Information Retrieval

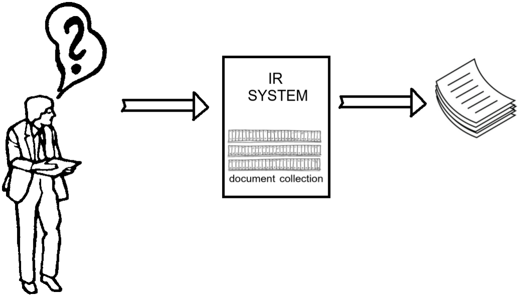
Behind the Scene of Web Search

Parantapa Goswami

Viseo R&D
Grenoble, France
parantapa.goswami@viseo.com

November 24, 2017

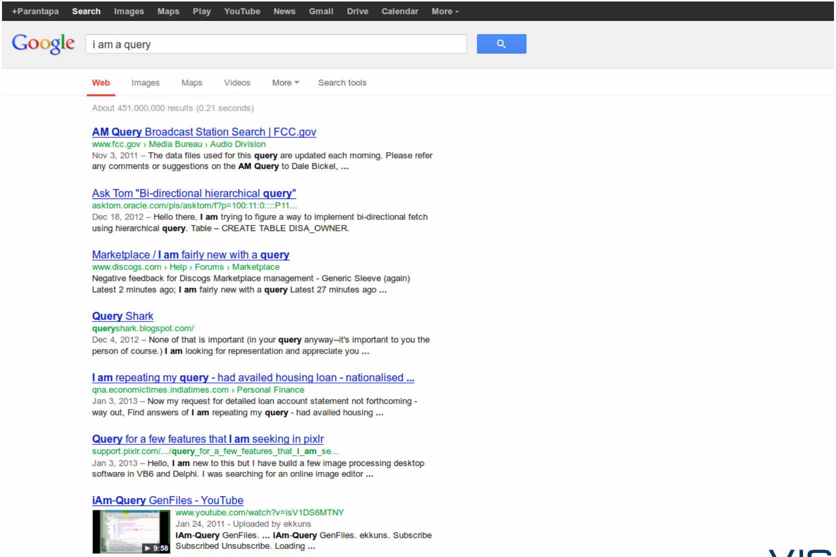
What is Information Retrieval?



A Definition

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

An Example



i am a query

About 451,000,000 results (0.21 seconds)

AM Query Broadcast Station Search | FCC.gov
www.fcc.gov › Media Bureau › Audio Division
Nov 3, 2011 – The data files used for this **query** are updated each morning. Please refer any comments or suggestions on the **AM Query** to Dale Bickel, ...

Ask Tom "Bi-directional hierarchical **query**"
asktom.oracle.com/pls/asktom/f?p=100:11:0::::P11...
Dec 18, 2012 – Hello there, **I am** trying to figure a way to implement bi-directional fetch using hierarchical **query**. Table – CREATE TABLE DISA_OWNER.

Marketplace / **I am** fairly new with a **query**
www.discogs.com › Help › Forums › Marketplace
Negative feedback for Discogs Marketplace management - Generic Sleeve (again) Latest 2 minutes ago; **I am** fairly new with a **query** Latest 27 minutes ago ...

Query Shark
queryshark.blogspot.com/
Dec 4, 2012 – None of that is important (in your **query** anyway--it's important to you the person of course.) **I am** looking for representation and appreciate you ...

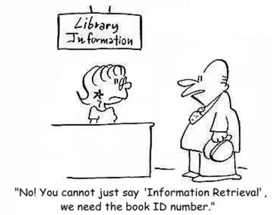
I am repeating my **query** - had availed housing loan - nationalised ...
qna.economictimes.indiatimes.com › Personal Finance
Jan 3, 2013 – Now my request for detailed loan account statement not forthcoming - way out, Find answers of **I am** repeating my **query** - had availed housing ...

Query for a few features that **I am** seeking in pixlr
support.pixlr.com/.../**query**_for_a_few_features_that_**i**_**am**_se...
Jan 3, 2013 – Hello, **I am** new to this but I have build a few image processing desktop software in VB6 and Delphi. I was searching for an online image editor ...

iAm-Query GenFiles - YouTube
www.youtube.com/watch?v=isV1DS6MTNY
Jan 24, 2011 - Uploaded by ekkuns
IAm-Query GenFiles. ... **IAm-Query** GenFiles. ekkuns. Subscribe Subscribed Unsubscribe. Loading ...

Databases vs IR: getting a book from library

Structured Queries

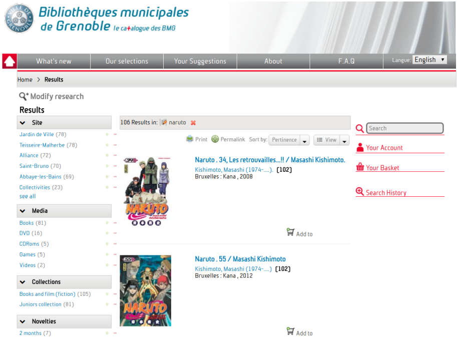


"No! You cannot just say 'Information Retrieval', we need the book ID number."

Databases are fully structured, and require structured queries for retrieval.

```
SELECT Title, Authors
FROM books
WHERE BookID="978-2-505"
```

Searching by Words



Results

Site
- Jardin de Ville (78)
- Teisseire-Malherbe (78)
- Alliance (72)
- Saint-Bruno (70)
- Abbaye-les-Bains (69)
- Collectivities (23)
- see all

Media
- Books (81)
- DVD (16)
- CDRoms (5)
- Games (5)
- Videos (2)

Collections
- Books and film (fiction) (105)
- Juniors collection (81)

Novelties
- 2 months (7)

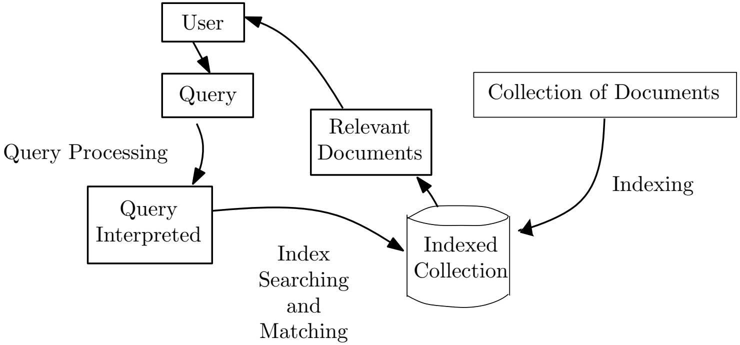
106 Results in: naruto

Naruto . 34, Les retrouvailles...!! / Masashi Kishimoto.
Kishimoto, Masashi (1974-....). [102]
Bruxelles : Kana , 2008

Naruto . 55 / Masashi Kishimoto
Kishimoto, Masashi (1974-....) [102]
Bruxelles : Kana , 2012

IR systems require words to find results containing those words.

IR Procedure



User

Query

Query Processing

Query Interpreted

Index Searching and Matching

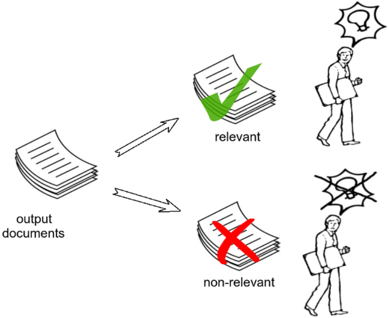
Indexed Collection

Relevant Documents

Collection of Documents

Indexing

How Good is a System: Relevant and Non-relevant



All that an IR system returns may not be relevant and required.

Precision

$$Precision = \frac{\text{number of relevant documents retrieved}}{\text{number of all documents retrieved}}$$

Ranked Retrieval

Non-ranked Retrieval

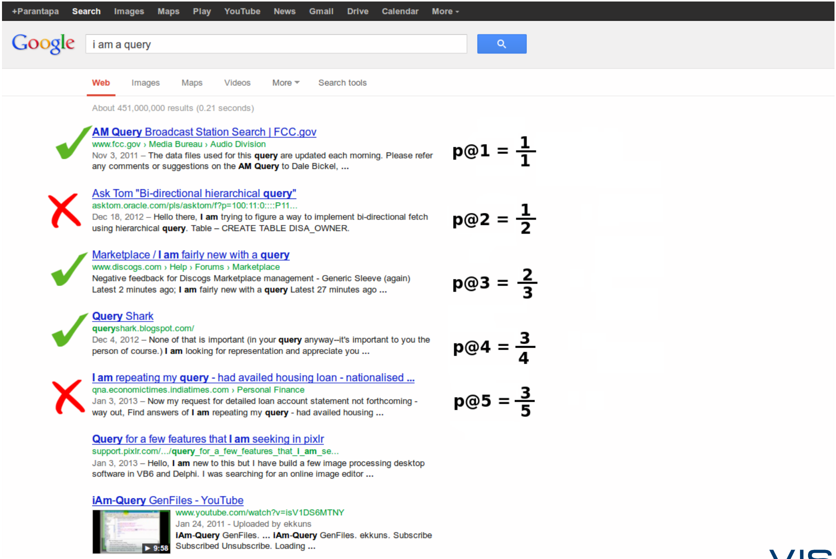
The IR system returns a bunch of documents, without any ordering.

Ranked Retrieval

- The IR system returns a ranked document set.
- Documents are ranked based on score.
- Scores are given by different *scoring models*.

VISEO

Evaluation of Ranked Retrieval Sets



About 451,000,000 results (0.21 seconds)

✓ **AM Query** Broadcast Station Search | FCC.gov
www.fcc.gov › Media Bureau › Audio Division
Nov 3, 2011 – The data files used for this **query** are updated each morning. Please refer any comments or suggestions on the **AM Query** to Dale Bickel, ...

$$p@1 = \frac{1}{1}$$

✗ Ask Tom "Bi-directional hierarchical **query**"
asktom.oracle.com/pls/asktom/f?p=100:11:0::::P11...
Dec 18, 2012 – Hello there, **I am** trying to figure a way to implement bi-directional fetch using hierarchical **query**. Table – CREATE TABLE DISA_OWNER.

$$p@2 = \frac{1}{2}$$

✓ Marketplace / **I am** fairly new with a **query**
www.discogs.com › Help › Forums › Marketplace
Negative feedback for Discogs Marketplace management - Generic Sleeve (again) Latest 2 minutes ago; **I am** fairly new with a **query** Latest 27 minutes ago ...

$$p@3 = \frac{2}{3}$$

✓ **Query** Shark
queryshark.blogspot.com/
Dec 4, 2012 – None of that is important (in your **query** anyway--it's important to you the person of course.) **I am** looking for representation and appreciate you ...

$$p@4 = \frac{3}{4}$$

✗ **I am** repeating my **query** - had availed housing loan - nationalised ...
qna.economictimes.indiatimes.com › Personal Finance
Jan 3, 2013 – Now my request for detailed loan account statement not forthcoming - way out, Find answers of **I am** repeating my **query** - had availed housing ...

$$p@5 = \frac{3}{5}$$

Query for a few features that **I am** seeking in pixlr
support.pixlr.com/.../**query**_for_a_few_features_that_**I**_**am**_se...
Jan 3, 2013 – Hello, **I am** new to this but I have build a few image processing desktop software in VB6 and Delphi. I was searching for an online image editor ...

i**Am-Query** GenFiles - YouTube
www.youtube.com/watch?v=isV1DS6MTNY
Jan 24, 2011 - Uploaded by ekkuns
IAm-Query GenFiles. ... **IAm-Query** GenFiles. ekkuns. Subscribe Subscribed Unsubscribe. Loading ...

Evaluation of Ranked Retrieval Sets

Average Precision

$$AP_{q_j} = \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk})$$

- q_j → given query
- $D = \{d_1, \ldots d_{m_j}\}$ → set of retrieved documents
- R_{jk} → is the set of ranked retrieval results from the top result until you get to document d_k

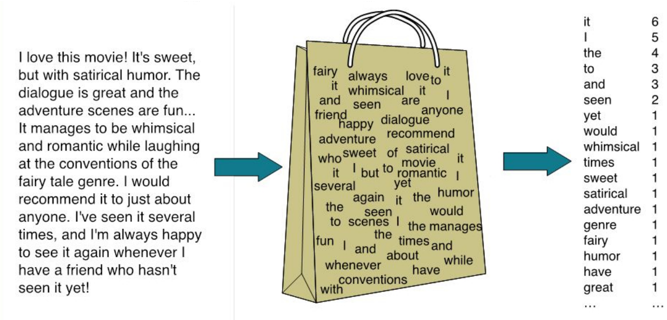
Mean Average Precision (MAP)

Let the query set is Q.

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} AP_{q_j}$$

Representing Text: Bag of Words

- A simplifying representation used in natural text processing
- A text (such as a sentence or a document) is represented as the bag of its words:
  - disregards: grammar, word ordering
  - maintains: multiplicity



I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Text to Matrix: Count Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
anthony	157	73	0	0	0	1
brutus	4	157	0	2	0	0
caesar	232	227	0	2	1	0
calpurnia	0	10	0	0	0	0
cleopatra	57	0	0	0	0	0
mercy	2	0	3	8	5	8
worser	2	0	1	1	1	5
...						

Each document is now represented as a count vector $\in \mathbb{N}^{|V|}$.

Vocabulary

Vocabulary is the set of unique words/terms present in the collection.
For example: $\{$anthony, brutus, caesar...$\}$

Text to Matrix: Count Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
anthony	157	73	0	0	0	1
brutus	4	157	0	2	0	0
caesar	232	227	0	2	1	0
calpurnia	0	10	0	0	0	0
cleopatra	57	0	0	0	0	0
mercy	2	0	3	8	5	8
worser	2	0	1	1	1	5
...						

Each document is now represented as a count vector $\in \mathbb{N}^{|V|}$.

Vocabulary

Vocabulary is the set of unique words/terms present in the collection.
For example: {anthony, brutus, caesar...}

Term Frequency

- The term frequency $\text{tf}_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.
- A document with $\text{tf} = 10$ occurrences of a term is more relevant than a document with $\text{tf} = 1$ occurrence of the term.
  - Relevance does not increase proportionally with term frequency.
- The frequency of the term in the collection...
  - Some terms are rare (e.g. `EPISTEMOPHOBIA`).
  - Some terms are frequent (e.g. `INFORMATION`)
  - Rare terms are more informative than frequent terms.
  - Rare terms should have higher weights than frequent terms.

Document Frequency

- We want:
  - higher weights for rare terms like `EPISTEMOPHOBIA`.
  - lower (but positive) weights for frequent terms like `INFORMATION`.
- The document frequency df_t of a term t is the number of documents in the collection that t occurs in.
- Rare terms will have lower df than frequent terms.
- df_t is an inverse measure of the informativeness of term t.

Inverse Document Frequency

- Inverse document frequency idf_t of term t is:

  $$Idf_t = \log \frac{N}{df_t}$$

  (N is the number of documents in the collection.)
- idf_t is a measure of the informativeness of the term.
- In the query "epistemophobia information", idf weighting increases the relative weight of `EPISTEMOPHOBIA` and decreases the relative weight of `INFORMATION`.

TF-IDF Scoring

- The tf-idf weight of a document d w.r.t a query term t is the product of its tf$_{t,d}$ and idf$_{d}$.

$$score_{tf-idf}(t,d) = (tf_{t,d}).\left(\log\frac{N}{df_t}\right)$$

- The tf-idf score:
  - increases with the number of occurrences of t within d. (term frequency)
  - increases with the rarity of the term in the collection. (inverse document frequency)
- One of the classical weighting scheme in information retrieval.

Weight Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
anthony	157	73	0	0	0	1
brutus	4	157	0	2	0	0
caesar	232	227	0	2	1	0
calpurnia	0	10	0	0	0	0
cleopatra	57	0	0	0	0	0
mercy	2	0	3	8	5	8
worser	2	0	1	1	1	5
...						

Weight Matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
anthony	5.25	3.18	0.0	0.0	0.0	0.35
brutus	1.21	6.10	0.0	1.0	0.0	0.0
caesar	8.59	2.54	0.0	1.51	0.25	0.0
calpurnia	0.0	1.54	0.0	0.0	0.0	0.0
cleopatra	2.85	0.0	0.0	0.0	0.0	0.0
mercy	1.51	0.0	1.90	0.12	5.25	0.88
worser	1.37	0.0	0.11	4.15	0.25	1.95
...						

Each document is now represented as a real-valued vector of tf-idf weights $\in \mathbb{R}^{|V|}$.

$V \implies$ the vocabulary of the collection under consideration

Vector Space Model

- The Vector Space Model is an algebraic model for representing text documents as vectors
- In this model:
  - Each document d of a collection $\mathcal{C}$ is represented with a vector $\vec{d} \in \mathbb{R}^{|V|}$
  - Dimension of the vector $\vec{d}$ is the size of the vocabulary $|V|$
  - Vector space constructed is a space of terms in which each dimension is associated with a term of the collection
- To score documents w.r.t queries:
  1. Each query q is similarly represented by a vector $\vec{q} \in \mathbb{R}^{|V|}$ of same dimension
  2. Rank documents according to their similarity to the query.
- Cosine similarity is used to measure similarity between query and document.

$$score_{vspace}(q, d) = \cos(\vec{q}, \vec{d}) = \frac{\vec{q}.\vec{d}}{|\vec{q}|.|\vec{d}|}$$

BM25

Key Idea

A term appearing in a small distinct number of documents with high frequency is far more important than the terms appearing in a large number of documents with very low frequency.

Elite Terms for a Document

Terms having a high term frequency in the document in question but low document frequency.

To score a document d for a word t in query q:

- the importance of term t in document d is proportional to the term frequency of t in d
- the importance of term t is inversely proportional to the document frequency of t
- the score is weighted by the number of occurrences of the term t inside the query q itself

BM25

Key Idea

A term appearing in a small distinct number of documents with high frequency is far more important than the terms appearing in a large number of documents with very low frequency.

Elite Terms for a Document

Terms having a high term frequency in the document in question but low document frequency.

BM25 equation

$$score_{BM25}(q, d) = \sum_{t \in q \cap d} \underbrace{\frac{(k_1 + 1) \times tf_t^d}{k_1 \left((1 - b) + b \frac{l_d}{l_{avg}} \right) + tf_t^d}}_{\text{term frequency}} \underbrace{ln \left(\frac{N - df_t + 0.5}{df_t + 0.5} \right)}_{\text{document frequency}} \underbrace{\frac{(k_3 + 1) \times tf_t^q}{k_3 + tf_t^q}}_{\text{key frequency}}$$

VISEO

Probabilistic Approach

- Given a query and a collection of documents, a system must determine how well the documents satisfy the query.
  - An IR system makes an uncertain guess of whether a document satisfies the query.
- Probability theory provides a principled foundation for such reasoning under uncertainty.
  - Probabilistic models exploit this foundation to estimate how likely it is that a document is relevant to a query.

Probabilistic Approach

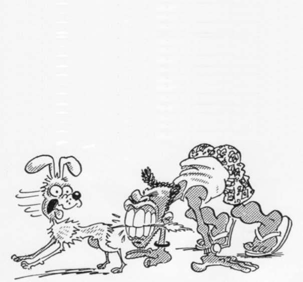
- Assume binary notion of relevance: $R_{d,q}$ is a random variable, such that
  - $R_{d,q} = 1$ if document d is relevant w.r.t query q
  - $R_{d,q} = 0$ otherwise
- Probabilistic ranking orders documents decreasingly by their estimated probability of relevance w.r.t. query: $P(R = 1|q, d)$
- Assume that the relevance of each document is independent of the relevance of other documents.
- The score is:

$$score_{prob}(q, d) = \log \prod_{t \in q} P(R = 1|q, d) = \sum_{t \in q} \log P(R = 1|q, d)$$

VISEO

Information-Based Models

A snake has bitten a man.



A man has bitten a snake.

Surprisal of a message

Measure of information content called the self-information or *surprisal* of a message m:

$$I(m) = -\log[P(m)]$$

VISEO

Information-Based Model

The more a word deviates in a document from its average behavior in the collection, the more likely it is *significant* for this particular document.

word behaves in a document as expected in collection $\implies$ word has high probability (p) of occurrence in the document $\implies$ information it brings to the document ($-\log p$) is small

Similarly, if a word has a low probability of occurrence in the document $\implies$ high information content.

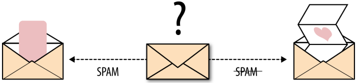
Information Model Equation

$$score_{info}(q,d) = \sum_{t \in q \cap d} tf_t^q \{-\log P(X_t \geq tf_t^d | \lambda_t)\}$$

VISEO

Using Machine Learning

- Document classification: a mail $\vec{d} \rightarrow \{spam, good\}$



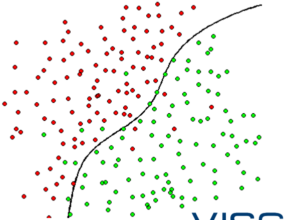
  - $\vec{d}$: count vector or tf.idf vector
- Can we do same for retrieval? $\vec{d} \rightarrow \{relevant, not\}$
  - assume we have labels for learning, i.e. if $\vec{d}$ is relevant or not

- Wrong feature set:
  - word or feature "obama"
    - highly relevant for query "44th US president"
    - useless for query "area of a parallelogram"
  - can classify for one particular query
  - solution? separate classifier for every query?

Learning to Rank

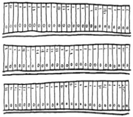
- Utilize the query, along with documents, as feature: a consistent feature set across all the queries
- Transform feature set: $f(q \times d) = \vec{X} \rightarrow \{relevant, not\}$
  - example features that have the same effect for any query:
    - BM25 score of query q against document d
    - tf.idf score of query q against document d
    - various count statistics of query words within documents: e.g. number of query words in the title
    - and more...
  - Refer to LETOR features
- Training data: $\{R, X\}$
  - $R = \{-1, +1\}$: is d is relevant to q
  - X: is (d, q) pairs
- Learn any standard classifier
- Three types of learning to rank:

  pointwise, pairwise, listwise



Experimental Setup

Test Collection



set of documents

set of queries

relevancy judgement or 'qrels'

Testing Platform

- Different modules (e.g. indexing, query processing matching) are implemented.
- Some popular scoring models are implemented.
- Anyone can customize the platform at code level to implement any new developments.
- Mostly open source and free.

Some Information...

Popular IR Platforms

- **Terrier IR Platform.** Written in Java, and is developed at the School of Computing Science, University of Glasgow.
  Ref: `http://terrier.org/`.
- **Indri (Lemur).** Written C++, and is developed at the Center for Intelligent Information Retrieval, University of Massachusetts and the Language Technologies Institute, Carnegie Mellon University.
  Ref: `https://www.lemurproject.org/indri/`.

Popular Test Collections

- **Text REtrieval Conference (TREC).**
  Ref: `http://trec.nist.gov/`.
- **Cross-Language Evaluation Forum (CLEF)**.
  Ref: `http://www.clef-initiative.eu/`

It's the beginning

Further Readings:

- Introduction to Information Retrieval. *Manning, Raghavan, Schütz*
  - Freely available: `https://nlp.stanford.edu/IR-book/`
- Recherche d'information: Applications, modles et algorithmes - Data mining, dcisionnel et big data. *Massih-R Amini, Eric Gaussier*

Thank you!

Questions, Comments...