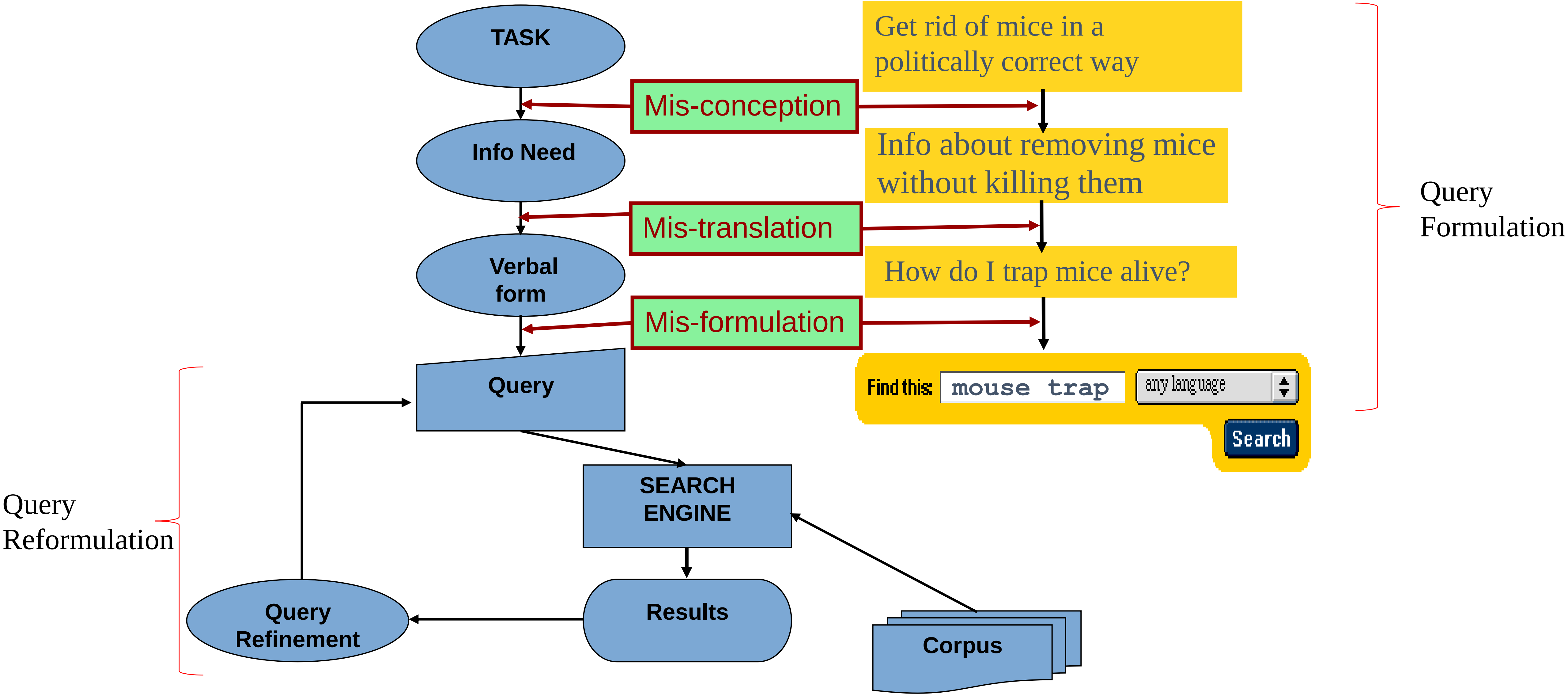
Lecture 2. Evolution of Retrieval Systems

Search Scenario



Find this: mouse trap

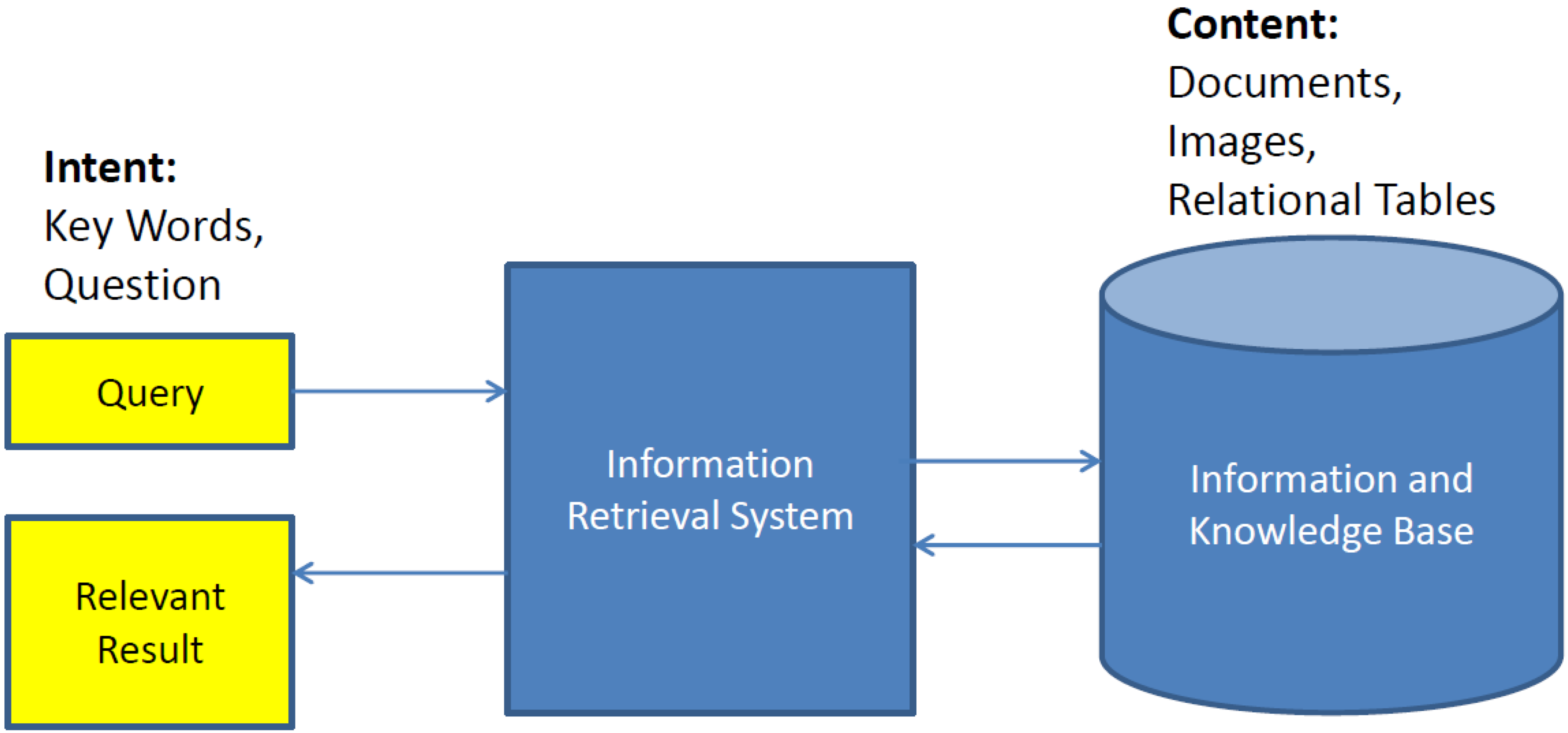
any language

Elements in the Search Scenario

- Users
  - Users of Different Backgrounds
  - Users of Different Expression Capabilities
  - Users in Different Contexts
  - Naïve Users vs. Expert Users
- Queries
  - Task vs. information need vs. query
  - Gaps between Information need and queries
  - Same information need expressed in different queries
  - Depend on users of different backgrounds and different expression capabilities, and in different contexts
  - Query Formulation vs. Query Reformulation
  - Manual vs. Automatic (term suggestion, concept recommendation, relevance feedback)

Elements in the Retrieval Scenario

- IR Systems
  - Evaluation Metrics
  - Efficiency and Effectiveness (from Users' points of views)
  - Different users have different feelings on Efficiency and Effectiveness
- Results
  - Determined by users submitting the queries or global users
  - Clicked vs. Non-Clicked
  - Interesting vs. Relevance
  - Non-Clicked vs. Non-Relevance
  - Relevance vs. Diversity



Intent:
Key Words,
Question

Query

Relevant Result

Information Retrieval System

Content:
Documents,
Images,
Relational Tables

Information and Knowledge Base

Key Questions: How to Represent Intent and Content, How to Match Intent and Content

(Li and Lu, SIGIR 2016 Tutorial)

Problems in IR

- Search
  - Matching between query and document
- Question Answering from Documents
  - Matching between question and answer
- Question Answering from Knowledge Base
  - Complicated matching between question and fact in knowledge base
- Generation-based Question Answering
  - Generating answer to question based on facts in knowledge base
- Question Answering over Personal Knowledge Base
- Unanswerable Question Resolution

Question Answering from Documents

問題

誰創立諾貝爾獎

文件檢索

諾貝爾獎（瑞典語：Nobelpriset，挪威語：Nobelprisen），是根據瑞典化學家阿佛烈·諾貝爾的遺囑於1901年開始頒發的獎項。諾貝爾獎分設物理、化學、生理學或醫學、文學、和平和經濟學六個獎項（經濟學獎於1968由瑞典中央銀行增設，全稱「瑞典銀行紀念諾貝爾經濟科學獎」，通稱「諾貝爾經濟學獎」）。諾貝爾獎普遍被認為是所頒獎的領域內最重要的獎項。

阿佛烈·伯恩哈德·諾貝爾（瑞典語：Alfred Bernhard Nobel，1833年10月21日－1896年12月10日）是瑞典化學家、工程師、發明家、軍工裝備製造商和矽藻土炸藥的發明者。他曾擁有Bofors軍工廠，主要生產軍火；還曾擁有一座鋼鐵廠。在他的遺囑中，他利用他的巨大財富創立了諾貝爾獎，各種諾貝爾獎項均以他的名字命名。

諾貝爾獎（瑞典語：Nobelpriset，挪威語：Nobelprisen）是一項由瑞典皇家科學院、瑞典學院、卡羅琳學院和挪威諾貝爾委員會頒發給對化學、物理、文學、和平和生理及醫學這五方面有著傑出貢獻的人士或組織的獎項[1]。諾貝爾獎是根據阿佛烈·諾貝爾在1895年的遺囑而設立的，並由諾貝爾基金會管理阿佛烈·諾貝爾的遺產及諾貝爾獎的頒發。

答案選擇

代名詞
指涉

在他的遺囑中，他利用他的巨大財富創立了諾貝爾獎，各種諾貝爾獎項均以他的名字命名。

諾貝爾獎是根據阿佛烈·諾貝爾在1895年的遺囑而設立的，並由諾貝爾基金會管理阿佛烈·諾貝爾的遺產及諾貝爾獎的頒發。

Question Answering from Knowledge Base

- QA over KB
  - Systems that are based on KB to find appropriate triples for answering question
- Simple questions
  - Questions can be answered by exactly one triple in the KB.
  - “Who is the author of Harry Potter?” → (Harry Potter, author, J.K. Rowling)
- Complex questions
  - Questions involve two or more triples in the KB.
  - Other semantic constraint:
  - “What is the name of the first Harry Potter novel?”

(Hsiao, Huang, and Chen, IJCNLP2017)

Generation-based Question Answering

- The representation of the query results from knowledge base is not intuitive for normal users

  (Jay Chou, people.person.profession, singer)
  (Jay Chou, music.artist.genre, pop)
  (Jay Chou, music.artist.album, Fantasy)

Slot	Value
people.person.profession	singer
music.artist.genre	pop
music.artist.album	Fantasy

范特西

v.s.

Jay Chou is a pop singer known for his album “Fantasy.”

(Yeh, Huang, and Chen, WI2018)

Goal

- The goal is to covert KB triples of a target entity into a natural language sentence.

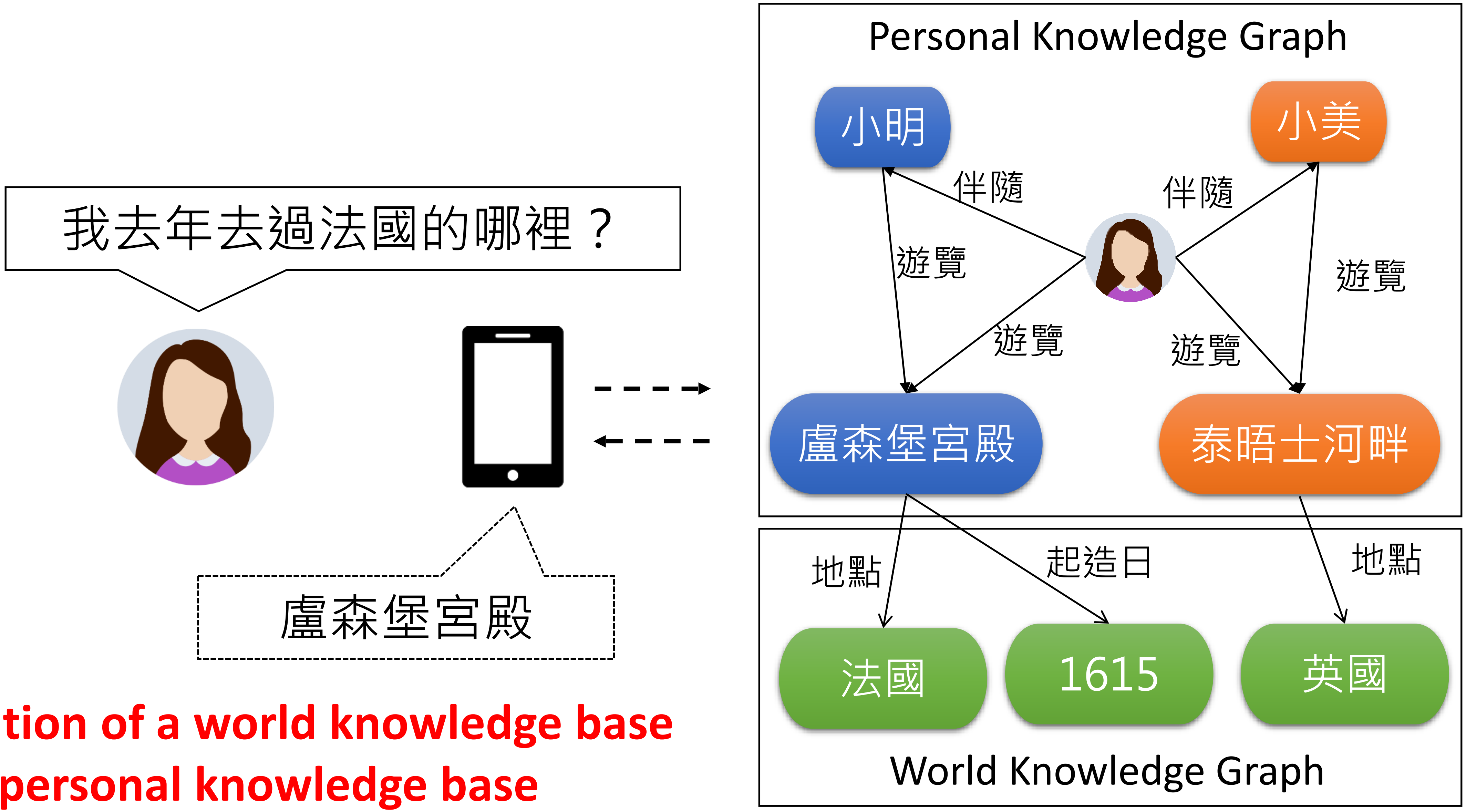
- Formally, given a target entity and a frame of $(slot, value)$ pairs, of which the slots are **unique**, generate a **one-sentence** description such that it conveys **exactly** the meaning of the frame.

Slot	Value
people.person.profession	singer
music.artist.genre	pop
music.artist.album	Fantasy

→

Jay Chou is a pop singer known for his album "Fantasy."

Question Answering over Personal Knowledge Base



Integration of a world knowledge base with a personal knowledge base

(Yen, Huang, Chen, SIGIR 2019)

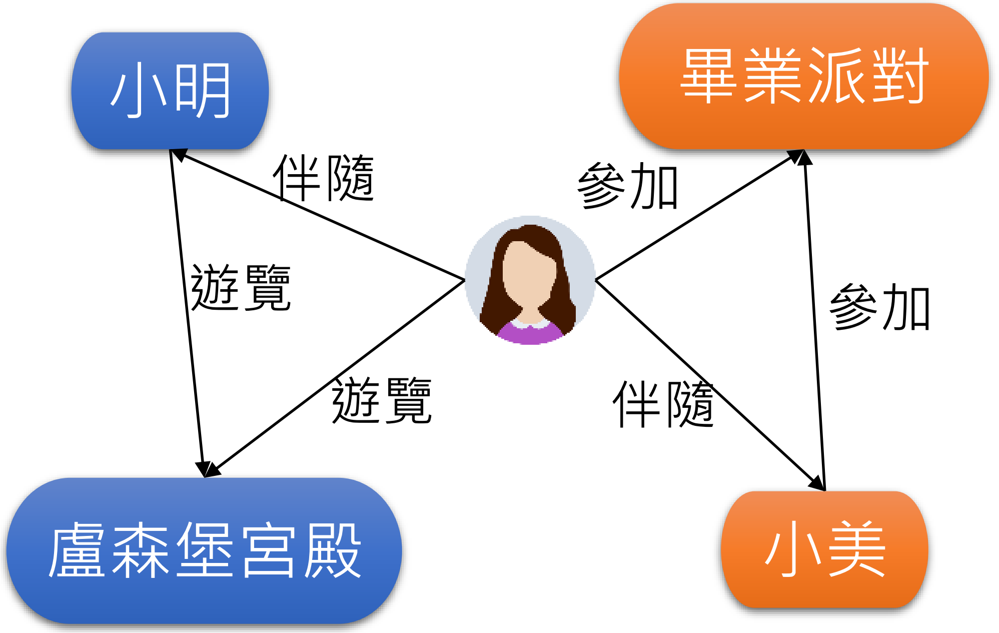
Challenging Issues for QA over PKB

1. KB is incomplete to cover all facts
2. The user-generated questions are ill-formed
3. Questions are ambiguous
4. Entities and/or relations in the question may not be mapped to the facts in the KB directly

How to answer unanswerable questions?

我去年和小美去什麼宮殿？

Personal Knowledge Graph



您的問題是指“我去年和小明去什麼宮殿？”
或是“我去年和小美參加什麼派對？”

Recall both Answers and Questions

(Yen, Huang, Chen, AAAI 2021)

IR Models

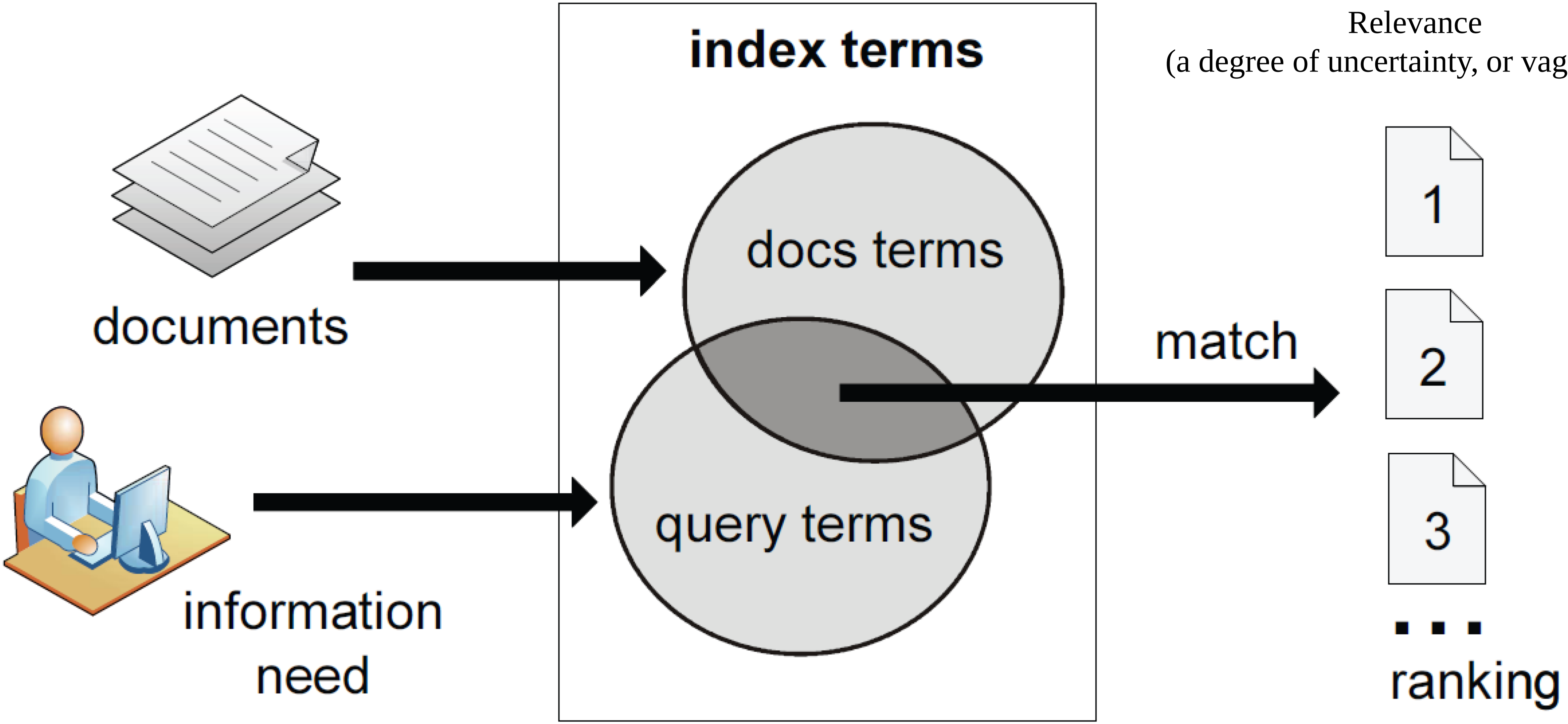
- **Modeling** in IR is a complex process aimed at producing a ranking function
  - **Ranking function**: a function that assigns scores to documents with regard to a given query
- This process consists of two main tasks:
  - The conception of a logical framework for representing documents and queries
  - The definition of a ranking function that allows quantifying the similarities among documents and queries

Modeling and Ranking

- IR systems usually adopt **index terms** to index and retrieve documents
- Index term:
  - In a restricted sense: it is a keyword that has some meaning on its own; usually plays the role of a noun
  - In a more general form: it is any word that appears in a document
- Retrieval based on index terms can be implemented efficiently
- Index terms are simple to refer to in a query
- Simplicity is important because it reduces the effort of query formulation

Information Retrieval Process

Relevance
(a degree of uncertainty, or vagueness)



index terms

docs terms

query terms

documents

information need

match

1

2

3

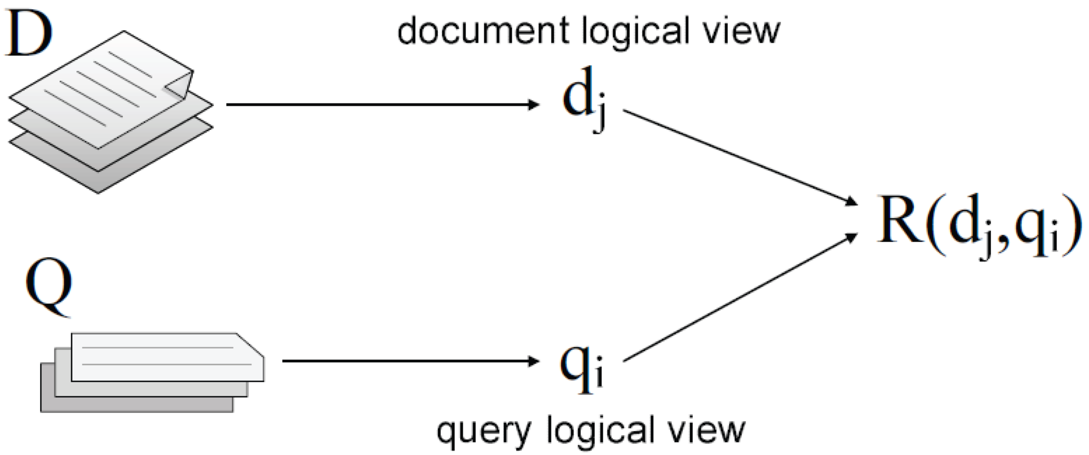
. . .

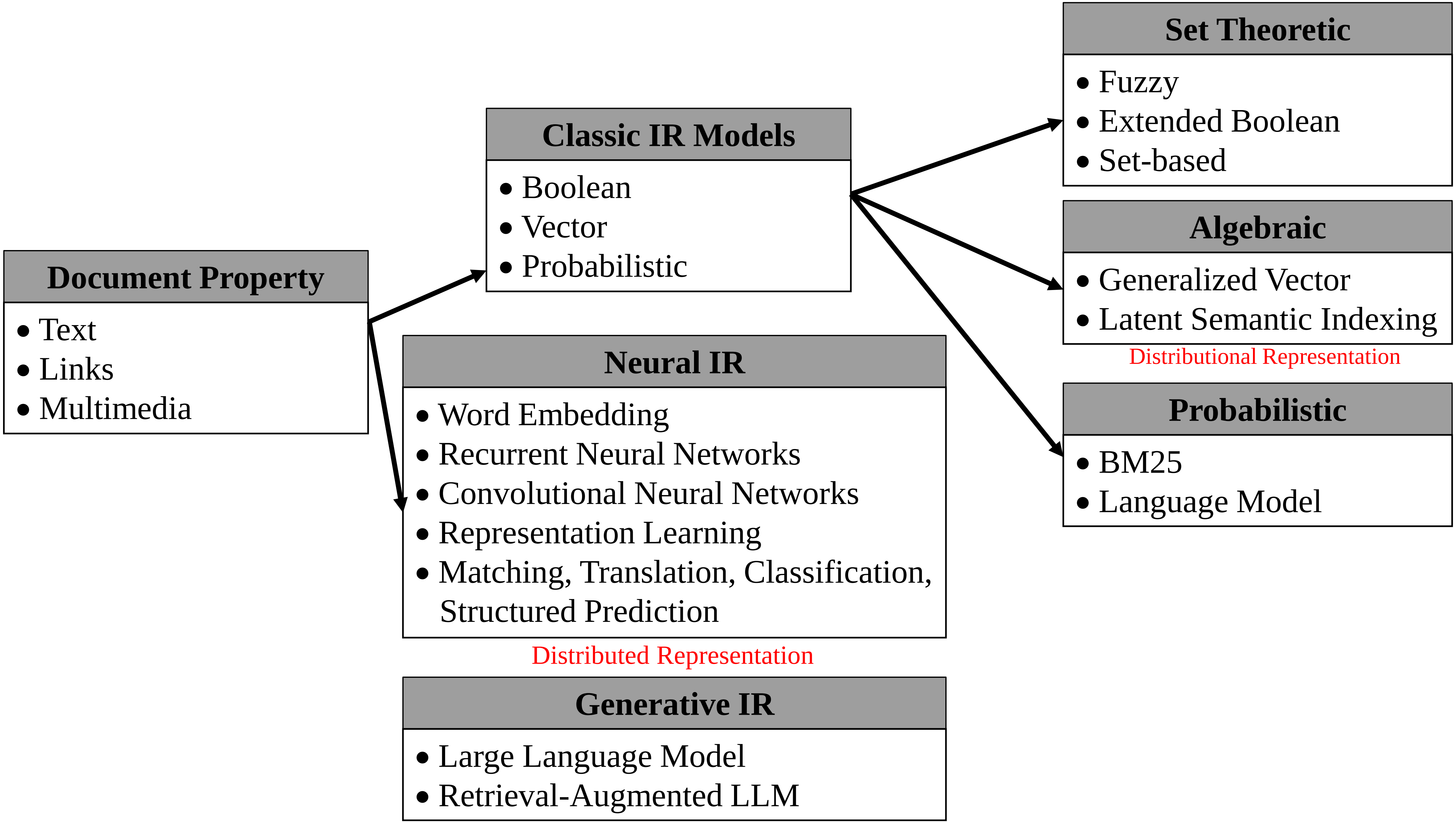
ranking

IR Models

An **IR model** is a quadruple $[\mathbf{D}, \mathbf{Q}, \mathcal{F}, R(q_i, d_j)]$ where

1. $\mathbf{D}$ is a set of logical views for the documents in the collection
2. $\mathbf{Q}$ is a set of logical views for the user queries
3. $\mathcal{F}$ is a framework for modeling documents and queries
4. $R(q_i, d_j)$ is a ranking function





Document Property

- Text
- Links
- Multimedia

Classic IR Models

- Boolean
- Vector
- Probabilistic

Set Theoretic

- Fuzzy
- Extended Boolean
- Set-based

Algebraic

- Generalized Vector
- Latent Semantic Indexing

Distributional Representation

Probabilistic

- BM25
- Language Model

Neural IR

- Word Embedding
- Recurrent Neural Networks
- Convolutional Neural Networks
- Representation Learning
- Matching, Translation, Classification, Structured Prediction

Distributed Representation

Generative IR

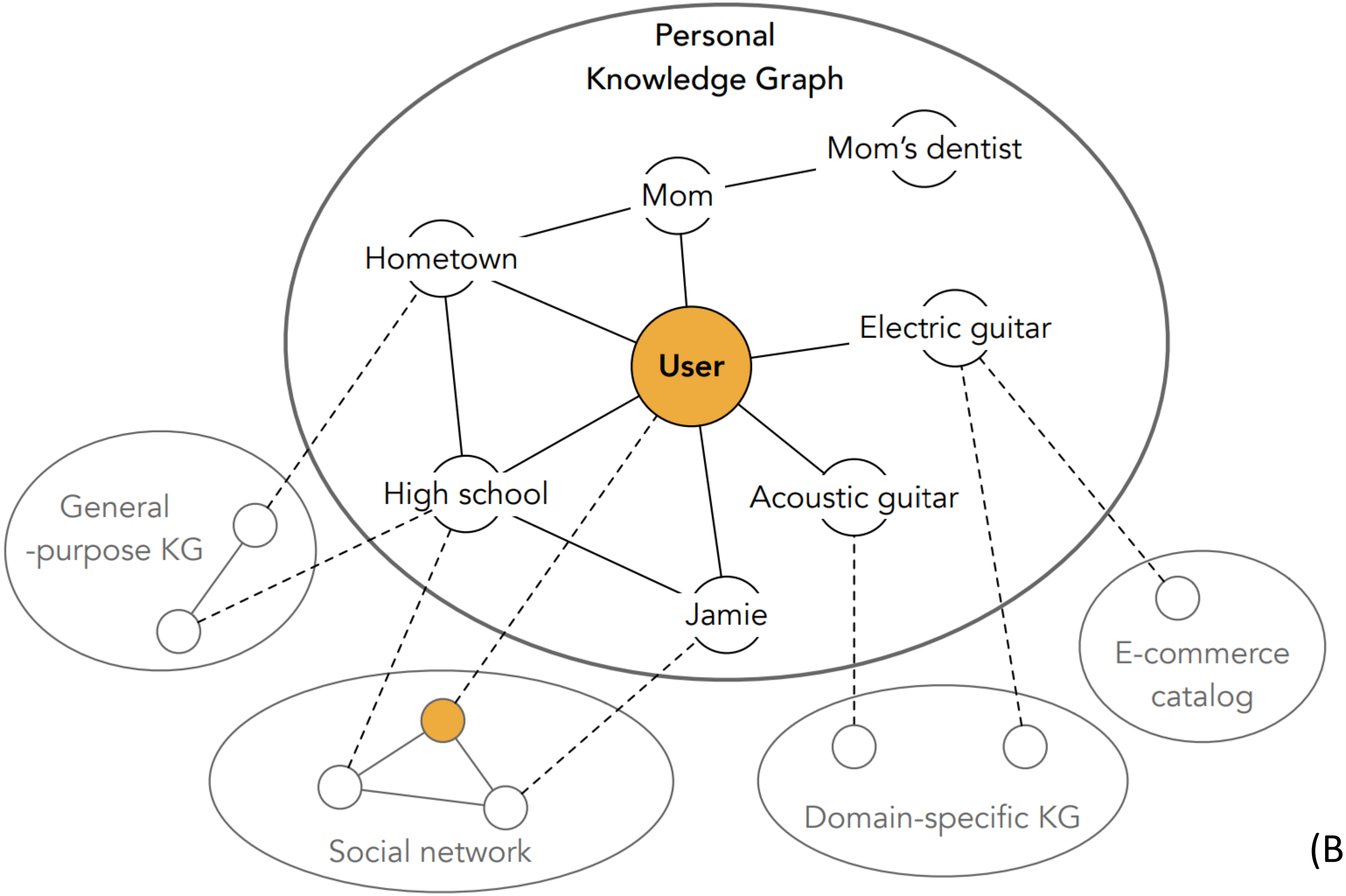
- Large Language Model
- Retrieval-Augmented LLM

Li and Lu’s View in IR

- “Easy” Problems in IR
    - Search
    - Question Answering from Documents
    - Deep Learning may not help so much
- “Hard” Problems in IR
    - Image Retrieval
    - Question Answering from Knowledge Base
    - Generation-based Question Answering
    - Deep Learning can make a big deal

(Li and Lu, SIGIR 2016)

Personal Knowledge Graph: A Research Agenda



(Balog & Kenter, ICTIR 2019)

Key Aspects of Personal Knowledge Graph

- Three key aspects of PKGs that separate them from general KGs
- (1) PKGs include entities of personal interest to the user
- (2) PKGs have a distinctive shape (“spiderweb” layout), where the user is always in the center
- (3) integration with external data sources is an inherent property of PKGs.

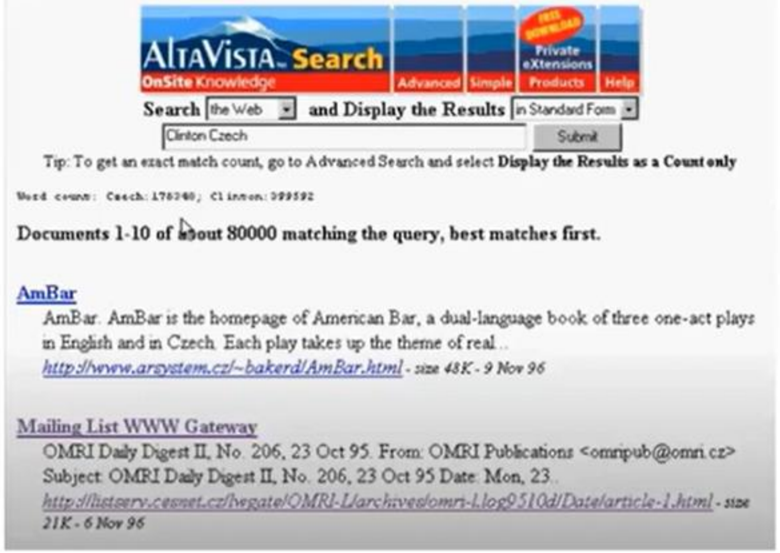
The following slides are selected from keynote speech by
Marc Najork at SIGIR 2023

Classical Retrieval Systems

Retrieval systems have followed the same paradigm for the past 70 years:

- Frame an information need as a query
- Submit the query to the retrieval system
- Get references to documents that may satisfy your information need
    - References may be rank-ordered by decreasing relevance
    - Results may contain relevant excerpts from each document ("snippets")

User experience of web search engines from 1996 to 2023



Tip: To get an exact match count, go to Advanced Search and select **Display the Results as a Count only**

Word count: Czech: 178340; Clinton: 399592

Documents 1-10 of about 80000 matching the query, best matches first.

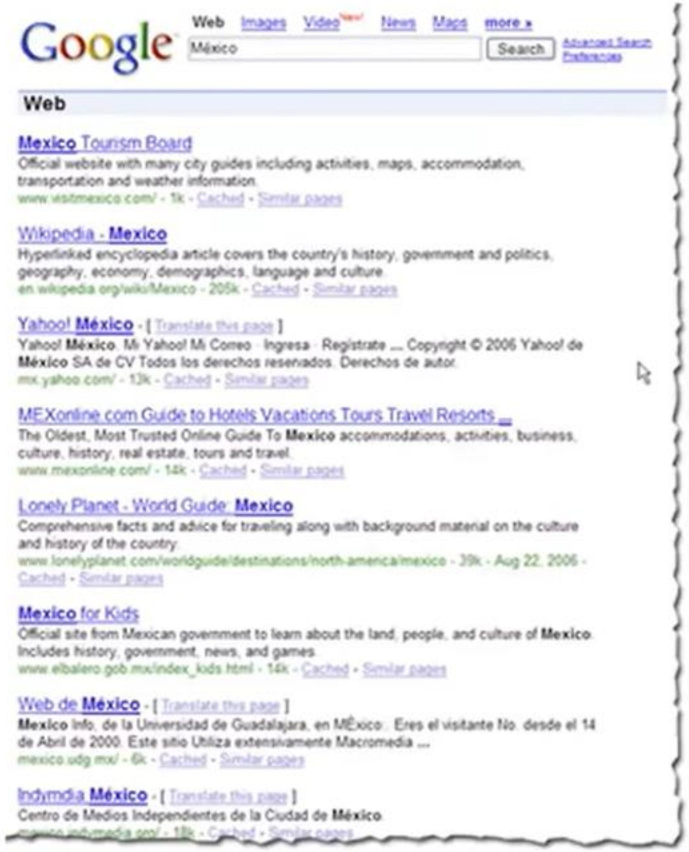
AmBar

AmBar. AmBar is the homepage of American Bar, a dual-language book of three one-act plays in English and in Czech. Each play takes up the theme of real...
http://www.arsystem.cz/~bakerd/AmBar.html - size 48K - 9 Nov 96

Mailing List WWW Gateway

OMRI Daily Digest II, No. 206, 23 Oct 95. From: OMRI Publications <omripub@omri.cz> Subject: OMRI Daily Digest II, No. 206, 23 Oct 95 Date: Mon, 23...
http://listserv.cesnet.cz/lwgate/OMRI-L/archives/omri-l.log9510d/Date/article-1.html - size 21K - 6 Nov 96

Altavista search results in 1996
https://www.youtube.com/watch?v=X5rjFLwXgew



Mexico Tourism Board
Official website with many city guides including activities, maps, accommodation, transportation and weather information.
www.visitmexico.com/ - 1k - Cached - Similar pages

Wikipedia - **Mexico**
Hyperlinked encyclopedia article covers the country's history, government and politics, geography, economy, demographics, language and culture.
en.wikipedia.org/wiki/Mexico - 205k - Cached - Similar pages

Yahoo! **México** - [Translate this page]
Yahoo! **México**. Mi Yahoo! Mi Correo · Ingresa · Regístrate ... Copyright © 2006 Yahoo! de **México** SA de CV Todos los derechos reservados. Derechos de autor.
mx.yahoo.com/ - 13k - Cached - Similar pages

MEXonline.com Guide to Hotels Vacations Tours Travel Resorts ...
The Oldest, Most Trusted Online Guide To **Mexico** accommodations, activities, business, culture, history, real estate, tours and travel.
www.mexonline.com/ - 14k - Cached - Similar pages

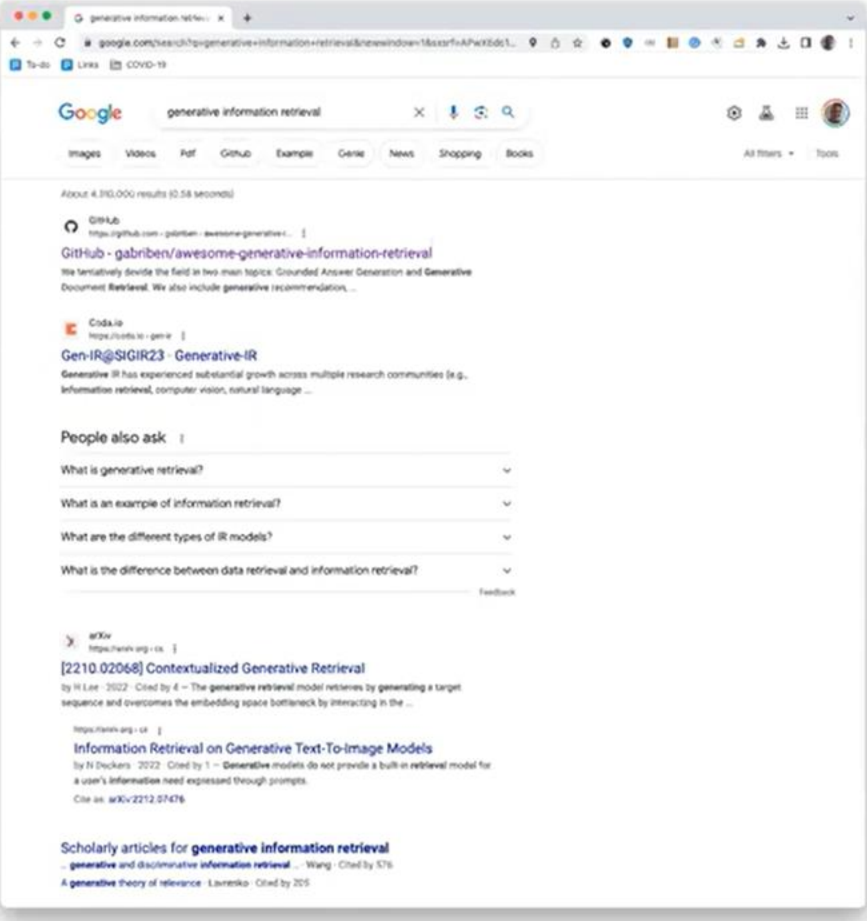
Lonely Planet - World Guide: **Mexico**
Comprehensive facts and advice for traveling along with background material on the culture and history of the country.
www.lonelyplanet.com/worldguide/destinations/north-america/mexico - 39k - Aug 22, 2006 - Cached - Similar pages

Mexico for Kids
Official site from Mexican government to learn about the land, people, and culture of **Mexico**. Includes history, government, news, and games.
www.elbalero.gob.mx/index_kids.html - 14k - Cached - Similar pages

Web de **México** - [Translate this page]
Mexico Info, de la Universidad de Guadalajara, en MÉxico . Eres el visitante No. desde el 14 de Abril de 2000. Este sitio Utiliza extensivamente Macromedia ...
mexico.udg.mx/ - 6k - Cached - Similar pages

Indymdia **México** - [Translate this page]
Centro de Medios Independientes de la Ciudad de **México**.
mexico.indymedia.org/ - 18k - Cached - Similar pages

Google search results in 2006
https://developers.google.com/search/blog/2006/08/how-search-results-may-differ-based-on


generative information retrieval
google.com/search?q=generative+information+retrieval&newwindow=1&sxsrf=APwXEdc1...
To-do
Links
COVID-19
Google
generative information retrieval
Images
Videos
Pdf
Github
Example
Genie
News
Shopping
Books
All filters
Tools

About 4,310,000 results (0.58 seconds)

GitHub
https://github.com › gabriben › awesome-generative-i...
GitHub - gabriben/awesome-generative-information-retrieval
We tentatively devide the field in two main topics: Grounded Answer Generation and **Generative** Document **Retrieval**. We also include **generative** recommendation, ...

Coda.io
https://coda.io › gen-ir
Gen-IR@SIGIR23 · Generative-IR
Generative IR has experienced substantial growth across multiple research communities (e.g., **information retrieval**, computer vision, natural language ...

People also ask

- What is generative retrieval?
- What is an example of information retrieval?
- What are the different types of IR models?
- What is the difference between data retrieval and information retrieval?

Feedback

arXiv
https://arxiv.org › cs
[2210.02068] Contextualized Generative Retrieval
by H Lee · 2022 · Cited by 4 — The **generative retrieval** model retrieves by **generating** a target sequence and overcomes the embedding space bottleneck by interacting in the ...

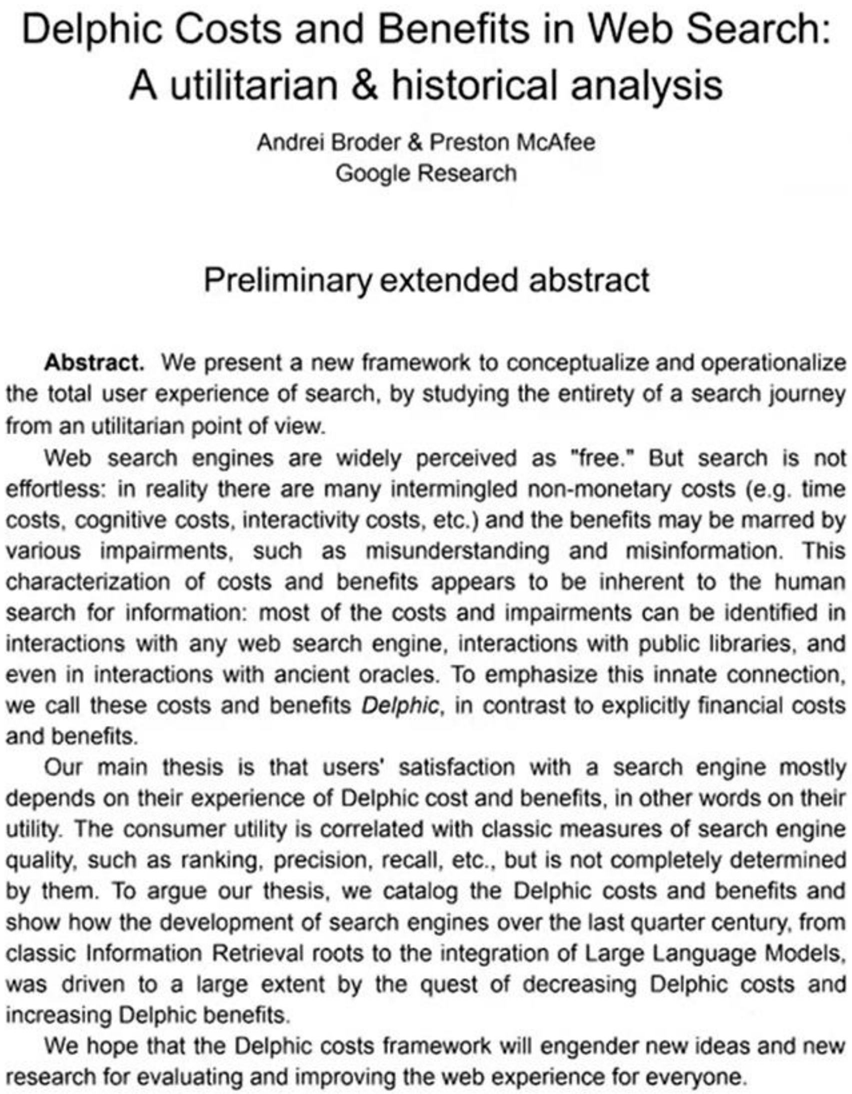
https://arxiv.org › cs
Information Retrieval on Generative Text-To-Image Models
by N Deckers · 2022 · Cited by 1 — **Generative** models do not provide a built-in **retrieval** model for a user's **information** need expressed through prompts.
Cite as: arXiv:2212.07476

Scholarly articles for **generative information retrieval**
... **generative** and discriminative **information retrieval** ... · Wang · Cited by 576
A **generative** theory of relevance · Lavrenko · Cited by 205

Google search results in 2023
Original screenshot

On the cognitive load of classic retrieval systems

Google DeepMind



Delphic Costs and Benefits in Web Search:
A utilitarian & historical analysis

Andrei Broder & Preston McAfee
Google Research

Preliminary extended abstract

Abstract. We present a new framework to conceptualize and operationalize the total user experience of search, by studying the entirety of a search journey from an utilitarian point of view.

Web search engines are widely perceived as "free." But search is not effortless: in reality there are many intermingled non-monetary costs (e.g. time costs, cognitive costs, interactivity costs, etc.) and the benefits may be marred by various impairments, such as misunderstanding and misinformation. This characterization of costs and benefits appears to be inherent to the human search for information: most of the costs and impairments can be identified in interactions with any web search engine, interactions with public libraries, and even in interactions with ancient oracles. To emphasize this innate connection, we call these costs and benefits *Delphic*, in contrast to explicitly financial costs and benefits.

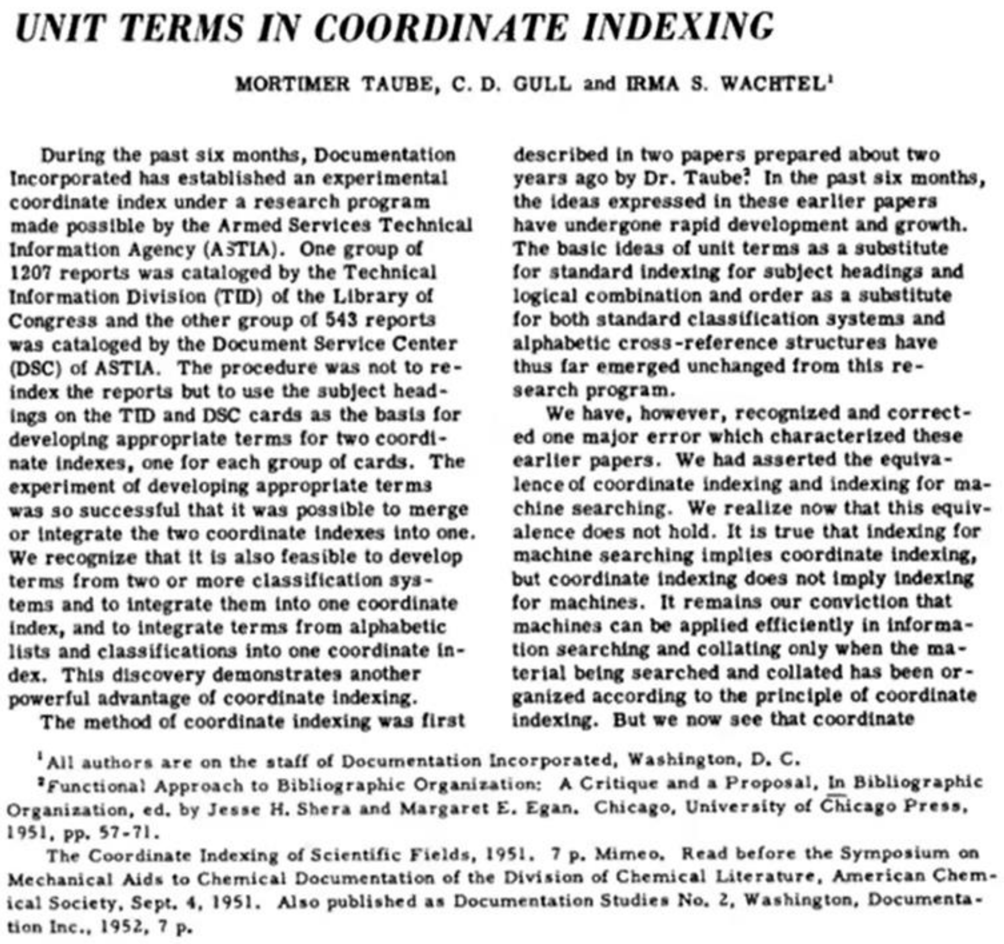
Our main thesis is that users' satisfaction with a search engine mostly depends on their experience of Delphic cost and benefits, in other words on their utility. The consumer utility is correlated with classic measures of search engine quality, such as ranking, precision, recall, etc., but is not completely determined by them. To argue our thesis, we catalog the Delphic costs and benefits and show how the development of search engines over the last quarter century, from classic Information Retrieval roots to the integration of Large Language Models, was driven to a large extent by the quest of decreasing Delphic costs and increasing Delphic benefits.

We hope that the Delphic costs framework will engender new ideas and new research for evaluating and improving the web experience for everyone.

- Information seeking activities have non-monetary costs: time investment, cognitive load, interactivity costs. Call them "Delphic Costs".
- User satisfaction with search engines is correlated with successful outcomes (did the user need get satisfied?) as well as Delphic costs (how difficult was it to find the information?)
- Much of IR research to date can be seen through lens of Delphic cost:
    - Synonym expansion lowers cognitive load of framing query
    - Ranking lowers time cost of scanning through results
    - Snippeting lowers time cost of highly-ranked irrelevant results (easier to skip)

arXiv paper forthcoming

Unit terms as document representations



UNIT TERMS IN COORDINATE INDEXING

MORTIMER TAUBE, C. D. GULL and IRMA S. WACHTEL[1]

During the past six months, Documentation Incorporated has established an experimental coordinate index under a research program made possible by the Armed Services Technical Information Agency (ASTIA). One group of 1207 reports was cataloged by the Technical Information Division (TID) of the Library of Congress and the other group of 543 reports was cataloged by the Document Service Center (DSC) of ASTIA. The procedure was not to re-index the reports but to use the subject headings on the TID and DSC cards as the basis for developing appropriate terms for two coordinate indexes, one for each group of cards. The experiment of developing appropriate terms was so successful that it was possible to merge or integrate the two coordinate indexes into one. We recognize that it is also feasible to develop terms from two or more classification systems and to integrate them into one coordinate index, and to integrate terms from alphabetic lists and classifications into one coordinate index. This discovery demonstrates another powerful advantage of coordinate indexing.

The method of coordinate indexing was first described in two papers prepared about two years ago by Dr. Taube[2] In the past six months, the ideas expressed in these earlier papers have undergone rapid development and growth. The basic ideas of unit terms as a substitute for standard indexing for subject headings and logical combination and order as a substitute for both standard classification systems and alphabetic cross-reference structures have thus far emerged unchanged from this research program.

We have, however, recognized and corrected one major error which characterized these earlier papers. We had asserted the equivalence of coordinate indexing and indexing for machine searching. We realize now that this equivalence does not hold. It is true that indexing for machine searching implies coordinate indexing, but coordinate indexing does not imply indexing for machines. It remains our conviction that machines can be applied efficiently in information searching and collating only when the material being searched and collated has been organized according to the principle of coordinate indexing. But we now see that coordinate

[1] All authors are on the staff of Documentation Incorporated, Washington, D. C.

[2] Functional Approach to Bibliographic Organization: A Critique and a Proposal, In Bibliographic Organization, ed. by Jesse H. Shera and Margaret E. Egan. Chicago, University of Chicago Press, 1951, pp. 57-71.

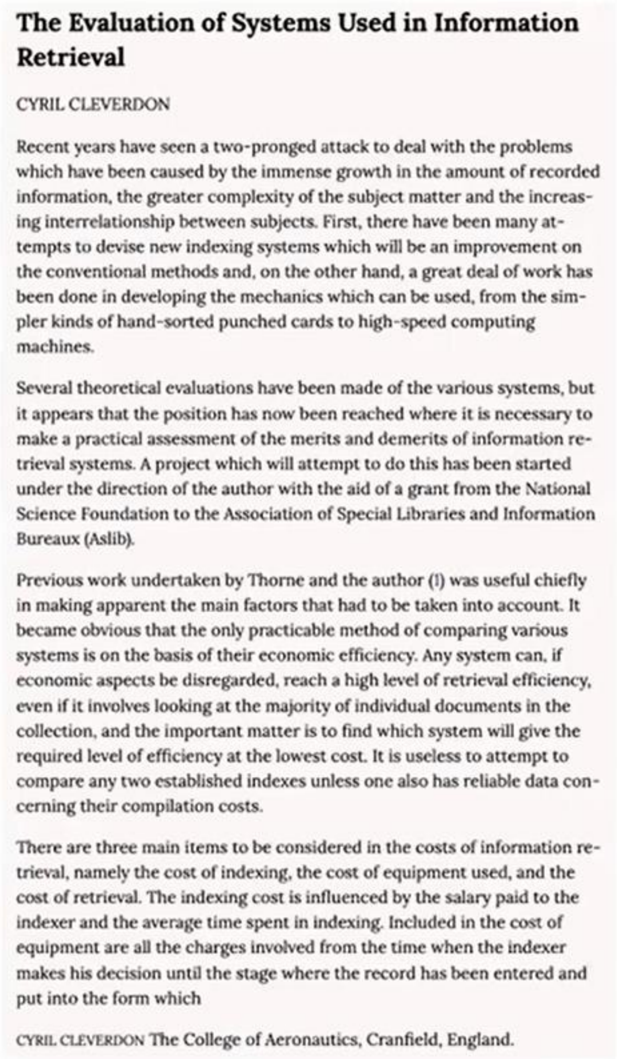
The Coordinate Indexing of Scientific Fields, 1951. 7 p. Mimeo. Read before the Symposium on Mechanical Aids to Chemical Documentation of the Division of Chemical Literature, American Chemical Society, Sept. 4, 1951. Also published as Documentation Studies No. 2, Washington, Documentation Inc., 1952, 7 p.

American Documentation 1952, https://doi.org/10.1002/asi.5090030404

Key findings:

- Documents relevant to an information need can be efficiently identified by the terms (words) they contain (in their subject heading, abstract, or body)
- The alternative approach is to assign a hierarchical subject descriptor (e.g. using the Dewey Decimal System) to each document and query using that descriptor

How did we get here?



The Evaluation of Systems Used in Information Retrieval

CYRIL CLEVERDON

Recent years have seen a two-pronged attack to deal with the problems which have been caused by the immense growth in the amount of recorded information, the greater complexity of the subject matter and the increasing interrelationship between subjects. First, there have been many attempts to devise new indexing systems which will be an improvement on the conventional methods and, on the other hand, a great deal of work has been done in developing the mechanics which can be used, from the simpler kinds of hand-sorted punched cards to high-speed computing machines.

Several theoretical evaluations have been made of the various systems, but it appears that the position has now been reached where it is necessary to make a practical assessment of the merits and demerits of information retrieval systems. A project which will attempt to do this has been started under the direction of the author with the aid of a grant from the National Science Foundation to the Association of Special Libraries and Information Bureaux (Aslib).

Previous work undertaken by Thorne and the author (1) was useful chiefly in making apparent the main factors that had to be taken into account. It became obvious that the only practicable method of comparing various systems is on the basis of their economic efficiency. Any system can, if economic aspects be disregarded, reach a high level of retrieval efficiency, even if it involves looking at the majority of individual documents in the collection, and the important matter is to find which system will give the required level of efficiency at the lowest cost. It is useless to attempt to compare any two established indexes unless one also has reliable data concerning their compilation costs.

There are three main items to be considered in the costs of information retrieval, namely the cost of indexing, the cost of equipment used, and the cost of retrieval. The indexing cost is influenced by the salary paid to the indexer and the average time spent in indexing. Included in the cost of equipment are all the charges involved from the time when the indexer makes his decision until the stage where the record has been entered and put into the form which

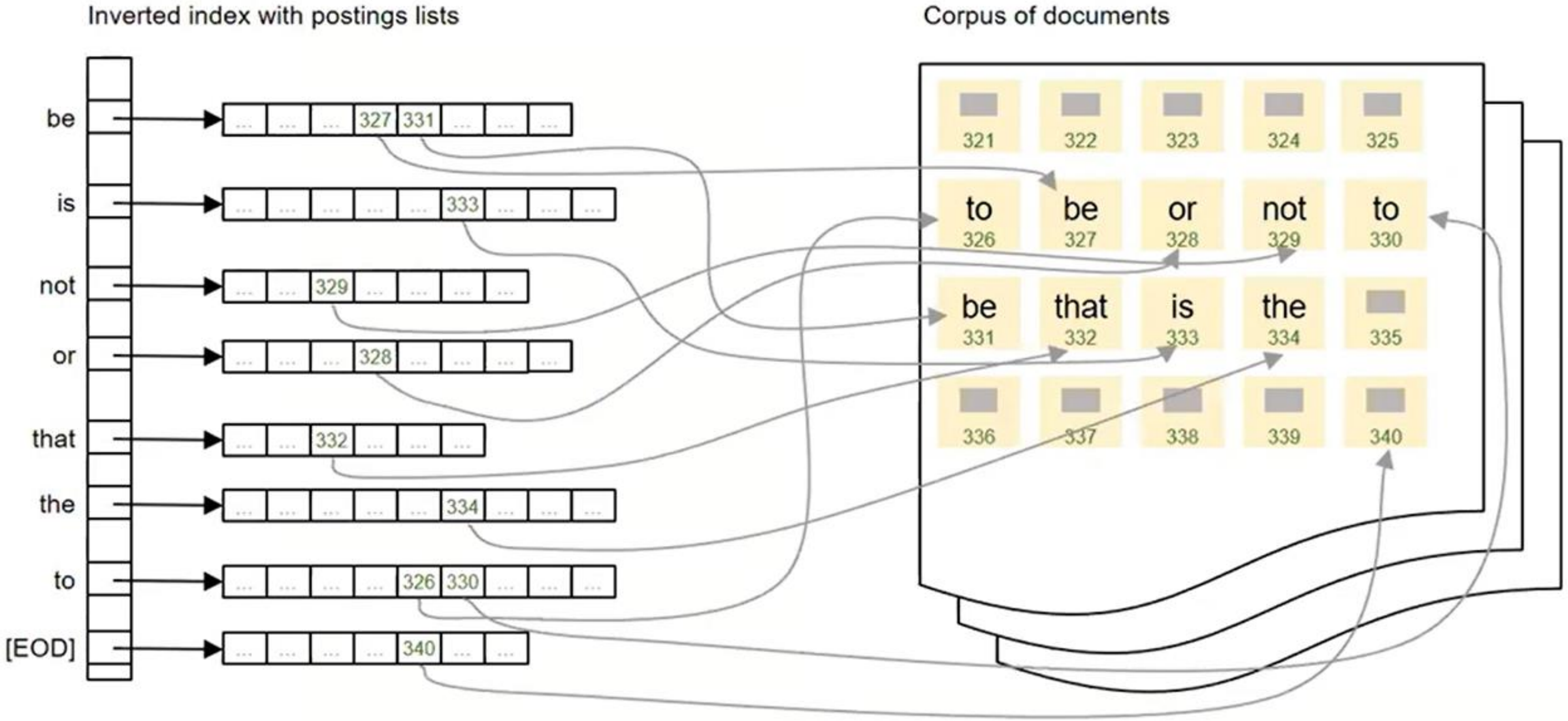
CYRIL CLEVERDON The College of Aeronautics, Cranfield, England.

Proc. of the International Conference on Scientific Information (1959), https://doi.org/10.17226/10866

Key finding:

- Uniterm-based systems perform at least as well as hierarchical document classification schemes
- This finding was initially highly controversial but was validated over the next few decades

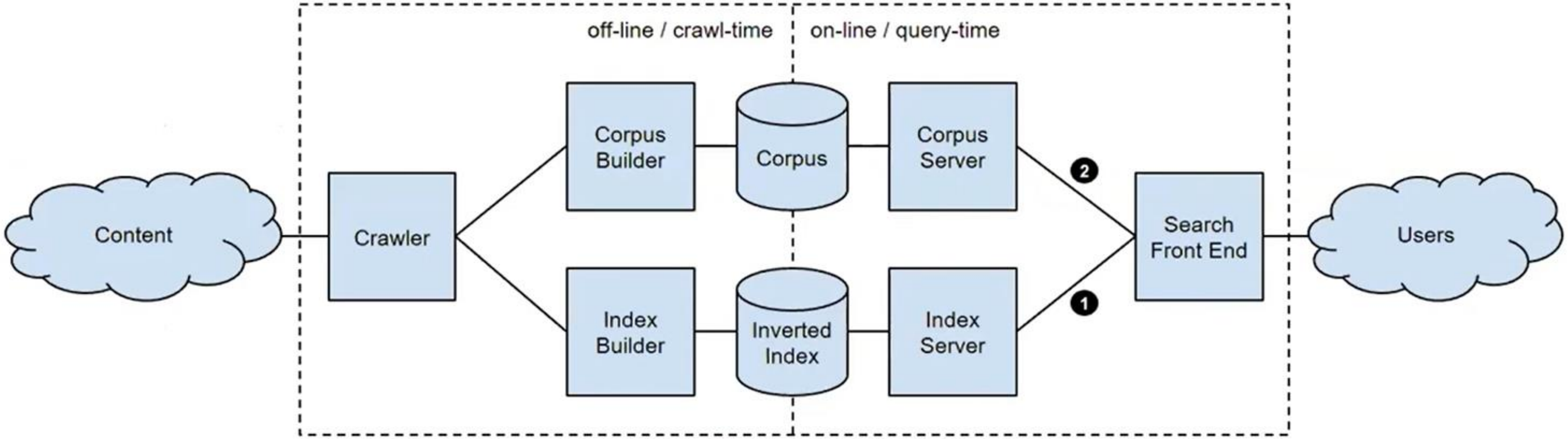
Structure of an inverted index



Using an inverted index shapes your view

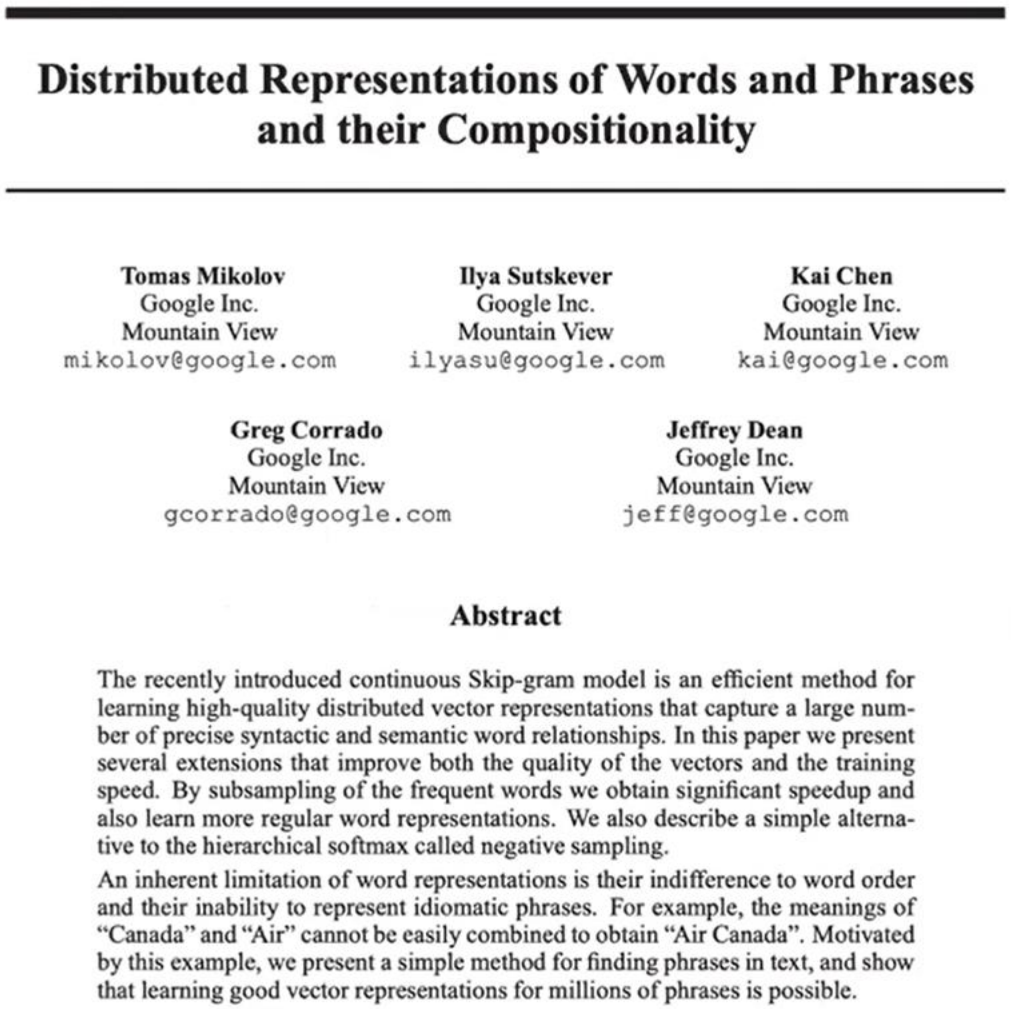
- Terms are uninterpreted tokens
- Synonymy is handled through query expansion
- AND, OR, NEAR can be done efficiently
- Fairly easy to scale (via sharding)
- Fairly easy to update (via tiering)
- Perfect match to Salton’s term vector model
- Term frequency easy to determine – tf.iDF, BM25 etc easy to implement
- Difficult to find semantically (as opposed to lexically) similar documents

Architecture of “classical” web retrieval system



“Classical” lexical retrieval system

The advent of word embeddings



Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

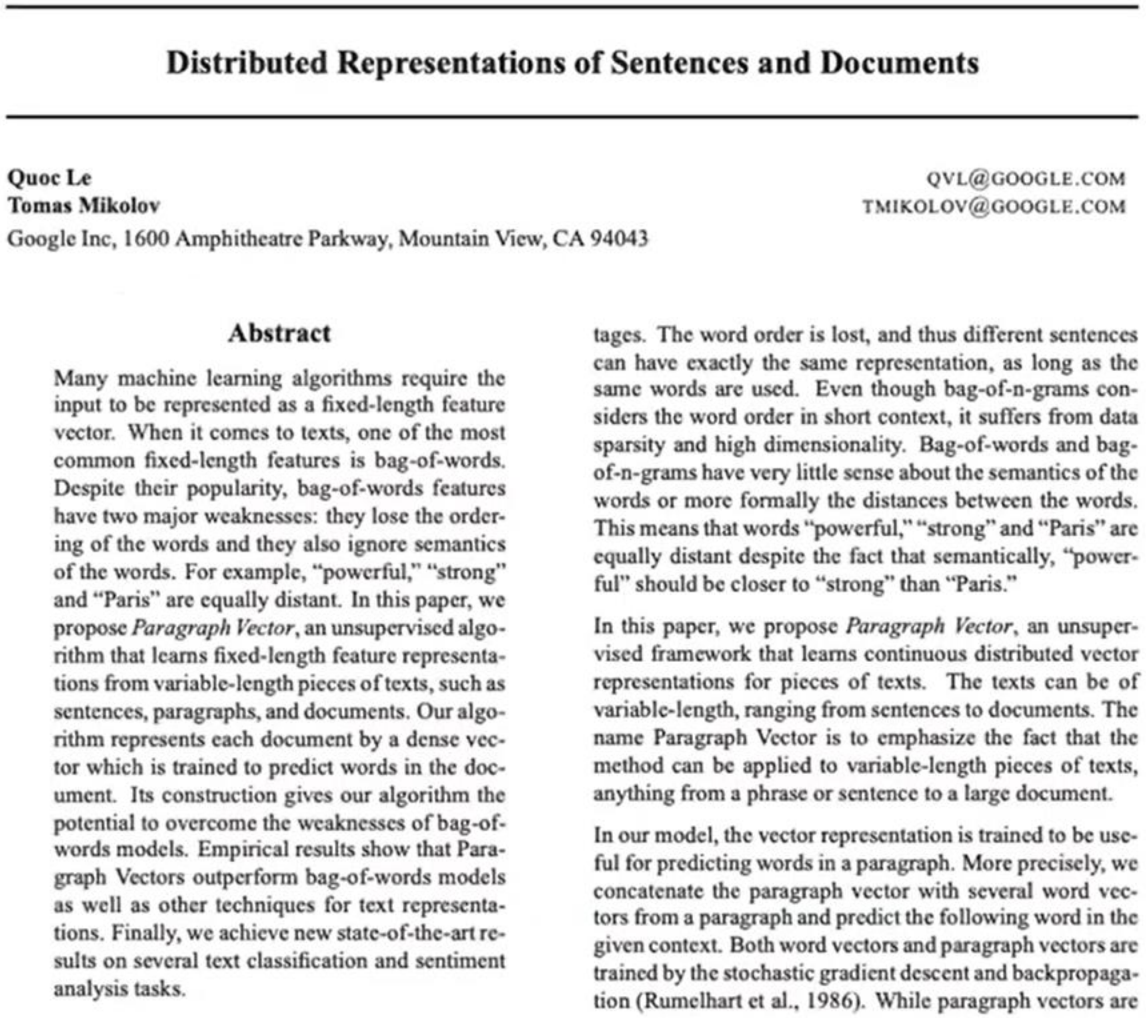
An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

Key concepts:

- Words represented by vectors of real numbers – an "embedding"
- Word embeddings are informed by neighboring words ("context")
- Semantically similar words are "close" in word embedding space
- Implementations: word2vec (Google), GloVe (Stanford), fastText (Meta), ...

NeurIPS 2013, https://papers.nips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

From word embeddings to document embeddings



Distributed Representations of Sentences and Documents

Quoc Le QVL@GOOGLE.COM
Tomas Mikolov TMIKOLOV@GOOGLE.COM
Google Inc, 1600 Amphitheatre Parkway, Mountain View, CA 94043

Abstract

Many machine learning algorithms require the input to be represented as a fixed-length feature vector. When it comes to texts, one of the most common fixed-length features is bag-of-words. Despite their popularity, bag-of-words features have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, "powerful," "strong" and "Paris" are equally distant. In this paper, we propose *Paragraph Vector*, an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. Our algorithm represents each document by a dense vector which is trained to predict words in the document. Its construction gives our algorithm the potential to overcome the weaknesses of bag-of-words models. Empirical results show that Paragraph Vectors outperform bag-of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classification and sentiment analysis tasks.

tages. The word order is lost, and thus different sentences can have exactly the same representation, as long as the same words are used. Even though bag-of-n-grams considers the word order in short context, it suffers from data sparsity and high dimensionality. Bag-of-words and bag-of-n-grams have very little sense about the semantics of the words or more formally the distances between the words. This means that words "powerful," "strong" and "Paris" are equally distant despite the fact that semantically, "powerful" should be closer to "strong" than "Paris."

In this paper, we propose *Paragraph Vector*, an unsupervised framework that learns continuous distributed vector representations for pieces of texts. The texts can be of variable-length, ranging from sentences to documents. The name Paragraph Vector is to emphasize the fact that the method can be applied to variable-length pieces of texts, anything from a phrase or sentence to a large document.

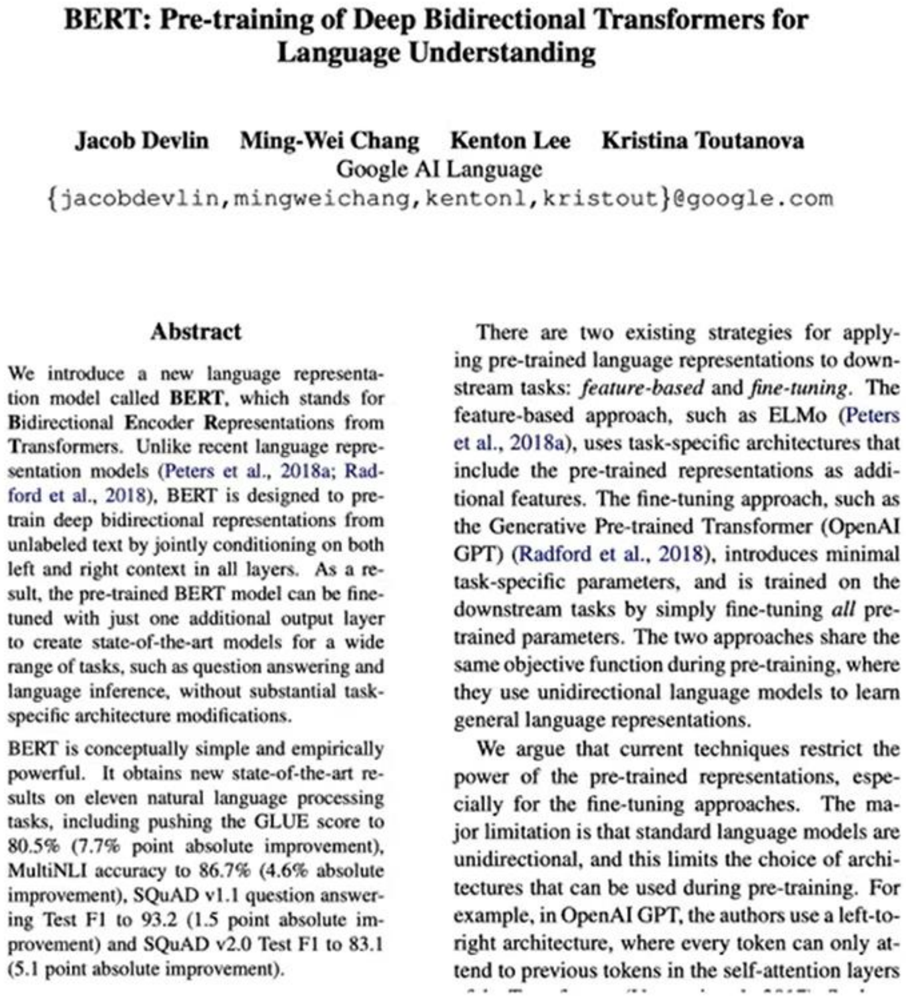
In our model, the vector representation is trained to be useful for predicting words in a paragraph. More precisely, we concatenate the paragraph vector with several word vectors from a paragraph and predict the following word in the given context. Both word vectors and paragraph vectors are trained by the stochastic gradient descent and backpropagation (Rumelhart et al., 1986). While paragraph vectors are

ICLR 2014, http://proceedings.mlr.press/v32/le14.pdf

Key ideas:

- Generalizes word embeddings to sentence / paragraph / document embeddings
- Variable-length passage represented by fixed-length vector
- Similar documents "close" in vector space (under e.g. cosine similarity)

Pre-trained language models



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

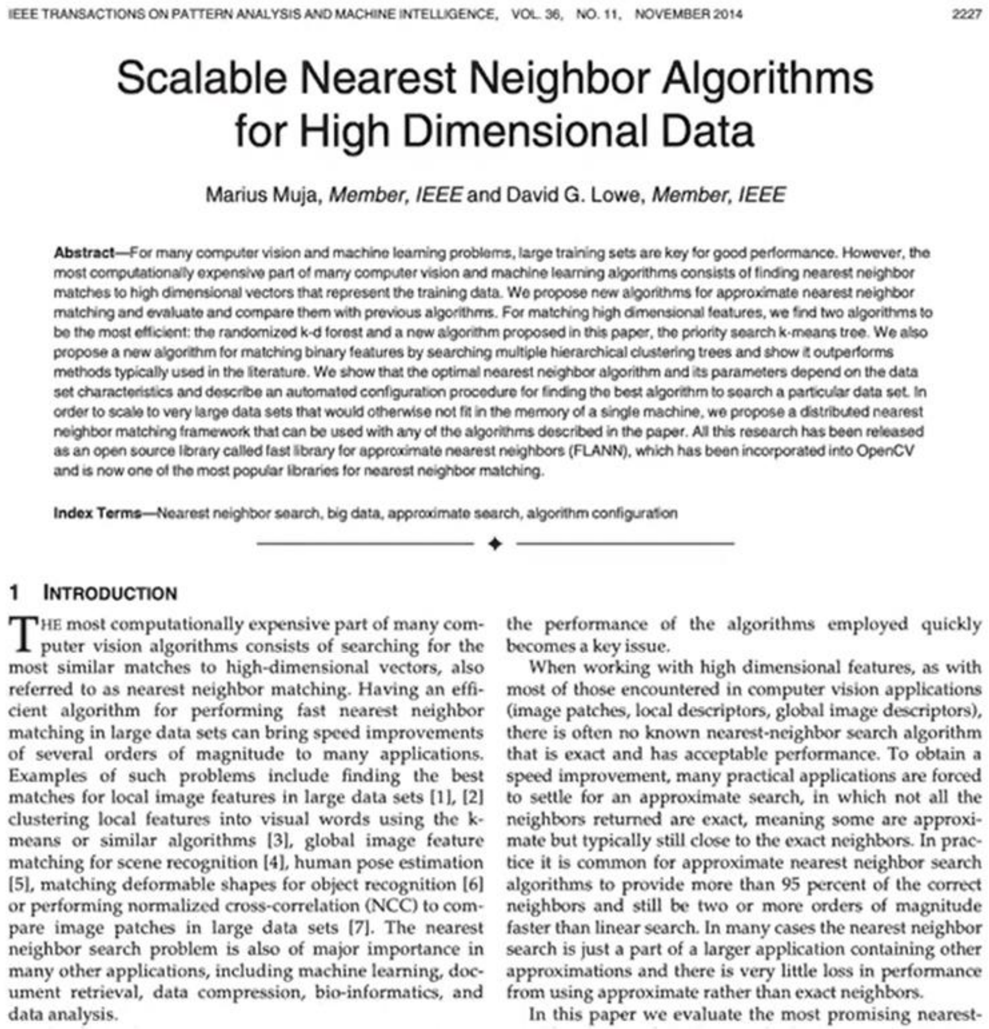
We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers

NAACL-HLT 2019, https://aclanthology.org/N19-1423

Key advances:

- Bi-directional sequence modeling.
- Contextual – homonyms have different representations
- More scalable than previous approaches due to using transformers as opposed to LSTMs used by ELMo
- Two training objectives: predicting masked-out-words (MLM or “cloze”), and predicting next word(s)
- Established new SotA for many tasks, including ranking

Term vector databases



IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 36, NO. 11, NOVEMBER 2014 2227

Scalable Nearest Neighbor Algorithms for High Dimensional Data

Marius Muja, *Member, IEEE* and David G. Lowe, *Member, IEEE*

Abstract—For many computer vision and machine learning problems, large training sets are key for good performance. However, the most computationally expensive part of many computer vision and machine learning algorithms consists of finding nearest neighbor matches to high dimensional vectors that represent the training data. We propose new algorithms for approximate nearest neighbor matching and evaluate and compare them with previous algorithms. For matching high dimensional features, we find two algorithms to be the most efficient: the randomized k-d forest and a new algorithm proposed in this paper, the priority search k-means tree. We also propose a new algorithm for matching binary features by searching multiple hierarchical clustering trees and show it outperforms methods typically used in the literature. We show that the optimal nearest neighbor algorithm and its parameters depend on the data set characteristics and describe an automated configuration procedure for finding the best algorithm to search a particular data set. In order to scale to very large data sets that would otherwise not fit in the memory of a single machine, we propose a distributed nearest neighbor matching framework that can be used with any of the algorithms described in the paper. All this research has been released as an open source library called fast library for approximate nearest neighbors (FLANN), which has been incorporated into OpenCV and is now one of the most popular libraries for nearest neighbor matching.

Index Terms—Nearest neighbor search, big data, approximate search, algorithm configuration

1 Introduction

THE most computationally expensive part of many computer vision algorithms consists of searching for the most similar matches to high-dimensional vectors, also referred to as nearest neighbor matching. Having an efficient algorithm for performing fast nearest neighbor matching in large data sets can bring speed improvements of several orders of magnitude to many applications. Examples of such problems include finding the best matches for local image features in large data sets [1], [2] clustering local features into visual words using the k-means or similar algorithms [3], global image feature matching for scene recognition [4], human pose estimation [5], matching deformable shapes for object recognition [6] or performing normalized cross-correlation (NCC) to compare image patches in large data sets [7]. The nearest neighbor search problem is also of major importance in many other applications, including machine learning, document retrieval, data compression, bio-informatics, and data analysis.

the performance of the algorithms employed quickly becomes a key issue.

When working with high dimensional features, as with most of those encountered in computer vision applications (image patches, local descriptors, global image descriptors), there is often no known nearest-neighbor search algorithm that is exact and has acceptable performance. To obtain a speed improvement, many practical applications are forced to settle for an approximate search, in which not all the neighbors returned are exact, meaning some are approximate but typically still close to the exact neighbors. In practice it is common for approximate nearest neighbor search algorithms to provide more than 95 percent of the correct neighbors and still be two or more orders of magnitude faster than linear search. In many cases the nearest neighbor search is just a part of a larger application containing other approximations and there is very little loss in performance from using approximate rather than exact neighbors.

In this paper we evaluate the most promising nearest-

IEEE TPAMI 2014, https://doi.org/10.1109/TPAMI.2014.2321376

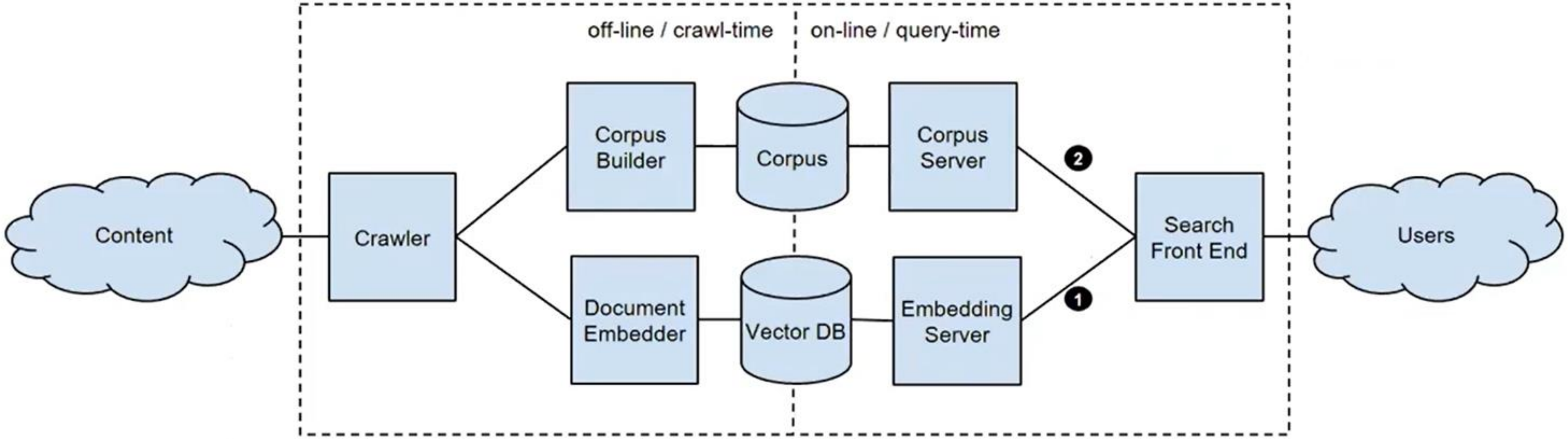
Main functionality: Given a vector, return k closest vectors in the database

Techniques: Hashing, graph-based, ...

Metrics: Euclidean distance, cosine similarity, ...

Many available implementations: FLANN (UBC), ScaNN (Google), FAISS (Facebook), Pinecone, Weaviate, ...

Architecture of semantic retrieval system



“Deep” semantic retrieval system

Contrasting lexical and semantic representations

Lexical representation (term vector):

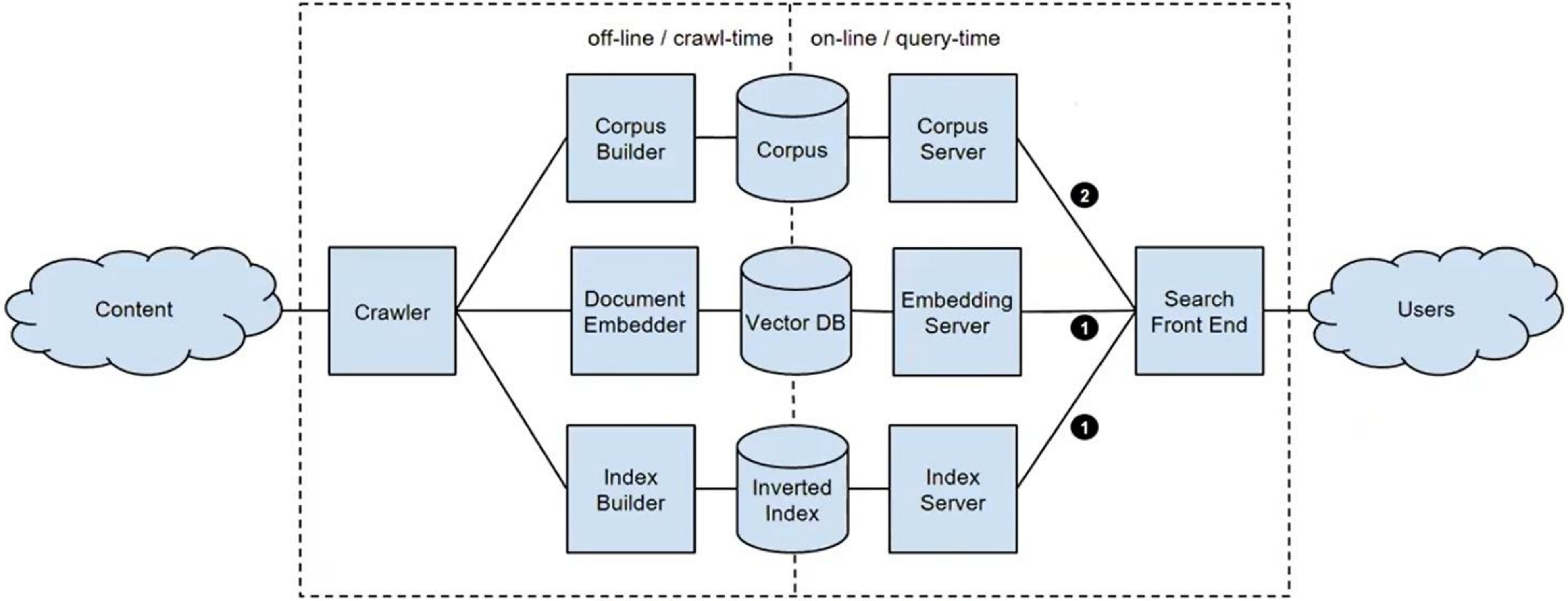
- High-dimensional (billions)
- Sparse
- Term vector model
- Words are uninterpreted tokens
- Computationally cheaper
- Easy to scale inverted index
- No explicit language model (caveat: pre & post retrieval)

Semantic representation (embedding):

- Low dimensional (hundreds)
- Dense
- Distributional semantics
- Words have meaning
- Computationally more expensive
- Harder to scale nn-index
- Updating language model very expensive

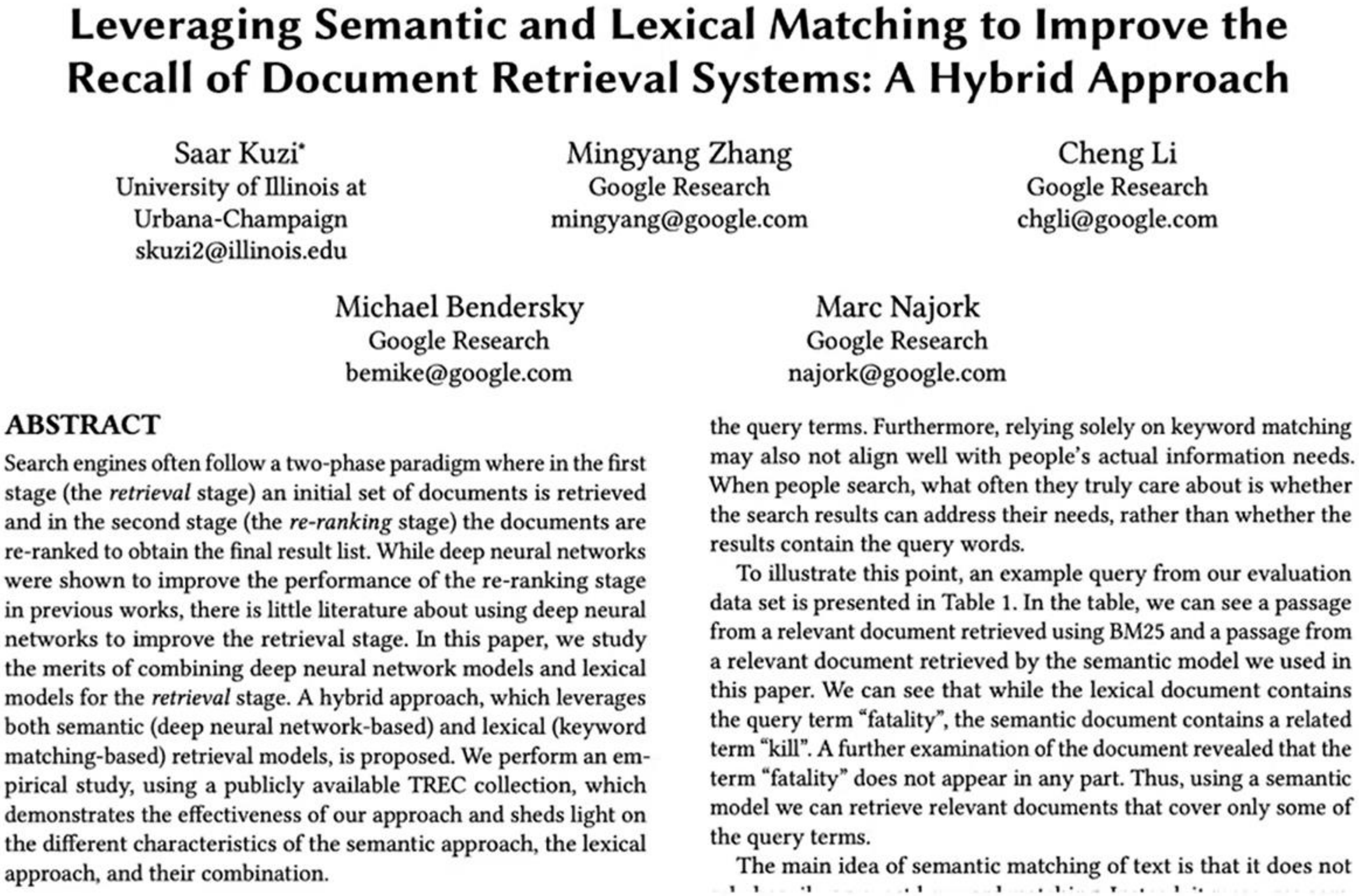
Architecture of hybrid lexical/semantic retrieval system

Google DeepMind



Hybrid lexical/semantic retrieval system

Semantic and lexical matching complement each other



Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach

Saar Kuzi*
University of Illinois at Urbana-Champaign
skuzi2@illinois.edu

Mingyang Zhang
Google Research
mingyang@google.com

Cheng Li
Google Research
chgli@google.com

Michael Bendersky
Google Research
bemike@google.com

Marc Najork
Google Research
najork@google.com

ABSTRACT

Search engines often follow a two-phase paradigm where in the first stage (the *retrieval* stage) an initial set of documents is retrieved and in the second stage (the *re-ranking* stage) the documents are re-ranked to obtain the final result list. While deep neural networks were shown to improve the performance of the re-ranking stage in previous works, there is little literature about using deep neural networks to improve the retrieval stage. In this paper, we study the merits of combining deep neural network models and lexical models for the *retrieval* stage. A hybrid approach, which leverages both semantic (deep neural network-based) and lexical (keyword matching-based) retrieval models, is proposed. We perform an empirical study, using a publicly available TREC collection, which demonstrates the effectiveness of our approach and sheds light on the different characteristics of the semantic approach, the lexical approach, and their combination.

the query terms. Furthermore, relying solely on keyword matching may also not align well with people's actual information needs. When people search, what often they truly care about is whether the search results can address their needs, rather than whether the results contain the query words.

To illustrate this point, an example query from our evaluation data set is presented in Table 1. In the table, we can see a passage from a relevant document retrieved using BM25 and a passage from a relevant document retrieved by the semantic model we used in this paper. We can see that while the lexical document contains the query term "fatality", the semantic document contains a related term "kill". A further examination of the document revealed that the term "fatality" does not appear in any part. Thus, using a semantic model we can retrieve relevant documents that cover only some of the query terms.

The main idea of semantic matching of text is that it does not

Lexical: BM25, Anserini (Lucene)

Semantic: BERT, ScaNN

Fusion: RM3

Data sets: TREC newswire

Hybrid approach has superior recall & higher precision

arXiv 2020, https://arxiv.org/abs/2010.01195