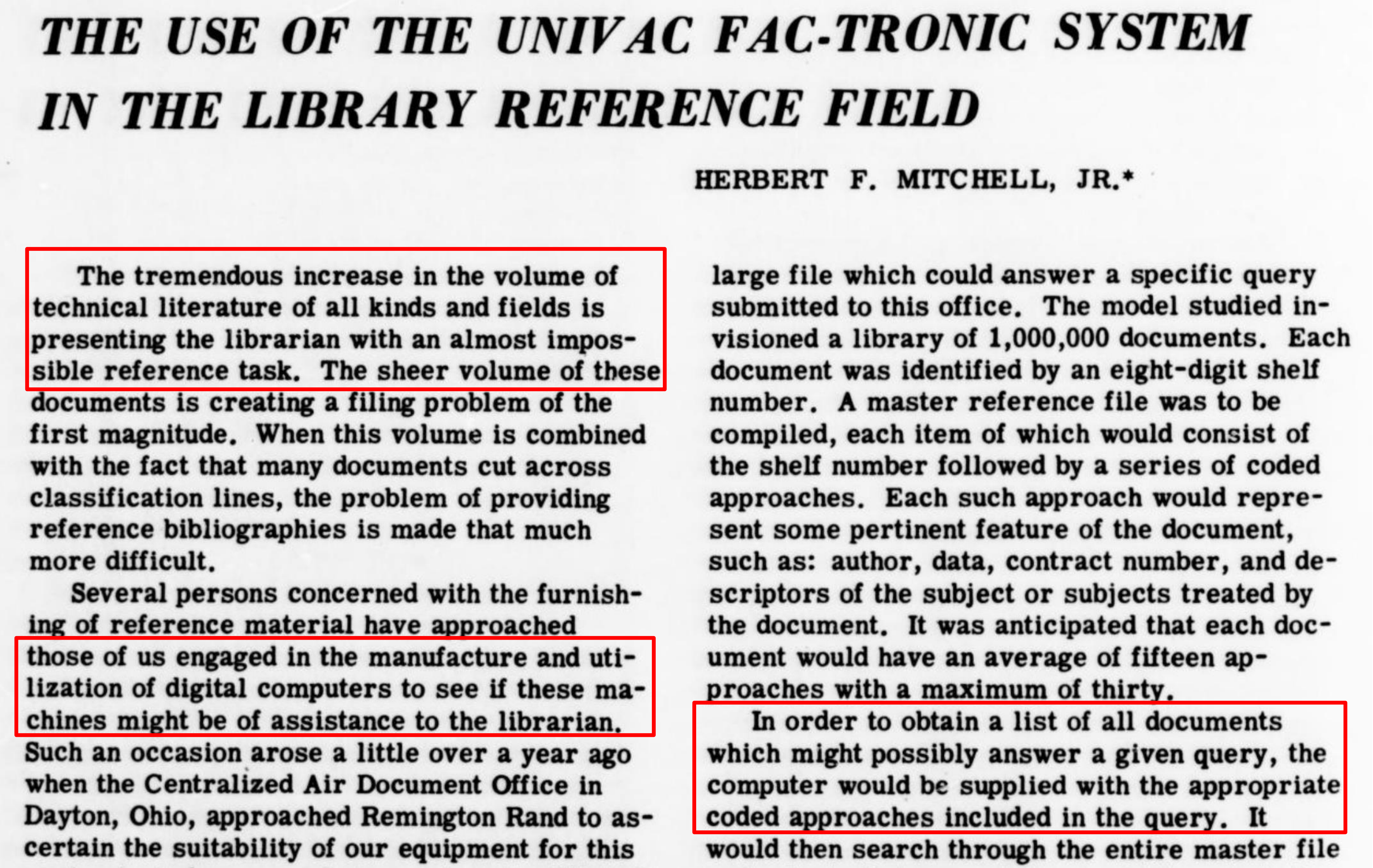
Information Retrieval and Extraction
資訊檢索與擷取 (CSIE5460)

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
hhchen@ntu.edu.tw

Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.

(Salton, 1968)



THE USE OF THE UNIVAC FAC-TRONIC SYSTEM IN THE LIBRARY REFERENCE FIELD

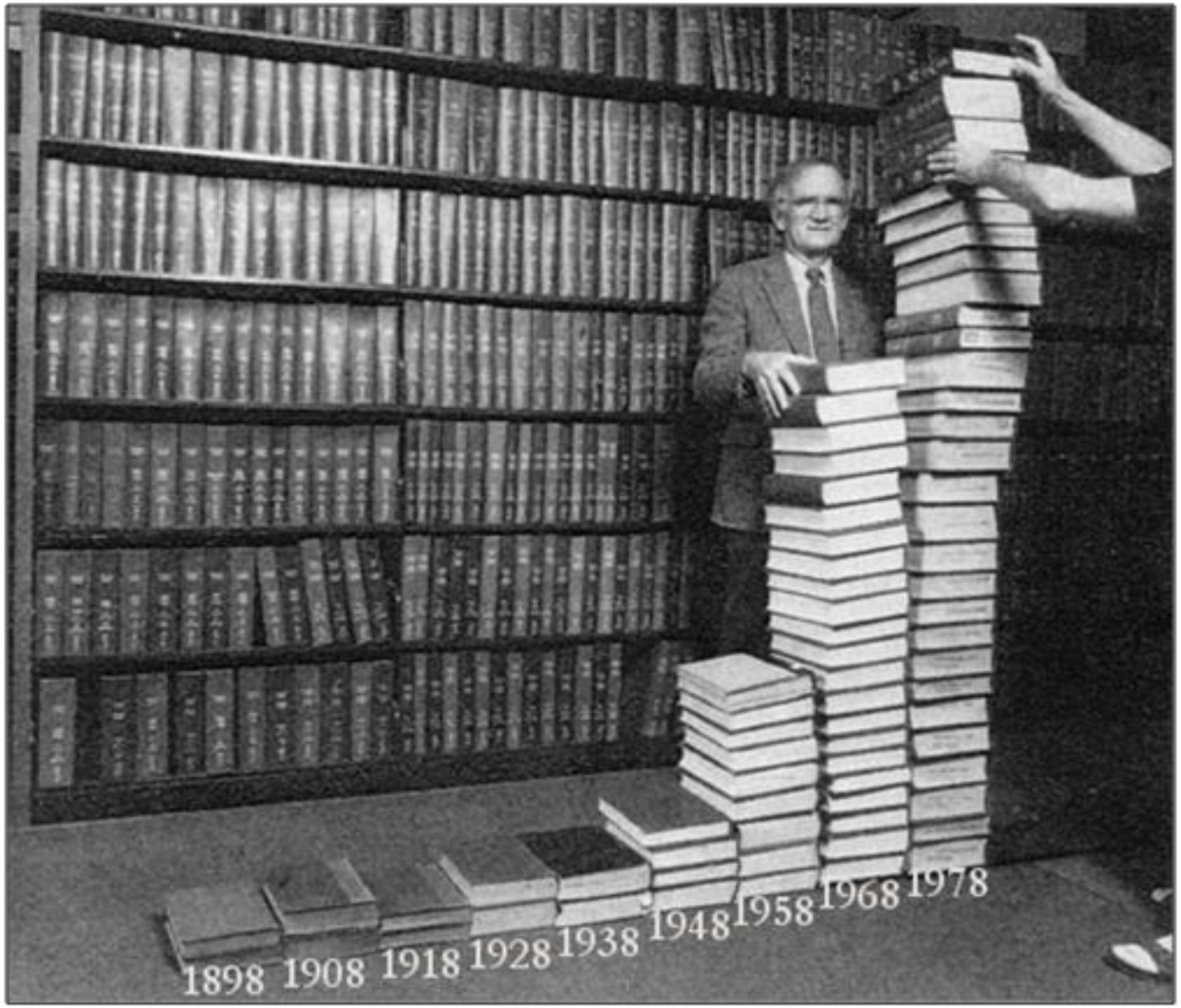
HERBERT F. MITCHELL, JR.*

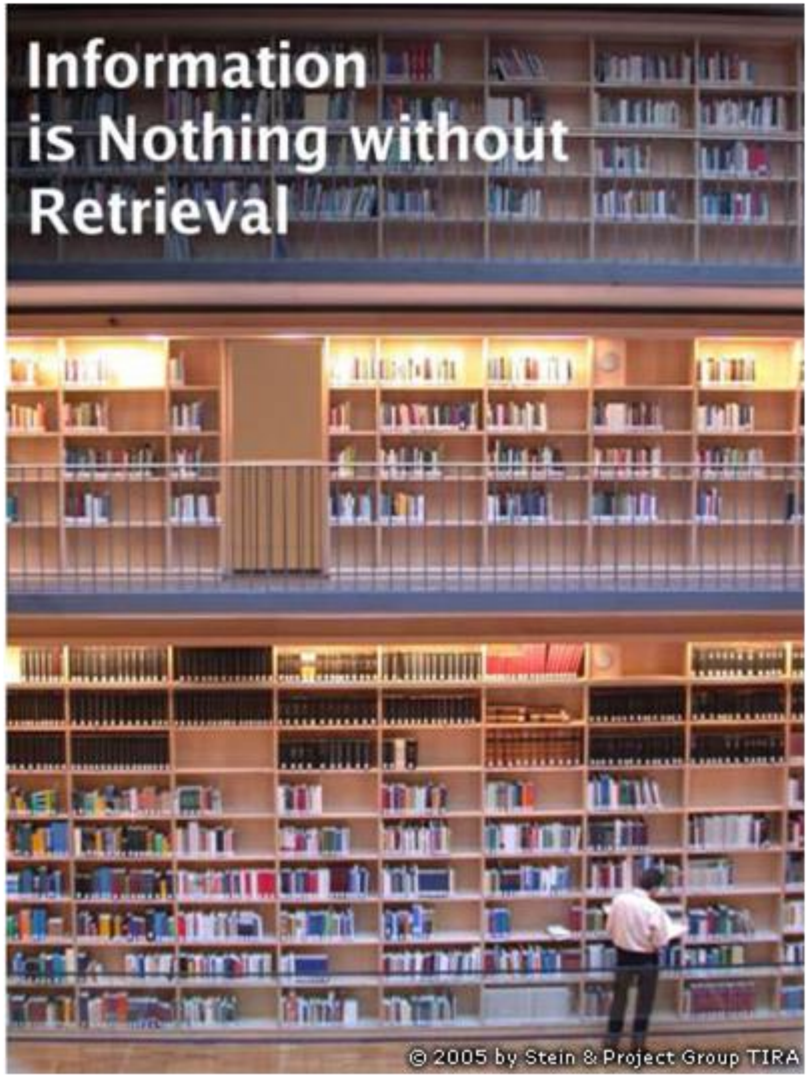
The tremendous increase in the volume of technical literature of all kinds and fields is presenting the librarian with an almost impossible reference task. The sheer volume of these documents is creating a filing problem of the first magnitude. When this volume is combined with the fact that many documents cut across classification lines, the problem of providing reference bibliographies is made that much more difficult.

Several persons concerned with the furnishing of reference material have approached those of us engaged in the manufacture and utilization of digital computers to see if these machines might be of assistance to the librarian. Such an occasion arose a little over a year ago when the Centralized Air Document Office in Dayton, Ohio, approached Remington Rand to ascertain the suitability of our equipment for this large file which could answer a specific query submitted to this office. The model studied invisioned a library of 1,000,000 documents. Each document was identified by an eight-digit shelf number. A master reference file was to be compiled, each item of which would consist of the shelf number followed by a series of coded approaches. Each such approach would represent some pertinent feature of the document, such as: author, data, contract number, and descriptors of the subject or subjects treated by the document. It was anticipated that each document would have an average of fifteen approaches with a maximum of thirty.

In order to obtain a list of all documents which might possibly answer a given query, the computer would be supplied with the appropriate coded approaches included in the query. It would then search through the entire master file

AMERICAN DOCUMENTATION, WINTER (JANUARY) 1953 , VOLUME IV, NUMBER 1, pp. 16-17


1898
1908
1918
1928
1938
1948
1958
1968
1978



From Physical Library to Digital Library

Google Books, …

Heterogeneous Resources, Devices, Context, …

Information Retrieval

- From search point of view
- Allow users to access information on heterogeneous sources efficiently and effectively
- Select the relevant information to meet users' information needs from heterogeneous sources

Information Retrieval vs. Information Filtering

- Two sides of the same coin
- 紐時攜手Alphabet 過濾不當網路留言 (中央社 2016.9.22)
  - 人身攻擊、淫穢、粗俗、褻瀆字眼、商業行為、罵人、冒名攻擊、顛三倒四、咆哮
  - 紐時目前僱用1個14人小組手動檢查每天約1萬1000則留言
  - 紐時目前僅1成文章開放留言
  - 提供多元社會多元討論的安全平台，讓每篇報導都聽得到讀者的聲音

Jeopardy人與電腦大戰
2011年2月14,15,16日

- DeepQA超級電腦(華生，Watson) 、簡寧斯(Ken Jennings)、洛特(Brad Rutter)

(美聯社)

- 第一名(華生)：一百萬美元，第二名(簡寧斯)：30萬美元，第三名(洛特)：20萬元

經典答錯題

- Q：指出一美國都市，此城市最大機場以二戰英雄命名，第二大機場以戰役命名

  A：芝加哥 (華生：多倫多)
- Q：老牌餅乾Oreo何時首次推出?

  A：1910年代 (簡寧斯： 20年代，華生：1920年代)

Is web search enough to win Jeopardy!

- To answer a *Jeopardy!* clue
  - go beyond the search result page
  - dig into documents that are likely to contain the answer
  - fetch them
  - read them
  - locate the precise and correct answer within them

http://www.research.ibm.com/deepqa/web_search.shtml

Jeopardy! Challenge

- Deeply analyze the question to figure out exactly what is being asked
- Deeply analyze the available content to extract precise answers
- Quickly compute a reliable confidence in light of whatever supporting or refuting information it finds.

DeepQA project at IBM

- The integration and advancement of Natural Language Processing, Information Retrieval, Machine Learning, Knowledge Representation and Reasoning, and massively parallel computation can drive open-domain automatic Question Answering technology to a point where it clearly and consistently rivals the best human performance

http://www.research.ibm.com/deepqa/deepqa.shtml

Related Topics

- Text Information Retrieval (文字)
- Multilingual Information Retrieval (語言)
- Multimedia Information Retrieval (多媒體)
- Web Information Retrieval (網路)
- Context-Aware Information Retrieval (情境)
- Social Media Retrieval (人、社群)
- Information Extraction (更細膩的資訊)

Text Information Retrieval

Information Retrieval

- Definition
  - “Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.” (Salton, 1968)
- Information need vs. Query
- Transformation Gap
- More precise questions, and more concrete answers

User Intention

Diversity and Popularity

Mention Disambiguation

Pitcher of New York Yankees

市委常委政法委書記

社會學學院教授

Reporter of
XinHua News Agency

CEO of Boeing China

王建民

Chien-Ming Wang

Research Fellow of
Chinese Academy of Social Sciences

Japanese Department Chair

Associate Fellow
of Academia Sinica

分院院長

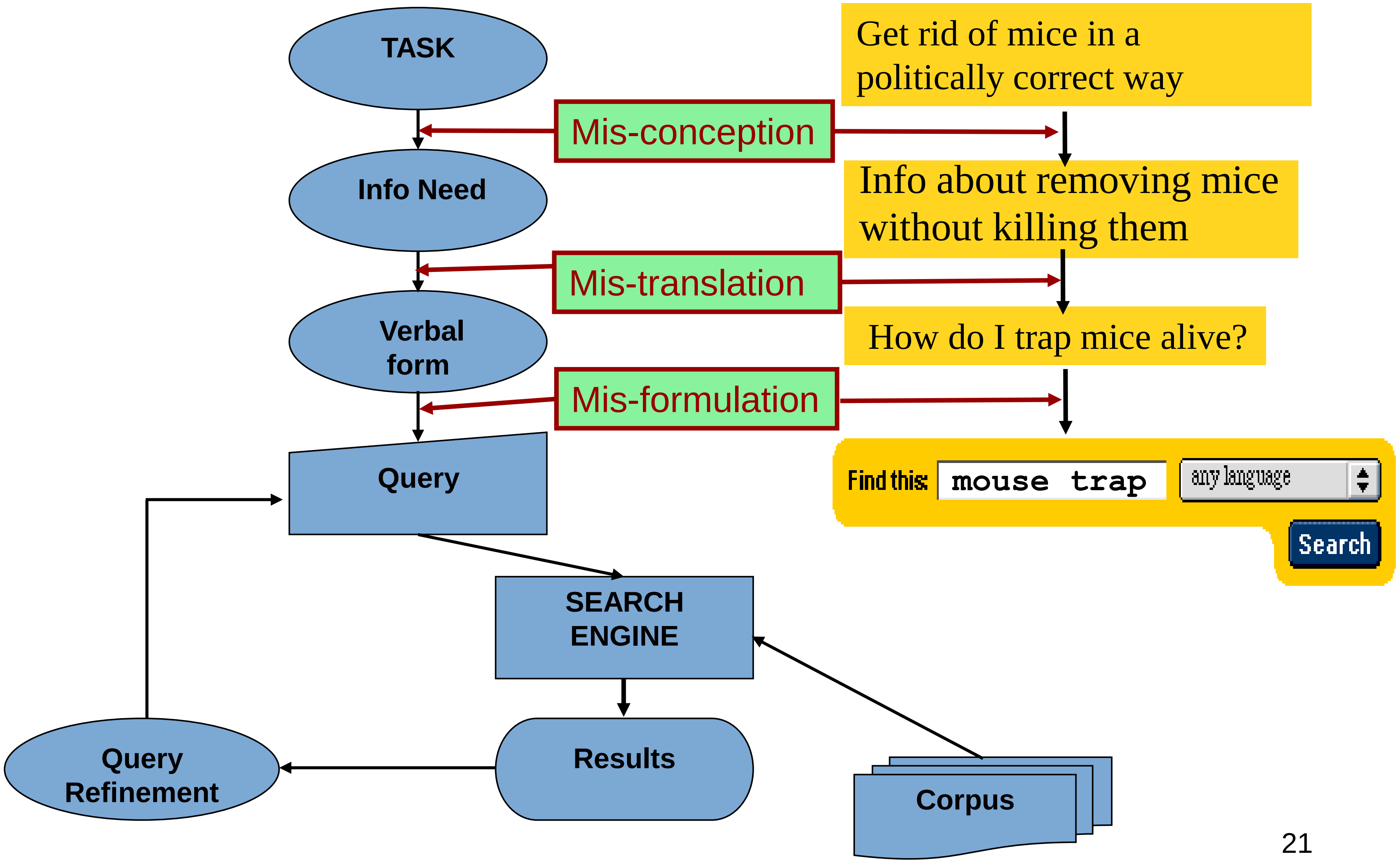
(Work done in 2006)

Entity Recognition and Disambiguation Challenge@ACM SIGIR 2014

- Recognize mentions of entities in a given text, disambiguate them, and map them to the entities in a given entity collection or knowledge base
  - Short track: queries
  - Long track: documents
  - Organizers: Google and Microsoft
- Winner of the first prize@ShortTrack: **NTUNLP**:
  Hsin-Hsi Chen, Yang-Yin Lee, Yong-Siang Shih, Chih-Chieh Shao, Yen-Pin Chiu, Sheng-Lun Wei and Ming-Lun Cai. National Taiwan University.
- The 4th place@LongTrack

http://web-ngram.research.microsoft.com/erd2014/Default.aspx

The Classic Search Model



TASK

Info Need

Verbal form

Query

SEARCH ENGINE

Results

Query Refinement

Corpus

Mis-conception

Mis-translation

Mis-formulation

Get rid of mice in a politically correct way

Info about removing mice without killing them

How do I trap mice alive?

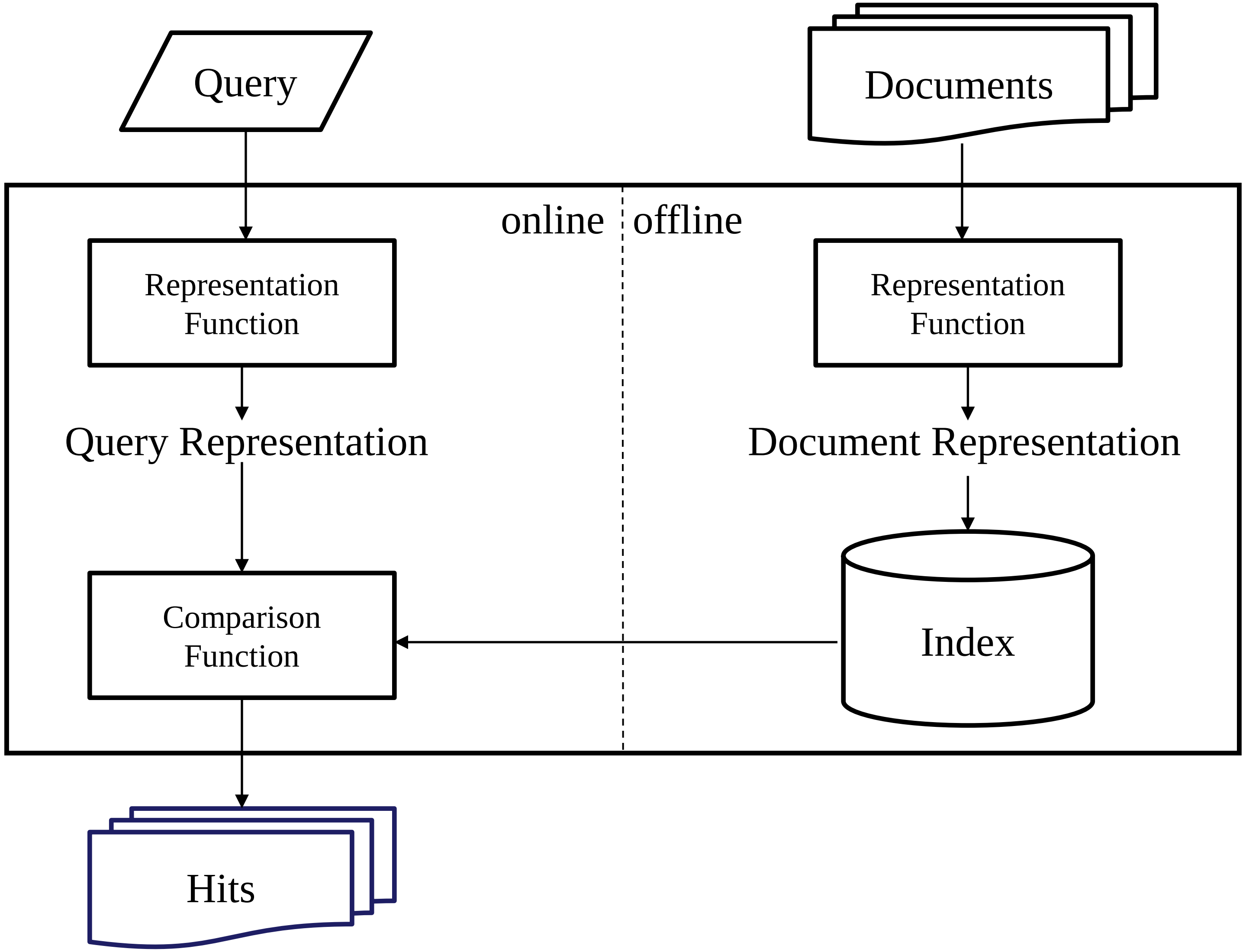
Find this: mouse trap

any language

Search

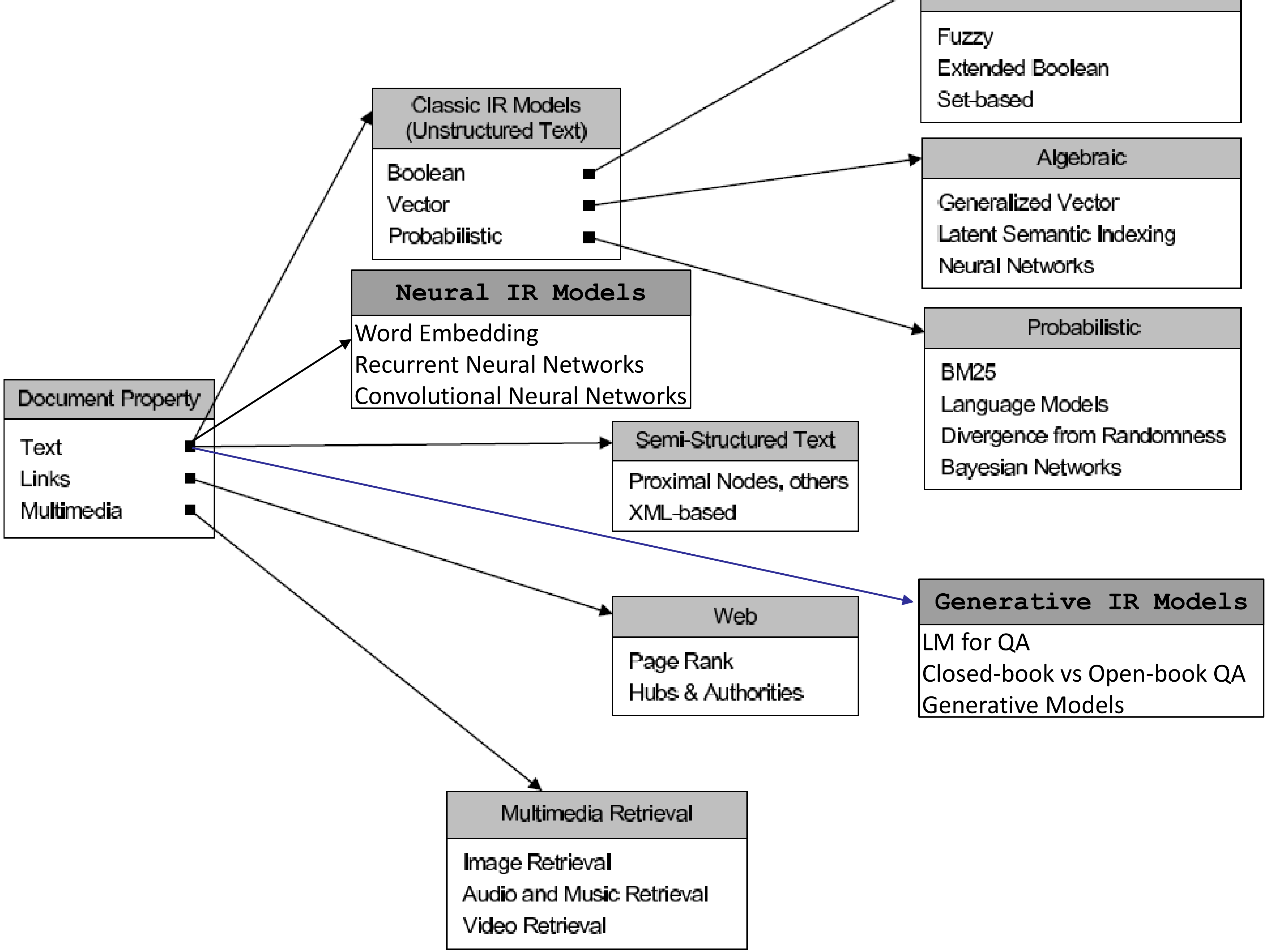
<from Manning>

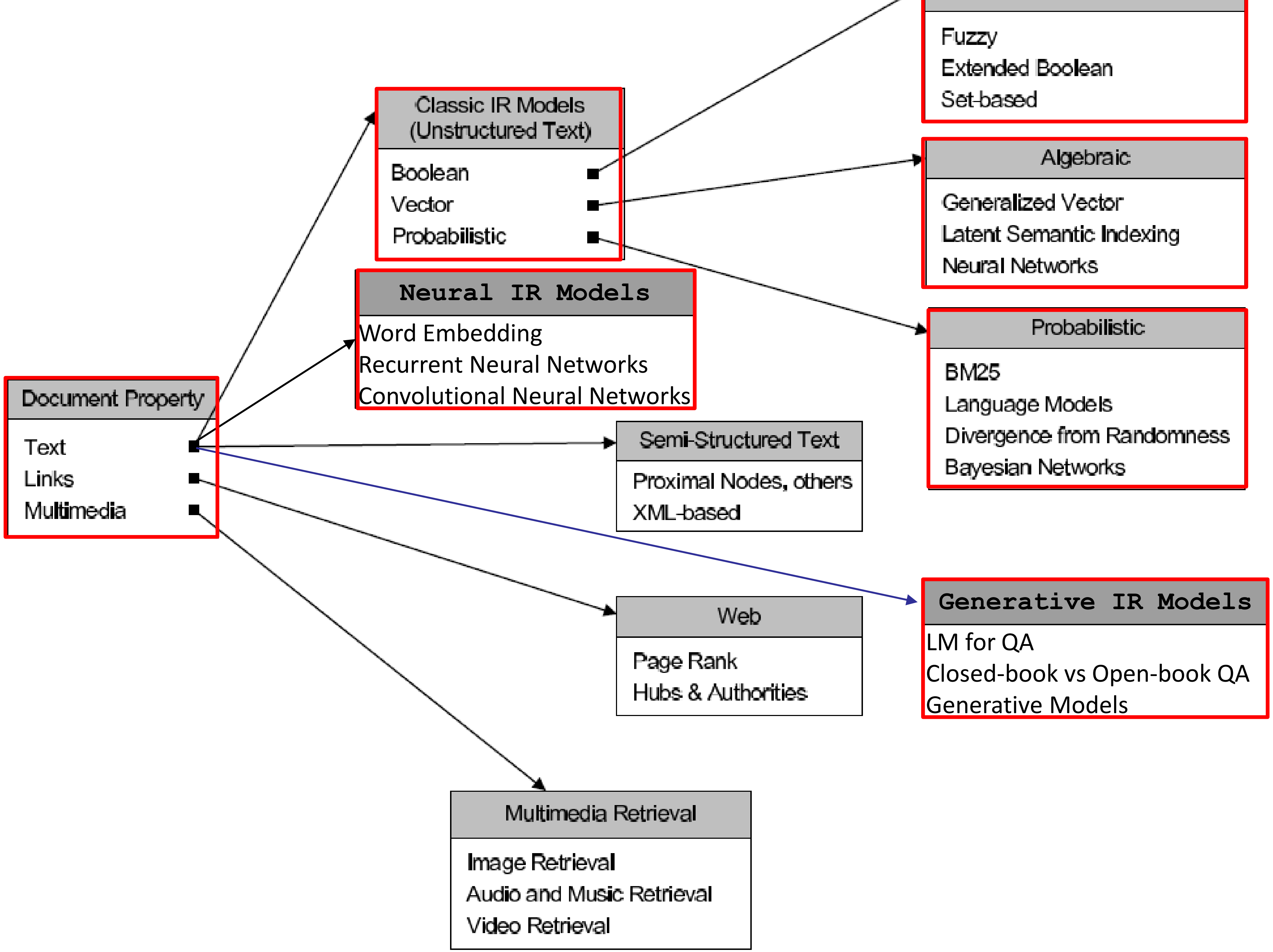
Architecture of IR Systems



Information Retrieval Models

- Boolean model
- Vector model
- Probabilistic model
- Language Model
- Neural IR Model
- Generative IR Model


Fuzzy
Extended Boolean
Set-based
Classic IR Models (Unstructured Text)
Boolean
Vector
Probabilistic
Algebraic
Generalized Vector
Latent Semantic Indexing
Neural Networks
Neural IR Models
Word Embedding
Recurrent Neural Networks
Convolutional Neural Networks
Probabilistic
BM25
Language Models
Divergence from Randomness
Bayesian Networks
Document Property
Text
Links
Multimedia
Semi-Structured Text
Proximal Nodes, others
XML-based
Generative IR Models
LM for QA
Closed-book vs Open-book QA
Generative Models
Web
Page Rank
Hubs & Authorities
Multimedia Retrieval
Image Retrieval
Audio and Music Retrieval
Video Retrieval



Document Property

Text

Links

Multimedia

Classic IR Models (Unstructured Text)

Boolean

Vector

Probabilistic

Fuzzy

Extended Boolean

Set-based

Algebraic

Generalized Vector

Latent Semantic Indexing

Neural Networks

Probabilistic

BM25

Language Models

Divergence from Randomness

Bayesian Networks

Neural IR Models

Word Embedding

Recurrent Neural Networks

Convolutional Neural Networks

Semi-Structured Text

Proximal Nodes, others

XML-based

Generative IR Models

LM for QA

Closed-book vs Open-book QA

Generative Models

Web

Page Rank

Hubs & Authorities

Multimedia Retrieval

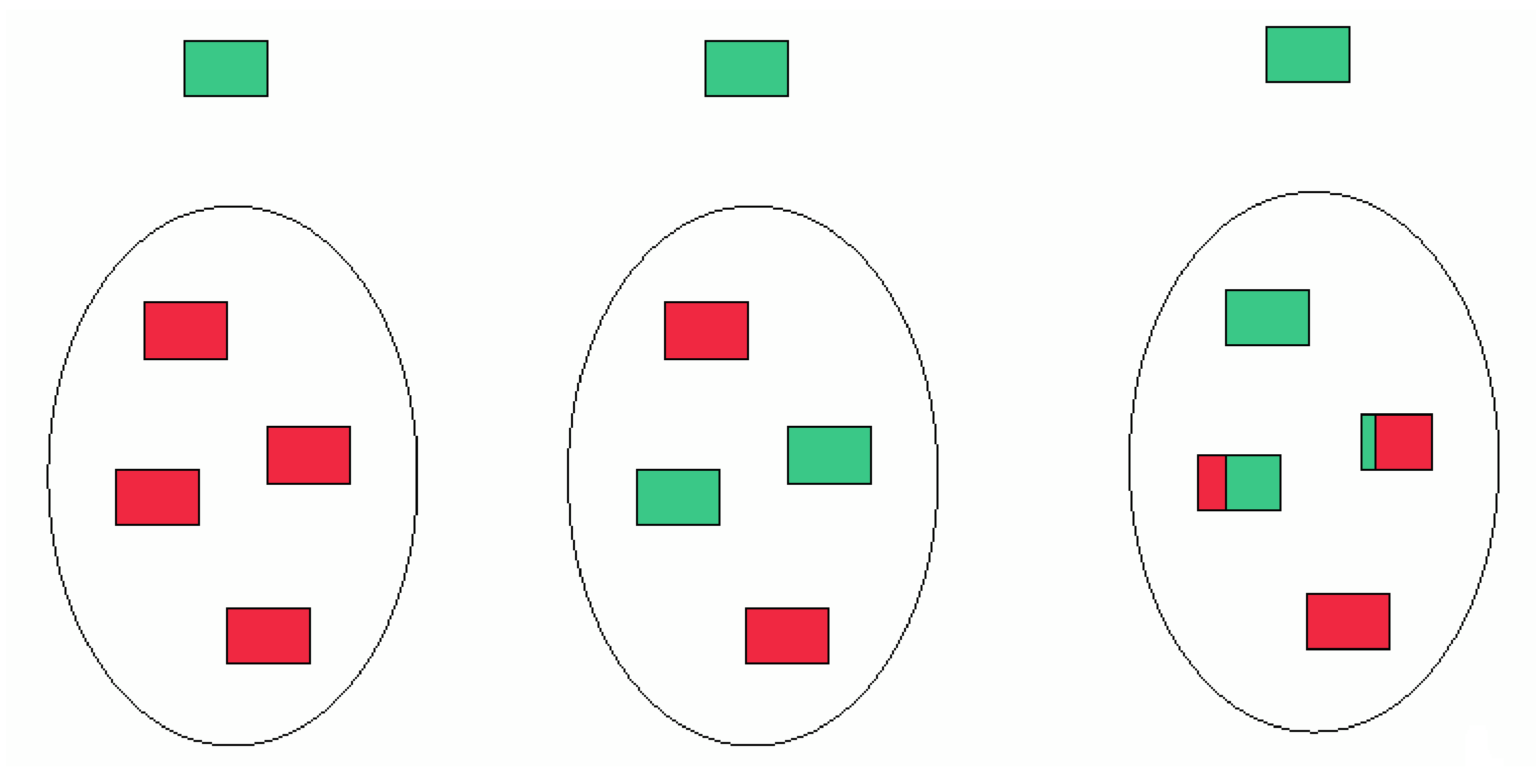
Image Retrieval

Audio and Music Retrieval

Video Retrieval

Multilingual Information Retrieval

Multi- & Cross- Lingual Information Access



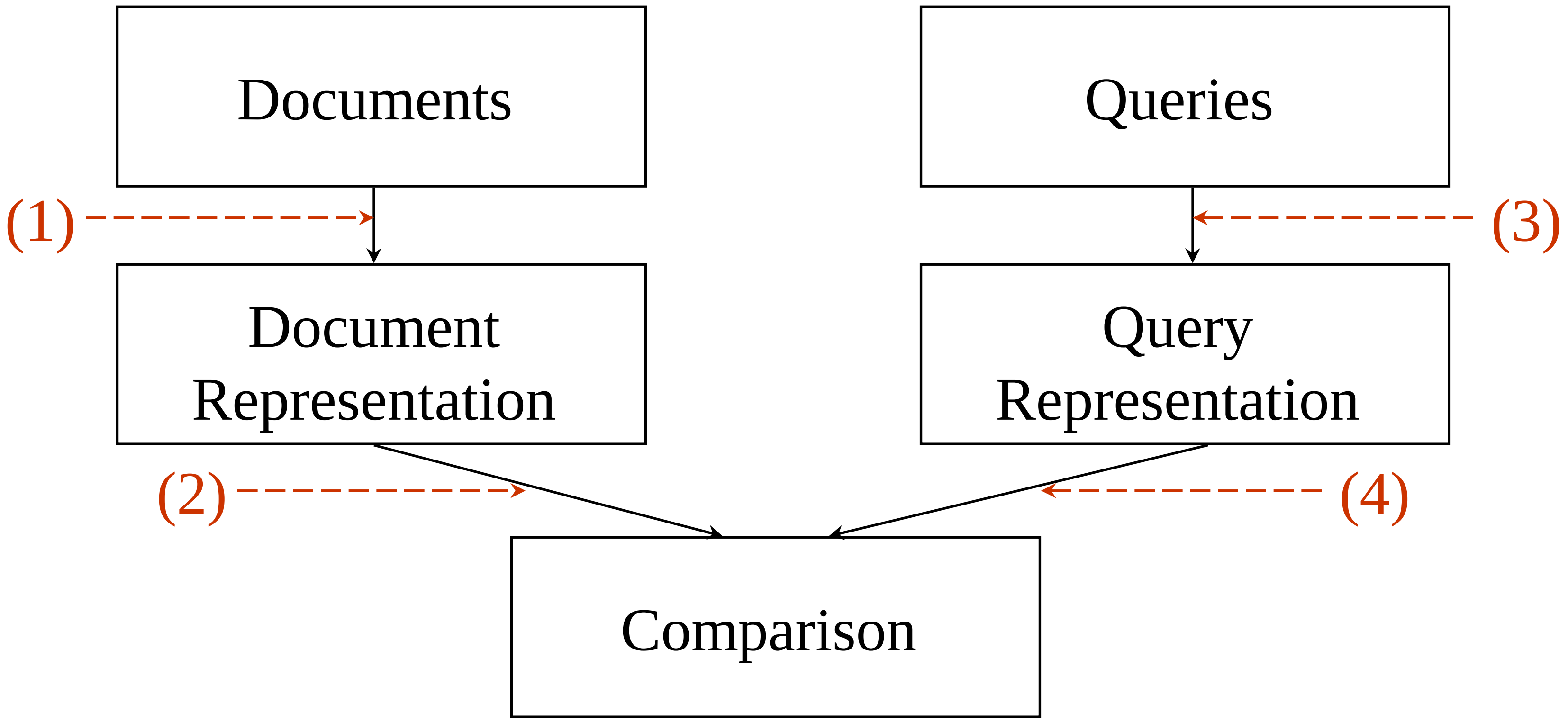
CLIR MLIR MLIR

What are the Problems?

- Ambiguous terms (e.g., bank)
- Multiword phrases may correspond to single-word phrases (e. g. South Africa => 南非， Südafrika)
- Coverage of the vocabulary
- There is not a one-to-one mapping between two languages
- Translating queries automatically (lack of syntax)
- Translating documents automatically (performance, …)
- Computing mixed result lists

Enhancing Traditional Information Retrieval Systems

- Which part(s) should be modified for CLIR?



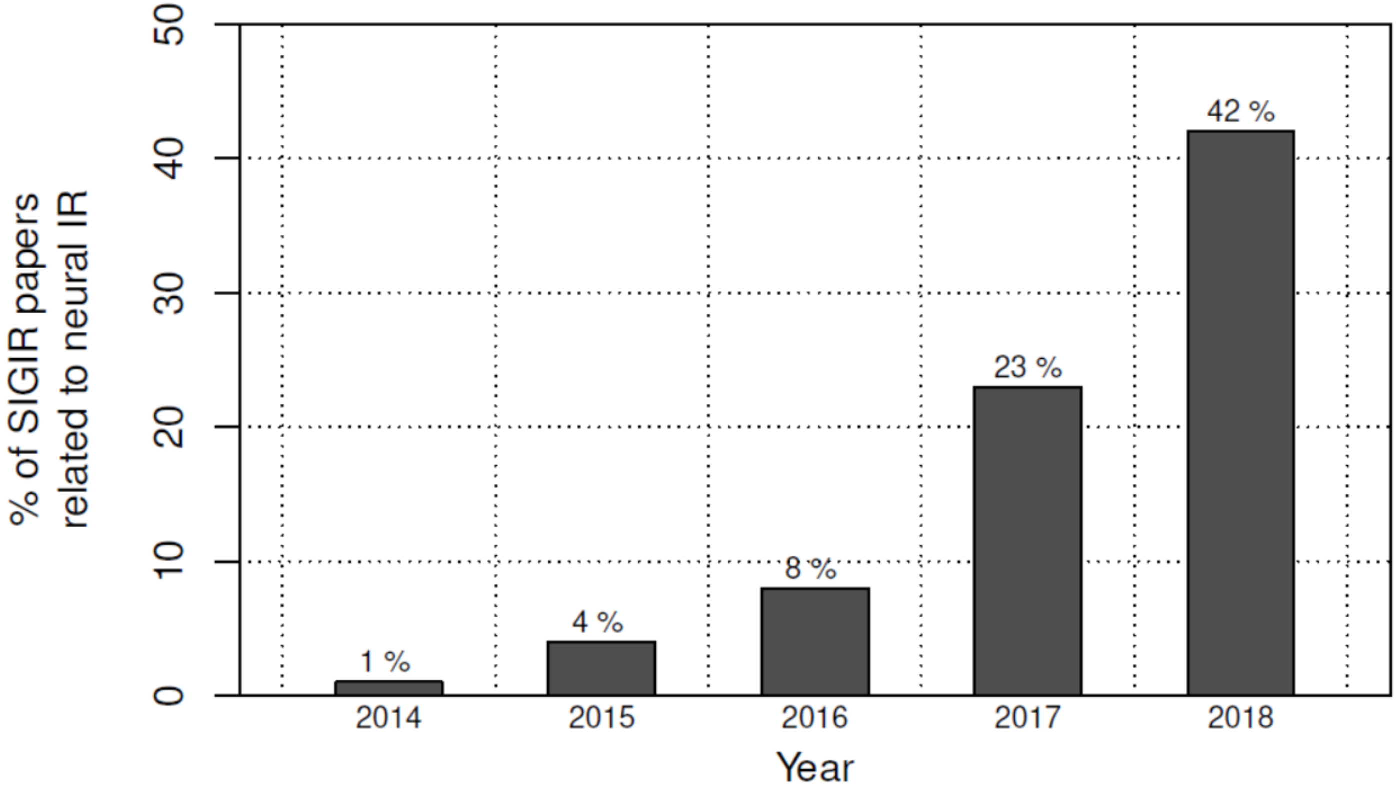
Enhancing Traditional Information Retrieval Systems (*Continued*)

- (1): text translation
- (2): vector translation
- (3): query translation
- (4): term vector translation
- (1) and (2), (3) and (4): interlingual form

Neural IR Approach

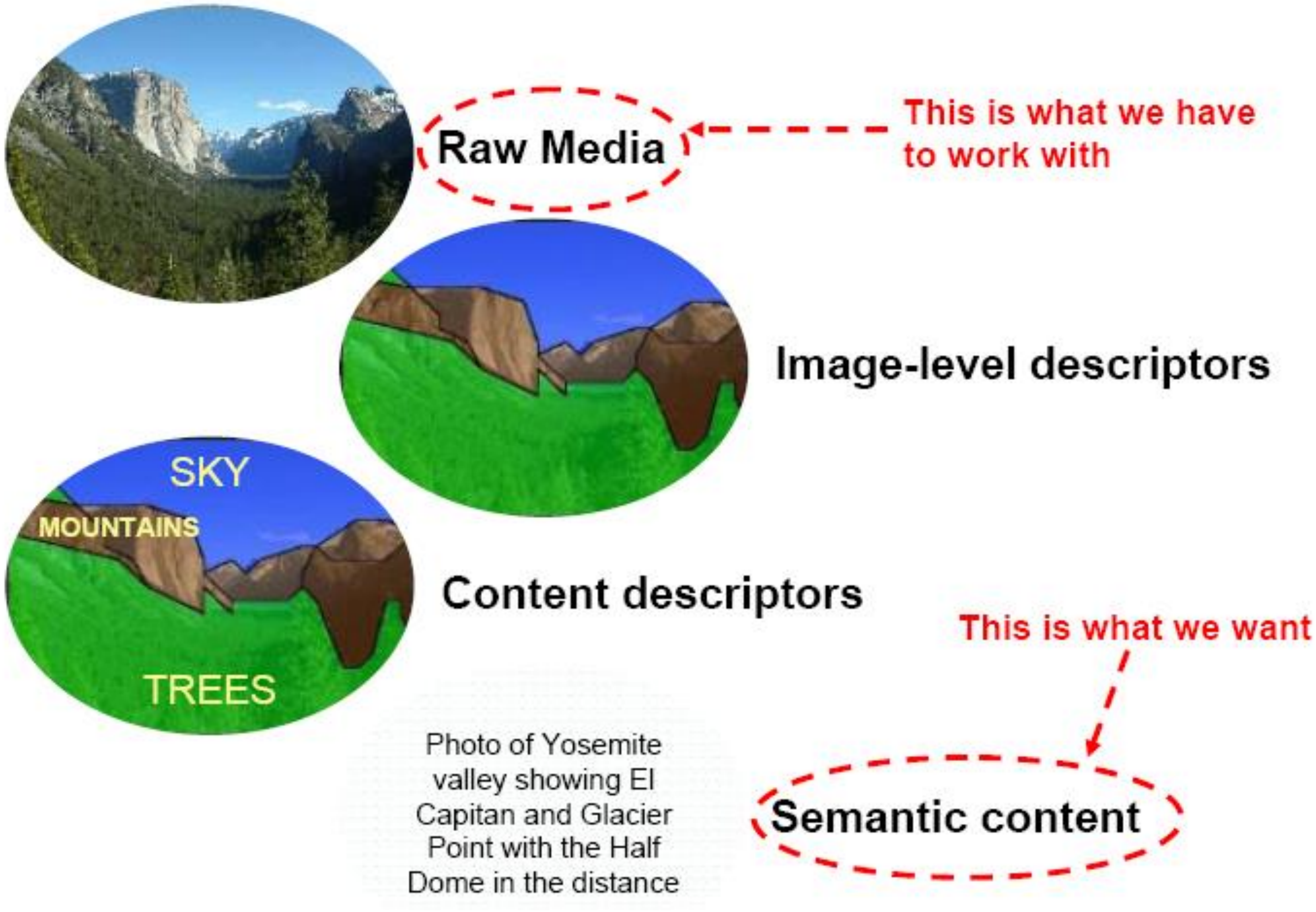
- Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings (SIGIR 2015)
- Represent words in different languages on the same space

% of Neural IR Papers at SIGIR



Multimedia Information Retrieval

Semantic Gap



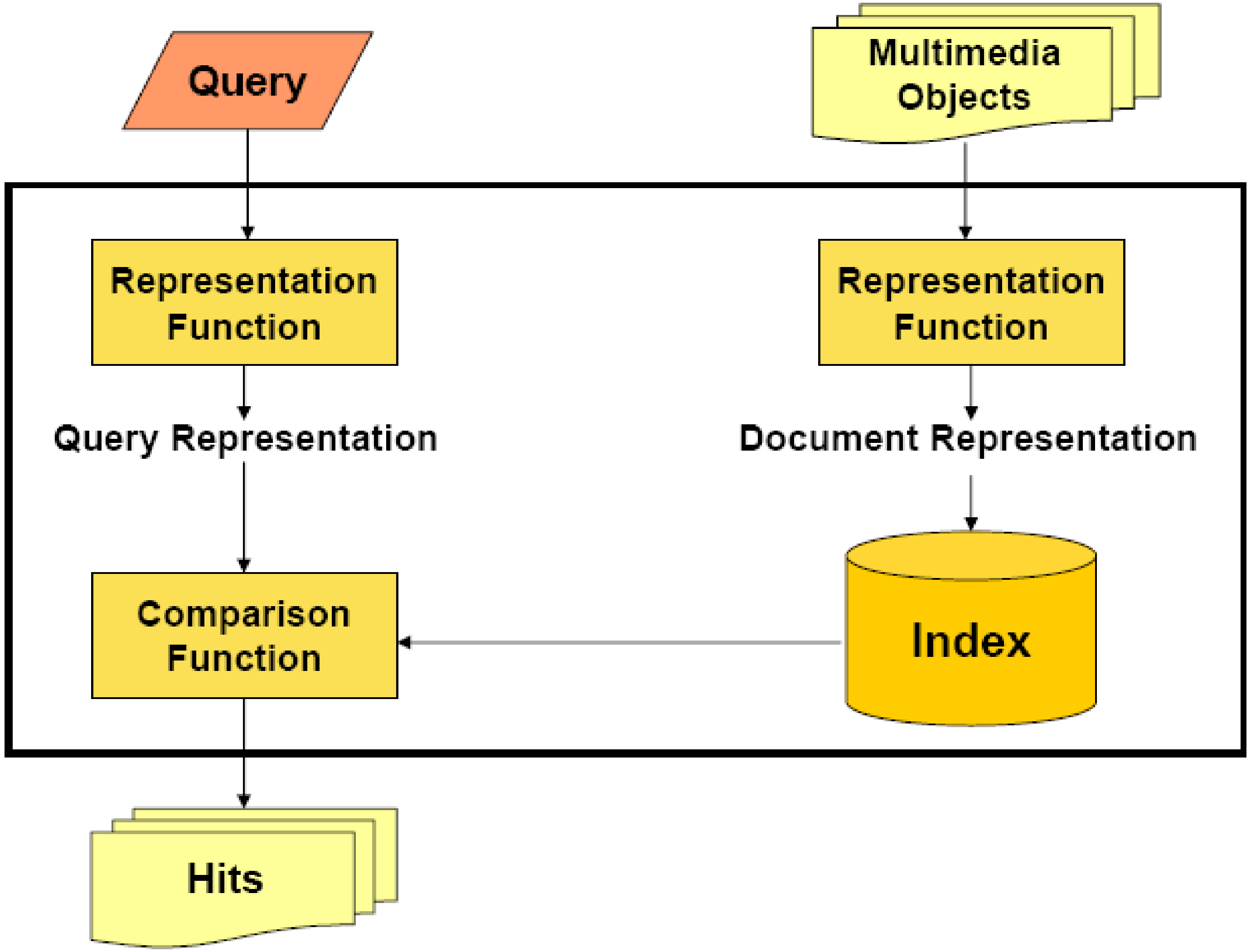
Raw Media
This is what we have to work with
Image-level descriptors
SKY
MOUNTAINS
TREES
Content descriptors
This is what we want
Photo of Yosemite valley showing El Capitan and Glacier Point with the Half Dome in the distance
Semantic content

relationships

Semantic Gap

- Content-based retrieval often fails due to the gap between information extractable automatically from the visual data (feature-vectors) and the interpretation a user may have for the same data.

Image Retrieval Black Box



Web Information Retrieval

The big challenge

Meet the user needs

given

the heterogeneity of web pages

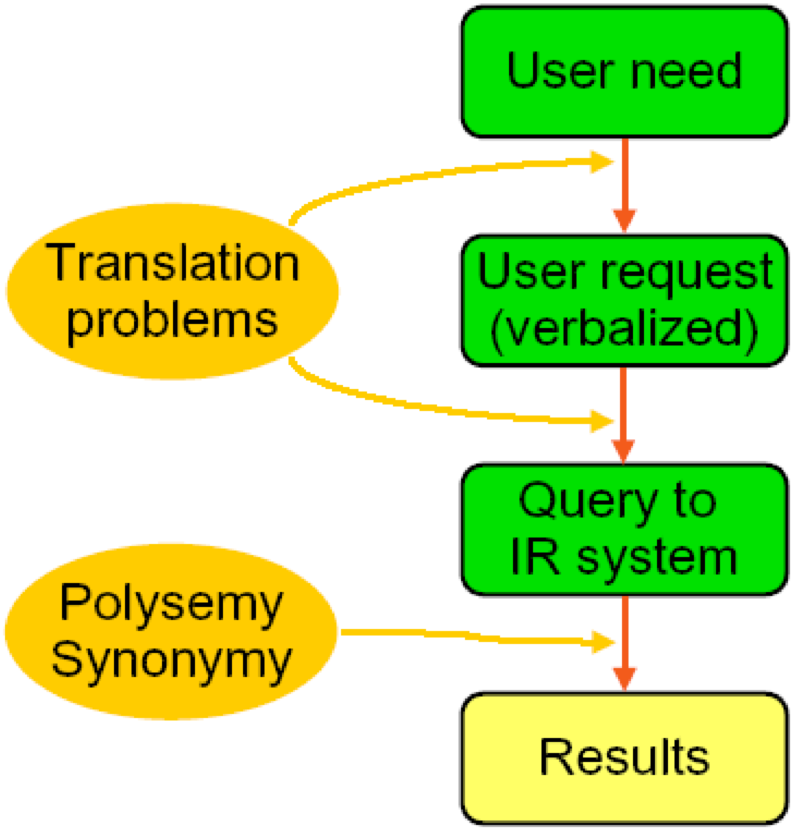
What’s different about the Web?
-- Users --

- Make poor queries
  - Short (2.35 terms avg)
  - Imprecise terms
  - Sub-optimal syntax
    (80% queries without operator)
  - Low effort
- Wide variance in
  - Needs
  - Knowledge
  - Bandwidth
- Specific behavior
  - 85% look over one result screen only
  - 78% of queries are not modified
  - Follow links
  - See various user studies in CHI, Hypertext, SIGIR, etc.

The bigger challenge

Meet the user needs
given
the heterogeneity of web pages
and
the poorly made queries

Why don’t the users get what they want?



User need

User request (verbalized)

Query to IR system

Results

Translation problems

Polysemy Synonymy

Example

I need to get rid of mice in the basement

What’s the best way to trap mice alive?

`mouse trap`

Software, toy cars, inventive products, etc

Context-Aware Information Retrieval

Definitions of Context for IR

- IR: location and delivery of documents which satisfy a user information need.
- IR takes place in “context”, but this context is generally ignored in IR models and system design.
- The definition of context in IR is widely interpreted.

Context in Mobile and Ubiquitous Information Access

- Physical context data can be made available via personal and environmental sensors.
- Some context information can be used directly, e.g. current location.
- Other may need to be aggregated, e.g., to decide on audio information delivery while the user is driving.

Social Media Retrieval

Applications of Social Media

- Brand monitoring
  - Social media measurement
- Communication
  - Blogs (Blogger, Livejournal, …)
  - Microblogs (Plurk, Twitter, …)
- Collaboration/authority building
  - Wikis (Wikipedia), Social bookmarking (Delicious), …
- Entertainment
  - Online gaming (World of Warcraft, 魔獸世界), …

Applications of Social Media

- Multimedia
  - – Photo sharing (Flickr), Video sharing (YouTube), …
- Reviews and Opinions
  - – Community Q&A (ask.com, Yahoo! Answers, …),
  - – Product reviews (eopinions.com, MouthShut.com)

Information Extraction

Information Extraction

- Automatic extraction of structured information from unstructured sources
  - entities
  - relationships between entities, and
  - attributes describing entities

What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

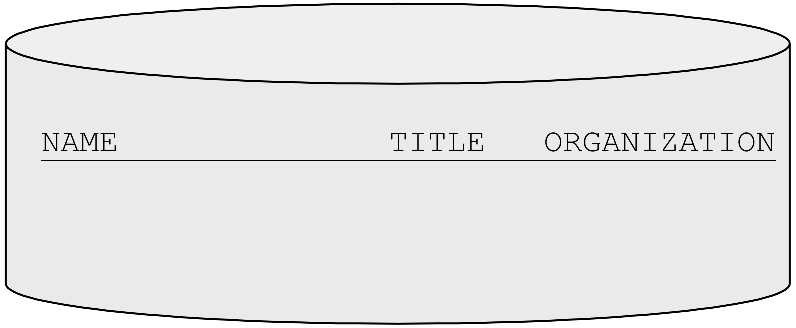
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION

What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

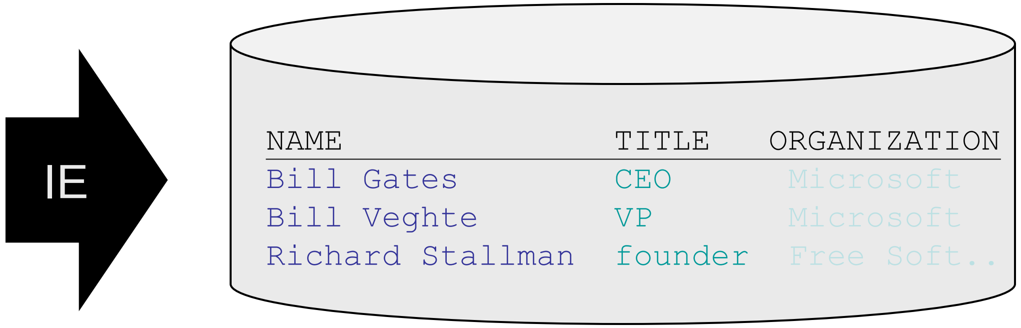
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is “Information Extraction”

As a task: Filling slots in a database from sub-segments of text.

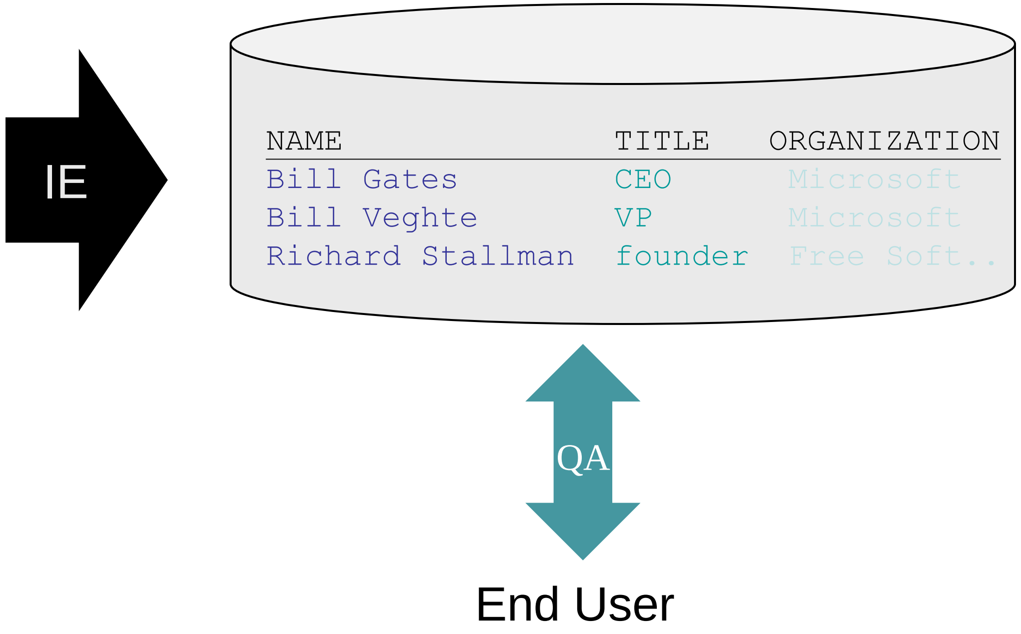
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

What is “Information Extraction”

As a family
of techniques:

Information Extraction =
segmentation + classification + clustering + association

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

aka “named entity extraction”

What is “Information Extraction”

As a family
of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family
of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

What is “Information Extraction”

As a family
of techniques:

Information Extraction =
segmentation + classification + association + clustering

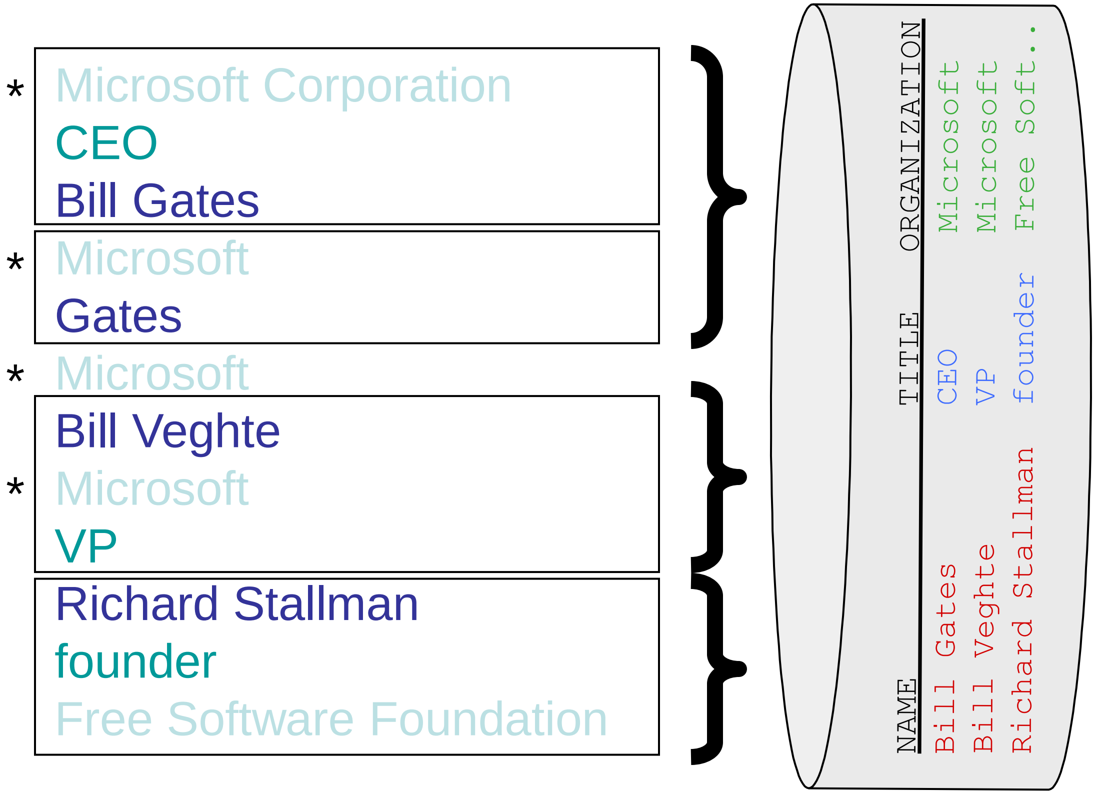
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…



* Microsoft Corporation
CEO
Bill Gates

* Microsoft
Gates

* Microsoft

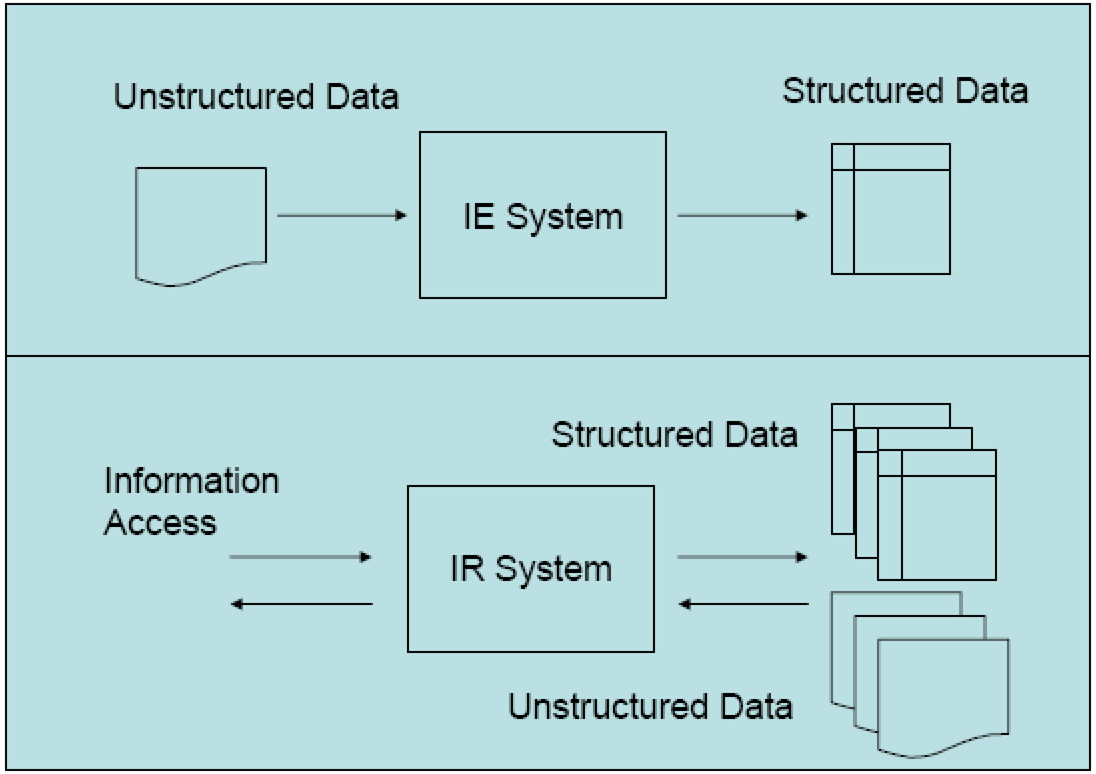
Bill Veghte

* Microsoft
VP

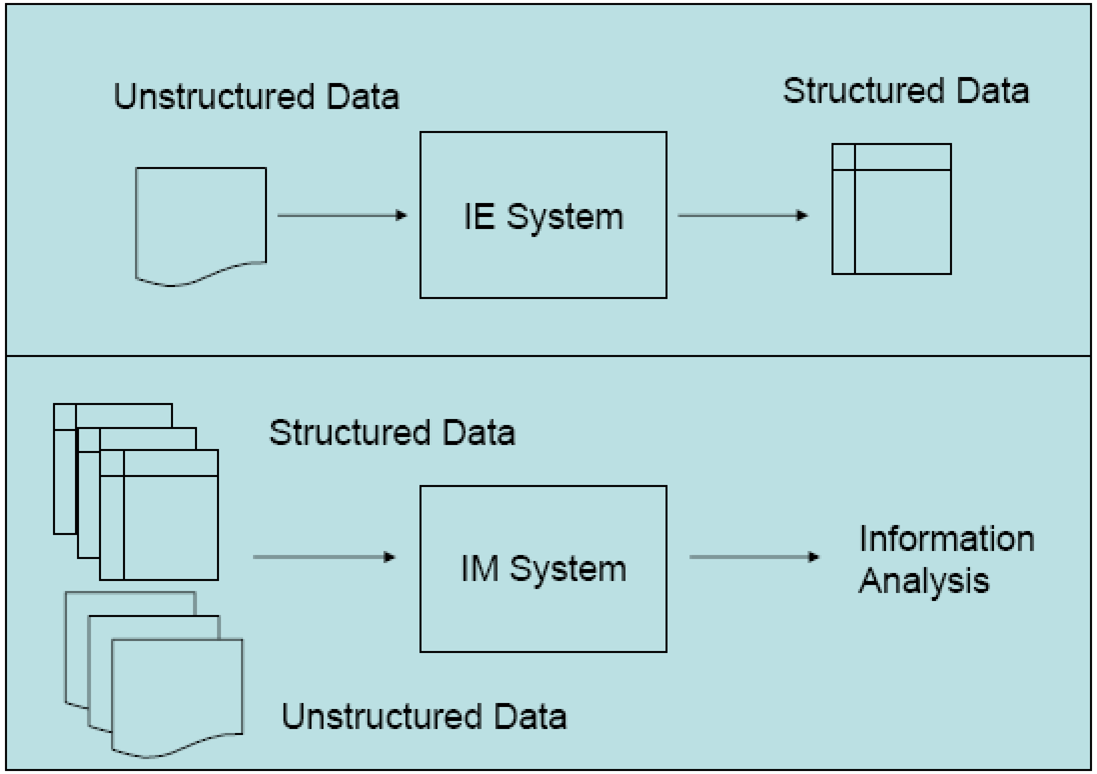
Richard Stallman
founder
Free Software Foundation

NAME	TITLE	ORGANIZATION
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	founder	Free Soft..

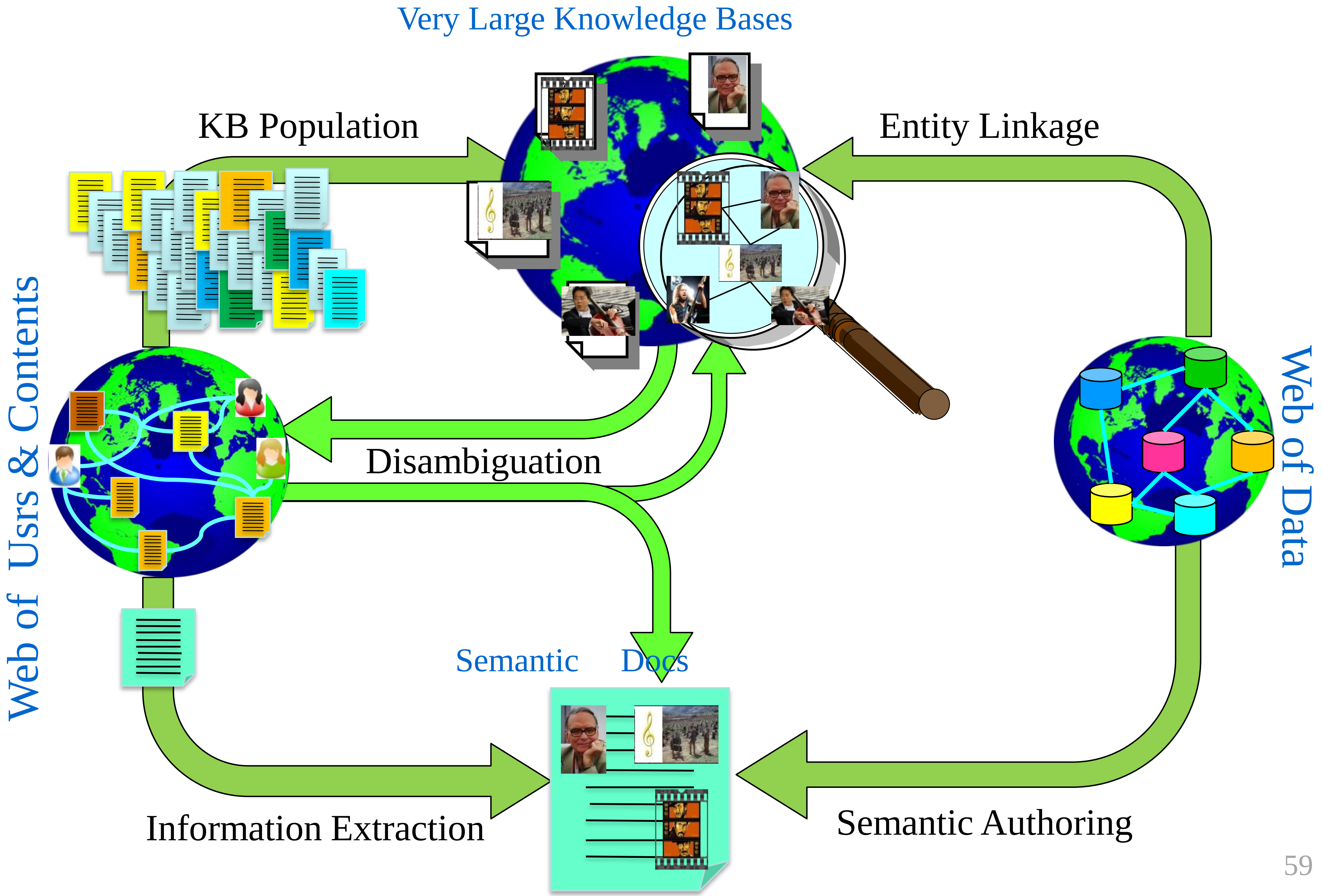
Information Extraction vs. Information Retrieval



Information Extraction vs. Information Mining



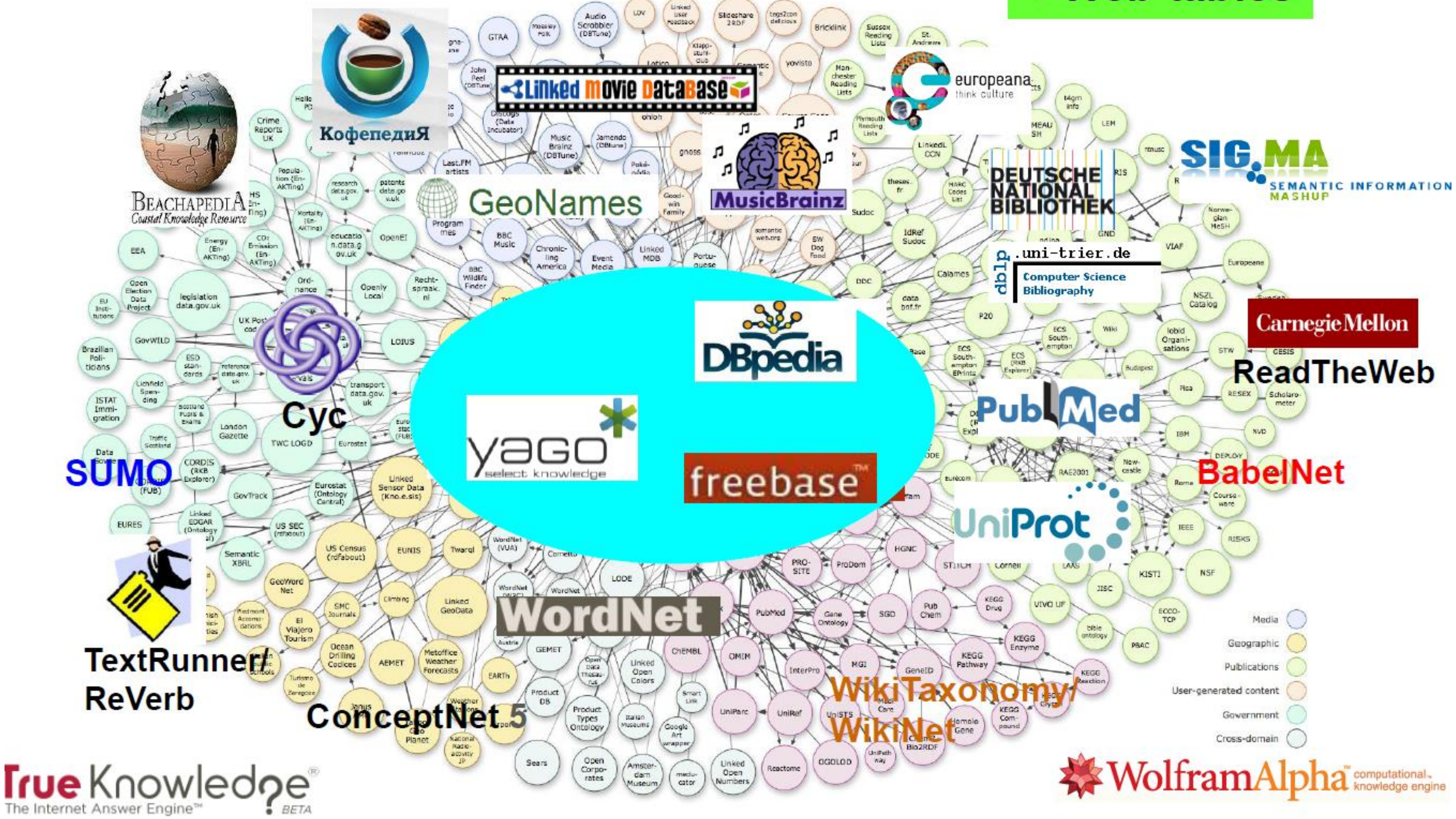
Turn Web into Knowledge Base



Web of Data & Knowledge

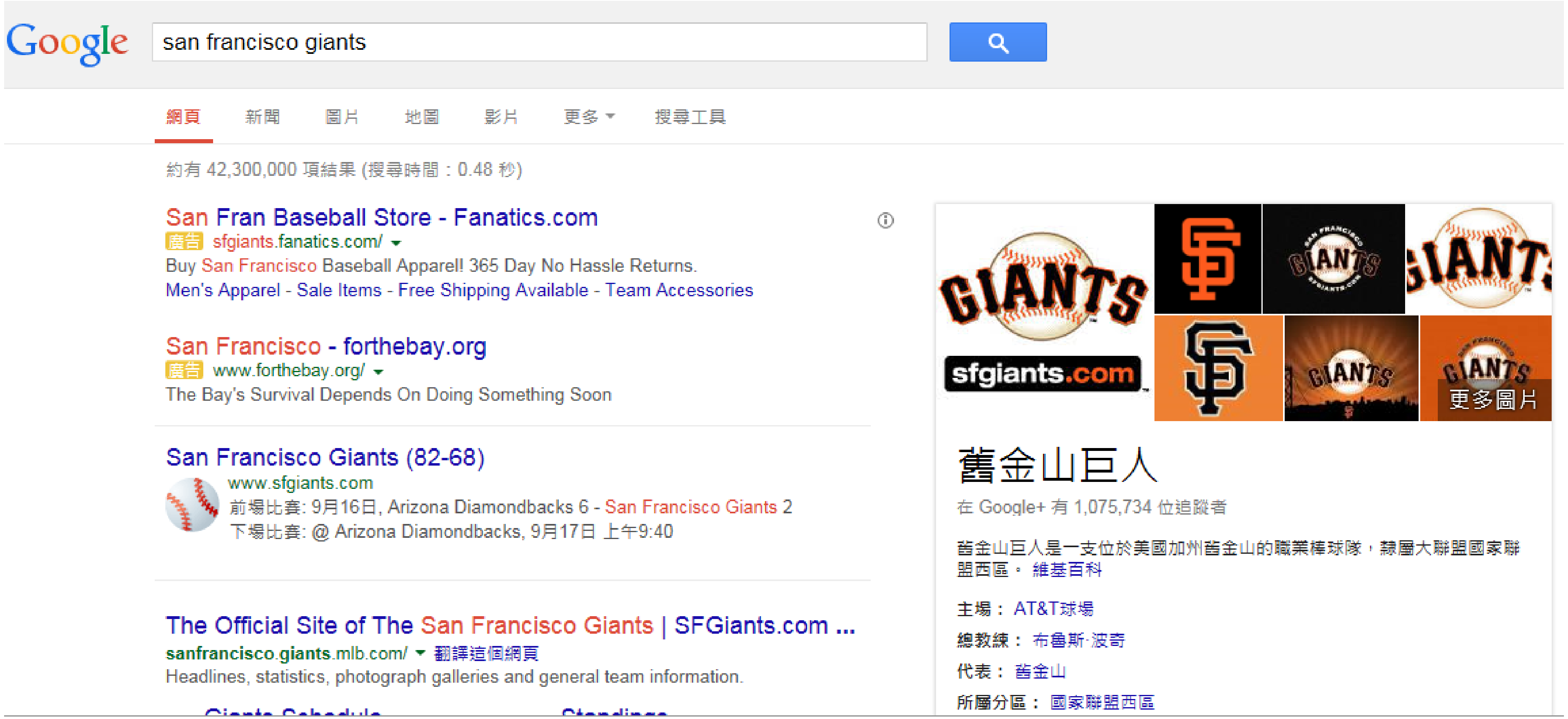
62 Bio. SPO triples (RDF) from 870 sources, and growing

+ Web tables



(Gerhard Weikum, 2013)

Summarize those facts about things



約有 42,300,000 項結果 (搜尋時間：0.48 秒)

San Fran Baseball Store - Fanatics.com
廣告 sfgiants.fanatics.com/
Buy San Francisco Baseball Apparel! 365 Day No Hassle Returns.
Men's Apparel - Sale Items - Free Shipping Available - Team Accessories

San Francisco - forthebay.org
廣告 www.forthebay.org/
The Bay's Survival Depends On Doing Something Soon

San Francisco Giants (82-68)
www.sfgiants.com
前場比賽: 9月16日, Arizona Diamondbacks 6 - San Francisco Giants 2
下場比賽: @ Arizona Diamondbacks, 9月17日 上午9:40

The Official Site of The San Francisco Giants | SFGiants.com ...
sanfrancisco.giants.mlb.com/ 翻譯這個網頁
Headlines, statistics, photograph galleries and general team information.
Giants Schedule Standings

舊金山巨人

在 Google+ 有 1,075,734 位追蹤者

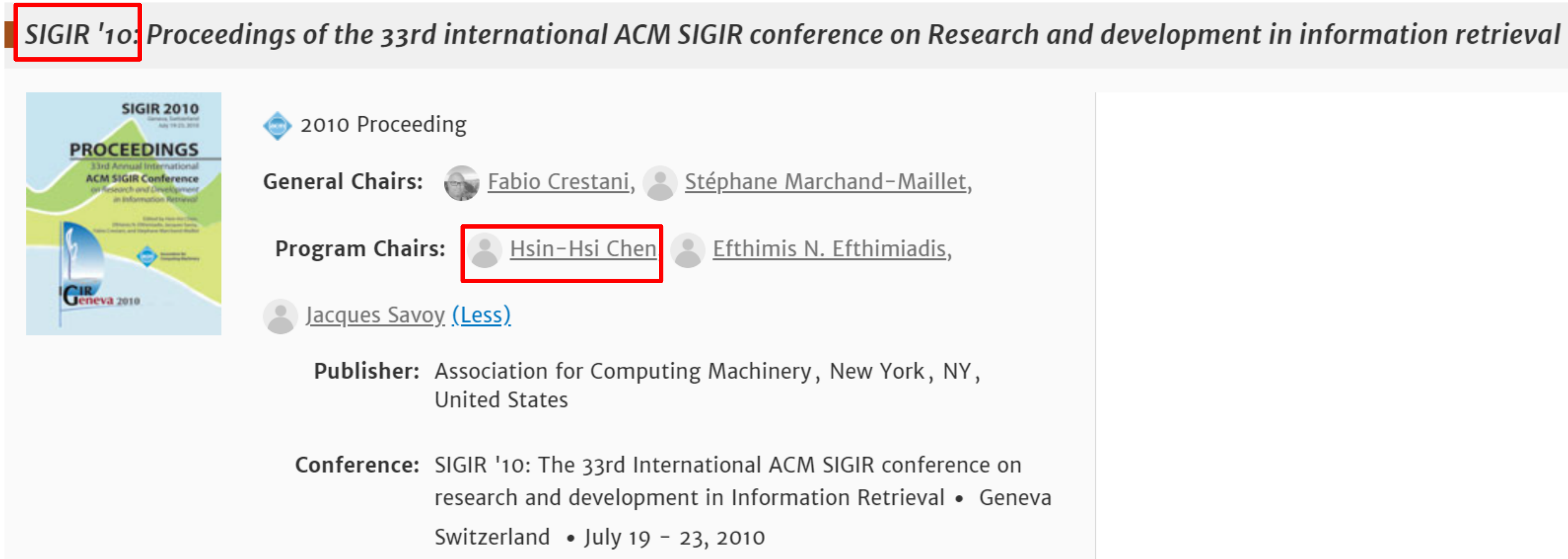
舊金山巨人是一支位於美國加州舊金山的職業棒球隊，隸屬大聯盟國家聯盟西區。 維基百科

主場： AT&T球場

總教練： 布魯斯·波奇

代表： 舊金山

所屬分區： 國家聯盟西區



SIGIR '10: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval

2010 Proceeding

General Chairs: Fabio Crestani, Stéphane Marchand-Maillet,

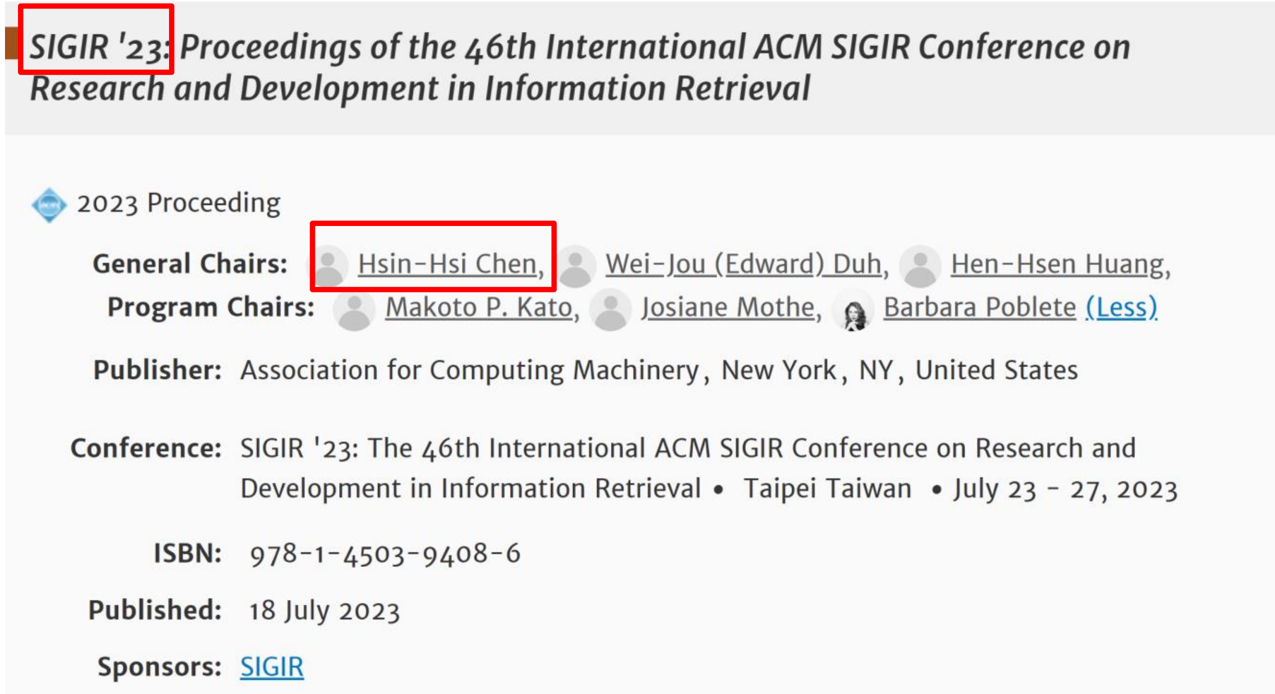
Program Chairs: Hsin-Hsi Chen Efthimis N. Efthimiadis,

Jacques Savoy (Less)

Publisher: Association for Computing Machinery, New York, NY, United States

Conference: SIGIR '10: The 33rd International ACM SIGIR conference on research and development in Information Retrieval • Geneva Switzerland • July 19 - 23, 2010

SIGIR 2010
PROCEEDINGS
33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval
Geneva 2010



SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

2023 Proceeding

General Chairs: Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang,

Program Chairs: Makoto P. Kato, Josiane Mothe, Barbara Poblete (Less)

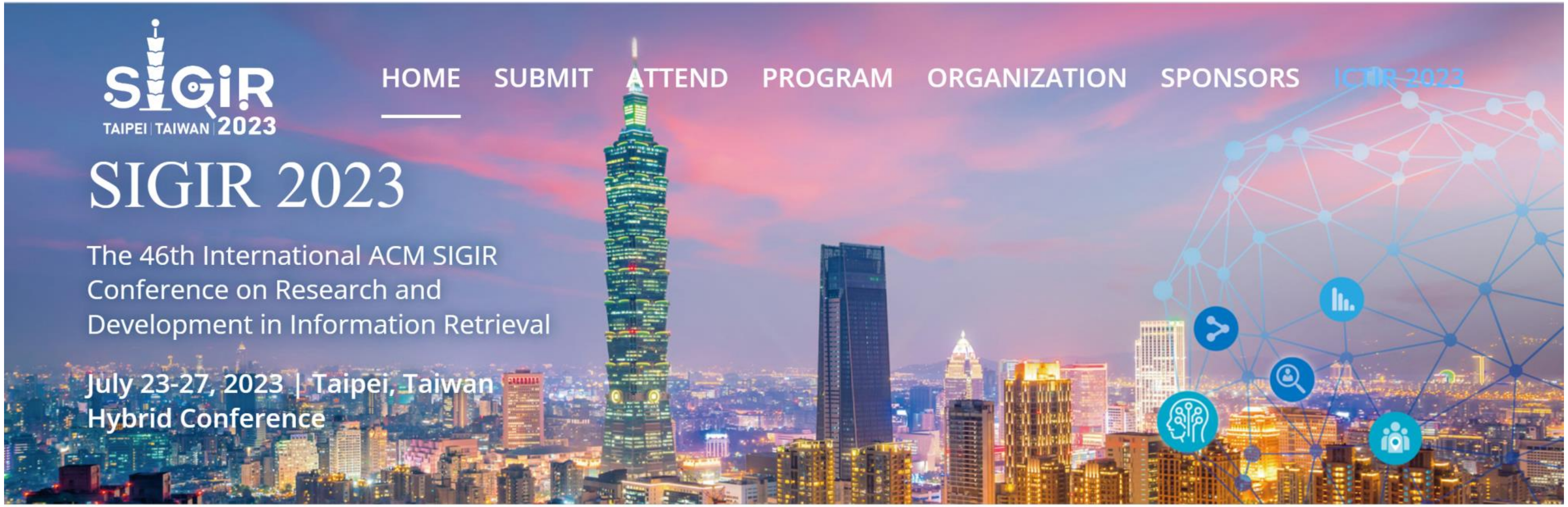
Publisher: Association for Computing Machinery, New York, NY, United States

Conference: SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval • Taipei Taiwan • July 23 - 27, 2023

ISBN: 978-1-4503-9408-6

Published: 18 July 2023

Sponsors: SIGIR



SIGIR 2023

The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

July 23-27, 2023 | Taipei, Taiwan
Hybrid Conference

“ The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval will be held from 23-27 July, 2023 in Taipei. Please

Hybrid Conference in 2023

SIGIR
TAIPEI TAIWAN 2023
July 23-27

SIGIR
TAIPEI TAIWAN 2023
July 23-27
The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval

SIGIR
TAIPEI TAIWAN 2023
The 46th International ACM SIGIR Conference on Research and Development in Info ation Retri

SIGIR 2023 vs. SIGIR 2010

- **Search and Ranking.** Research on core IR algorithmic topics, including IR at scale.
  - Queries and Query Analysis (e.g., query intent, query understanding, query suggestion and prediction, query representation and reformulation, spoken queries).
  - Web Search (e.g., ranking at Web scale, link analysis, sponsored search, search advertising, adversarial search and spam, vertical search).
  - Retrieval Models and Ranking (e.g., ranking algorithms, learning to rank, language models, retrieval models, combining searches, diversity, aggregated search, dealing with bias).
  - Efficiency and Scalability (e.g., indexing, crawling, compression, search engine architecture, distributed search, metasearch, peer-to-peer search, search in the cloud).

https://sigir.org/sigir2023/submit/call-for-full-papers/

SIGIR 2023 vs. SIGIR 2010

- **Search and Ranking.** Research on core IR algorithmic topics, including IR at scale.
  - Theoretical models and foundations of information retrieval and access (e.g., new theory, fundamental concepts, theoretical analysis).

SIGIR 2023 vs. SIGIR 2010

- **Search Recommendation & Content Analysis for Search and Recommendation.** Research focusing on recommender systems, rich content representations and content analysis.
  - Filtering and Recommendation (e.g., content-based filtering, collaborative filtering, recommender systems, recommendation algorithms, zero-query and implicit search, personalized recommendation).
  - Document Representation and Content Analysis for search or recommendation (e.g., cross- and multi-lingual search, NLP: summarization, text representation, linguistic analysis, readability, opinion mining and sentiment analysis, clustering, classification, topic models for search and recommendation).
  - Knowledge acquisition (e.g., information extraction, relation extraction, event extraction, query understanding, human-in-the-loop knowledge acquisition).

SIGIR 2023 vs. SIGIR 2010

- **Machine Learning and Natural Language Processing for Search and Recommendation.** Research bridging ML, NLP, and IR.
  - Core ML (e.g., deep learning for IR, embeddings, intelligent personal assistants and agents, unbiased learning).
  - Question Answering (e.g., factoid and non-factoid question answering, interactive question answering, community-based question answering, question answering systems).
  - Conversational systems (e.g., conversational search interaction, dialog systems, spoken language interfaces, intelligent chat systems).
  - Explicit semantics (e.g., semantic search, named-entities, relation and event extraction).
  - Knowledge representation and reasoning (e.g., link prediction, knowledge graph completion, query understanding, knowledge-guided query and document representation, ontology modeling).

SIGIR 2023 vs. SIGIR 2010

- **Human factors and interfaces.** Research into user-centric aspects of IR including user interfaces, behavior modeling, privacy, interactive systems.
  - Mining and Modeling Users (e.g., user and task models, click models, log analysis, behavioral analysis, modeling and simulation of information interaction, attention modeling).
  - Interactive Search (e.g., search interfaces, information access, exploratory search, search context, whole-session support, proactive search, personalized search).
  - Social search (e.g., social media search, social tagging, crowdsourcing).
  - Collaborative search (e.g., human in-the-loop, knowledge acquisition).
  - Information Security (e.g., privacy, surveillance, censorship, encryption, security).
  - User studies comparing theory to human behaviour for search and recommendation.

SIGIR 2023 vs. SIGIR 2010

- **Evaluation.** Research that focuses on the measurement and evaluation of IR systems.
  - User-centered evaluation (e.g., user experience and performance, user engagement, search task design).
  - System-centered evaluation (e.g., evaluation metrics, test collections, experimental design, evaluation pipelines, crowdsourcing).
  - Beyond Cranfield (e.g., online evaluation, task-based, session-based, multi-turn, interactive search).
  - Beyond labels (e.g., simulation, implicit signals, eye-tracking and physiological signals).
  - Beyond effectiveness (e.g., value, utility, usefulness, diversity, novelty, urgency, freshness, credibility, authority).
  - Methodology (e.g., statistical methods, reproducibility, dealing with bias, new experimental approaches, metrics for metrics)

- Beyond labels (e.g., simulation, implicit signals, eye-tracking and physiological signals).
- Beyond effectiveness (e.g., value, utility, usefulness, diversity, novelty, urgency, freshness, credibility, authority).
- Methodology (e.g., statistical methods, reproducibility, dealing with bias, new experimental approaches).

SIGIR 2023 vs. SIGIR 2010

- **Fairness, Accountability, Transparency, Ethics, and Explainability (FATE) in IR**. Research on aspects of fairness and bias in search and recommender systems.
  - Fairness, accountability, transparency (e.g. confidentiality, representativeness, discrimination and harmful bias)
  - Ethics, Economics, and Politics (e.g., studies on broader implications, norms and ethics, economic value, political impact, social good).
  - Two-sided search and recommendation scenarios (e.g. matching users and providers, marketplaces).

- **Domain-Specific Applications.** Research focusing on domain-specific IR challenges.
  - Local and Mobile Search (e.g., location-based search, mobile usage understanding, mobile result presentation, audio and touch interfaces, geographic search, location context in search).
  - Social Search (e.g., social networks in search, social media in search, blog and microblog search, forum search).
  - Search in Structured Data (e.g., XML search, graph search, ranking in databases, desktop search, email search, entity-oriented search).
  - Multimedia Search (e.g., image search, video search, speech and audio search, music search).
  - Multimedia search (e.g., image search, video search, speech and audio search, music search).
  - Education (e.g,. search for educational support, peer matching, info seeking in online courses).
  - Legal (e.g., e-discovery, patents, other applications in law).
  - Health (e.g., medical, genomics, bioinformatics, other applications in health).

- Knowledge graph applications (e.g. conversational search, semantic search, entity search, KB question answering, knowledge-guided NLP, search and recommendation).
- Other Applications and Domains (e.g., digital libraries, enterprise, expert search, news search, app search, new retrieval problems including applications of search technology for social good).

Topics

1. Introduction to Information Retrieval and Extraction
2. Boolean model
3. Vector model
4. Probabilistic model
5. Language Model
6. Neural IR Model
7. Generative IR Model
8. Retrieval Evaluation
9. Relevance Feedback and Query Expansion
10. Fundamental of Information Extraction

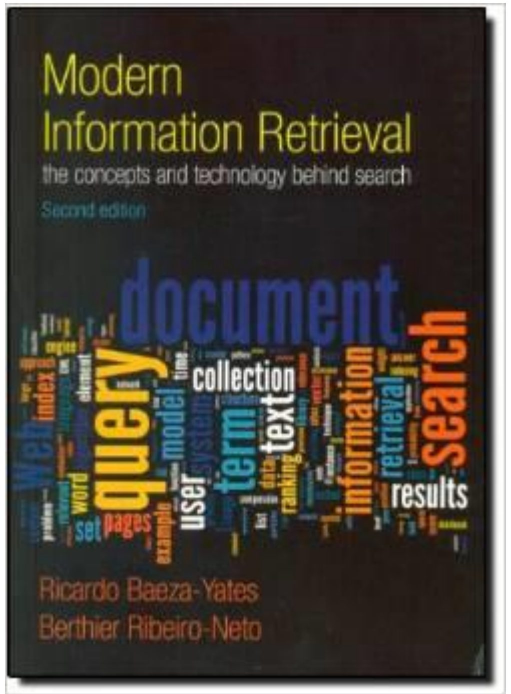
Topics (Continued)

11. Knowledge Graph for Information Retrieval
12. Information Retrieval for Knowledge Graph

Reference Materials

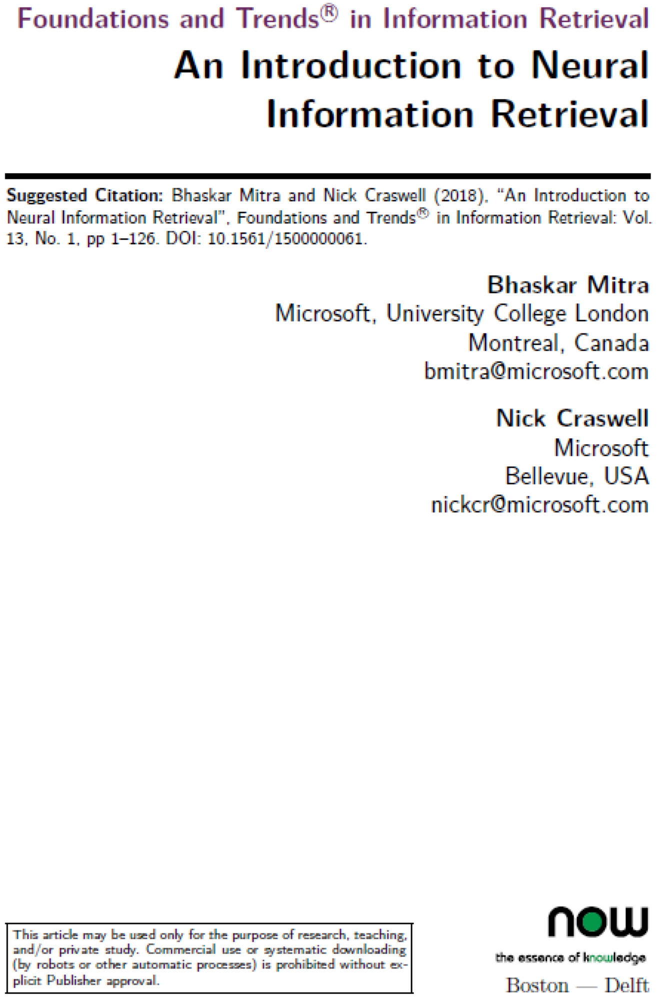
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology behind Search, Second edition, 2011.

  http://www.mir2ed.org/



Reference Materials

- An Introduction to Neural Information Retrieval



Foundations and Trends® in Information Retrieval

An Introduction to Neural Information Retrieval

Suggested Citation: Bhaskar Mitra and Nick Craswell (2018), "An Introduction to Neural Information Retrieval", Foundations and Trends® in Information Retrieval: Vol. 13, No. 1, pp 1–126. DOI: 10.1561/1500000061.

Bhaskar Mitra
Microsoft, University College London
Montreal, Canada
bmitra@microsoft.com

Nick Craswell
Microsoft
Bellevue, USA
nickcr@microsoft.com

now
the essence of knowledge
Boston — Delft

Reference Materials

- Pre-training Methods in Information Retrieval

Foundations and Trends® in Information Retrieval

Pre-training Methods in Information Retrieval

Suggested Citation: Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang and Jiafeng Guo (2022), "Pre-training Methods in Information Retrieval", Foundations and Trends® in Information Retrieval: Vol. 16, No. 3, pp 178–317. DOI: 10.1561/1500000100.

Yixing Fan
ICT, CAS
fanyixing@ict.ac.cn

Xiaohui Xie
Tsinghua University
xiexiaohui@mail.tsinghua.edu.cn

Yinqiong Cai
ICT, CAS
caiyinqiong18s@ict.ac.cn

Jia Chen
Tsinghua University
chenjia0831@gmail.com

Xinyu Ma
ICT, CAS
maxinyu17g@ict.ac.cn

Xiangsheng Li
Tsinghua University
lixsh6@gmail.com

Ruqing Zhang
ICT, CAS
zhangruqing@ict.ac.cn

Jiafeng Guo
ICT, CAS
guojiafeng@ict.ac.cn

now
the essence of knowledge
Boston — Delft

Reference Materials

- Knowledge Graphs: An Information Retrieval Perspective

Foundations and Trends® in Information Retrieval

Knowledge Graphs: An Information Retrieval Perspective

Suggested Citation: Ridho Reinanda, Edgar Meij and Maarten de Rijke (2020), "Knowledge Graphs: An Information Retrieval Perspective", Foundations and Trends® in Information Retrieval: Vol. 14, No. 4, pp 289–444. DOI: 10.1561/1500000063.

Ridho Reinanda
Bloomberg L.P.
UK
rreinanda@bloomberg.net

Edgar Meij
Bloomberg L.P.
UK
emeij@bloomberg.net

Maarten de Rijke
University of Amsterdam & Ahold Delhaize
The Netherlands
m.derijke@uva.nl

now
the essence of knowledge
Boston — Delft

Information Sources

- Conference Proceedings
  - ACM SIGIR Annual International Conference on Research and Development in Information Retrieval (SIGIR)
  - ACM International Conference on Web Search and Data Mining (WSDM)
  - The Web Conference (World Wide Web Conference (WWW))
  - ACM Conference on Information and Knowledge Management (CIKM)

Information Sources

(Continued)

  - European Conference on IR (ECIR)
  - INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA (ICWSM)
  - Text Retrieval Conference (TREC)
  - Joint ACM-IEEE Conference on Digital Libraries (JCDL)
  - European Conference on Digital Libraries (ECDL)
- Major Natural Language Processing Conferences
  - http://www.aclweb.org/anthology/

Information Sources

(Continued)

- Journals
  - ACM Transactions on Information Systems (TOIS)
  - Information Processing and Management (IP&M)
  - Journal of the American Society for Information Science and Technology (JASIST)
  - Information Systems
  - Information Retrieval
  - ACM Transactions on the Web
  - Knowledge and Information Systems (KAIS)
  - International Journal on Digital Libraries

Course Staffs

- Professor: Hsin-Hsi Chen
  Office: csie 311, E-mail: hhchen@ntu.edu.tw
- TAs: 陳建宏，潘淙軒
  Office: csie 301
- Course web site:
  https://cool.ntu.edu.tw/courses/28996

Grading

- Midterm Examination
- Final Term Examination
- 1 Term Project
- 1 Oral Presentation

加選登記