Information Retrieval

COMP90049 Knowledge Technologies

Information retrieval

Definition

History

Text collections

Information seeking

Information needs

Answers

Document matching

Boolean querying

Similarity

Principles & models

Evaluation

References

Information Retrieval

Jeremy Nicholson and Justin Zobel and Karin Verspoor, CIS

Semest

THE UNIVERSITY OF MELBOURNE

Defining “information retrieval”

2).

, file

What distinguishes IR from these other are

Data retrieval versus information retrieval

Conventional database systems, such as relational systems, are designed for data retrieval:

- Prior to storage, the data is transformed into a representation

s

⟨"Chambers", "Jill", "687651", 1

- The information is unambiguous.

- A typical information can not be repre anticipated at database-creation ti
- Queries are represented in an algebraic language.

```
select * from Student where Surname = "Chambers"
```

Data retrieval versus information retrieval

In IR systems:

- The stored documents are real-world objects that have been created for individual reasons. They do not have to have consistent
-
- Users may not agree on the value of a par relation to the same query.
- Documents are rich and ambiguous, automatic method for translating the
- Text in some kinds of collection has structured attributes, but these are only occasionally useful for searching. Examples include `<author>` tags and other metadata.

Data retrieval versus information retrieval

Thus a data retrieval system is used to retrieve items based on facts that describe them. For example:

- “Get articles from The Age dated 11/8/2017.”
- “Find articles that argue for better publi
- “Is Bosnia a good holiday destination
- “Get articles about different kinds of de

Or, more plausibly: “rural public transport”, “Bosnia holiday”, “dementia senility”.

Defining “information retrieval”

Information retrieval (IR) is “the subfield of computer science that deals with storage and retrieval of documents” (Frakes & Baeza-Yates, 1992)

This definition emphasises documents. Other fields (databases, file

- terized as mechanisms for finding documen individual.
- The meaning or content of a docu specific words used to express the me

IR systems are arguably the primary means of access to stored information in our society.

The past

The present

Search engines are a key part of the management of data such as web sites, legislation, corporate documentation, online retailers, digital libraries, and intelligence services.

In some applications – email management, personal document management – IR systems are beginning to replace file systems, and

Search engines are used to search over a wid

They are ubiquitous, with close integration
web – for example, help systems mix on-co
information.

Search is political: data access is a human rights issue.

Google handles several thousand million queries a day; when it was first successful, it was handling 10,000 queries a day. It has grown by 8% per month!

Text collections

Text collection	Size	
A single document	5 kB	0.05 MB
Complete text of Moby Dick	600 kB	0.6 MB
A researcher's papers – 10 years	10 MB	10 MB
All books in a small university library		
Govt web pages in English		
US Library of Congress, 2012		
Google, 2010	200 TB?	200,000,000 MB

Source for Library of Congress figures: `https://en.wikipedia.org/wiki/List_of_unusual_units_of_measurement#Data_volume`

PubMed: Biomedical Literature Repository

Statistical reports on MEDLINE/PubMed baseline data [Internet]. Bethesda (MD): National Library of Medicine (US), Bibliographic Services Division.

Document Collections

Typical kinds of document collection include web pages, newspaper articles, intranets, academic publications, company reports, all documents on a PC, research grant applications, parliamentary

il,

sages: an object that conveys information from one person to another.

In the context of IR, “documents” include tex

handwriting, video, and genomes.

There are practical or prototype IR systems for content-based retrieval on each of these kinds of data.

Information needs

The different kinds of IR system are linked by the concept of information need.

An IR system is used by someone because they have an information

-
- What are the best travel destinations i
- Do I want to move to Adelaide?
- Are arguments for a space program m

Many information needs cannot be described succinctly. For example, whether a travel destination is interesting depends on who is asking – some people like nightlife, other people like wildlife.

Searching

People search in a wide variety of ways. Perhaps the commonest mode is to:

- Issue an initial query.
-
-
-
- Use advanced querying features.

The purpose of many searches is to find a start

Casual users generally use only the first pag favorite search engine. Professionals use a range of search strategies and are prepared to view hundreds of potential answers. However, much the same IR techniques work for both kinds of searcher.

Searching ...

To resolve an information need using a search engine, a user chooses words and phrases that are intended to match appropriate documents, then use these words and phrases to construct a query.

If the query is unsuccessful, the user may reformulate it, thus many

different type of information need is meant in each case.

- Requests for information: “global wa
- Factoid questions: “what is the meltin
- Topic tracking: “what is the history of thi
- Navigational: “University of Melbourne”
- Service or transaction: “Macbook Air”
- Geospatial: “Carlton restaurant”

Searching ...

To resolve an information need using a search engine, a user chooses words and phrases that are intended to match appropriate documents, then use these words and phrases to construct a query.

If the query is unsuccessful, the user may reformulate it, thus many [illegible]

different type of information need is meant in each case.

- Informational: `global warming`
- Factoid: `melting point of lead`
- Topic tracking: `Trump administr`
- Navigational: `university of melbourne`
- Transactional: `Macbook Air`
- Geospatial: `carlton restaurants`

Some web queries (Excite, 2001)

action bible
texas state government

sacramento apartments
the fairmont chateau whistler
forbed global the quiet american
four models of public relations
unlock mobile phone

centeroid galleries
excalibur 1981

lam

drive pcmcia scsi
ball busting
brass instu[illegible]
algebra links
horrible news

Information retrieval in a nutshell

A simple solution, ‘grepping’

Imagine we wish to search through the texts of Project Gutenburg for **Pangolin**

- t

-
- What about handling more complex queries?
  - **Pangolin** AND **ant-eater**
  - **Pangolin** OR **ant-eater**
  - **Pangolin** NEAR **ant-eater**
  - **Pang*in**

Answers

An answer to a query could be defined as a document that matches the query according to formal criteria: if it contains all the query words, for example, then it could be described as a match.

unreliable. For example, documents often contain information such as a title or date, but not in a consistent way, and su

helpful for retrieval.

What is required is that the document should

the user is seeking.

That is, the document should be relevant.

Answers

The relevance of a document to an information need cannot be determined computationally.

- The information need is knowledge held by the user, and is not written down.
-
-

“Enron is bankrupt” is relevant, even t

Relevance can be defined as: a document is r
right topic) if it contains knowledge that help
information need.

There are many other kinds of relevance: consider searches for a particular fact, or a particular document, or a particular individual or organization.

Answers

Answers

Fundamentally, a response from a search engine is a list of documents of potential relevance.

- ______

  specific to the query .)
- Duplicates are pruned, or aggregate
- A single source might only contribute a
- Answer types may be augmented wit

Approaches to retrieval

Consider the criteria that a human might use to judge whether a document should be returned in response to a query. They would:

- Try and guess what the query might be inspired by, and what kind of information or document is being sought.
-
-

the terms.

- Be ready to consider a document even i
completely different.

That is, a human would see the query as representative of a topic, and evaluate documents accordingly.

There is no computational way of approximating this process. Instead, we have to develop methods that use other forms of evidence to make a guess as to whether a document is relevant.

Boolean querying

Until about 1994, all retrieval systems used Boolean querying (and professional searchers) to identify matches.

Documents match if they contain the terms, NOT terms.

There is no ordering; matching is yes/no.

Boolean Querying

- For the query diabetes AND risk
- Take the bit representations diabetes = **110** risk = **011**
- Perform bitwise AND, ,
-

	doc1	doc2	doc3
juvenile	1	0	0

To support:

- disjunction, simply use bitwise OR,
- negation, use bitwise complement, ˆ

diabetes AND ((NOT risk) OR juvenile)
110 AND ((NOT **011**) OR **100**) = **100**

Boolean querying

Boolean querying is still the method of choice for legal and biomedical search:

- It is repeatable, auditable, and controllable.
- Boolean queries allow expression of complex concepts.

dozens of clauses.

- The time investment in developing pr
perceived to be compensated for by re
(also months).

For general querying, Boolean querying is unsatisfactory in several respects: there is no ranking and no control over result set size, and it is difficult to incorporate useful heuristics. And it is remarkably difficult to do well.

How does ranking work?

In principle, the idea of ranked retrieval is simple. A query is matched to a document by looking for evidence in the document that it is on the same topic as the query (or the same topic as an information need that the query might represent).

- [illegible]
- What is the probability that the document is relevant to the query?
- Are the document and query o

The more similar or likely a document is, relati in the collection, the higher its rank.

For the commonest IR activity, text search, there are many kinds of evidence of similarity.

Evidence of similarity

Some matches to the query “active south american volcano”:

Expedition Chile
. . . highest mountain in Chile and also the highest active volcano in the world,

VolcanoWorld Monthly Contest
. . . October 1999. The last eruption of this South Ame
1999. This is a North American stratovolcano . . . As

Volcanic Activity On The Rise In Central America
A volcano erupted near here, and another crater . . . officials in the two Central American countries said Thursday they had no . . .

Evidence of similarity

Why might these documents have been ranked highly?

- Choose documents with words in common with the query.

or

making effective use of such statistics is a cor

n each of the four matches, the word 'volcan
certainly this is the most significant word. In a c
of web data:

word	active	south	american	volcano
occurrences	185,876	425,912	591,652	16,336

Evidence of similarity

Evidence in addition to word-match can be used to select documents.

- Choose documents with the query terms in the title.
-

- Choose documents that were created recently.
- Attempt to translate between langua
- Choose authoritative, reliable docu

Incorporating these concepts involves varying difficulty.

Heuristics in Similarity measurement

Effective similarity measures for IR combine information about queries and documents so that three observations are enforced:

- Less weight is given to a term that appears in many documents.

- ars
- Less weight is given to a document that has many terms.

The intention is to bias the score towards rele
favouring terms that seem to be discriminat
of terms that seem to be randomly distribute

A model that incorporates these ideas is known as a “TF-IDF” model.

Principles

The observation that word matching and word counts can be used to find answers provides a basis for ad-hoc development of retrieval algorithms, but such a piecemeal approach is hard to justify.

[illegible] queries are made up of terms or tokens.

(In early IR these might have been manually
web IR they could include many things in addi
content.)

A mathematical model can then be used as the basis of a similarity measure.

The vector-space model (quick recap)

Suppose there are n distinct indexed terms in the collection. Then each

d,t n d.

(Most $w_{d,t}$ values will be zero, because m

tiny proportion of a collection's terms.)

The vector-space model (quick recap)

For example:

$$\langle \text{a}, \text{aardvark}, \ldots, \text{band}, \ldots, \text{brothers}, \ldots, \text{few}, \ldots, \text{happy}, \ldots \rangle$$

$\langle 0, 0, \ldots, 1, \ldots, 1$

The vector-space model

A vector locates a document (or, equivalently in this context, a query) as a point in *n*-space.

Consequently, documents with a similar distribution of terms have similar angles in the space. Typical proble

- It isn't clear how to (best) choose the wei
- Typical formulations of the vector sp sian); there is much evidence that this is incorrect, but there are no clearly better alternatives

The vector-space model

Some typical information which might appear in a similarity calculation:

- $f_{d,t}$, the frequency of term t in document d.
- $f_{q,t}$, the frequency of term t in the query.
- f_t, the number of documents containing term t.
-
-
- tion
- $F = \quad_t F_t$, the number of occurrences in the collection.

These statistics are sufficient for computat
underlying highly effective search engine

To link back to our heuristics: we wish to find documents d that have

- Terms t with low f_t, that is, are rare;
- But t has high $f_{d,t}$, that is, is common in the document;
- And $|d|$ is low, that is, the document is short.

The cosine measure

______________________, not to formally solve the mathematical problem

The cosine measure

Many possible choices for a TF-IDF model, consistent with our heuristics!

For example,

-
-
- Length: $|r| = \overline{{}_i\, w_{r,t}^2}$

Cosine with this TF-IDF weighting model:

$$S(q,d) = \frac{{}_t\, w_{d,t} \times w_{q,t}}{|q||d|}$$

The cosine measure

Many possible choices for a TF-IDF model, consistent with our heuristics!

For example,

-
- $-$
- Length: $|r| = \sqrt{{}_i\, w_{r,t}^2}$

Cosine with this TF-IDF weighting model:

$$S(q,d) = \frac{{}_t\, w_{d,t} \times w_{q,t}}{|q||d|}$$

The cosine measure

Alternative formulation:

- –
-

$$S(q, d) = \frac{}{}$$

Cosine Example

Term–document matrix (vector space model)

	doc1	doc2	doc3
			0
			0
	0	3	1
		1	2

TF: $w_{d,t} = f_{d,t}$; IDF: $w_{q,t} = \frac{N}{f_t}$

$$S(q,d) = \frac{q \cdot d}{|q||d|}$$

$$S(q,d_1) = \frac{\langle 0, \frac{3}{2}, \frac{3}{2}, 0 \rangle \cdot \langle 2,1,0,0 \rangle}{\sqrt{0^2 + \frac{3}{2}^2 + \frac{3}{2}^2 + 0^2}\sqrt{2^2 + 1^2 + 0^2 + 0^2}}$$

$$S(q,d_1) = \frac{1.5}{(2.12)(2.24)} \approx 0.316$$

Cosine Example

Term–document matrix (vector space model)

	doc1	doc2	doc3
			0
			0
	0	3	1
		1	2

TF: $w_{d,t} = f_{d,t}$; IDF: $w_{q,t} = \frac{N}{f_t}$

$$S(q,d) = \frac{q\,d}{|q||d|}$$

$$S(q,d_2) = \frac{\langle 0,\frac{3}{2},\frac{3}{2},0\rangle \cdot \langle 0,2,3,1\rangle}{\sqrt{0^2+\frac{3}{2}^2+\frac{3}{2}^2+0^2}\sqrt{0^2+2^2+3^2+1^2}}$$

$$S(q,d_2) = \frac{7.5}{(2.12)(3.74)} \approx 0.945$$

Cosine Example

Term–document matrix (vector space model)

	doc1	doc2	doc3
			0
			0
	0	3	1
		1	2

TF: $w_{d,t} = f_{d,t}$; IDF: $w_{q,t} = \frac{N}{f_t}$

$$S(q,d) = \frac{q \cdot d}{|q||d|}$$

$$S(q,d_3) = \frac{\langle 0, \frac{3}{2}, \frac{3}{2}, 0 \rangle \cdot \langle 0,0,1,2 \rangle}{\sqrt{0^2 + \frac{3}{2}^2 + \frac{3}{2}^2 + 0^2}\sqrt{0^2+0^2+1^2+2^2}}$$

$$S(q,d_3) = \frac{1.5}{(2.12)(2.24)} \approx 0.316$$

Cosine Example

Term–document matrix (vector space model) — weighted by TF-IDF

	doc1	doc2	doc3
	0	3 1 2	2

TF-IDF: $w_{d,t} = f_{d,t} \times \frac{N}{f_t}$

$$S(q,d) = \frac{\sum_{t \in q} w_{d,t}}{|d|}$$

$$S(q,d_1) = \frac{1 \times \frac{3}{2} + 0}{\sqrt{6^2 + 1.5^2 + 0^2 + 0^2}} \approx 0.242$$

Cosine Example

Term–document matrix (vector space model) — weighted by TF-IDF

	doc1	doc2	doc3
[illegible]	[illegible]	[illegible]	[illegible]

TF-IDF: $w_{d,t} = f_{d,t} \times \frac{N}{f_t}$

$$S(q,d) = \frac{\sum_{t \in q} w_{d,t}}{|d|}$$

$$S(q,d_2) = \frac{2 \times \frac{3}{2} + 3 \times \frac{3}{2}}{\sqrt{0^2 + 3^2 + 2.25^2 + 1.5^2}} \approx 1.86$$

Cosine Example

Term–document matrix (vector space model) — weighted by TF-IDF

	doc1	doc2	doc3
	0	3	
		1	
		$\frac{}{2}$	$\frac{}{2}$

TF-IDF: $w_{d,t} = f_{d,t} \times \frac{N}{f_t}$

$$S(q,d) = \frac{\sum_{t \in q} w_{d,t}}{|d|}$$

$$S(q,d_3) = \frac{0+1\times\frac{3}{2}}{\sqrt{0^2+0^2+1.5^2+3^2}} \approx 0.447$$

Evaluation Metrics

Recall evaluation in Approximate String Search:

-
-
- ”

  (the intended word)

  - Accuracy
  - Precision
  - Recall

Evaluation Metrics

Evaluation in Information Retrieval:

-
- m the
- vant (meet the user’s information need)
  - ~~Accuracy~~
  - Precision
  - Recall

Evaluation Metrics

Some differences between evaluation in the two applications:

-
  -

- e
  -
  -
  - Accuracy isn’t meaningful
- IR results are ranked, Approx. Searc
  - Boolean querying typically more lik
  - Approx. Search could be ranked, bu

Evaluation Metrics in an IR Context

k

(Recall at *k* usually not meaningful)

Evaluation Metrics in an IR Context

–

Typically averaged over many querie

The TREC experiments

NIST established the large-scale TREC framework in 1992 to compare search engines in a systematic, unbiased way. (The twenty-fifth TREC was held last year.)

year. Most of the document collections wer

The largest current TREC collection is half a t
pages). About 100 groups participate eac

Tasks have included video and bioinformatic retrieval as well as different languages and different aspects of text retrieval (named pages, home pages, topic coverage).

The TREC experiments

- Define relevance carefully (topic search, named-page search, multi-aspect search . . .)

- es answers for each query, which are then combined into per-query pools.

- Assess the documents in each pool for rel It is reasonable (most of the time) to assume th pool are irrelevant.
- Compare the ability of engines to find these pages.

The TREC experiments

In a typical year, 1998,

- The document pools were (a) 2 gigabytes of newswire-type data, or about 0.5 million documents, and (b) 100 gigabytes of web data

-
- systems, each reporting the top 1000 documents for each query.
- The top 100 answers for each system w 3,000 documents per query or 150,00

- Humans assessed each of the 150,0 the queries, finding an average of about 70 relevant documents per query.

The TREC experiments

The appearance of effective web scale search systems would have been delayed without the evaluation framework given by a large volume

There are now several other “TRECs”, incl
TREC Legal, TREC Biomedical, INEX for X
cross-linguage information retrieval, TD
tracking, and the Japanese NTCIR for Asian languages.

Summary

- Text search is a key computational technology.
- Search is much broader than the web and is used on vastly different scales. Specific search tasks require specific tools.
- Queries are distinct from information needs; the former are the

- tion of a collection.
- Ranking involves assessment of evi of documents but in particular term sig

- There are many models for encapsul TF-IDF weighting for the vector-space model.
- Measurement of effectiveness depends on the concept of relevance, and requires large-scale assessment of queries and documents.

Background Readings

ütze (2008). "Introduction to Information Retri Cambridge University Press.