Information Retrieval System Evaluation
October 3, 2012

1. **Module name**: Information Retrieval System Evaluation
2. **Scope**: The module introduces the evaluation in information retrieval. It focuses on the standard measurement of system effectiveness through relevance judgments.
3. **Learning objectives**
   Students should be able to:
   a. Understand relevance judgments and the techniques applied to evaluate unranked and ranked IR systems.
   b. Evaluate an IR system given document collection with information needs and interpret the results.
4. **5S characteristics of the module**
   a. Streams: Relevance judgment results are "ground truths" of test collections. Lucene's "quality" package takes "ground truth", queries and indexed search results as inputs to produce search quality results.
   b. Structures: The built-in formats of queries and topics in the Lucene benchmark is based off of the format of the TREC corpus. Documents in LucidWorks are indexed in XML format.
   c. Spaces: The ADI documents are located on the server running LucidWorks software.
   d. Scenarios: Relevance judgments are required to build test collections given document collection and information needs. Test collections are then used to evaluate information retrieval systems.
   e. Society: Students studying the evaluation approaches of information retrieval.
5. **Level of effort required**

   This module should take at least 4 hours to complete.
   a. **Out-of-class:** students are expected to spend at least 4 hours to complete the module and exercises. Time should be spent reading the material from the textbook and Lucene chapters, as well as revisiting the relevant lecture [12e, 12f, 12d].
   b. **In-class:** students will have the opportunity to ask and discuss exercises with their teammates.
6. **Relationships with other modules** (flow between modules)
   a. Overview of Lucidworks big data software module

      Lucidworks overview module introduces the software and provides instructions to learn. This module requires the Lucidworks software to perform the exercises.

b. LucidWorks: Searching with cURL module

   LucidWorks searching module introduces the searching with cURL using LucidWorks. This module requires the students to search documents in test collection using LucidWorks.

7. Prerequisite knowledge/skills required (what the students need to know prior to beginning the module; completion optional; complete only if prerequisite knowledge/skills are *not* included in other modules)

a. Knowledge of probability in statistics

b. Background knowledge of standard relevance benchmarks

8. Introductory remedial instruction

The following relevance benchmarks are widely used as the most standard test collections.

a. TREC (*Text Retrieval Conference*)

   i. TREC ad hoc tracks: 50 information needs

   ii. TREC web track

b. GOV2

   i. GOV2 web page collection: 25 million page

   ii. The largest Web collection easily available

c. NTCIR (*NII Test Collections for IR Systems*)

   i. East Asian languages and cross-language information retrieval

d. CLEF (*Cross Language Evaluation Forum*)

   i. European languages and cross-language information retrieval

9. Body of knowledge

If students wish to study more advanced work, see section (f).

a. Measuring the effectiveness of IR systems

   A test collection is needed to measure information retrieval effectiveness. It consists of three things:

   i. A document collection

   ii. A test suite of information needs, expressible as queries

      E.g., **Information Need**: “Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine”. **Query**: “wine AND red AND white AND heart AND attack AND effective”.

   iii. A set of relevance judgments

      Relevance is assessed relative to an information need, **not** a query, i.e. a document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query.

b. Evaluation of unranked retrieval sets

i. **Precision**: fraction of retrieved documents that are relevant

$$P = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

ii. **Recall**: fraction of relevant documents that are retrieved

$$R = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})} = P(\text{retrieved}|\text{relevant})$$

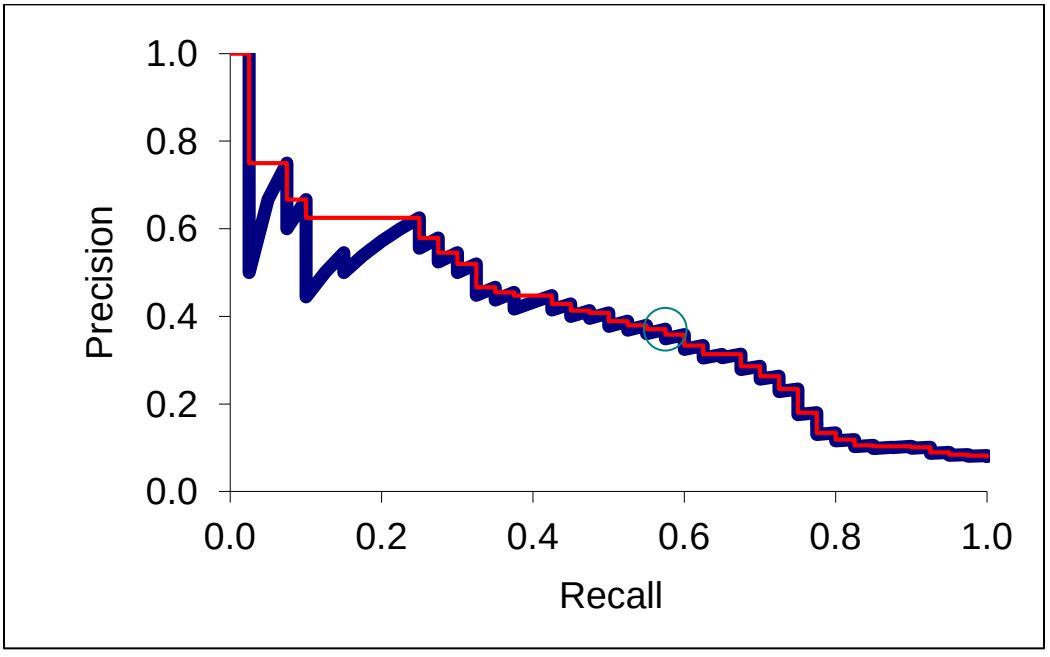
iii. **F measure**: weighted harmonic mean of precision and recall

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \text{ where } \beta^2 = \frac{1-\alpha}{\alpha}, \alpha \in [0,1]$$

c. Evaluation of ranked retrieval results

i. Precision-recall curve

Precision-recall curves plot precision and recall against each other. They have a distinctive saw-tooth shape: if the *(k+1)*th document retrieved is nonrelevant then recall is the same as for the top *k* documents, but precision has dropped. If it is relevant, then both precision and recall increase, and the curve jags up and to the right, shown in blue below. It is often useful to remove these jiggles and the standard way to do this is with an interpolated precision. **Interpolated precision** at certain recall level *r* is defined as the highest precision found for any recall level: $r' \geq r: p_{interp}(r) = \max_{r' \geq r} p(r')$, show in red below.



ii. 11-point interpolated average precision

Take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation, and average them.

iii. Mean average precision (**MAP**)

Average of the precision value obtained for the top *k* documents, each time a relevant doc is retrieved.

$$\text{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\text{Precision}(R_{jk})$$

iv. ROC curve

Plot the true positive rate or sensitivity against the false positive rate.

v. R-precision

If have known set of relevant documents of size *Rel*, then calculate precision of top *Rel* documents returned.

vi. Normalized Discounted Cumulative Gain (NDCG)

For a set of queries *R*, let *R(j, d)* be the relevance score assessors gave to document *d* for query *j*. Then,

$$\text{NDCG}(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}Z_{kj}\sum_{m=1}^{k}\frac{2^{R(j,m)}-1}{\log_2(1+m)}$$

Where Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at query *j* is 1.

d. Accessing relevance

i. Pooling

Relevance is assessed over a subset of the collection that is formed from the top *k* documents returned by a number of different IR systems.

ii. **Kappa statistic**

A common measure for agreement between judges. It is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement.

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Where *P(A)* is the proportion of the times the judges agreed, and *P(E)* is the proportion of the times they would be expected to agree by chance.

iii. Marginal relevance

Whether a document still has distinctive usefulness after the user has looked at certain other documents.

e. Refining a deployed system

i. A/B testing

Purpose: Test a single innovation

Prerequisite: You have a large search engine up and running

Have most users use old system

Divert a small proportion of traffic to the new system including the innovation

Evaluate with an "automatic" measure to see if the innovation has a positive or negative effect.

f. Evaluating search quality using Lucene benchmark

10. Exercises / Learning activities

a. **Relevance Judgments.** You are given a document collection and a set of information needs. Please do the relevance judgments with the members in your team and provide answers to the following questions.

The ADI test collection is a relatively small benchmark containing 82 documents and 35 information needs. The documents collection is available at

http://fox.cs.vt.edu/VAD1/DOWN/IRCOLLS/ADI.ALL

It is also loaded in the LucidWorks software:

http://fetcher.dlib.vt.edu:8888/solr/#/test_collection_vt

i. Each student judges the relevance of all of the 82 documents in the ADI collection independently. The information need is: "What is information science? Give definitions where possible." Please list the judgment results of each student in your team.

ii. Fill in the following table with your judgment results and calculate the kappa statistics. (Note: for a team consisting of more than two students, calculate an average pair-wise kappa value.)

	Judge "name1" relevance			
Judge "name2" relevance		Yes	No	Total
	Yes			
	No			
	Total			

iii. Based on the kappa statistic result from step iii, please discuss if your team has achieved a good/fair/bad agreement and the reasons why the disagreement exists.

iv. In the test collection, the number of relevant documents for this information need is 3, whose ids are 43, 45, and 60. Compare your own relevance judgments with this and summarize the differences if any.

b. **Unranked retrieval sets evaluation.** Please use your own relevance judgment results to answer the following questions.

i. Search in the ADI document collection using the following query. Get all of the documents that can be retrieved.

information AND (science OR definition)

Hint: (The following command retrieves all the documents in the collection that matches this query.)

curl -u username:password -verbose -X POST -H 'Content-type: application/json' -d '{"query":{"q":"information AND (science OR definition)","rows":9}}' http://fetcher.dlib.vt.edu:8341/sda/v1/client/collections/test_collection_vt/documents/retrieval | python -mjson.tool

This query expresses the information need in Exercise a.

Please list ids of the retrieved documents.

ii. What is the precision of the system on this search, and what is its recall?

iii. What is the balanced F measure?

c. **Ranked retrieval results evaluation**. (Exercises are derived from Exercise 8.8 and 8.9 in the textbook)

i. Consider an information need for which there are 4 relevant documents in the collection. Contrast two systems run on this collection. Their top 10 results are judged for relevance as follows (the leftmost item is the top ranked search result):

System 1 R N R N N N N N R R

System 2 N R N N R R R N N N

What is the MAP of each system? Which has higher MAP?

What is the R-precision of each system? (Does it rank the systems the same as MAP?)

ii. The following list or Rs and Ns represents relevant (R) and nonrelevant (N) returned documents in a ranked list of 20 documents retrieved in response to a query from a collection of 10,000 documents. The top of the ranked list (the document the system thinks is the most likely to be relevant) is on the left of the list. The list shows 6 relevant documents. Assume that there are 8 relevant documents in total in the collection.

R R N N N N N N R N R N N N R N N N N R

What is the precision of the system on top 20?

What is the F_1 on the top 20?

What is the uninterpolated precision of the system at 25% recall?

What is the interpolated precision at 33% recall?

Assume that these 20 documents are the complete result set of the system. What is the MAP for the query?

11. Evaluation of learning objective achievement

a. Relevance judgments exercise will be evaluated on the correct procedure instead of correct relevance results.

b. Detailed steps for each question. The answer should not just be a number result.

c. Accurate fetching of documents using LucidWorks.

12. Resources

a. http://fox.cs.vt.edu/VAD1/DOWN/IRCOLLS/ADI.ALL

b. http://fetcher.dlib.vt.edu:8888/solr/#/test_collection_vt/

c. http://lucidworks.lucidimagination.com/display/bigdata/Search

d. Ch8CS5604F20120920.mp4 lecture located in CS5604/Resources/MP4Fall2012 on Scholar

e. Manning, C., Raghavan, P., and Schutze, H. (2008). Chapter 8: Evaluation in information retrieval. In *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

f. McCandless, M., Hatcher, E., and Gospodnetic, O. (2010). Appendix C: Lucene/contrib benchmark. In *Lucene in Action* (2nd Ed.). Stamford: Manning Publications Co.

13. Glossary

a. **Relevant and nonrelevant:** A document is relevant if it addresses the stated information need, not because it just happens to contain all the words in the query.

b. **Precision:** fraction of retrieved documents that are relevant. P(relevant|retrieved)

c. **Recall:** fraction of relevant docs that are retrieved. P(retrieved|relevant)

d. **F measure:** weighted harmonic mean of precision and recall.

e. **Interpolated precision**: at certain recall level *r* is defined as the highest precision found for any recall level *r'*≥*r*:$p_{interp}(r) = \max_{r' \geq r} p(r')$

f. **Mean Average Precision (MAP):** average of the precision value obtained for the top *k* documents, each time a relevant doc is retrieved.

g. **Kappa measure**: A common measure for agreement between judges

14. Additional useful links

15. Contributors

Authors: Shiyi Wei (wei@vt.edu), Victoria Suwardiman (vsuwardi@vt.edu), Anand Swaminathan (anand90@vt.edu)
Reviewers: Dr. Edward A. Fox, Kiran Chitturi, Tarek Kanan
Class: CS 5604: Information Retrieval and Storage. Virginia Polytechnic Institute and State University