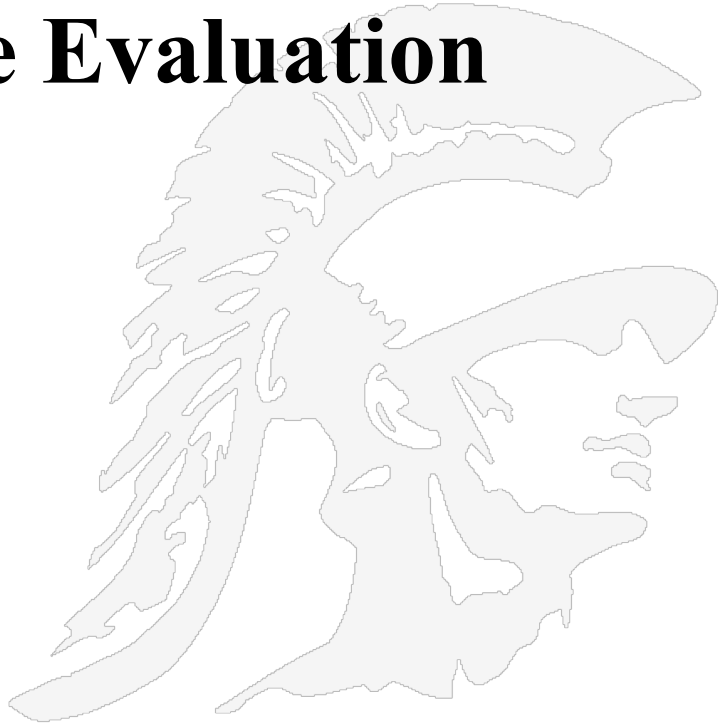
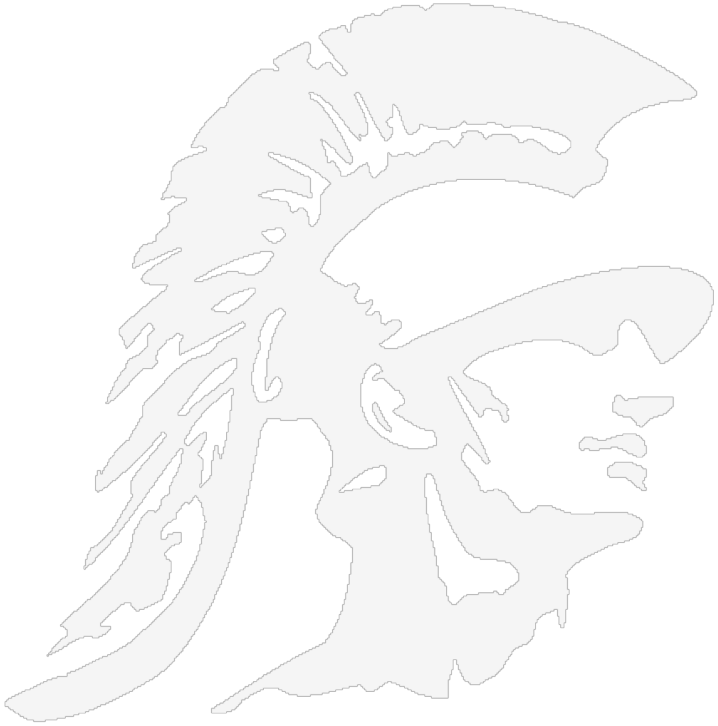
Search Engine Evaluation

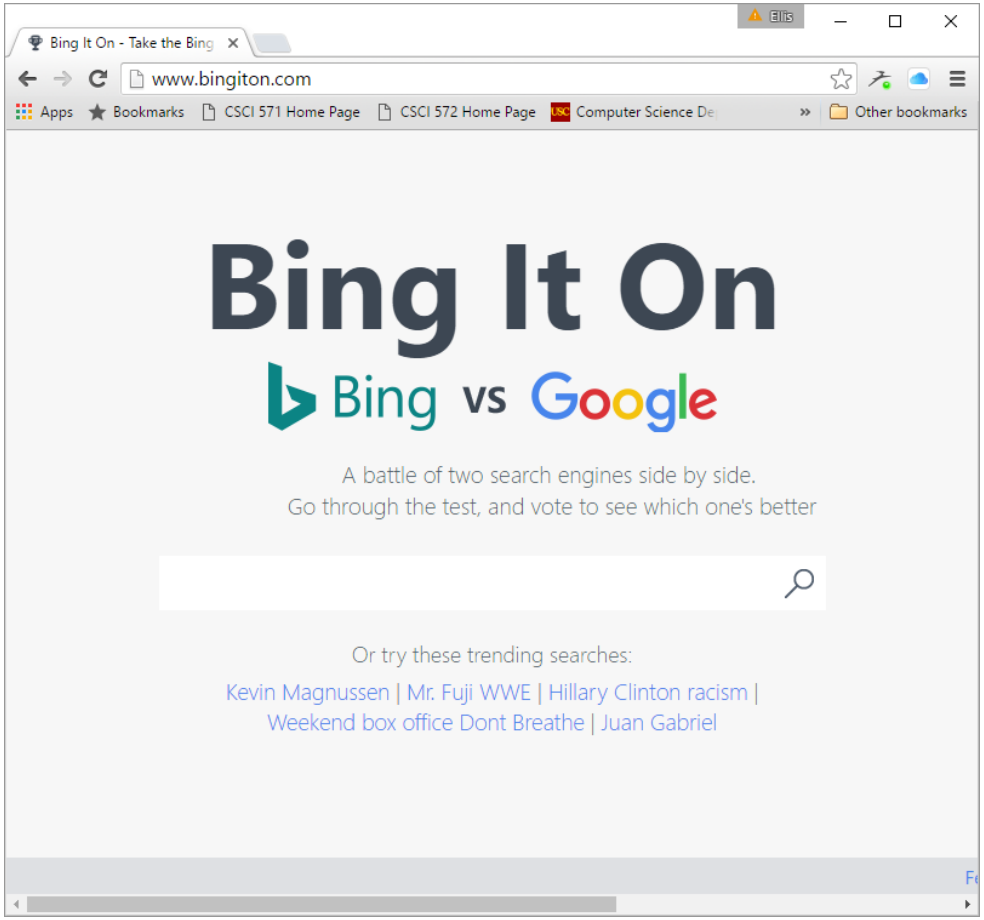


Outline

- Defining precision/recall
- Mean Average Precision
- Harmonic Mean and F Measure
- Discounted Cumulative Gain
- Elements of Good Search Results
- Google's Search Quality Guidelines
- Using log files for evaluation
- A/B Testing



Comparing Bing and Google



- This site is no longer active, but we can simulate the experiment

Try it yourself!

here are some queries:

- ac versus dc current
- best bottled water
- worst hotel in Santa Monica
- how many gears do I need on a bicycle
- Clint Eastwood's best movie

Precision/Recall

- How do we measure the quality of search engines?
- Precision = #(relevant items retrieved)
  divided by
  #(all retrieved items)

- Recall = #(relevant items retrieved)
  divided by
  #(all relevant items)

Formalizing Precision/Recall

	Relevant	Nonrelevant
Retrieved	True positive (tp)	False positive (fp)
Not retrieved	False negative (fn)	True negative (tn)

$$\text{Precision} = tp/(tp + fp)$$

$$\text{Recall} = tp/(tp + fn)$$

- The accuracy of an engine is defined as:

the fraction of these classifications that are correct

$$(tp + tn) / (tp + fp + fn + tn)$$

For web applications,
Precision is more important than Recall

Precision/Recall Using Set Notation

A is set of relevant documents,
B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

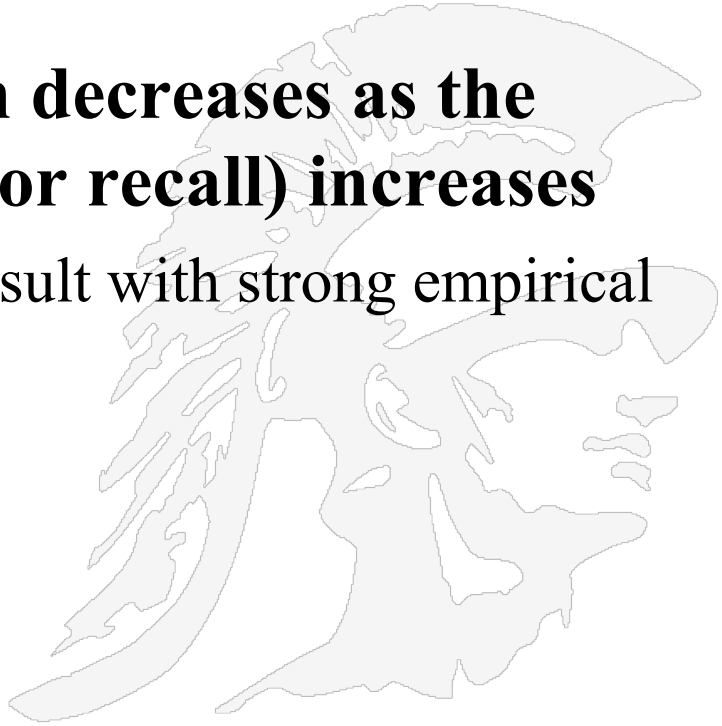
you may not be able to see them, but A and B have a bar over them and it denotes the complement set

$$Recall = \frac{|A \cap B|}{|A|}$$

$$Precision = \frac{|A \cap B|}{|B|}$$

Precision/Recall Two Observations

- **You can get high recall (but low precision) by retrieving all docs for all queries!**
  - a rather foolish strategy
- **In a good system, precision decreases as the number of docs retrieved (or recall) increases**
  - This is not a theorem, but a result with strong empirical confirmation



Harmonic Mean

- **There are three Pythagorean means**
  - 1. arithmetic mean, 2. geometric mean, 3. harmonic mean
  - of course we all know how to compute the arithmetic mean
  - the geometric mean is the *n*th root of the product of *n* numbers
- **The harmonic mean tends strongly toward the least element of the list making it useful in analyzing search engine results**
- **To find the harmonic mean of a set of *n* numbers**
  1. add the reciprocals of the numbers in the set
  2. divide the sum by *n*
  3. take the reciprocal of the result
- **e.g. for the numbers 3, 6, 9, and 12**
  - **The arithmetic mean is: (3+6+9+12)/4 = 7.5**
  - **The geometric mean is: nth-root(3*6*9*12) = 4^{th}-root(1944) = 6.64**
  - **The harmonic mean is: (1/3+1/6+1/9+1/12)=(.33+.16+.11+.08)/4=0.17 and 1/0.17 = 5.88**

F Measure

- The harmonic mean of the precision and the recall is often used as an aggregated performance score for the evaluation of algorithms and systems: called the

  F-score (or F-measure).

- *Harmonic mean* of recall and precision is defined as

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R}+\frac{1}{P})} = \frac{2RP}{(R+P)}$$

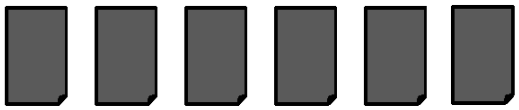
  - harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is affected more by outliers that are unusually large

- More general form of F-Measure
  - β is a parameter that controls the relative importance of recall and precision

$$F_\beta = (\beta^2 + 1)RP/(R + \beta^2 P)$$

USC Viterbi School of Engineering

Calculating Recall/Precision at Fixed Positions



= the relevant documents

Steps

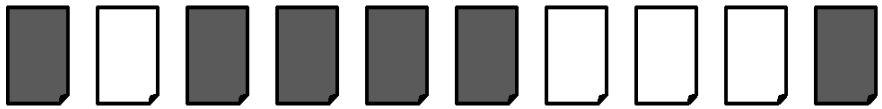
Recall:
1/6 = 0.17
Precision:
1/1

Recall:
1/6 = 0.17
Precision:
1/2 = 0.5

Recall:
2/6 = 0.33
Precision:
2/3 = 0.67

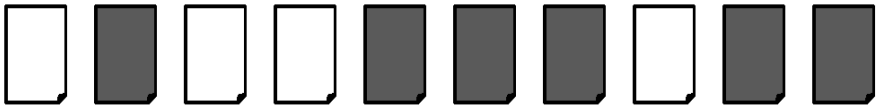
Recall:
3/6 = 0.5
Precision:
3/4 = 0.75

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

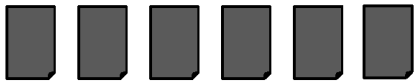
Ranking #2



Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

Recall=#RelevItemsRetr/allRelevItems
Prec=#RelevItemsRetr/allItemsRetr

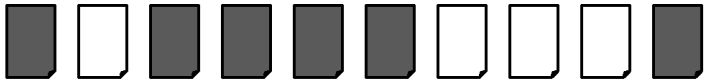
Average Precision of the Relevant Documents



= the relevant documents

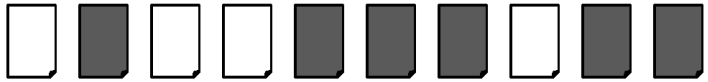
computes the sum of the precisions of the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2



Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

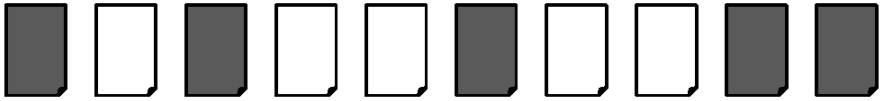
$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

Conclusion: Ranking #1 for this query is best

Averaging Across Queries


= relevant documents for query 1

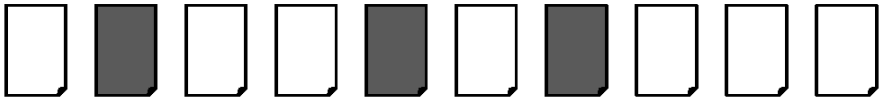
Ranking #1



Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5


= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

Average precision across the two queries for relevant docs is:
(1 + .67 + .5 + .44 + .5 + .5 + .4 + .43)/8 = 0.55

Averaging Across Queries

- ***Mean average precision (MAP)*** for a set of queries is the mean of the average precision scores for each query.

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

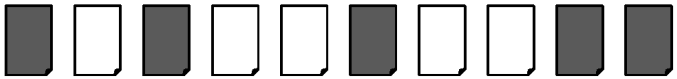
where *Q* is the number of queries

- **Summarize rankings from multiple queries by averaging average precision**
- **This is the *most commonly used* measure in research papers**
- **Assumes user is interested in finding many relevant documents for each query**
- **Requires many relevance judgments in text collection**

Mean Average Precision Example


= relevant documents for query 1

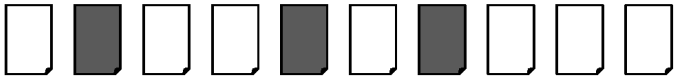
Ranking #1



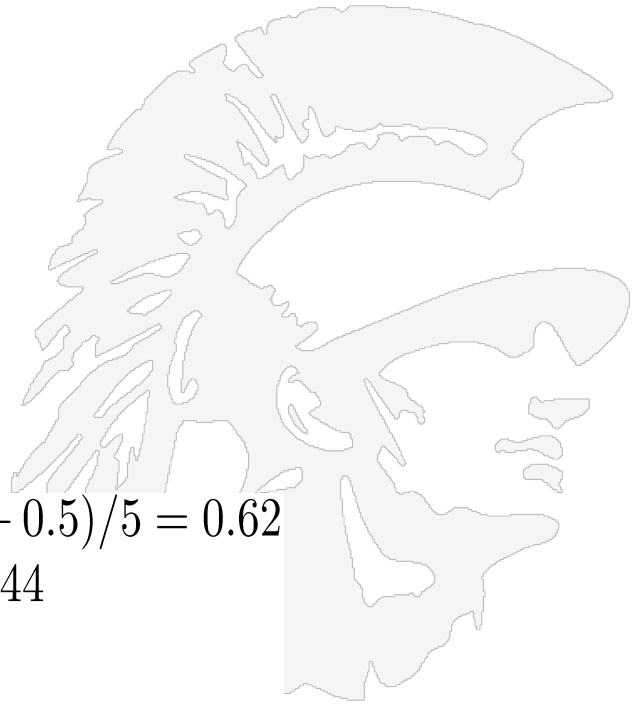
Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5


= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

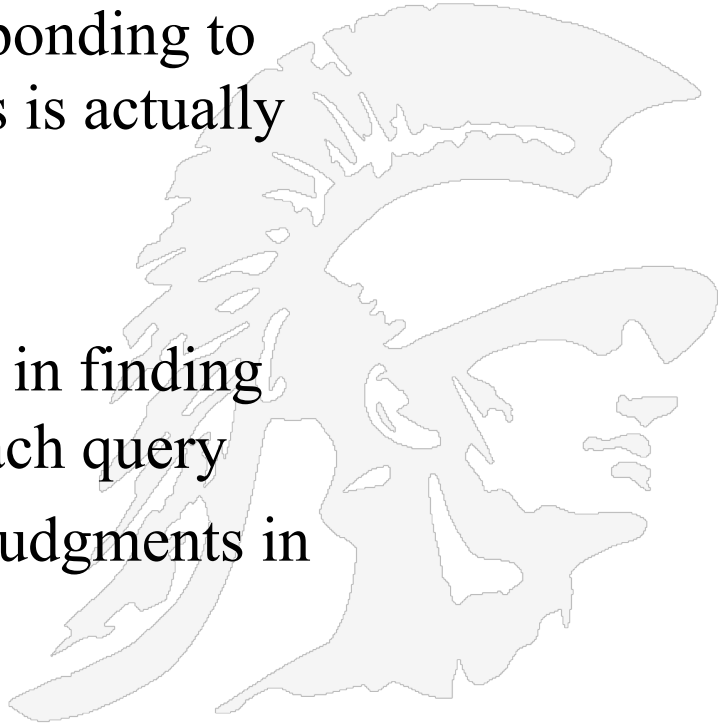


$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

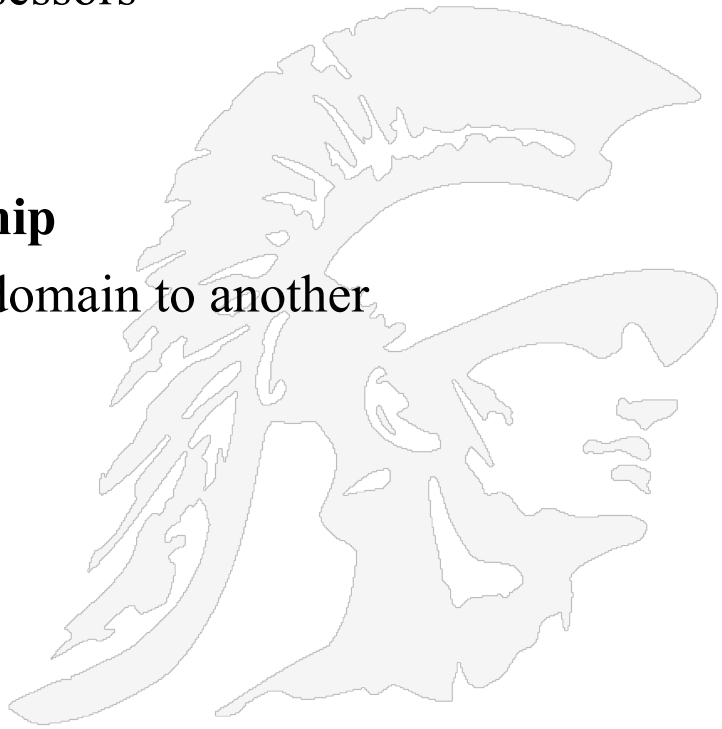
More on Mean Average Precision Calculation

- **Mean Average Precision (MAP)**
  - Some negative aspects
  - If a relevant document never gets retrieved, we assume the precision corresponding to that relevant doc to be zero (this is actually reasonable)
  - Each query counts equally
  - MAP assumes user is interested in finding many relevant documents for each query
  - MAP requires many relevance judgments in the document collection



Difficulties in Using Precision/Recall

- **Should average over large document collection and query ensembles**
- **Need human relevance assessments**
  - But people aren't always reliable assessors
- **Assessments have to be binary**
  - Nuanced assessments?
- **Heavily skewed by collection/authorship**
  - Results may not translate from one domain to another



A Final Evaluation Measure: Discounted Cumulative Gain

- The premise of DCG is that highly relevant documents appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the position of the result.
- The discounted CG accumulated at a particular rank position *p* is defined as

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

where rel_i is the graded relevance of the result at position *i*

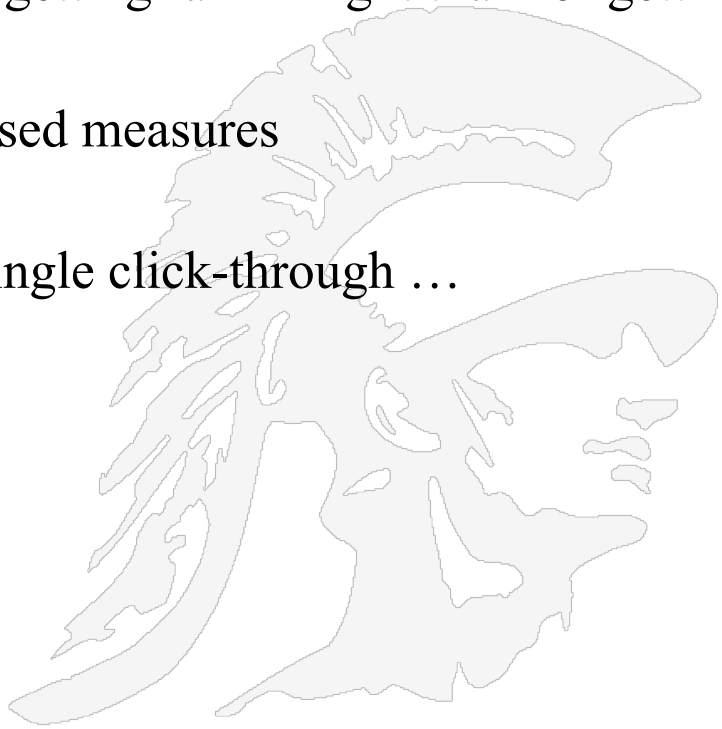
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/log *(rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3
- An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

How Evaluation is Done at Web Search Engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top *k* positions, e.g., *k = 10*
- . . . or measures that reward you more for getting rank *1* right than for getting rank *10* right.
- Search engines also use non-relevance-based measures
  - Click-through on first result
    - Not very reliable if you look at a single click-through …

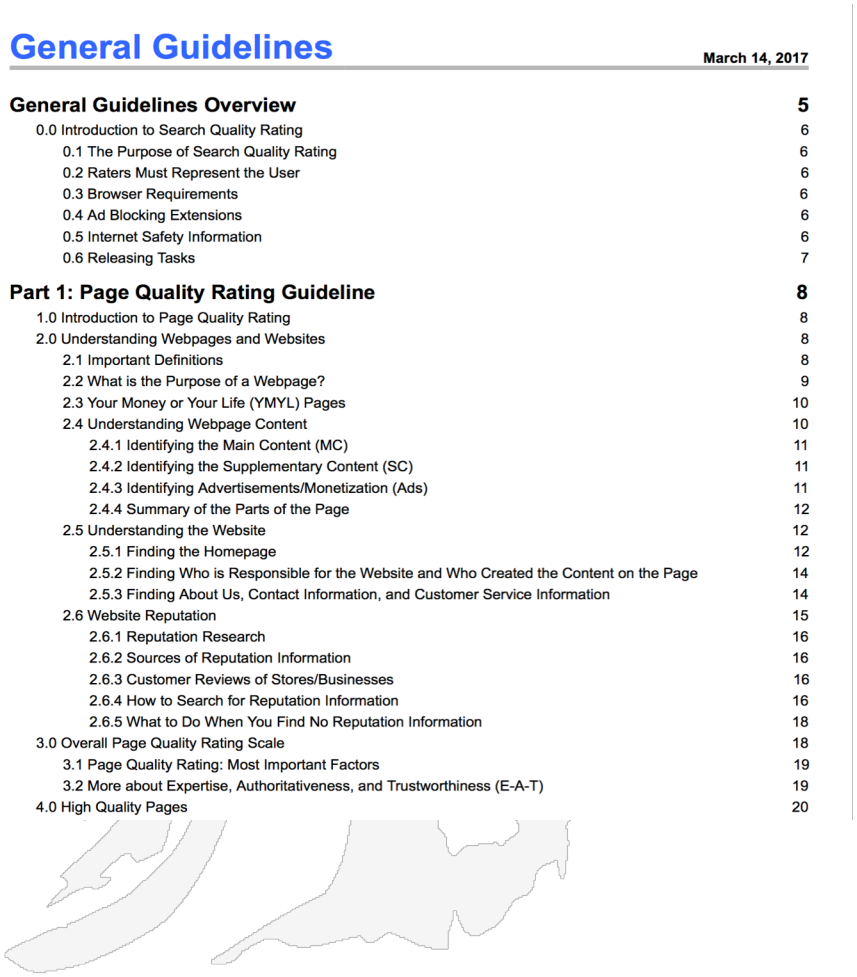
    but pretty reliable in the aggregate.
  - Studies of user behavior in the lab
  - A/B testing



Google's Search Quality Rating Guidelines Document

- **Google relies on raters, working in many countries and languages around the world**
- **The data they generate is rolled up statistically to give**
  - a view of the quality of search results and search experience over time, and
  - an ability to measure the effect of proposed changes to Google's search algorithms

http://www-scf.usc.edu/~csci572/papers/searchqualityevaluatorguidelines.pdf

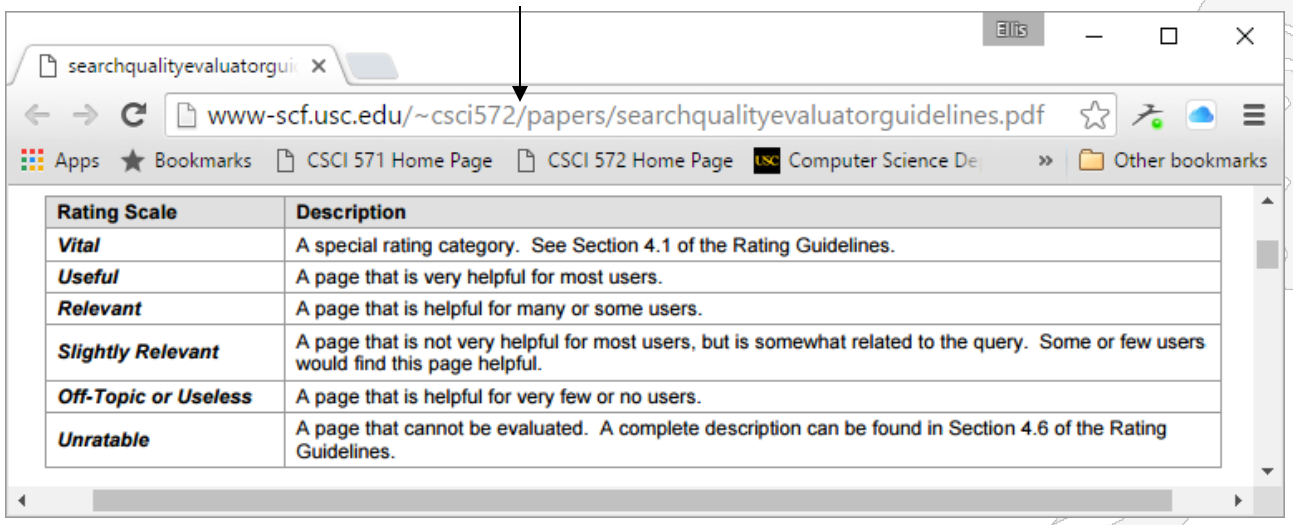


General Guidelines March 14, 2017

Google's Search Quality Ratings Guidelines Document

- This document gives evaluators examples and guidelines for appropriate ratings.
- the evaluator looks at a search query and a result that could be returned. They rate the relevance of the result for that query on a scale described within the document.

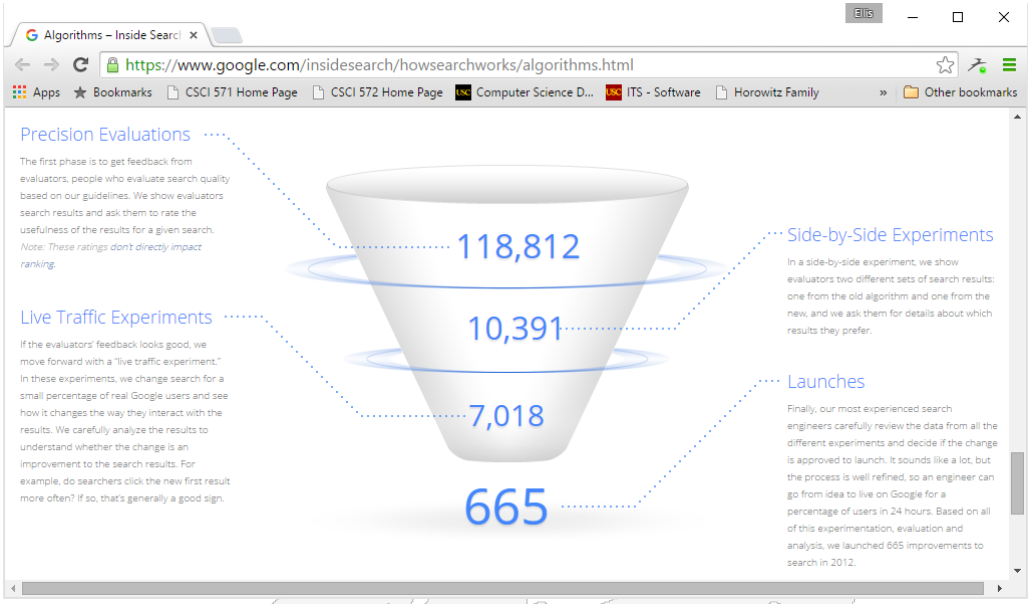
The six rating scale categories



Rating Scale	Description
Vital	A special rating category. See Section 4.1 of the Rating Guidelines.
Useful	A page that is very helpful for most users.
Relevant	A page that is helpful for many or some users.
Slightly Relevant	A page that is not very helpful for most users, but is somewhat related to the query. Some or few users would find this page helpful.
Off-Topic or Useless	A page that is helpful for very few or no users.
Unratable	A page that cannot be evaluated. A complete description can be found in Section 4.6 of the Rating Guidelines.

Google's 4-Step Process for Changing Their Search Algorithm

1. **Precision Evaluations**
   People use the Guidelines to rate search results
2. **Side-by-Side Experiments**
   people are shown two different sets of search results and asked which they prefer
3. **Live Traffic Experiments**
   the search algorithm is altered for a small number of actual users
4. **Full Launch**
   A final analysis by Google engineers and the improvement is released



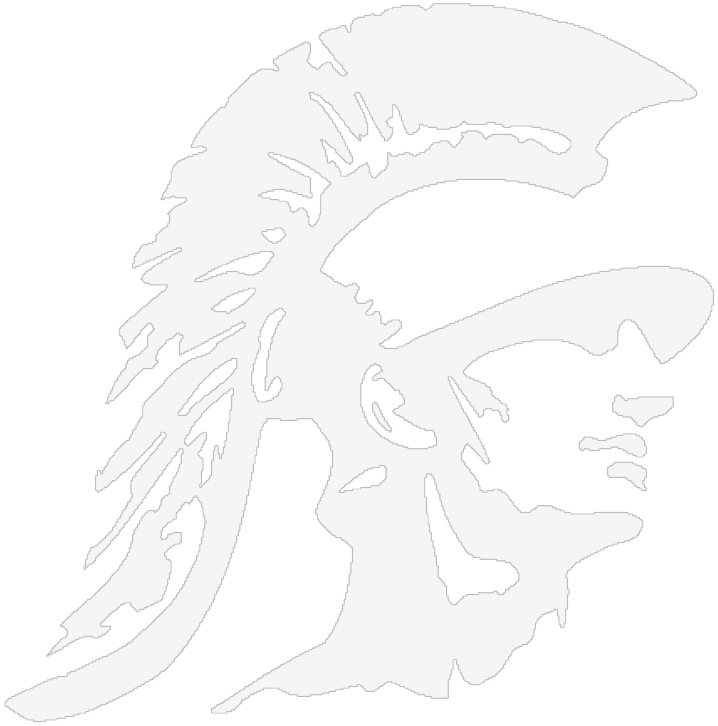
A/B Testing at Web Search Engines

- **A/B testing** is comparing two versions of a web page to see which one performs better. You compare two web pages by showing the two variants (let's call them **A** and **B**) to similar visitors at the same time. The one that gives a better conversion rate, wins!

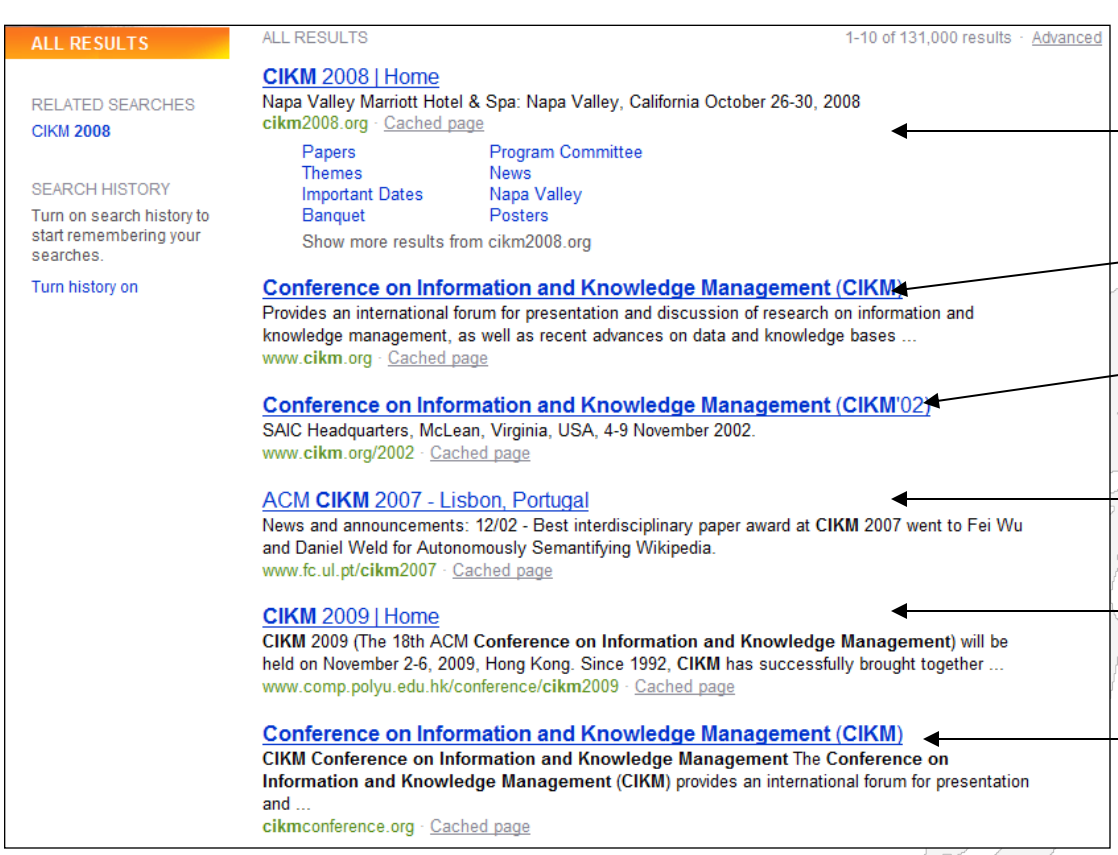
1. **Purpose:** Test a single innovation
2. **Prerequisite:** You have a large search engine up and running.
3. **Have most users use old system**
4. **Divert a small proportion of traffic (e.g., 1%) to an experiment to evaluate an innovation**
5. **Evaluate with an automatic measure like click through on first result**

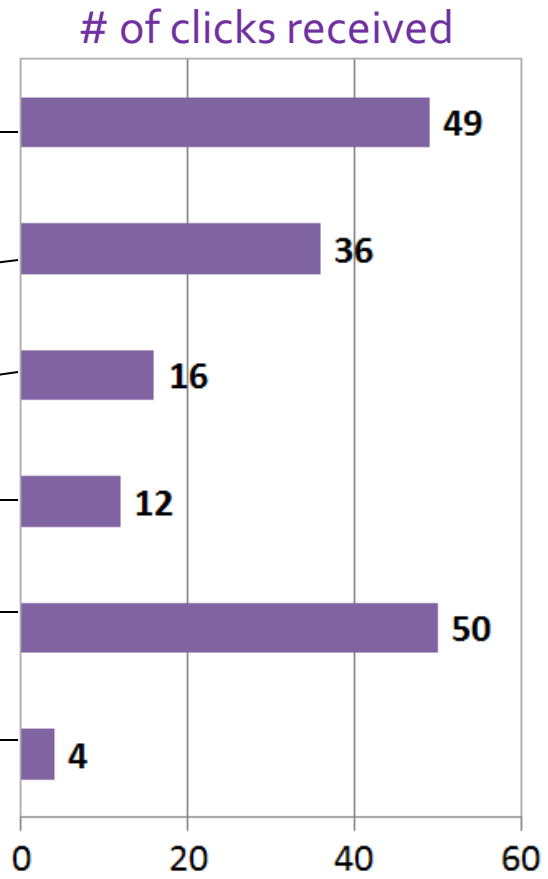
- we directly see if the innovation does improve user happiness
- *This is the evaluation methodology large search engines trust the most*

USING USER CLICKS FOR EVALUATION



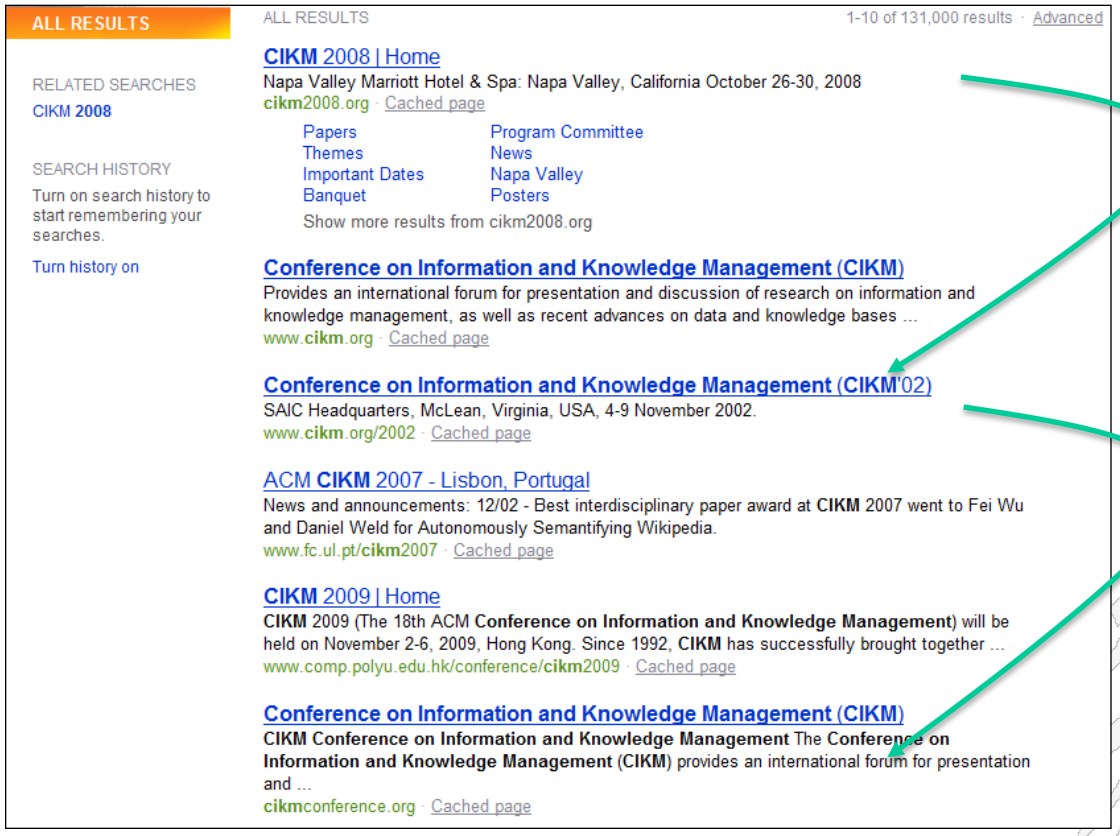
What Do Clicks Tell Us?





There is strong position bias, so absolute click rates unreliable

Relative vs Absolute Ratings



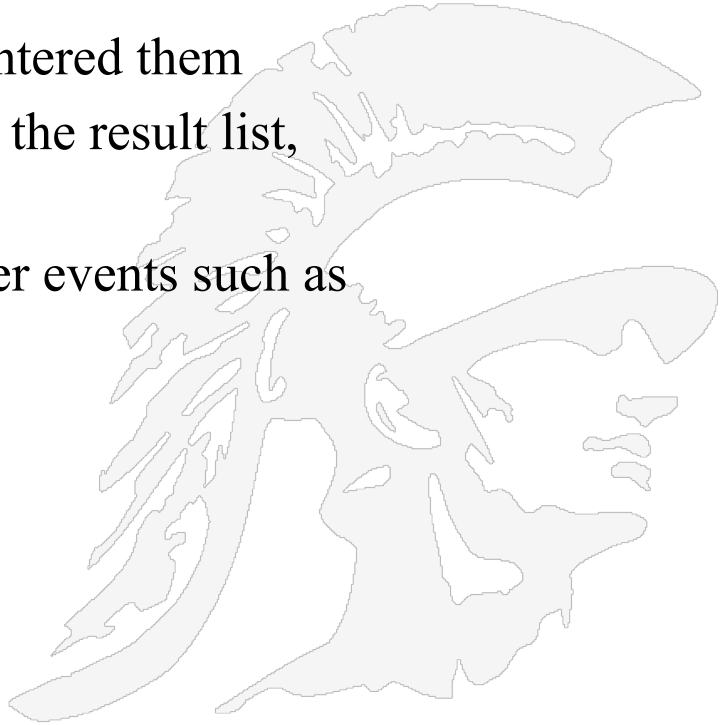
User's click sequence

Hard to conclude Result1 > Result3

Probably can conclude Result3 > Result2

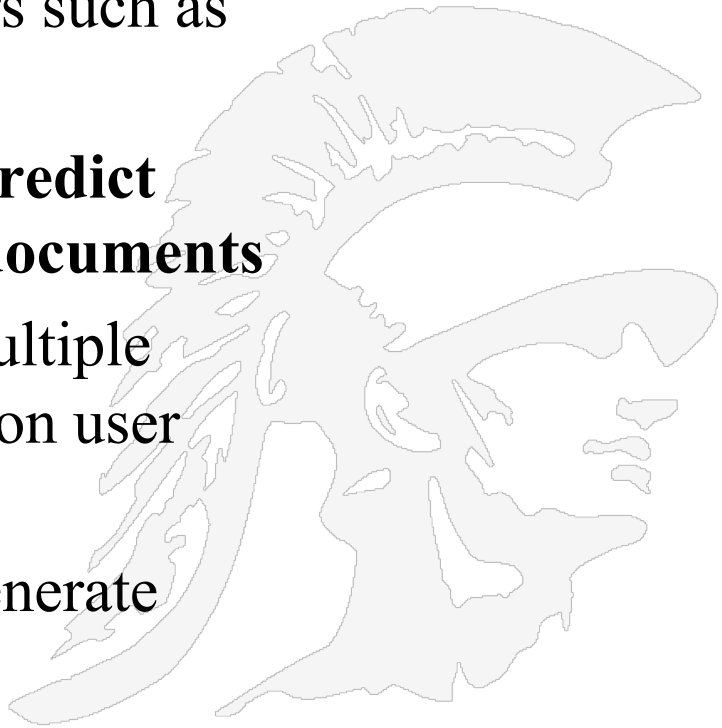
Query Logs

- **Used for both tuning and evaluating search engines**
  - also for various techniques such as query suggestion
- **Typical contents of the query log files**
  - User identifier or user session identifier
  - Query terms - stored exactly as user entered them
  - List of URLs of results, their ranks on the result list, and whether they were clicked on
  - Timestamp(s) - records the time of user events such as query submission, clicks



How Query Logs Can Be Used

- **Clicks are not relevance judgments**
  - although they are correlated
  - biased by a number of factors such as rank on result list
- **Can use clickthough data to predict *preferences* between pairs of documents**
  - appropriate for tasks with multiple levels of relevance, focused on user relevance
  - various "policies" used to generate preferences



A Final Thought
Google's Enhancements for Good Search Results

Google's list of Search Result elements includes:

- immediate answers
- autocomplete
- results from books
- freshest information
- Google instant
- images in results
- knowledge graph
- design for mobile
- latest news from newspapers
- query understanding
- advanced search
- safe search
- search by image
- search by voice
- site and page quality
- snippets
- spelling
- synonyms
- translations
- universal search
- user context
- videos

The page below discusses the many aspects that go into producing search results at Google

https://www.google.com/insidesearch/howsearchworks/algorithms.html