Information Retrieval
Part 2

Günter Neumann
LT lab, DFKI

(Using slides from Raymond Mooney's IR course
http://www.cs.utexas.edu/users/mooney/ir-course/)

Evaluation

What to Evaluate?

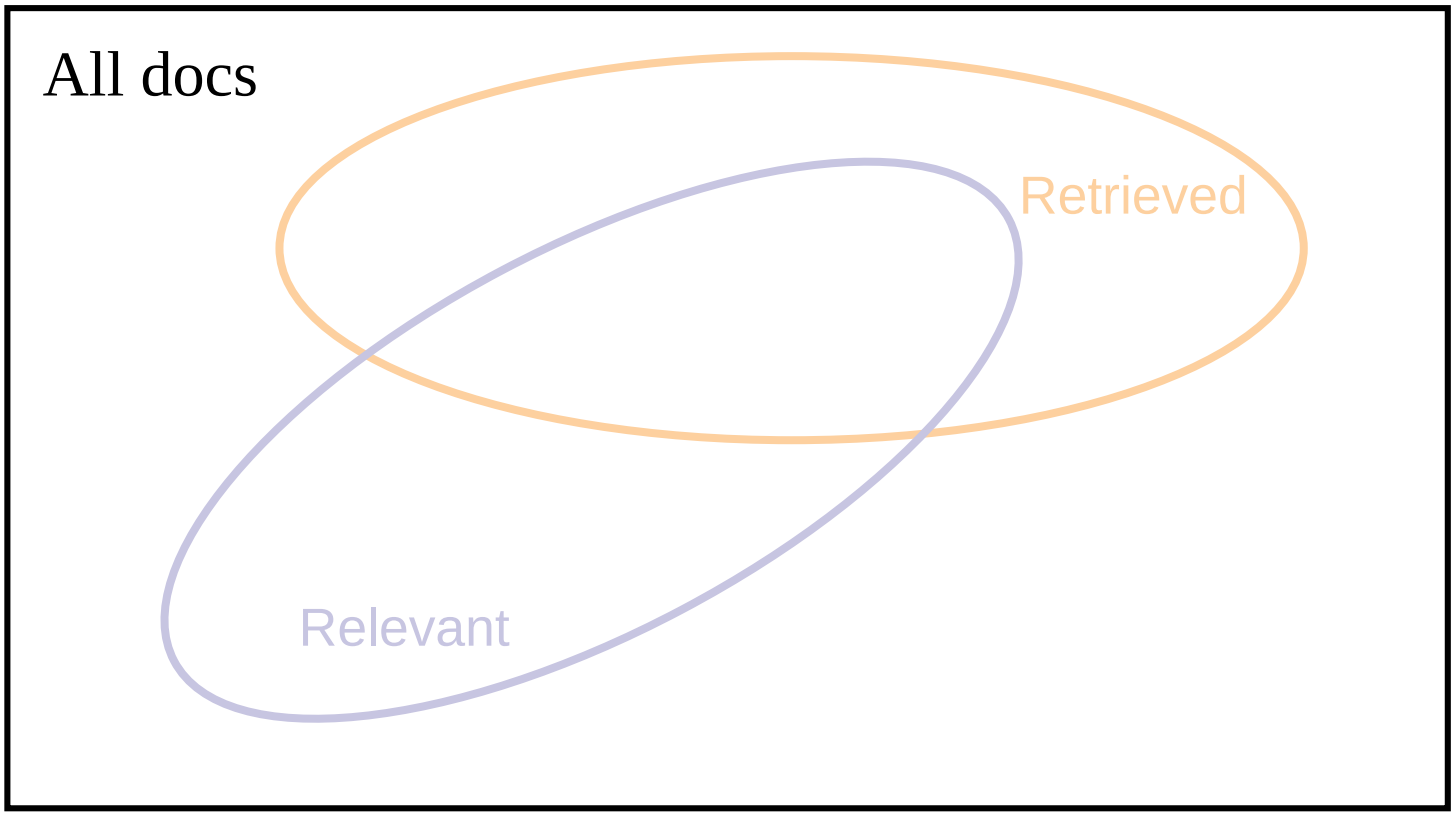
What can be measured that reflects users' ability to use system? (Cleverdon 66)

- Coverage of Information
- Form of Presentation
- Effort required/Ease of Use
- Time and Space Efficiency

effectiveness

- Recall
  - proportion of relevant material actually retrieved
- Precision
  - proportion of retrieved material actually relevant

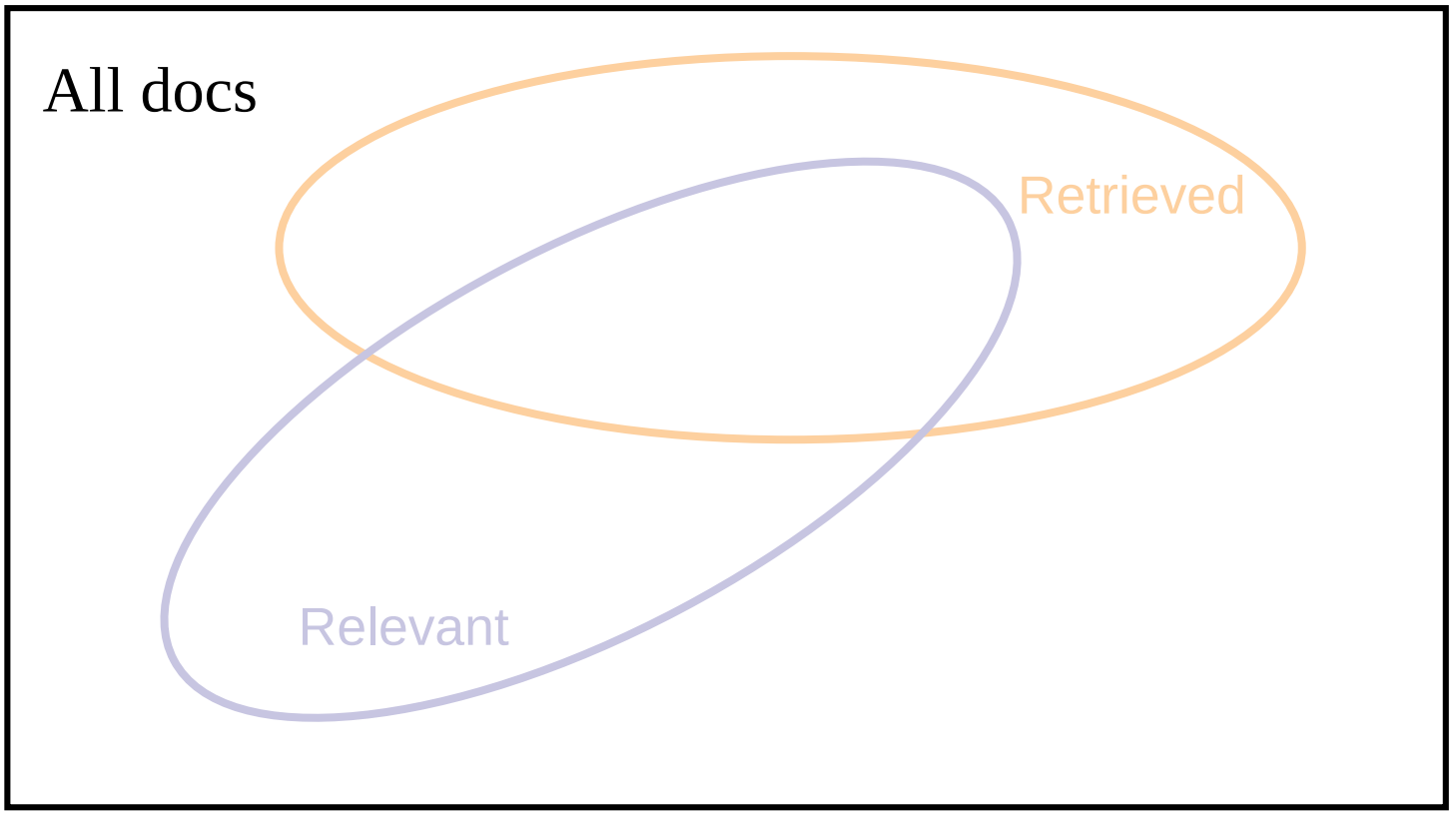
Relevant vs. Retrieved



Precision vs. Recall

$$\text{Precision} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Retrieved}\,|}$$

$$\text{Recall} = \frac{|\,\text{RelRetrieved}\,|}{|\,\text{Rel in Collection}\,|}$$

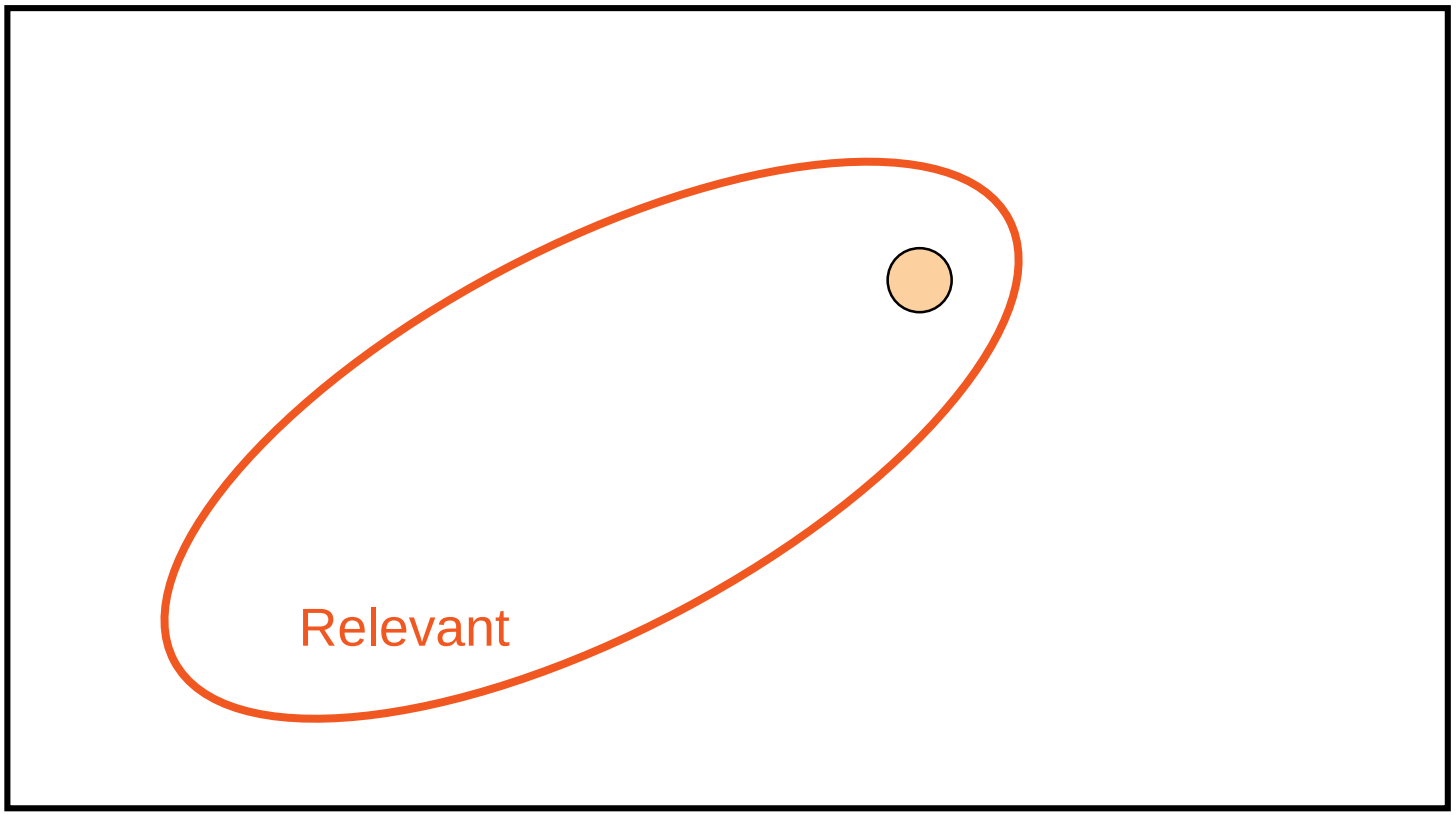


Why Precision and Recall?

Get as much good stuff while at the same time getting as little junk as possible.

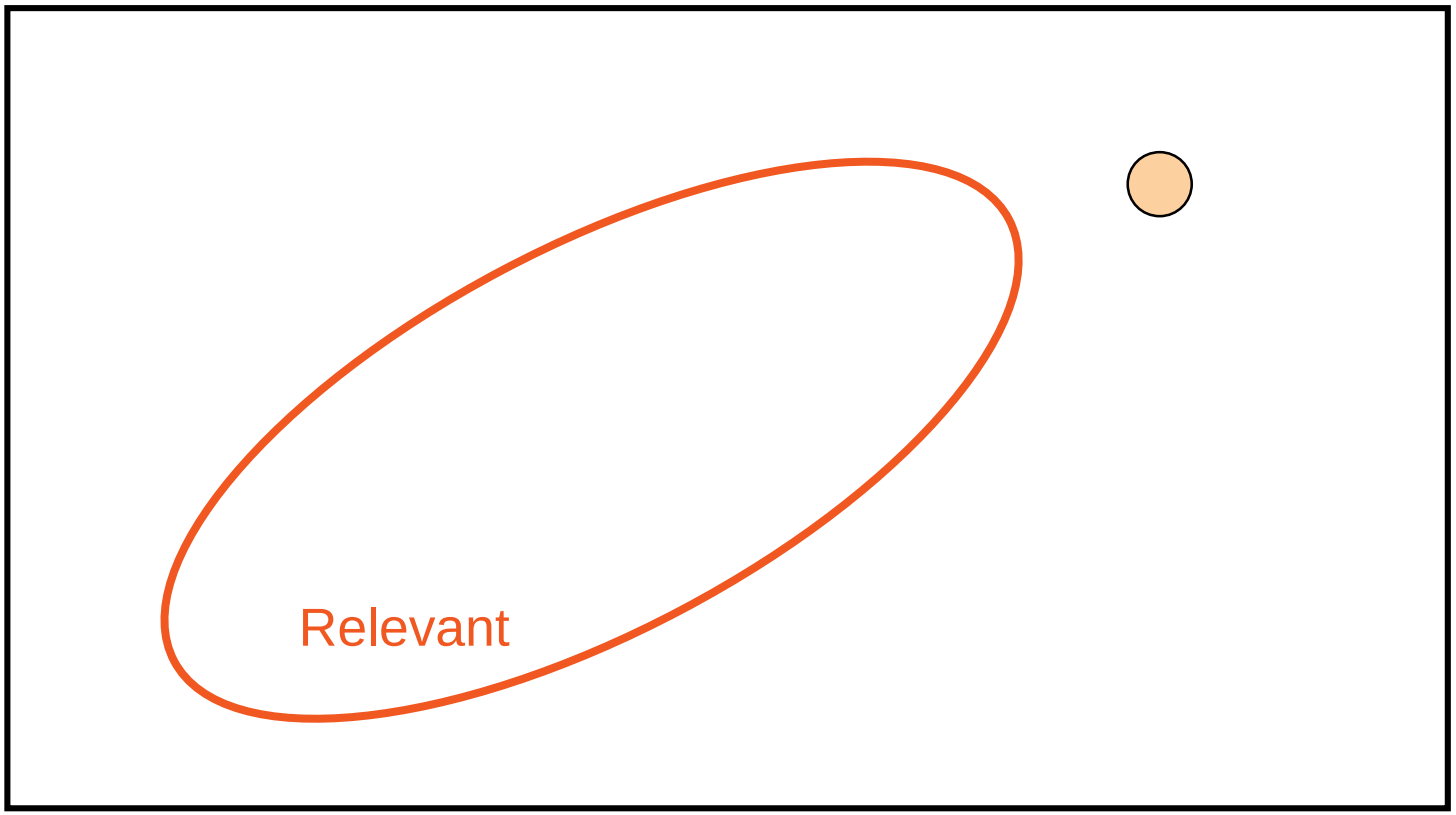
Retrieved vs. Relevant Documents

Very high precision, very low recall



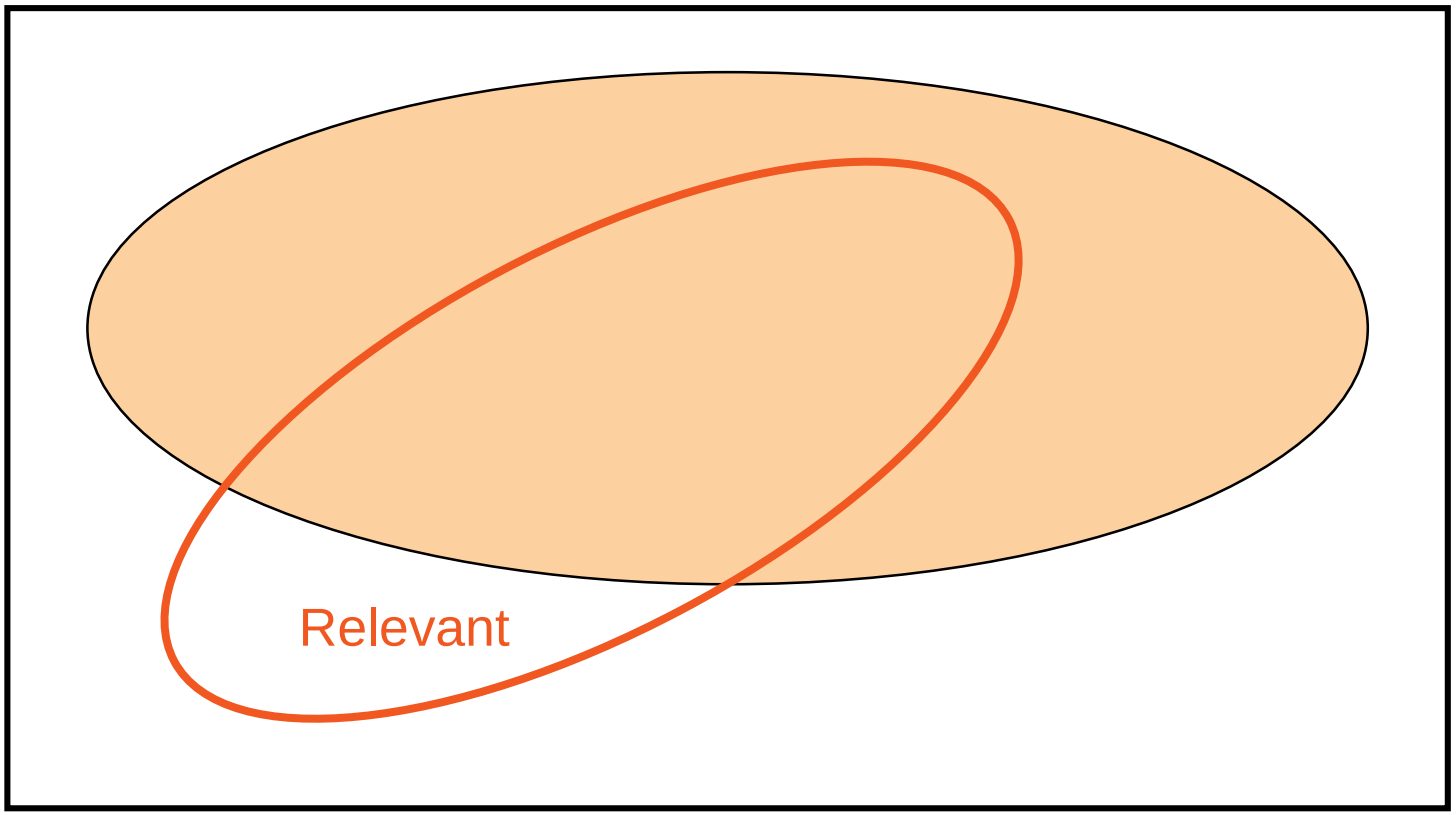
Retrieved vs. Relevant Documents

Very low precision, very low recall (0 in fact)



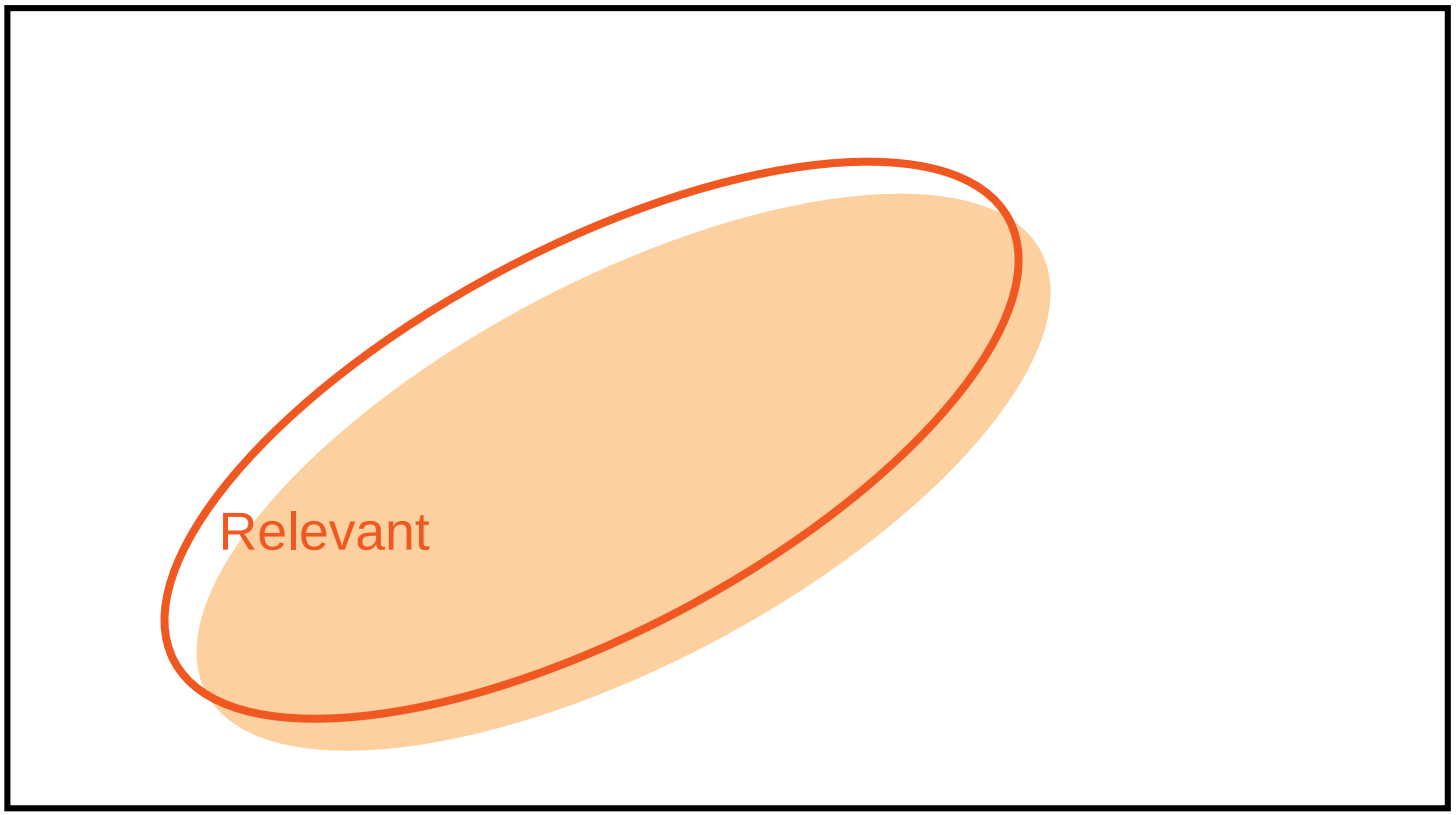
Retrieved vs. Relevant Documents

High recall, but low precision



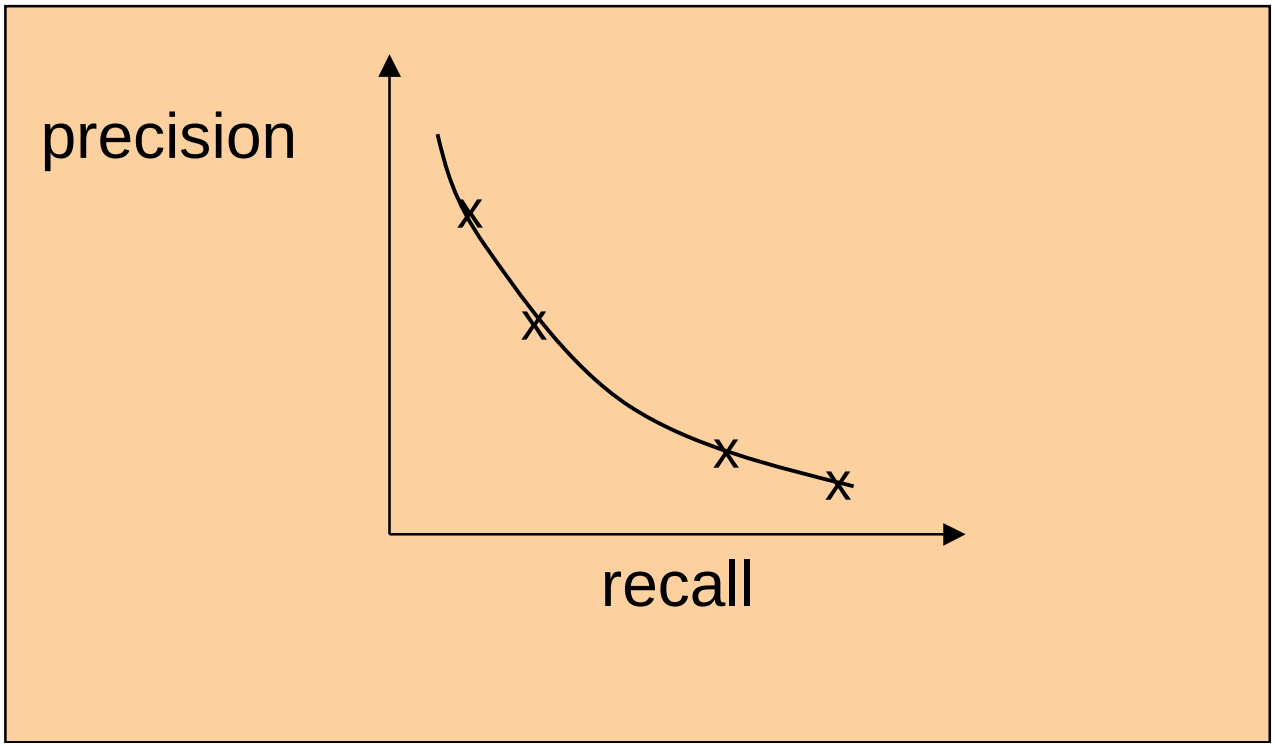
Retrieved vs. Relevant Documents

High precision, high recall (at last!)



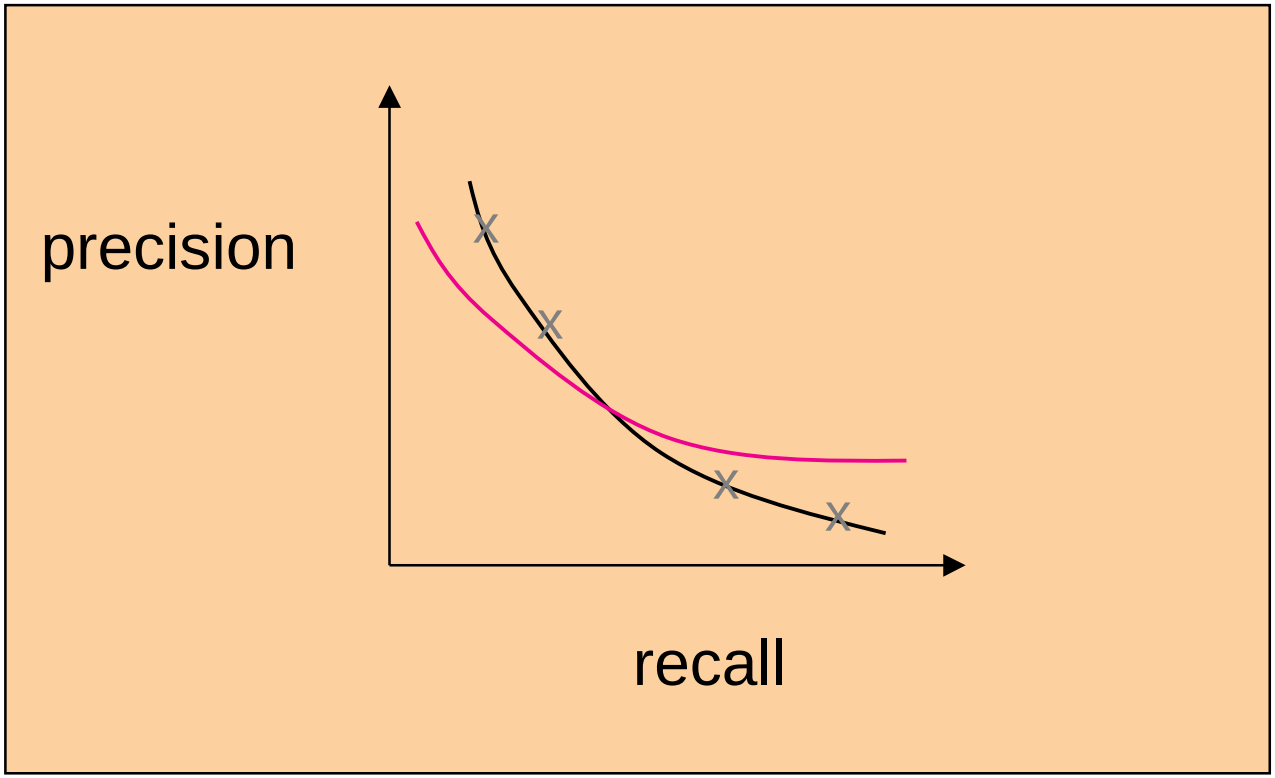
Precision/Recall Curves

- There is a tradeoff between Precision and Recall
- So measure Precision at different levels of Recall
- Note: this is an AVERAGE over MANY queries

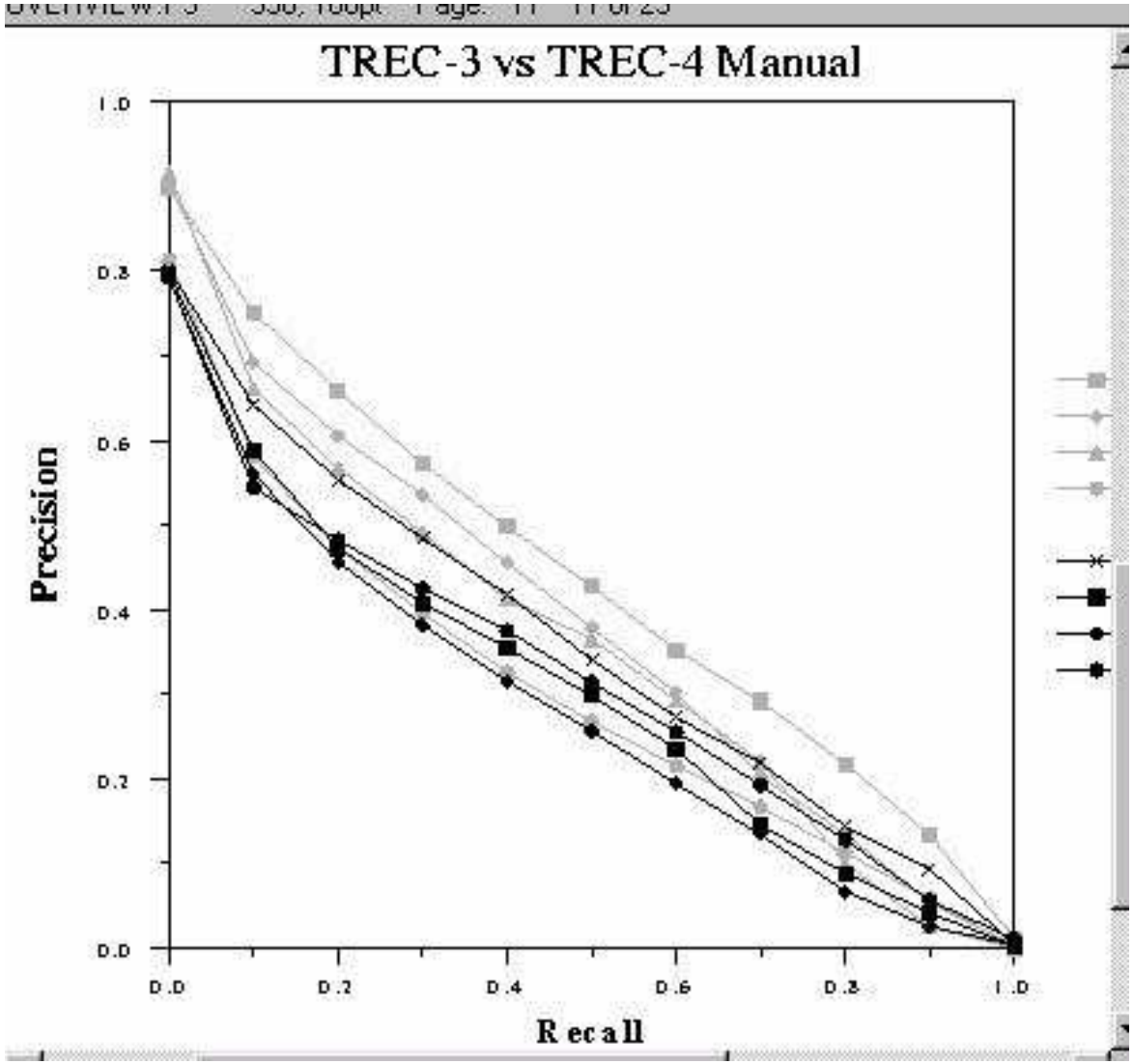


Precision/Recall Curves

- Difficult to determine which of these two hypothetical results is better:



Precision/Recall Curves



Document Cutoff Levels

- Another way to evaluate:
  - Fix the number of documents retrieved at several levels:
    - top 5
    - top 10
    - top 20
    - top 50
    - top 100
    - top 500
  - Measure precision at each of these levels
  - Take (weighted) average over results
- This is a way to focus on how well the system ranks the first k documents.

Problems with Precision/Recall

- Can’t know true recall value
  - except in small collections
- Precision/Recall are related
  - A combined measure sometimes more appropriate
- Assumes batch mode
  - Interactive IR is important and has different criteria for successful searches
- Assumes a strict rank ordering matters.

Relation to Contingency Table

	Doc is Relevant	Doc is NOT relevant
Doc is retrieved	a	b
Doc is NOT retrieved	c	d

- Accuracy: (a+d) / (a+b+c+d)
- Precision: a/(a+b)
- Recall: ?
- Why don't we use Accuracy for IR?
  - (Assuming a large collection)
  - Most docs aren't relevant
  - Most docs aren't retrieved
  - Inflates the accuracy value

F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

E Measure (parameterized F Measure)

- A variant of F measure that allows weighting emphasis on precision over recall:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- Value of β controls trade-off:
  - β = 1: Equally weight precision and recall (E=F).
  - β > 1: Weight recall more.
  - β < 1: Weight precision more.

Issues Regarding Keyword-Based Searches

- Information Overload
- Multiple Vocabularies
- Synonymy
- Polysemy

Issues Regarding Keyword-Based Searches (Cont.)

Information Overload

- The need for *effective information retrieval systems* becomes increasingly important as computer-based information repositories grow larger and more diverse.
- Information overload is when the users of system are overwhelmed with the amount of current information available, the constant updating of new information, and usually *not enough knowledge about the subject and system* required to access the information

Issues Regarding Keyword-Based Searches (Cont.)

Multiple Vocabularies

- Users also must deal with multiple vocabularies, which arise from the *varying backgrounds and expertise of the users accessing the system*. This often leads to poor recall, where the set of data that the user is searching on is to large, or to small, which in turn leads to low precision on the information the user wants.

Issues Regarding Keyword-Based Searches (Cont.)

Synonymy

- Synonymy deals with the *equivalence of meaning* of words, where different words mean the same thing, or synonyms.
- An example would be searching through the yellow pages for a particular type of business.

Issues Regarding Keyword-Based Searches (Cont.)

Polysemy

- Polysemy deals with the polar opposite problem, where single words have *multiple meanings* depending on the context in which they are used.

- An example would be searching for the word “pen”, which depending on the context used could mean a writing instrument or jail

Alternative To Keyword-Based Search Engines

- Concept-based information retrieval looks at the increasing problem with keyword-based searches with an intuitive approach, which first *sorts the data by their relationships*, and then searches the data for specific information.
- In sorting the data first by how they are interrelated, we can get a *good recall* on the set of data that we want, and in searching on this set of related data for specific information, we can get high precision.
- The research into concept-based IR (information retrieval) as a viable alternative is *still in its infancy* though

Concept-Based Searches: Two Distinct Fields

- Concept-based IR is based upon two distinct fields: Taxonomy and Ontology
- Taxonomy: Division into *ordered groups of categories*. The science, laws, or principles of classification.
- Ontology: The branch of metaphysics that deals with the nature of being. The *relationships between objects* in the real world.

Concept-Based Searches: Two Distinct Fields (Cont.)

- Taxonomy deals with the *classification of data*, where the different objects belong under certain categories depending on their characteristics.
- An example would be the term *laptop*, which would fall under the category of computers→hardware→portable devices.

Concept-Based Searches: Two Distinct Fields (Cont.)

- Ontology deals with the relationship between objects in the real world, such as aggregation (“part-of” relationship such as “an engine is part of an automobile”) and inheritance (“is-a” relationship such as “a human is a mammal”).
- Ontology answers the question, “What kinds of objects exist in one or another domain of the real world and how are they interrelated?”.

Concept-Based Searches: Two Distinct Fields (Cont.)

- The two basic concepts behind ontology are *types and roles*, where types are instances of an object which always exhibit certain features, while roles are instances of an object which change depending on circumstance.
- An example of a type would be a plant, which exhibits certain features such that it will always be a plant its entire lifetime.
- A role would be a student at a university, where even after they graduate, after they cease to be a student, they are still an individual.

Concept-Based Searches: Two Distinct Fields (Cont.)

- It is important to separate these two fields due to the fact that objects that belong to a certain concept can be classified in very different ways depending on the viewpoint, where the object can be looked at in different ways depending on the user
- The viewpoint corresponds to a role of the object, so there is a one to one mapping between the user and the function of the object.

Concept-Based Searches: Two Distinct Fields (Cont.)

- The way to sort the data in a way which has the advantages of both an ontology and taxonomy is to create a concept map of the current information.
- A concept map is a visual knowledge representation technique, which is used to express relationships between ideas.
- Examples of where concept maps are utilized include brainstorming, planning, documentation, presentation and software blueprints.

Concept-Based Searches

- A practical example of the concept mapping technique (used for the purpose of relationships between objects) would be the semantic network, where the nodes in the directed graph are the ideas and the links are the relationships between them
- Building such a conceptual network of ideas and objects would give the user a good recall on the information requested (based on idea clusters), and from the information given back to the user, they would choose which of the branches are most relevant to them. Then the user would search the relevant branch of concepts, where the information in the branches is stored in a categorical fashion

Concept-Based Searches (Cont.)

- Currently, there is an effort to sort data on the web (defined and linked) in a way that *it can be used by machines* - not just for display purposes, but for using it in various applications.
- The name of the effort is “The Semantic Web” (http://www.semanticweb.org/)
- They are employing both the fields of ontology and taxonomy to solve the problems

Concept-Based Searches (Cont.)

- Another example of a concept mapping technique (used for the purpose of relationships between objects) in a search engine would be the Information Mapping Project at Stanford.
- http://www-csli.stanford.edu/semlab/infomap.html (Homepage)
- http://infomap.stanford.edu/webdemo (search engine)

Conclusion

- The volume of information will only increase, so there must be methods established to harvest the data into a coherent form, where a user can find the relevant and interesting information they are looking for with accurate results.
- Combining the two fields of ontology and taxonomy is a good approach to the concept-based model of information retrieval.