Google Search. Artificial Intelligence as Question Answering. IBM Watson on Jeopardy
Announcements

- Average on Midterm was 75
- Tried to make it so no-one would get below a 50, so if you did, study for final!
- Last Programming assignment due in 10 days
  - Creativity & Pair Programming: make a game
  - The two creativity assignments are 20% of your grade
- Final is on Wed March 21st at 4PM
  - will cover the material since the midterm through next week
  - should be about the same difficulty as the midterm, just a tiny bit longer
Important questions: What you should learn

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does a search engine decide what gets listed first?
- If you try another search engine will you get the same result? If so, which is right? Which is better? Which is more authoritative?
- Are sponsored links better than “organic” links? Is the advertising necessary?
- What is the role of government? What should it be?
Search engines offer unprecedented access to information.

Search engines place the power to shape what we see into the hands of a few companies.

Search engines continue to evolve.
- Recently adding in Google Plus
- Startups on indexing twitter etc

Question Answering is the next new thing!
Watson: In Winter 2011 it won at Jeopardy!

- Let’s try it.
Watson: Let’s try it


The Watson Trivia Challenge
Play against I.B.M.’s question-answering supercomputer. [Read Article](#) | [Play Video](#)
Watson playing Jeopardy

**Watson Does**
- Use IBM’s TTS engine to
  - Speak Clue Selections
  - Speak Responses
- Physically press the button to Ring-in
- Gets the clue electronically when humans see it

**Watson Does NOT (today)**
- Hear
- See
- Process Audi/Visual Clues
  - These are excluded from the contest
- Have a life-line
  - It is a completely **self-contained** HW/SW system.
  - Highlighting the human vs. machine comparison at this task
What is the underlying technology called?
Question Answering (QA)

This is a Natural Language Processing Technology
How long have people been working on it?

- TREC = Text REtrieval Conferences
  - Series of annual evaluations, started in 1992
  - Organized into “tracks”
- Test collections are formed by “pooling”
  - Gather results from all participants
  - Corpus/topics/judgments can be reused
- TREC has had a QA Track since 1999.
Roots of Question Answering

- Information Retrieval (IR)
- Information Extraction (IE)
Information Retrieval (IR)

- The first versions of search engines
- IR systems
  - Use statistical methods
  - Rely on frequency of words in query, document, collection
  - Retrieve complete documents
  - Return ranked lists of “hits” based on relevance
- Limitations
  - Answers questions indirectly (with a ranked list of documents that may or may not have the answer)
  - Does not attempt to understand the “meaning” of user’s query or documents in the collection
Information retrieval technology was the start for web search
Web search: It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.
Information retrieval technology was the start for web search

Will review Web Search and continue from Lecture 14 after Q/A Artificial Intelligence

Q/A is what web search might be like in 5 years (possibly less. You heard it here first!)
Contrast IR/Search with Asking an Expert

**Decision Maker**
- Has Question
- Distills to 2-3 Keywords
- Reads Documents, Finds Answers
- Finds & ASKs

**Search Engine**
- Finds Documents containing Keywords
- Delivers Documents based on
  - Experts

**Decision Maker**
- Asks NL Question
- Considers Answer & Evidence

**Expert**
- Understands Question
- Produces Possible Answers & Evidence
- Analyzes Evidence, Computes Confidence
- Delivers Response, Evidence & Confidence
What does Jeopardy Host do?

Informed Decision Making

Decision Maker
- Asks NL Question
- Considers Answer & Evidence

Expert
- Understands Question
- Produces Possible Answers & Evidence
- Analyzes Evidence, Computes Confidence
- Delivers Response, Evidence & Confidence

Computer
- Understands Question
- Produces Possible Answers & Evidence
- Analyzes Evidence, Computes Confidence
- Delivers Response, Evidence & Confidence

Jeopardy! Host
- Asks NL Question
- Judges Answer Correct
Information Extraction (IE)

- IE systems (*usually... but recent advances*)
  - Identify documents of a specific type
  - Extract information according to pre-defined templates
  - Place the information into frame-like database records

<table>
<thead>
<tr>
<th>Weather disaster:</th>
<th>Type</th>
<th>Damage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Date</td>
<td>Deaths</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Location</td>
<td>...</td>
</tr>
</tbody>
</table>

- Templates = pre-defined questions
- Extracted information = answers
- Limitations
  - Templates are domain dependent and not easily portable
Types of Question Answering

- http://trec.nist.gov/data/qa/T8_QAdata/development.qa

Number: 10000
What date in 1989 did East Germany open the Berlin Wall?
LA012890-0072
LA122489-0101
Nov 9

Number: 10001
Who was Johnny Mathis' high school track coach?
LA053189-0069
Lou Vasquez

Number: 10002
What is the shape of a porpoises' tooth?
LA120390-0087
spade-shaped

Number: 10003
What is the number of buffaloes thought to have been living in North America when Columbus landed in 1492?
LA030289-0053
60 million
Types of Question Answering


- **Factoid**
  - Who discovered oxygen?
  - When did Hawaii become a state?
  - Where is Ayers Rock?
  - What team won the World Series in 1992?

- **List**
  - What countries export oil?
  - Name U.S. cities that have a “Shubert” theater.

- **Definition**
  - Who is Aaron Copland?
  - What is a quasar?
Central Idea of Factoid QA

- Determine the semantic type of the expected answer
  “Who won the Nobel Peace Prize in 1991?” is looking for a PERSON

- Retrieve documents that have keywords from the question
  Retrieve documents that have the keywords “won”, “Nobel Peace Prize”, and “1991”

- Look for named-entities of the proper type near keywords
  Look for a PERSON near the keywords “won”, “Nobel Peace Prize”, and “1991”
An Example

Who won the Nobel Peace Prize in 1991?

But many foreign investors remain sceptical, and western governments are withholding aid because of the Slorc's dismal human rights record and the continued detention of Ms Aung San Suu Kyi, the opposition leader who won the Nobel Peace Prize in 1991.

The military junta took power in 1988 as pro-democracy demonstrations were sweeping the country. It held elections in 1990, but has ignored their result. It has kept the 1991 Nobel peace prize winner, Aung San Suu Kyi - leader of the opposition party which won a landslide victory in the poll - under house arrest since July 1989.

The regime, which is also engaged in a battle with insurgents near its eastern border with Thailand, ignored a 1990 election victory by an opposition party and is detaining its leader, Ms Aung San Suu Kyi, who was awarded the 1991 Nobel Peace Prize. According to the British Red Cross, 5,000 or more refugees, mainly the elderly and women and children, are crossing into Bangladesh each day.
Generic QA Architecture

- **NL question** (Golden color) → **Question Analyzer**
- **IR Query** (Yellow color) → **Document Retriever**
- **Documents** (Yellow color) → **Passage Retriever**
- **Passages** (Yellow color) → **Answer Extractor**
- **Answers** (Yellow color) → **Answer Type**
Question analysis

- Question word cues
  - Who → person, organization, location (e.g., city)
  - When → date
  - Where → location
  - What/Why/How → ??

- Head noun cues
  - What city, which country, what year...
  - Which astronaut, what blues band, ...

- Scalar adjective cues
  - How long, how fast, how far, how old, ...
Answer Type Hierarchy

- Top
  - Reason
  - Time
  - Product
  - Organization
  - Manner
  - Nationality
  - Alphabet
  - Mammal
  - Reptile
  - Author
  - Landmark
  - Date
  - Money
  - Location
  - Language
  - Person
  - Quotation
  - Numerical
  - Game
  - Dogbreed
  - Definition
  - Value
  - Degree
  - Rate
  - Percentage
  - Amount
  - Speed
  - Dimension
  - Duration
  - Count
  - Temperature

- University
  - Country
  - Continent
  - City
  - Province
  - Other loc

- Scientist
  - Biologist
  - Botanist
  - Linguist
  - Cosmographer
  - Performer
  - Actor
  - Dancer
  - Musician
  - Actress
  - Oboist

- European
  - Italian
  - Romanian

- Swede

- Guardian, defender
  - Bodyguard
  - Escort
  - Custodian
  - Girth

- Philosopher
  - Yogi
  - Realist
  - Existentialist
  - Pragmatist
Extracting Named Entities

Person:  Mr. Hubert J. Smith, Adm. McInnes, Grace Chan
Title:  Chairman, Vice President of Technology, Secretary of State
Country:  USSR, France, Haiti, Haitian Republic
City:  New York, Rome, Paris, Birmingham, Seneca Falls
Province:  Kansas, Yorkshire, Uttar Pradesh
Business:  GTE Corporation, FreeMarkets Inc., Acme
University:  Bryn Mawr College, University of Iowa
Organization:  Red Cross, Boys and Girls Club
Wikipedia: What knowledge can we get from Wikipedia?

Using WordNet: Online Thesaurus.

- [http://wordnetweb.princeton.edu/](http://wordnetweb.princeton.edu/)
- What is the service ceiling of a U-2?
- Can access it FROM a program (not just this interface).
Using WordNet: Online Thesaurus.

- What is the service ceiling of a U-2
- Can access it as a program.

http://wordnetweb.princeton.edu/

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: ceiling

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) ceiling (the overhead upper surface of a covered space) "he hated painting the ceiling"
- S: (n) ceiling ((meteorology) altitude of the lowest layer of clouds)
- S: (n) ceiling, roof, cap (an upper limit on what is allowed) "he put on the number of women who worked for him"; "there was a roof on salaries"; "they established a cap for prices"
- S: (n) ceiling (maximum altitude at which a plane can fly (under specified conditions))
  - direct hypernym / full hypernym
  - direct hypernym / inherited hypernym / sister term
  - S: (n) altitude, height (elevation especially above sea level or above the earth's surface) "the altitude gave her a headache"
    - S: (n) level (height above ground) "the water reached ankle level"; "the pictures were at the same level"
    - S: (n) ceiling (meteorology) altitude of the lowest layer of clouds
    - S: (n) ceiling (maximum altitude at which a plane can fly (under specified conditions))
More Named Entities

Currency: 400 yen, $100, DM 450,000
Linear: 10 feet, 100 miles, 15 centimeters
Area: a square foot, 15 acres
Volume: 6 cubic feet, 100 gallons
Weight: 10 pounds, half a ton, 100 kilos
Duration: 10 day, five minutes, 3 years, a millennium
Frequency: daily, biannually, 5 times, 3 times a day
Speed: 6 miles per hour, 15 feet per second, 5 kph
Age: 3 weeks old, 10-year-old, 50 years of age
Watson: In Winter 2011 it won at Jeopardy!

- Let’s try it.
The Turing Test

- Turing in 1950 published a philosophical paper designed to stop people arguing about whether or not machines could think.
- He proposed that the question be replaced with a test, which is what is now called the Turing Test.

![Turing's paper cover](image)
I believe that in about fifty years’ time it will be possible to program computers….to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning…… I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.
Artificial Intelligence’s Great Quest: How did Watson do it?

Courtesy of Jennifer Chu-Carroll IBM Research
IBM’s big team: Jennifer “harder than I’ve ever worked in my life”

Core team funded by DARPA under TREC for more than 5 years before IBM decided to do this
Open Domain Question Answering

Automatic Open-Domain Question Answering
A Long-Standing Challenge in Artificial Intelligence to Emulate Human Expertise

- **Given**
  - Rich Natural Language Questions
  - Over a Broad Domain of Knowledge

- **Deliver**
  - **Precise Answers:** Determine what is being asked & give precise response
  - **Accurate Confidences:** Determine likelihood answer is correct
  - **Consumable Justifications:** Explain why the answer is right
  - **Fast Response Time:** Precision & Confidence in <3 seconds
Jeopardy seemed like it could be possible

The Jeopardy! Challenge
A palpable, compelling and notable way to drive the technology of Question Answering along 5 Key Dimensions

- Broad/Open Domain
- Complex Language
- High Precision
- Accurate Confidence
- High Speed

$200
If you're standing, it's the direction you should look to check out the wainscoting

$400
A round type of this structure, with the silo inside it, was popular in the 1900s; it's rare now

$1000
Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

$600
1948: Johns Hopkins scientists find that this antihistamine alleviates motion sickness

$800
This title character was the crusty & tough city editor of the Los Angeles Tribune

Enabling Technologies – The Time Was Right

Natural Knowledge
- Large volumes natural language electronic text (e.g., news, wikis, reference, web, etc.)
- Encodes knowledge and greater linguistic contexts to better resolve intended meaning

Semi-Structured Knowledge
- Large volumes of Thesauri, Dictionaries, Folksonomies, Linked Data, and the Semantic Web
- Rapid, community-based construction
- Across many domains – Specialized and General

NLP (Text Analysis)
- Entity and Relation Detection, Syntactic & Semantic Parsing
- Statistical NLP - Broader coverage, lower cost Information Extraction
- Statistical Paraphrasing: Learn ways to express same meaning

Compute Power
- Massive parallel compute power
- 1000s of compute cores working simultaneously
- TBs of globally addressable main memory
Answer Type Hierarchy

- Top
  - Reason
  - Time
  - Product
  - Organization
  - Manner
  - Nationality
  - Alphabet
  - Mammal
  - Reptile
  - Author
  - Landmark
  - Date
  - Money
  - Location
  - Language
  - Person
  - Quotation
  - Numerical
  - Game
  - Dog breed
  - Definition
  - Value
  - Degree
  - Rate
  - Percentage
  - Amount
  - Speed
  - Dimension
  - Duration
  - Count
  - Temperature

- University
- Country
- Continent
- City
- Province
- Other Loc

- Scientist
- Biologist
- Botanist
- Cosmographer
- Performer
- Actor
- Dancer
- Musician
- Actress
- Oboist
- European
- Italian
- Romanian
- Swede
- Guardian
- Defender
- Bodyguard
- Escort
- Custodian
- Girth
- Philosopher
- Yogi
- Realist
- Existentialist
- Pragmatist
Answer Types: Cannot Anticipate!!

Broad Domain

We do NOT attempt to anticipate all questions and build specialized databases.

In a random sample of 20,000 questions we found 2,500 distinct types*. The most frequent occurring <3% of the time. The distribution has a very long tail.

And for each these types 1000’s of different things may be asked.

Even going for the head of the tail will barely make a dent

*13% are non-distinct (e.g., it, this, these or NA)

Our Focus is on reusable NLP technology for analyzing vast volumes of as-is text. Structured sources (DBs and KBs) provide background knowledge for interpreting the text.
Interesting tradeoff: Knowledge, Precision, Open Domain

Structured KB approach delivers high confidence if questions can be precisely mapped to existing & reliable sources. Turns out to be rarely the case and confidence and accuracy drop off quickly.

Basic Text Search approach never delivers high confidences but quickly reaches and maintains a peak accuracy at about 30%.

Must combine deep and shallow semantic analysis over structured & unstructured content to drive up precision, recall and confidence.

© 2017 IBM Corporation
Does Wordnet know about Chocolate Cake?

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: chocolate cake

Display Options: (Select option to change) Change
Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations
Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) chocolate cake (cake containing chocolate)
  - direct hyponym / full hyponym
  - S: (n) devil's food, devil's food cake (very dark chocolate cake)
  - direct hypernym / inherited hypernym / sister term
Chocolate Decadence Cake?: The Web knows all!

Chocolate Decadence Cake Recipe - Allrecipes.com
allrecipes.com/recipe/chocolate-decadence-cake-i/
4.5 stars 65 reviews - 425 cal
This is the richest chocolate cake ever! Any questions?
Recipes like this - Recipe Reviews - Photos of this recipe

Chocolate Decadence Recipe - Allrecipes.com
allrecipes.com/recipe/chocolate-decadence/
4.5 stars 10 reviews - 4 hrs 45 mins - 250 cal
Six dainty cups of deeply rich chocolate are gently baked and served cooled to a small grateful crowd.

Chocolate Decadence Cake III Recipe - Allrecipes.com
allrecipes.com/recipe/chocolate-decadence-cake-iii/
4 stars 6 reviews - 1 hr - 324 cal
This is a very rich cake, a small piece goes a long way. The best way to describe it is that it's a cross between a moist brownie and fudge. Be sure to use good...
How good does it have to be to win?

Our Analysis Reveals the Winner's Cloud

Top human players are remarkably good.

Each dot – actual historical human Jeopardy! games

Winning Human Performance

Grand Champion Human Performance

2007 QA Computer System

More Confident

Less Confident
How good does it have to be to win?

Our Analysis Reveals the Winner’s Cloud

Each dot – actual historical human Jeopardy! games

Winning Human Performance

In 2007, we committed to making a Huge Leap!

Grand Champion Human Performance

Computers? Not So Good.

2007 QA Computer System

More Confident

Less Confident
Incremental Progress

A Few Guiding Principles

- Specific large hand-crafted models won’t cut it
  - Too slow, too narrow, too brittle, too biased
  - Need to acquire and analyze information from as-is knowledge sources

- Intelligence from many diverse methods
  - Many diverse algorithms must be combined: No single one is expected to solve the whole problem. Each addressing different weaknesses.
  - Relative impact of many overlapping methods must be learned

- Massive Parallelism is a Key Enabler
  - Pursue many competing independent hypotheses over large data
  - Efficiency will demand simultaneous threads of evidence evaluation

**DeepQA: The Technology Behind Watson**

Massively Parallel Probabilistic Evidence-Based Architecture

Generates and scores many hypotheses using a combination of 100's of **Natural Language Processing, Information Retrieval, Machine Learning** and **Reasoning Algorithms**. These gather, evaluate, weigh and balance different types of evidence to deliver the answer with the best support it can find.
This is not easy!

Example Question

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE
More NLP: techniques to learn knowledge from texts, invented in the last 10 years

Inducing Frames

Volumes of Text → Syntactic Frames → Semantic Frames

Parsing & Entity Detection → Generalization & Statistical Aggregation

Syntactic Frames:
- subject
- verb
- object

Semantic Frames:
- Inventors patent inventions (.8)
- Officials Submit Resignations (.7)
- People earn degrees at schools (.9)
- Fluid is a liquid (.6)
- Liquid is a fluid (.5)
- Vessels Sink (0.7)
- People sink 8-balls (0.5) (in pool/0.8)
Evaluating possibilities and their Evidence

In cell division, mitosis splits the nucleus & cytokinesis splits this **liquid cushioning** the nucleus.

- Many candidate answers (CAs) are generated from many different searches
- Each possibility is evaluated according to **different dimensions of evidence**.
- **Just One** piece of evidence is if the CA is of the right type. In this case a “liquid”.

```
ls(“Cytoplasm”, “liquid”) = 0.2
ls(“organelle”, “liquid”) = 0.1
ls(“vacuole”, “liquid”) = 0.2
ls(“plasma”, “liquid”) = 0.7
```

“Cytoplasm is a **fluid surrounding** the nucleus…”

[[Wordnet ➔ ls_a(Fluid, Liquid) ➔ ?]]

[[Learned ➔ ls_a(Fluid, Liquid) ➔ yes.]]
Big Data: So much stuff on the web!

In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

In May, Gary arrived in India after he celebrated his anniversary in Portugal.

Evidence suggests “Gary” is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence.
What is good evidence? Wikipedia?

- Stronger evidence can be much harder to find and score.
Grouping Features => “Evidence Profiles”

**Clue:** In 1698, this comet discoverer took a ship called the Paramour Pink on the first purely scientific sea voyage.

**Peter Sellers comes into the picture because he played a character whose wife is a paramour of the Phantom in “The Pink Panther.”**

**Positive Evidence**

**Negative Evidence**
Creative Question Types: Edible Rhyme

Question Decomposition: Parallel

A long, tiresome speech delivered by a frothy pie topping

Diatribes
Harangue

Whipped Cream
Meringue

Answer: Meringue, Harangue

Category: Edible Rhyme Time
Decomposing into phrases

Question Decomposition: Nested

Lyndon B Johnson

In 1968 this man was U.S. president.

When "60 Minutes" premiered this man was U.S. president.

The DeepQA architecture attempts different decompositions and recursively applies the QA algorithms

Must identify and solve sub-questions from different sources to answer the top level question
Incremental Progress


Playing in the Winners Cloud

Graph showing the incremental progress in answering precision from 6/2007 to 11/2010, with specific dates marked for significant improvements.
Watson: In Winter 2011 it won at Jeopardy!

- Let’s try it.
Potential Business Applications

**Healthcare / Life Sciences:** Diagnostic Assistance, Evidenced-Based, Collaborative Medicine

**Tech Support:** Help-desk, Contact Centers

**Enterprise Knowledge Management and Business Intelligence**

**Government:** Improved Information Sharing and Security
The Library vs the Information Bazaar

- “Yellow pages”, directories, and catalogues
- The “Web” is not hierarchical
  - no structure like a library
- Catalogues are out - search engines are in.
- But - search engines control what you see
“The search tools that help us find needles in the digital haystack have become the lenses through which we view the digital landscape. Businesses and governments use them to distort our picture of reality.”

Blown to Bits - pg 110
“For the user, search is the power to find things, and for whoever controls the engine, search is the power to shape what you see.”

Blown to Bits pg 112
Web search: It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.
Its all free?? : Well no. Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.

- And it matters who pays cause it affects how it works
Page Rank Algorithms

- The “crown jewels” of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles
  - keyword only in the body text
  - site is “trustworthy”
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)
Web search: It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.
4. Understand the Query

- Steps 1-3 happen in “the background”
- Not much “understanding” in today’s search engines but that could change soon.
- Advanced search engine features help
Boolean Queries

Search Engine words are independent

- Words don’t have to occur together
- Use Boolean queries and quotes
- Logical Operators: AND, OR, NOT
  
  monet AND water AND lilies
  “van gogh” OR gauguin
  vermeer AND girl AND NOT pearl
- When Google gets the query
- It “ands” the two lists together, finding URLs that are on both lists
- It counts them up, records time, shows 10 hits
Tarta de chocolate de mi amiga Carol Kotkin y otra del NYT

Carol Kotkin es una maestra de cocina de Miami con quien cogí clases de cocina allá y luego cuando abrí Kitchen World la tuvimos como maestra invitada. Carol es una señora encantadora, la pasamos muy bien cuando estuvo aquí. Hace muchos años que no sé de ella. Hace unos años escribió un libro de cocina, el cual me mandó de regalo, donde resalta no tan solo la cocina judía/americana, sino la gran influencia latina que ha tenido esa cocina en el Sur de la Florida. Hizo muy buenas recetas en Kitchen World, pero para mí la mejor fue ésta.

Es una tarta, no un bizcocho porque en vez de harina, lleva nueces. Carol es judía y siempre pensé que esta receta era judía. Pero en el New York Times Magazine del domingo pasado, encontré otra receta similar que suena buenísima también. Se las voy a traducir aquí, aunque todavía no la he hecho pero tengo la intención de hacerla este fin de semana.

El artículo está muy interesante pues habla del libro Chocolate Decadence de Janice Feuer el cual compré hace muchísimos años en un viaje a Williamsburgh, Virginia. Allí almorzamos en un restaurante y de postre Patricia, por supuesto, pidió el bizcocho de chocolate que se llamaba "Death by Chocolate". Era tan chocolatoso que mi esposo le dijo que si no iba a tener un "chocolate overdose" a lo que ella respondió: "you can never have a chocolate overdose"... ahí entendimos la fascinación de
5. Determine Relevance

- “Recall” == what percentage of relevant documents are returned by the search?
- Simple relevance calculation
  - count the number of times each search word appears in the document, add them all up
- Long documents get higher scores.
- Uninteresting words like “the” contribute to the score.
- All word occurrences are not equal (title words should count more).
- Relevance is an IR term. Page Rank a Search idea
6. Determine Ranking: Page Rank

- Which of the relevant documents should be displayed first?
- Simple solution - put one with highest relevance score first.
- What if many have the same score?
- Are ones with the highest relevance score really the most important? What about the source of the document (e.g. NY Times vs some random blog post).
Page Rank Algorithms

- The “crown jewels” of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles
  - keyword only in the body text
  - site is “trustworthy”
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)
Page Rank

- You want the most likely hits ... how does Google show you what you want?
- Page Rank – a mechanism to estimate the “importance” of a page; pages are listed by page rank, highest to lowest
Google Short Index & Long Index

- Known as the short barrels and the long barrels..
- Short index:
  - store the words in link texts that point to a page (inbound links!!)
  - the words in a page’s title, and one or two other special things.
- The link text words are attributed to the target page, and not to the page that the link is on.
  - In other words, if my page links to your page, using the link text “Miami hotels”, then the words “Miami” and “hotels” are stored in the short index as though they appeared in your page, but they belong to my page. If 100 pages link to your page, using those same words as link text, then your page will have a lot of entries in the short index for those particular words.
- The long index is used to store all the other words on a page – its actual content.
Processing a query: short and long indices

- First try to get enough results from the short index.
- If you can’t get enough results, then use the long index to add to what they have.
  - It means that, if they can get enough results from the short index – that’s the index that contains words in link texts and page titles – then they don’t even look in the long index where the actual contents of pages are stored. Page content isn’t even considered if they can get enough results from the link texts and titles index – the short index.
- Thus: link texts are very powerful for Google rankings.
  - Much more powerful than page titles, because a page can have the words from only one title in the short index, but it can have the words from a great many link texts in there..
  - Page titles and meta descriptions were the second most powerful ranking factors, because they are stored in the short index.
What does Google use to search?

- An inbound link is simply a hyperlink that can also have an href description: the **anchor** text.
- Last Thurs: To put in a link, use 2 tags

```html
<a href=http://users.soe.ucsc.edu/~maw>Prof. Walker’s</a>
```

Hyper-text reference – the link  Anchor  End

- Inbound links are important because of the way that Google stores a page’s data, and the way that they process a search query.
Google’s PageRank Algorithm

- If lots of pages point TO this page, this must be a “more important” page
- Tweaking the page rank algorithm can make or break a small business.
**DATA and SOFTWARE DOWNLOADS.**

**PROFESSIONAL EXPERIENCE**

**Professor of Computer Science**, Natural Language and Dialogue Systems Lab, University of California, Santa Cruz, 2009 to present

**Professor of Computer Science**, Head of Cognitive Systems Group, University of Sheffield, Sheffield, England, 2003 to 2009


**Research Scientist**, Mitsubishi Electric Research Laboratories, Cambridge, Ma., Interactive Learning and Entertainment, 1993 to 1996

**Consultant**, Hewlett Packard Laboratories, Bristol, England, on dialogue systems, speech technology, and personal information systems: 1989-1993

**Researcher**, Dialogue Modeling Department, Electrotechnical Laboratory, Tsukuba City, Japan: Summer 1991


**EDUCATION**


Who points to a page?

Site-specific searches
You can use special searches (called operators) to find pages that are similar to or link to a specific URL.

Learn more about these and other site-specific searches in the Webmaster Tools Help Center.

Search for pages that link to a URL
To search for web pages that link to a URL, use the "link:" operator. For example, to find pages that link to www.google.com, use [link:google.com]. You can also search for links to specific pages or directories, e.g. [link:google.com/webmasters].

Search for pages that are similar to a URL
The "related:" operator displays pages similar to a URL. For example, to find pages similar to nytimes.com, search for [related:nytimes.com]. We look at similarities between the link structures as well as other factors to help determine similar pages.

The "related:" operator returns the same results as clicking the Similar link within a search result's Instant Preview.
Seeing what pages link to a page. Use Link:
## Personalization search/ads across apps

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Search</strong></td>
<td>Search billions of web pages</td>
</tr>
<tr>
<td><strong>iGoogle</strong></td>
<td>Add news, games and more to your Google homepage</td>
</tr>
<tr>
<td><strong>Bookmarks</strong></td>
<td>Access your bookmarks and starred items</td>
</tr>
<tr>
<td><strong>Chrome</strong></td>
<td>A browser built for speed, simplicity and security</td>
</tr>
<tr>
<td><strong>Toolbar</strong></td>
<td>Add a search box to your browser</td>
</tr>
</tbody>
</table>

### Mobile

- **Mobile**: Get Google products on your mobile phone
- **Search for mobile**: Search Google wherever you are

### Media

- **YouTube**: Watch, upload and share videos
- **Books**: Search the full text of books
- **News**: Search thousands of news stories
- **Picasa**: Find, edit and share your photos

### Geo

- **Google Music**: Access your music collection instantly anywhere
- **Image Search**: Search for images on the web
- **Video Search**: Search for videos on the web
- **Picnik**: Fast, easy and fun photo-editing
2. Keep Copies

- Spider downloads the page as part of the “visit” in order to create the index.
- Search engine may “cache” the copy.
- Is this legal? What about copyright?
- But wait, browsing requires copying as well.

“(AFP) – Sep 15, 2011

NEW YORK — Google and publishers told a US judge Thursday they are close to settling a lawsuit over the Internet giant's controversial book-scanning project…”
Lilly and Zyprexa

- treatment for schizophrenia
- possible serious side effects
- lawyer leaked a confidential internal Lilly memo
- Judge ruled lawyer acted improperly but “the bits had escaped and could not be recaptured,” (Blown to bits pg 116)
7. Presenting Results

- Mostly just a list.
- Maybe there are better forms.
- Those sponsored links…
Search Engines … A Summary

- A search engine has two parts
  - Crawler, to index the data
  - Query Processor, to answer queries based on index
- In the case of many hits, a query processor must rank the results; page rank does that by
  - “using data differentially” ... not all associations are equivalent; anchors and file names count more
  - “noting relationship of pages” ... a page is more important if important pages link to it

Google, Bing, Yahoo and other Search Engines Use All of These Ideas
Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.
Placements, Clicks, and Auctions

- Buy higher position in the ranking - FTC said don’t do it without flagging it as such.
- Banner ads displayed when search included certain terms.
  - pay for view or pay for click throughs?
- Companies bid for popular terms.
- Companies exercise editorial power (censorship?) by refusing certain ads.
Manipulating Search Results

- White text on white background with words that will raise your rank.
- Google Bombing - “miserable failure” search in 2000 yielded white house biography of George Bush
- Companies that will help you move up in the ranking with changes to your web site.
Google Bombing

- Companies **purchasing** inbound links to increase their page rank
- If it's not on the first page it might as well not be there
- Google has begun trying to change the page rank algorithm to move away from inbound links.
- However, this was their core competency from 1996-2007, what differentiated them from AltaVista, Yahoo, Netscape etc.
Important questions: What you should learn

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does a search engine decide what gets listed first?
- If you try another search engine will you get the same result? If so, which is right? Which is better? Which is more authoritative?
- Are sponsored links better than “organic” links? Is the advertising necessary?
- What is the role of government? What should it be?
Summary

- Search engines offer unprecedented access to information.
- Search engines place the power to shape what we see into the hands of a few companies.
- Search engines continue to evolve.
  - Recently adding in Google Plus
  - Startups on indexing twitter etc
- Question Answering is the next new thing!
The Information Retrieval Cycle

Source Selection

Query Formulation

Search

Selection

Examination

Delivery

Resource

Query

Ranked List

Documents

Documents

query reformulation, vocabulary learning, relevance feedback

donece source reselection
Supporting the Search Process

- Source Selection
- Query Formulation
- Search
- Selection
- Examination
- Delivery
- Acquisition
- Indexing