INLS 509: Introduction to Information Retrieval

Jaime Arguello
jarguell@email.unc.edu

January 11, 2016
Outline

Introductions

What is information retrieval (IR)?

What is a search engine?

Why is information retrieval difficult?

How do search engines predict relevance?

How good is a search engine?
Introductions

- Hello, my name is ______.
- However, I’d rather be called ______. (optional)
- I’m in the ______ program.
- I’m taking this course because I want to ______.
What is Information Retrieval?

- Information retrieval (IR) is the science and practice of developing and evaluating systems that match information seekers with the information they seek.
What is Information Retrieval?

• This course mainly focuses on search engines
• Given a query and a corpus, find relevant items
  
  query: a user’s expression of their information need
  corpus: a repository of retrievable items
  relevance: satisfaction of the user’s information need
What is Information Retrieval?

- Gerard Salton, 1968:

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

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the SMART Information Retrieval System, which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in Nuremberg, Germany. He received a Bachelor’s (1950) and Master’s (1952) degree in mathematics from Brooklyn College, and a Ph.D. from Harvard in Applied Mathematics in 1958, the last of Howard Aiken's doctoral students, and taught there until 1965, when he joined Cornell University and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used Vector Space Model for Information Retrieval. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by Karen Sparck-Jones.) Later in life, he became interested in automatic text summarization and analysis, as well as automatic hypertext generation.

Salton was editor-in-chief of the Communications of the ACM and the Journal of the ACM, and chaired SIGIR. He was an associate editor of the ACM Transactions on Information Systems. He was an ACM Fellow (elected 1995), received an Award of Merit from the American Society for Information Science (1989), and was the first recipient of the SIGIR Award for outstanding contributions to study of information retrieval (1983) -- now called the Gerard Salton Award.

References

Gerard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the SMART Information Retrieval System, which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlimann on March 8, 1927 in Nuremberg, Germany. He received a Bachelor's (1950) and Master's (1952) degree in mathematics from Brooklyn College, and a Ph.D. from Harvard in Applied Mathematics in 1958, the last of Howard Aiken's doctoral students, and taught there until 1965, when he joined Cornell University and co-founded its department of Computer Science.

Salton was perhaps most well known for developing the now widely used Vector Space Model for Information Retrieval[1]. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by Karen Sparck-Jones[2].) Later in life, he became interested in automatic text summarization and analysis[3], as well as automatic hypertext generation[4]. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the Communications of the ACM and the Journal of the ACM, and chaired SIGIR. He was an associate editor of the ACM Transactions on Information Systems. He was an ACM Fellow (elected 1995), received an Award of Merit from the American Society for Information Science (1989), and was the first recipient of the SIGIR Award for outstanding contributions to study of information retrieval (1983) -- now called the Gerard Salton Award.

References

However, the main content of the page is in the form of natural language text, which has little structure that a computer can understand.
However, the main content of the page is in the form of natural language text, which has little structure that a computer can understand. As it turns out, it’s not necessary for a computer to understand natural language text for it to determine that this document is likely to be relevant to a particular query (e.g., “Gerard Salton”).
Information Retrieval

collection structure
Information Retrieval
analysis: classification and information extraction
Information Retrieval
organization: cataloguing

http://www.dmoz.org
Information Retrieval
analysis and organization: reading-level

blues clues
About 12,800 results (0.18 seconds)

Reading level → Advanced
Results by reading level for blues clues - View results for all reading levels
Basic 66%
Intermediate 30%
Advanced 4%

Blue - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Blue - Cached
Blue is a colour, the perception of which is evoked by light having a spectrum dominated by energy with a wavelength of roughly 440–490 nm. ...

Images for blues clues - Report images

Researching Blue's Clues: Viewing behavior and impact
www.cmch.tv/research/fullrecord.asp?id=1773 - Cached

Do children learn how to watch television? The impact of extensive ...
www.cmch.tv/research/fullrecord.asp?id=1932 - Cached
The impact of extensive experience with Blues Clues on preschool children’s ...
Information Retrieval
organization: recommendations

http://www.yelp.com/biz/cosmic-cantina-chapel-hill
(not actual page)
What is Information Retrieval?

- Gerard Salton, 1968:

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

storage

How might a web search engine view these pages differently in terms of storage?
Information Retrieval
retrieval

- **Efficiency**: retrieving results in this lifetime (or, better yet, in 0.18 seconds)
- **Effectiveness**: retrieving results that satisfy the user’s information need (more on this later)
- We will focus more on effectiveness
- However, we **will** also discuss in some detail how search engines retrieve results as fast as they do
Outline

Introductions

What is information retrieval (IR)?

What is a search engine?

Why is information retrieval difficult?

How do search engines predict relevance?

How good is a search engine?
Many Types of Search Engines
Many Types of Search Engines

- Pandora
- Bing
- Twitter
- Facebook
- eHarmony
- Google
- Yahoo!
- MapQuest
- Yahoo! Answers
- ACM
- LinkedIn
- Westlaw
- The New York Times
- Yelp
- YouTube
- Flickr
- FINDER
- Picasa
- JSTOR
- LexisNexis
- Broadcast Yourself
The Search Task

- Given a query and a corpus, find relevant items
  - **query**: user’s expression of their information need
  - **corpus**: a repository of retrievable items
  - **relevance**: satisfaction of the user’s information need
Search Engines

web search

query

facebook and productivity

results

corpus

web pages
Search Engines
digital library search

query

facebook productivity

results

1. Effective teaching practices using free Google services: conference tutorial
   Paul Gestwicki, Brian McNely
   October 2010 Journal of Computing Sciences in Colleges, Volume 26 Issue 1
   Publisher: Consortium for Computing Sciences in Colleges
   Full text available: PDF (22.76 KB)
   Bibliometrics: Downloads (6 Weeks): 2, Downloads (12 Months): 48, Downloads (Overall): 48,
   In this 90-minute tutorial, we will share our experiences using free Web services from Google to improve teaching effectiveness. Participants will engage with these services as part of the tutorial. We have used and studied these technologies, ...

2. Model-Based Engineering of Software: Three Productivity Perspectives
   Shawn A. Bohner, Sriram Mohan
   Publisher: IEEE Computer Society
   Full text available: Publisher Site
   Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0
   Evolving software products is a tricky business, especially when the domain is complex and changing rapidly. Like other fields of engineering, software engineering productivity advances have come about largely through abstraction, reuse, process, and ...
   Keywords: Agent-Based Software Systems, Model-Driven Architecture, Model-Driven Development, Model-Based Software Development, Model-Based Software Engineering

   Michael Miller
   Publisher: Que Publishing Company
   Bibliometrics: Downloads (6 Weeks): n/a, Downloads (12 Months): n/a, Downloads (Overall): n/a, Citation Count: 0
   Everything casual users need to know to get the most out of their new Windows 7 PCs, software, and the Internet. The best-selling beginner's guide, now completely updated for Windows 7 and today's most popular Internet tools -
Search Engines

news search

query

emily

results

Tropical Storm Emily Heads Toward Haiti, Dominican Republic
Voice of America - 23 minutes ago
August 03, 2011 Tropical Storm Emily Heads Toward Haiti, Dominican Republic VOA News Tropical storm warnings and watches are posted for parts of the ...

Emily could reach hurricane strength
Samarinda Herald-Tribune (blog)
Blog: Mubarak Triat Begin's Tropical Storm Emily Threatens East Coast NPR (blog)
The Guardian - Fox News
all 1804 news articles »

Emily Heads For Hispaniola
WAVY-TV (blog) - Jannay Wheeler - 2 hours ago
Emily has had little change in strength since yesterday. She has winds of 50mph. The pressure is down a little to 1003 mb (millibars of pressure). ...

Tropical Storm Emily remains weak, hits Haiti tonight Examiner.com
August 2, 2011 Midday Tropics Update: T3 Emily Now Moving More ... Wikulla.com
Emily the Elfinv晨报 - Caribbean Hurricane Network
all 4 news articles »

'Snow White' Writer to Pen Universal's 'Emily the Strange' (Exclusive)
Hollywood Reporter - Boys Kit - 13 hours ago
Melissa Wallack, who wrote the script that became Relativity's high-profile Snow White project, has been drafted by Universal to pen 'Emily the Strange'. ...

Emily the Strange to Darken Big Screen; Vore News Moviefone (blog)
Writer Melissa Wallack Will Follow SNOW WHITE WITH EMILY THE STRANGE Collider.com
Details on the upcoming 'Emily the Strange' movie Examiner.com
ComingSoon.net - 411mania.com
all 7 news articles »

High heat in Midwest and South
Reuters - Tim Sharp - Kevin Murphy - 22 hours ago
By Wendell Marsh WASHINGTON (Reuters) - Record-breaking heat continued to broil central and southern states on Tuesday as Tropical Storm Emily threatened to ...

Corinna

Corral

Mexico wins the Soccer World Cup in Japan-Korea 2002!
Search Engines
local business search

query

results

corpus

curated/synthesized business listings
Search Engines
desktop search

query

results

corpus

files in my laptop
Search Engines
micro-blog search

query

results

query: twitter and productivity

corpus

tweets

Twitter search results for 'twitter and productivity':

- Neen James (Neen James): Productivity tip: Follow ppl on Twitter that inspire, challenge and inform you - delete the clutter! (4 minutes ago)

- Jason Pitts (mr_Ostentatious): Took a day off from twitter to increase my productivity and ended up having a productive day! (1 hour ago)

- Adam Wiebe (adamwiebe): Social media at work is here. Be wary of what is and is not productive. http://Inkd.in/DW3z8J (3 hours ago)

- Gert van der Linde (ViggosDaddy): A brief look: To tweet, or not to tweet? - How does Twitter affect our productivity, influence and how informe... http://tinyurl.com/3wbz3rn (6 hours ago)

- Raghavender (IncorrectMystic): #productivity day - So going be off twitter and other social networks till work is over :) bye tweeple for a while (6 hours ago)
Information Retrieval Tasks and Applications

digital library search
desktop search
web search
question-answering
enterprise search
federated search
news search
social search
local business search
expert search
image search
product search
video search
patent search
(micro-)blog search
recommender systems
community Q&A search
opinion mining
The Search Task
in this course

• Given a query and a corpus, find relevant items
  
query: user’s expression of their information need
  ‣ a textual description of what the user wants
corpus: a repository of retrievable items
  ‣ a collection of textual documents
relevance: satisfaction of the user’s information need
  ‣ the document contains information the user wants
Outline

Introductions

What is information retrieval (IR)?

What is a search engine?

Why is information retrieval difficult?

How do search engines predict relevance?

How good is a search engine?
Why is IR Difficult?

- Information retrieval is an uncertain process
  - users don’t know what they want
  - users don’t know how to convey what they want
  - computers can’t elicit information like a librarian
  - computers can’t understand natural language text
  - the search engine can only guess what is relevant
  - the search engine can only guess if a user is satisfied
  - over time, we can only guess how users adjust their short- and long-term behavior for the better
AOL apologizes for release of user search data

By Dawn Kawamoto and Elinor Mills
Staff Writers, CNET News
Last modified: August 7, 2006 2:30 PM PDT

AOL apologized on Monday for releasing search log data on subscribers that had been intended for use with the company's newly launched research site.

The randomly selected data, which focused on 658,000 subscribers and posted 10 days ago, was among the tools intended for use on the recently launched AOL Research site. But the Internet giant has since removed the search logs from public view.

"This was a screw-up, and we're angry and upset about it. It was an innocent enough attempt to reach out to the academic community with new research tools, but it was obviously not appropriately vetted, and if it had been, it would have been stopped in an instant," AOL, a unit of Time Warner, said in a statement. "Although there was no personally identifiable data linked to these accounts, we're absolutely not defending this. It was a mistake, and we apologize. We've launched an internal investigation into what happened, and we are taking steps to ensure that this type of thing never happens again."

Although AOL had used identification numbers rather than names or user IDs when listing the search logs, that did not quell concerns of privacy advocates, who said that anyone among the 658,000 could easily be identified based on the searches each individual conducted.

"We think it's a major privacy concern, and we're glad to see AOL is taking it seriously," said Ari Schwartz, deputy director of the Center for Democracy and Technology. "Companies that deal in search results have to understand that they carry very sensitive information, even if it doesn't have what we would traditionally consider to be personally identifiable information involved."
Queries and Relevance

- soft surroundings
- trains interlocking dog sheets
- belly dancing music
- christian dior large bag
- best western airport sea tac
- www.bajawedding.com
- marie selby botanical gardens
- big chill down coats
- www.magichat.co.uk
- marie selby botanical gardens
- broadstone raquet club
- seadoo utopia
- seasons white plains condo
- priority club.com
- aircat tools
- epicurus evil
- instructions
- hinds county city of jackson
- last searches on aol a to z
- riverbank run

(AOL query-log)
Queries and Relevance

- A query is an impoverished description of the user’s information need
- Highly ambiguous to anyone other than the user
Queries and Relevance

- **Query 435**: curbing population growth

- **Description**: What measures have been taken worldwide and what countries have been effective in curbing population growth? A relevant document must describe an actual case in which population measures have been taken and their results are known. Reduction measures must have been actively pursued. Passive events such as decease, which involuntarily reduce population, are not relevant.

(from TREC 2005 HARD Track)
Queries and Relevance

• Query 435: curbing population growth

• Description:

(from TREC 2005 HARD Track)
Queries and Relevance

- **Query 435**: curbing population growth
- Can we imagine a relevant document without all these query terms?
Queries and Relevance

• **Query 435:** curbing population growth

• The same concept can be expressed in different ways

---

**Human population control**

From Wikipedia, the free encyclopedia

*Human population control* is the practice of artificially altering the rate of growth of a human population.

Historically, human population control has been implemented by limiting the population's birth rate, usually by government mandate, and has been undertaken as a response to factors including high or increasing levels of poverty, environmental concerns, religious reasons, and overpopulation. While population control can involve measures that improve people's lives by giving them greater control of their reproduction, some programs have exposed them to exploitation.[1]

Worldwide, the population control movement was active throughout the 1960s and 1970s, driving many reproductive health and family planning programs. In the 1980s, tension grew between population control advocates and women's health activists who advanced women's reproductive rights as part of a human rights-based approach.[2] Growing opposition to the narrow population control focus led to a significant change in population control policies in the early 1990s.[3]
Queries and Relevance

• **Query 435**: curbing population growth

• Can we imagine a *non-relevant* document with all these query terms?
Queries and Relevance

• Query 435: curbing population growth

• The query concept can have different “senses”

*Dengue in Philippines taken seriously at one hotel in Clark Philippines*

*Philippines government department of Health has issued warning of Dengue epidemic in July 2010 which is expected to last until the end of the year.*

*While most hotels and resorts are slow to take steps to address this health and safety issues, one hotel in Clark Philippines is already well on its way to implementing proactive measures to prevent the spread of Dengue by curbing population growth of mosquito.*
Queries and Relevance

- This is why IR is difficult (and fascinating!)
- Croft, Metzler, & Strohman:
  Understanding how people compare text and designing computer algorithms to accurately perform this comparison is at the core of information retrieval.
- IR does not seek a deep “understanding” of the document text
- It uses statistical properties of the text to predict whether a document is relevant to a query
  - easier and often times sufficient
Outline

Introductions

What is information retrieval (IR)?

What is a search engine?

Why is information retrieval difficult?

How do search engines predict relevance?

How good is a search engine?
Predicting Relevance

What types of evidence can we use to predict that a document is relevant to a query?

- **query-document evidence**: a property of the query-document pair (e.g., a measure of similarity)
- **document evidence**: a property of the document (same for all queries)
Query: bathing a cat

How to Bathe a Cat

Even though they mostly keep themselves clean, most cats need a bath every now and then. Everyone knows that cats hate getting wet. Cats generally keep themselves clean, and therefore should not be bathed any more often than is absolutely necessary. But they sometimes get especially dirty, get bombarded by fleas or ticks, or get into substances that are toxic or otherwise harmful. On such occasions, it is a good idea to bathe your cat. Here’s how to keep your feline fresh and lovely as painlessly as possible!

Steps

1. Decide, whether your cat really needs a bath. There may be other ways to clean your cat, such as brushing the cat, combing or even rubbing it down with a cloth.

2. Wear appropriate clothing. Not only is it important to wash your cat but it is important to be safe from any possible scratches, minor or major. A jumper or a long-sleeved shirt must be worn so that your cat has no bare skin to scratch. If you don’t have one, you could wear long-sleeved gloves. It is also a good idea to wear clothing that isn’t new.

3. Get at least two people involved in washing your cat, especially if your cat is rather strong and can kick and wriggle its way out of your hands. One person should hold all four legs and hold the cat’s jaw so it can’t open its mouth to bite you, but be sure you don’t hold it really tight so it can’t breathe. Hold your cat firmly so it cannot wriggle out from your grip.

16. Reward your cat. Give him/her their favorite canned food or catnip or treats, and he/she will come to realize that there is a good side to being bathed.
Query-Document Evidence

- **Query:** bathing a cat

- The important query terms occur frequently

- Both terms occur

- Terms occur close together

- Terms occur in the title

- Terms occur in the URL
  
  www.wikihow.com/bathe-your-cat

- Any other ideas?
Query-Document Evidence

- Terms occur in hyperlinks pointing to the page
- Same language as query
- Other terms semantically related to query-terms (e.g., feline, wash)
Query-Document Evidence

- Does not contain "\text{.com}"
- Not one of the most popular queries
- Does not contain the term "\text{news}"
We can also use previous user interactions, e.g.:

The query is similar to other queries associated with clicks on this document

The document is similar to other documents associated with clicks for this query
• Lots of in-links (endorsements)
• Non-spam properties:
  ‣ grammatical sentences
  ‣ no profanity
• Has good formatting
• Anything other ideas?
Document Evidence

- Author attributes
- Peer-reviewed by many
- Reading-level appropriate for user community
- Has pictures
- Recently modified (fresh)
- Normal length
- From domain with other high-quality documents
Predicting Relevance

- IR does not require a deep “understanding” of information.
- We can get by using shallow sources of evidence, which can be generated from the query-document pair or just the document itself.
Outline

Introductions

What is information retrieval (IR)?

What is a search engine?

Why is information retrieval difficult?

How do search engines predict relevance?

How good is a search engine?
**The Search Task**

- **Output:** a ranking of items in descending order of predicted relevance (simplifies the task)

- **Assumption:** the user scans the results from top to bottom and stops when he/she is satisfied or gives up
So, how good is a particular ranking?

Suppose we know which documents are truly relevant to the query...
Evaluating a Ranking

- Which ranking is better?
In general, a ranking with all the relevant documents at the top is best (A is better than B)
Evaluating a Ranking

- Which ranking is better?
Evaluating a Ranking

A

B

• Oftentimes the (relative) quality of a ranking is unclear and depends on the task
Evaluating a Ranking

- Web search: ??????
Evaluating a Ranking

• Web search: A is better than B
• Many documents (redundantly) satisfy the user; the user doesn’t want all of them, the higher the first relevant document, the better
Evaluating a Ranking

A

B

- Patent search: ?????
• Patent search: B is better than A
• User wants to see everything in the corpus that is related to the query (high cost in missing something)
• **Exploratory search:** ??????
• **Exploratory search:** A is better than B
• **Satisfying the information need requires information found in different documents**
Evaluating a Ranking

evaluation metrics

• Given a ranking with known relevant/non-relevant documents, an evaluation metric outputs a quality score

• Many, many metrics

• Different metrics make different assumptions

• Choosing the “right one” requires understanding the task

• Often, we use several (sanity check)
Summary

- The goal of information retrieval is to match information-seekers with the information they seek.
- IR involves analysis, organization, storage, and retrieval.
- There are many types of search engines.
- There is uncertainty at every step of the search process.
- Simple heuristics don’t work, so IR systems make predictions about relevance!
- IR systems use “superficial” evidence to make predictions.
- Users expect different things, depending on the task.
- Evaluation requires understanding the user community.
- My goal is convince you that IR is a fascinating science.