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

Assignment Project Exam Help

COMP90049
Knowledge Technologies

https://powcoder.com

Jeremy Nicholson and Justin Zobel and Karin Verspoor, CIS

Add WeChat powcoder

Semester 2, 2017

THE UNIVERSITY OF MELBOURNE

Defining “information retrieval”

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

Assignment Project Exam Help

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

https://powcoder.com

This definition emphasises documents. Other fields (databases, file structures, . . .) deal with storage and retrieval in general.

What distinguishes IR from these other areas?

Add WeChat powcoder

Data retrieval versus information retrieval

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

- Prior to storage, the data is transformed into a representation suitable for manipulation by an algebraic query language.
  For example, the information that "enrolled student Jill Chambers was born on 17 Mar 1989" might be represented in a relational database by
  ⟨"Chambers", "Jill", "687651", 1989, 3, 17⟩
- The information is unambiguous.
- Atypical information cannot be represented or queried unless it was anticipated at database-creation time.
- Queries are represented in an algebraic language.
  `select * from Student where Surname = "Chambers"`

Data retrieval versus information retrieval

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

In IR systems:

- The stored documents are real-world objects that have been created for individual reasons. They do not have to have consistent format, wording, language, length, ...
- The retrieval system is concerned with the document as originally created, not with a formal representation of the document (such as a list of keywords).
- Users may not agree on the value of a particular document, even in relation to the same query.
- Documents are rich and ambiguous, and there is no conceivable automatic method for translating them into an algebraic form.
- 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."
- "Fetch articles filed by Piotr Kulowsky in Kursograd."
- "Get the article entitled 'Alta Vista Searching for Success'."

An IR system is used to retrieve items based on their meaning.

- "Find articles that argue for better public transport in rural areas."
- "Is Bosnia a good holiday destination?"
- "Get articles about different kinds of dementia."

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Defining “information retrieval”

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 structures, . . .) deal with storage and retrieval in general.

What distinguishes IR from these other areas?

- There is an emphasis on the user. IR systems can be characterized as mechanisms for finding documents that are of value to an individual.
- The meaning or content of a document is of more interest than the specific words used to express the meaning.

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

The past

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

History of IR better left for a dedicated subject!

The present

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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 the traditional role of curator is being marginalized. Thus IR is an example of a unifying technology that is replacing a diversity of prior approaches.

Search engines are used to search over a wide range of scales of data.

They are ubiquitous, with close integration between the desktop and the web – for example, help systems mix on-computer with on-line 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	0.005 MB
Complete text of Moby Dick	600 kB	0.6 MB
A researcher's papers – 10 years	10 MB	10 MB
An individual's email – 10 years	100 MB	100 MB
All the web pages at one small university	1 GB	1,000 MB
A single-purpose digital library	20 GB	20,000 MB
All books in a small university library	100 GB	100,000 MB
Govt web pages in English	1 TB	1,000,000 MB
US Library of Congress, 2012	20 TB	20,000,000 MB
Google, 2010	200 TB?	200,000,000 MB

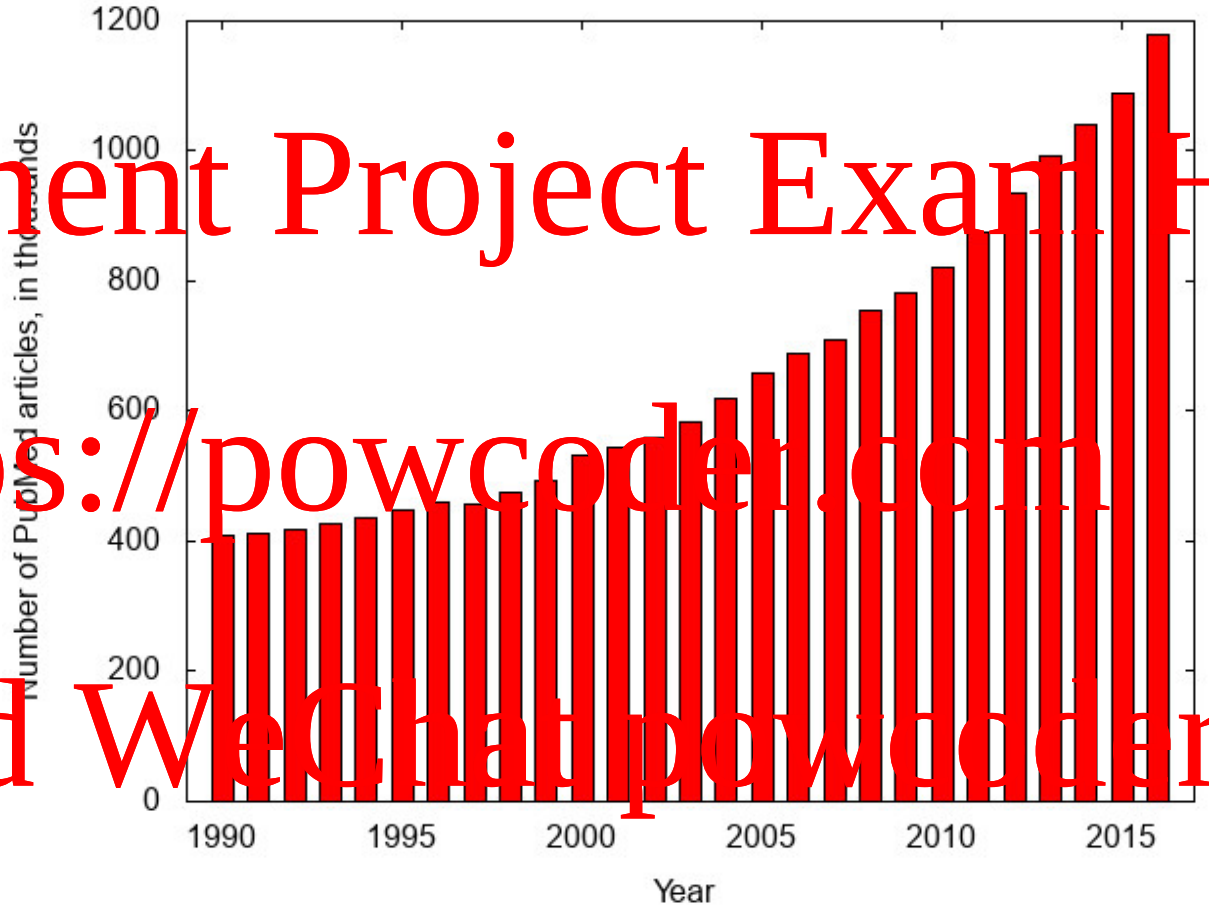
Source for Library of Congress figures: `https://en.wikipedia.org/wiki/List_of_unusual_units_of_measurement#Data_volume`

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

PubMed: Biomedical Literature Repository

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder



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

Document Collections

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Typical kinds of document collection include web pages, newspaper articles, intranets, academic publications, company reports, all documents on a PC, research grant applications, parliamentary proceedings, bibliographic entries, historical records, electronic mail, and court transcripts.

Documents aren't always text. They can be defined as messages: an object that conveys information from one person to another.

In the context of IR, "documents" include text, images, music, speech, handwriting, video, and genomes.

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

Information needs

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

Assignment Project Exam Help

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 need they wish to resolve. Information needs can be highly specific, but may be difficult to articulate or explain (to a human or a search system). For example:

https://powcoder.com

- When does the next train depart from Flinders St?
- What are the best travel destinations in Northumberland?
- Do I want to move to Adelaide?
- Are arguments for a space program mature or simplistic?

Add WeChat powcoder

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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

- Issue an initial query.
- Scan a list of suggested answers.
- Follow links to specific documents.
- Refine or modify the query.
- Use advanced querying features.

The purpose of many searches is to find a starting point for browsing.

Casual users generally use only the first page or so returned by their 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 ...

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

Assignment Project Exam Help

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 queries can represent the same information need.

https://powcoder.com

Consider the query "intel processor" under the web, news groups, images, video, shopping, and scholar tabs provided by Google. A different type of information need is meant in each case.

Add WeChat powcoder

- Requests for information: "global warming"
- Factoid questions: "what is the melting point of lead?"
- Topic tracking: "what is the history of this news story?"
- Navigational: "University of Melbourne"
- Service or transaction: "Macbook Air"
- Geospatial: "Carlton restaurant"

Searching ...

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

Assignment Project Exam Help

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 queries can represent the same information need.

https://powcoder.com

Consider the query "intel processor" under the web, news groups, images, video, shopping, and scholar tabs provided by Google. A different type of information need is meant in each case.

Add WeChat powcoder

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

Some web queries (Excite, 2001)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

action bible	centeroid galleries
texas state government	excalibur 1981
interior design institute	free url redirection
reversi othello	lamborghini dioblo
ruben hurrican cater the book	april erikson
toronto sun newspaper	cow hunter
sacramento apartments	drive pcmcia scsi
the fairmont chateau whistler	ball busting
forbed global the quiet american	brass instruments
four models of public relations	algebra links
unlock mobile phone	horrible news

Information retrieval in a nutshell

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

Assignment Project Exam Help

- User has an information need
- User formulates a query
- IR engine retrieves a set of documents

https://powcoder.com

Add WeChat powcoder

A simple solution, ‘grepping’

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

Assignment Project Exam Help

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

- Can simply use `grep` which performs a linear scan over the text searching for a match (via a regular expression)
- How will this scale to large collections?
- What about handling more complex queries?
  - **Pangolin** AND **ant-eater**
  - **Pangolin** OR **ant-eater**
  - **Pangolin** NEAR **ant-eater**
  - **Pang*in**

https://powcoder.com
Add WeChat powcoder

Answers

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

Assignment Project Exam Help

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.

But this does not mean that the document is a helpful response for that particular information need.

https://powcoder.com

Moreover, such matching criteria are likely to be simplistic and unreliable. For example, documents often contain information such as a title or date, but not in a consistent way, and such content is often not helpful for retrieval.

Add WeChat powcoder

What is required is that the document should contain information that the user is seeking.

That is, the document should be relevant.

Answers

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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.
- Identifying the topic of a document requires understanding of the text.
- The relevance may be implicit. For example, for the information need "will a US company take over BHP", a document that states "Enron is bankrupt" is relevant, even though BHP is not mentioned.

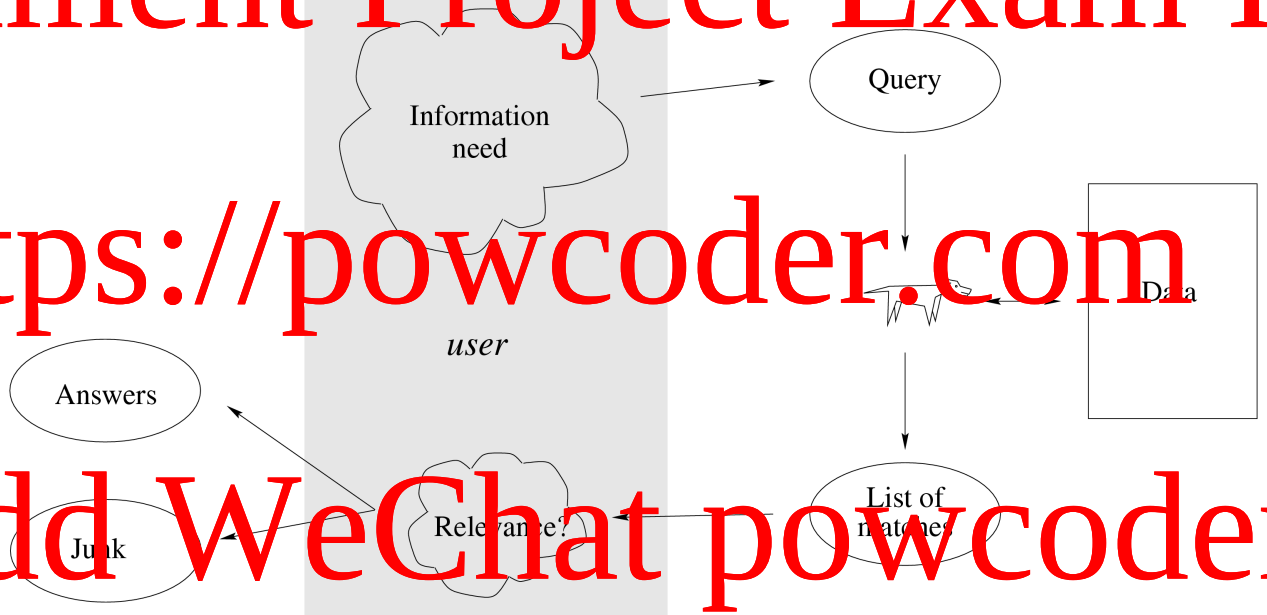
Relevance can be defined as: a document is relevant (that is, on the right topic) if it contains knowledge that helps the user to resolve the information need.

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

Answers

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder



Answers

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

Possible improvements:

- A snippet, which indicates which part(s) of the document is the basis of the answer. (This must be prepared on the fly, as it is specific to the query .)
- Duplicates are pruned, or aggregated into a single entry.
- A single source might only contribute a single answer.
- Answer types may be augmented with a map or other infobox.

Approaches to retrieval

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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.
- Consider current news or events, or cultural perspectives, or their own experience with the query terms.
- Approach the task of looking through the documents with expectations of what a match is that is based on much more than the terms.
- Be ready to consider a document even if the terminology is 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

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

Assignment Project Exam Help

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

A typical query might be

https://powcoder.com

```
diabetes & risk & factor & NOT juvenile
```

Documents match if they contain the terms, and don't contain the `NOT` terms.

Add WeChat powcoder

There is no ordering; matching is yes/no.

Boolean Querying

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- For the query `diabetes` AND `risk`
- Take the bit representations `diabetes` = **110** `risk` = **011**
- Perform bitwise AND, $\wedge$, resulting in **010**
- Therefore document 2 is the only match

	doc1	doc2	doc3
juvenile	1	0	0
diabetes	1	1	0
risk	0	1	1
factor	0	1	1

To support:

- disjunction, simply use bitwise OR, $\vee$
- negation, use bitwise complement, ˆ

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

Boolean querying

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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.

```
(randomized & controlled & trial)
or (clinical & study)
```

It is common for biomedical queries to contain hundreds of terms in dozens of clauses.

- The time investment in developing precise queries (months) is perceived to be compensated for by reduction in time spent reading (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?

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

Assignment Project Exam Help

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).

https://powcoder.com

There are several common terminologies for describing this:

- Is the query similar to the document?
- What is the probability that the document is relevant to the query?
- Are the document and query on the same topic?

Add WeChat powcoder

The more similar or likely a document is, relative to the other documents 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

Information Retrieval
COMPP90049 Knowledge Technologies
Information retrieval
Definition
History
Text collections
Information seeking
Information needs
Answers
Document matching
Boolean querying
Similarity
Principles & models
Evaluation
References

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Some matches to the query "active south american volcano":

Expedition Chile
. . . highest mountain in Chile and also the highest active volcano in the world, with active . . . We will only attempt this major South American peak . . .

Ray's Volcano Zone
. . . and Central American Volcanoes Images of South American Volcanoes Images of South . . . Images, maps, movies of Sicilian active . . .

VolcanoWorld Monthly Contest
. . . October 1999. The last eruption of this South American volcano was . . . 1999. This is a North American stratovolcano . . . Also, an active fumarole

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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Why might these documents have been ranked highly?

- Choose documents with words in common with the query.

This is obvious, but some words are more significant than others. The query "volcano" might well find relevant documents by itself, but the query "south" is highly unlikely to do so.

Significance can be estimated statistically. Investigation of methods for making effective use of such statistics is a core research activity in IR.

In each of the four matches, the word "volcano" is prominent – almost certainly this is the most significant word. In a collection of 45 gigabytes of web data:

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

Evidence of similarity

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

Assignment Project Exam Help

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

- Choose documents with the query terms in the title.
- Look for occurrences of the query terms as phrases.
  For example, the first match contains "active volcano" and "south america".
- Choose documents that were created recently.
- Attempt to translate between languages.
- Choose authoritative, reliable documents.

https://powcoder.com
Add WeChat powcoder

Incorporating these concepts involves varying difficulty.

Heuristics in Similarity measurement

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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. (Inverse document frequency or IDF.)
- More weight is given to a document where a query term appears many times. (Term frequency or TF.)
- Less weight is given to a document that has many terms.

The intention is to bias the score towards relevant documents by favouring terms that seem to be discriminatory, and reducing the impact of terms that seem to be randomly distributed.

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

Principles

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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.

Models are used throughout science to unify observations, make predictions, and provide direction. IR is no exception.

The basis of the effective IR models in use today is that documents and queries are made up of terms or tokens.

(In early IR these might have been manually assigned index terms. In web IR they could include many things in addition to full document content.)

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

The vector-space model (quick recap)

Information Retrieval
COMPP90049 Knowledge Technologies
Information retrieval
Definition
History
Text collections
Information seeking
Information needs
Answers
Document matching
Boolean querying
Similarity
Principles & models
Evaluation
References

Assignment Project Exam Help

Suppose there are n distinct indexed terms in the collection. Then each document d can be thought of as a vector

https://powcoder.com

$$\langle w_{d,1}, w_{d,2}, \ldots, w_{d,t}, \ldots, w_{d,n} \rangle$$

where $w_{d,t}$ is a weight describing the importance of term t in d.

(Most $w_{d,t}$ values will be zero, because most documents only contain a tiny proportion of a collection's terms.)

Add WeChat powcoder

The vector-space model (quick recap)

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

For example:

We few, we happy few, we band of brothers

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

$$\langle 0, 0, \ldots, 1, \ldots, 1, \ldots, 2, \ldots, 1, \ldots \rangle$$

The vector-space model

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

Documents with similar terms have points that are "nearby" in the space. In estimating topical similarity, the length of the vector is relatively unimportant.

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

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

The vector-space model

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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.
- N, the number of documents in the collection.
- n, the number of indexed terms in the collection.
- $F_t = \sum_d f_{d,t}$, the number of occurrences of t in the collection.
- $F = \sum_t F_t$, the number of occurrences in the collection.

These statistics are sufficient for computation of the similarity functions underlying highly effective search engines.

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

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

Assignment Project Exam Help

A typical (old) strategy is to find the cosine of the angle between two vectors: one defined by the document and one defined by the query.

https://powcoder.com

Remember: our goal is to find the most relevant documents, not to formally solve the mathematical problems!

Add WeChat powcoder

The cosine measure

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

For example,

- TF: $w_{d,t} = f_{d,t}$
- IDF: $w_{q,t} = \frac{N}{f_t}$, if $f_{q,t} > 0$, otherwise $w_{q,t} = 0$
- Length: $|r| = \sqrt{\sum_i w_{r,t}^2}$

Cosine with this TF-IDF weighting model:

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

The cosine measure

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

For example,

- TF: $w_{d,t} = 1 + \log_2 f_{d,t}$ if $f_{d,t} > 0$, otherwise $w_{d,t} = 0$
- IDF: $w_{q,t} = \log_2 1 + \frac{N}{f_t}$ if $f_{q,t} > 0$, otherwise $w_{q,t} = 0$
- Length: $|r| = \sqrt{\sum_i w_{r,t}^2}$

Cosine with this TF-IDF weighting model:

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

The cosine measure

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Alternative formulation:

- TF $\times$ IDF: $w_{d,t} = f_{d,t} \times \frac{N}{f_t}$
- Query is binary: $w_{q,t} \in \{0, 1\}$

"Cosine" with this TF-IDF weighting model:

$$s(q, d) = \frac{\sum_{t \in q} \text{TF-IDF}_{d,t}}{|d|}$$

Cosine Example

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Term–document matrix (vector space model)

	doc1	doc2	doc3
juvenile	2	0	0
diabetes	1	2	0
risk	0	3	1
factor	0	1	2

Query: `diabetes risk`

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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Term–document matrix (vector space model)

	doc1	doc2	doc3
juvenile	2	0	0
diabetes	1	2	0
risk	0	3	1
factor	0	1	2

Query: `diabetes risk`

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

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

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

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

Cosine Example

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Term–document matrix (vector space model)

	doc1	doc2	doc3
juvenile	2	0	0
diabetes	1	2	0
risk	0	3	1
factor	0	1	2

Query: `diabetes risk`

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

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

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

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

Cosine Example

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

	doc1	doc2	doc3
juvenile	$2 \times \frac{3}{1}$	0	0
diabetes	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$	0
risk	0	$3 \times \frac{3}{2}$	$1 \times \frac{3}{2}$
factor	0	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$

Query: `diabetes risk`

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

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

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

Cosine Example

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

	doc1	doc2	doc3
juvenile	$2 \times \frac{3}{1}$	0	0
diabetes	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$	0
risk	0	$3 \times \frac{3}{2}$	$1 \times \frac{3}{2}$
factor	0	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$

Query: `diabetes risk`

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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

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

	doc1	doc2	doc3
juvenile	$2 \times \frac{3}{1}$	0	0
diabetes	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$	0
risk	0	$3 \times \frac{3}{2}$	$1 \times \frac{3}{2}$
factor	0	$1 \times \frac{3}{2}$	$2 \times \frac{3}{2}$

Query: `diabetes risk`

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

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

Assignment Project Exam Help

Recall evaluation in Approximate String Search:

- We have one (or more) probably misspelled token(s) of interest
- Our system returns one (or more) item(s) from the dictionary
- We examine whether the returned dictionary item(s) are "correct" (the intended word)
  - Accuracy
  - Precision
  - Recall

https://powcoder.com
Add WeChat powcoder

Evaluation Metrics

Information Retrieval
COMPP90049 Knowledge Technologies
Information retrieval
Definition
History
Text collections
Information seeking
Information needs
Answers
Document matching
Boolean querying
Similarity
Principles & models
Evaluation
References

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Evaluation in Information Retrieval:

- We have one (or more) queries
- Our system returns one (or more) **many** documents from the collection
- We examine whether the returned documents are relevant (meet the user's information need)
  - Accuracy
  - Precision
  - Recall

Evaluation Metrics

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Some differences between evaluation in the two applications:

- Typically many more results in IR than Approx. Search
  - The collection is larger
- IR has multiple "correct" relevant results; Approx. Search only one
  - The collection is larger, and redundant
  - User's need can potentially be met in many different ways
  - Accuracy isn't meaningful
- IR results are ranked, Approx. Search typically not
  - Boolean querying typically more like Approx. Search evaluation
  - Approx. Search could be ranked, but typically many ties

Evaluation Metrics in an IR Context

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Precision: $\frac{\text{number of returned relevant results}}{\text{number of returned results}}$

Recall: $\frac{\text{number of returned relevant results}}{\text{total number of relevant results}}$ (often useless in an IR context)

Precision at k ($P@k$): $\frac{\text{number of returned relevant results in top } k}{k}$

(Recall at k usually not meaningful)

Evaluation Metrics in an IR Context

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

Assignment Project Exam Help

Average Precision: $\frac{1}{R}\sum_{k|d(k)\text{is relevant}} P@k$

where R is the total number of relevant documents for the query (denominator of Recall)

https://powcoder.com

Typically averaged over many queries: MAP (Mean Average Precision)

Add WeChat powcoder

The TREC experiments

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

Assignment Project Exam Help

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.)

The first year of TREC used two gigabytes of newswire – a huge volume of data for its day. (Two gigabytes of disk might have cost around $20,000.)

https://powcoder.com

Throughout the 1990s, an additional 50 queries were evaluated each year. Most of the document collections were re-used over several years.

Add WeChat powcoder

The largest current TREC collection is half a terabyte (25,000,000 web pages). About 100 groups participate each year.

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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- Define relevance carefully (topic search, named-page search, multi-aspect search . . .)
- Identify a set of systems that are to be compared.
- Given a set of queries, use pooling to find a set of interesting pages from a collection. In pooling, each system returns its top k answers for each query, which are then combined into per-query pools.
- Assess the documents in each pool for relevance – if the pool is large, it is reasonable (most of the time) to assume that documents outside the pool are irrelevant.
- Compare the ability of engines to find these pages.

The TREC experiments

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- 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 (massive at the time), or about 7 million documents.
  - On the newswire data there was 50 queries.
  - On the newswire data, about 30 groups participated with 61 systems, each reporting the top 1000 documents for each query.
  - The top 100 answers for each system were pooled, giving about 3,000 documents per query or 150,000 documents overall.
  - Humans assessed each of the 150,000 documents for relevance to the queries, finding an average of about 70 relevant documents per query.

The TREC experiments

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

Assignment Project Exam Help

The appearance of effective web scale search systems would have been delayed without the evaluation framework given by a large volume of shared and robust test data, and by the opportunity it provided to share knowledge about system implementation.

https://powcoder.com

In a typical year around 100 groups participate with hundreds of systems, each exploring new avenues towards improving retrieval.

There are now several other “TRECs”, including TRECVID for video, TREC Legal, TREC Biomedical, INEX for XML documents, CLEF for cross-language information retrieval, TDT for topic detection and tracking, and the Japanese NTCIR for Asian languages.

Add WeChat powcoder

Summary

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- 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 written approximation of the latter. Search is one component, but not the only one, of the task of resolving an information need.
- Search can be Boolean or ranked. Boolean search is only appropriate for heavyweight applications such as deep exploration of a collection.
- Ranking involves assessment of evidence, including many features of documents but in particular term significance.
- There are many models for encapsulating evidence, including the 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

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

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Zobel, Justin and Alistair Moffatt (2006). “Inverted Files for Text Search Engines”. ACM Computing Surveys 38 (2): 1–56. doi:10.1145/1132956.1132959

Manning, Christopher D., Prabhakar Raghavan, Heinrich Schütze (2008). “Introduction to Information Retrieval”. Chapters 1, 6. Cambridge University Press.