Relevance

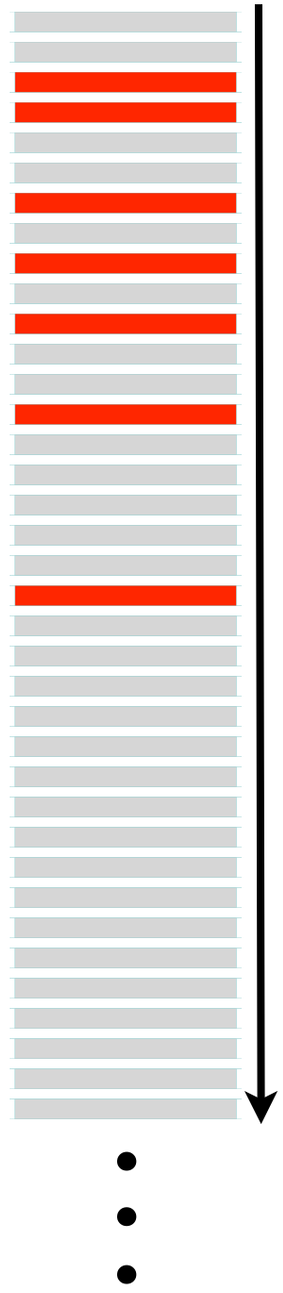
Jaime Arguello
INLS 509: Information Retrieval
jarguell@email.unc.edu

November 13, 2013

Let's Take a Step Back

Information Retrieval Task

- Given a query and a corpus, retrieve documents that are relevant
  - query: textual representation of the user's information need
  - corpus: collection of textual documents
  - relevance: satisfaction of the user's information need



Why Talk about Relevance?

- The goal of an IR system is to predict relevance in the same way that users judge relevance

- So, it seems natural to ask: what is relevance and how do "real" users judge relevance?

What is relevance?

- Relevance is the A of B existing between C and D as determined by E where,
  - A = measure, degree, estimate
  - B = correspondence, utility, fit
  - C = document, information provided, fact
  - D = query, request, information requirement
  - E = user, judge, information specialist

(Saracevic '07)

What is relevance?

- Relevance is a relation
- Relevance is judged (it is subjective)
- The ability to judge relevance is not learned; it is innate
- Judgements are made within a context
  - internal context: the user's knowledge, feelings, and expectations (about the information need, the corpus, and the system)
  - external context: the user's higher-level task at hand and their environment
- Context is dynamic, so relevance is dynamic across users and for the same user across time

(Saracevic '07)

Types of Relevance

- Algorithmic relevance: relation between the document and the query as determined by the system
- Topical relevance: relation between the subject of the document and the subject of the query
- Cognitive relevance: relation between the document and the user's state of knowledge
- Situational relevance: relation between the document and the user's external context
- Affective relevance: relation between the document and the user's goals, emotions, and motivations

(Saracevic '07)

How do users judge relevance?

- A survey of the literature reveals four major findings:

1. Users make relevance judgements based on different document attributes (content is one of them)

2. The attributes that matter most depend upon the user's internal and external context

3. Context varies across users, so relevance judgements vary across users

4. Context varies over time, so relevance judgements (for the same user) vary over time

(Saracevic '07)

Relevance Clues
document attributes

- Content attributes: topic, quality, depth, scope, freshness, readability, clarity
- Object attributes: organization, representation, format, availability, accessibility, cost
- Validity: accuracy, authority, trustworthiness, verifiability

(Saracevic ’07)

Relevance Clues

internal and external influences

- Situational match: appropriateness to situation or task, usability, urgency, value in use
- Cognitive match: understanding, novelty, mental effort
- Affective match: emotional responses to information
- Belief match: personal credence given to information

(Saracevic ’07)

Relevance Clues

the good news

- There is a limited number of document attributes that seem to strongly influence how users judge relevance
- There is a limited number of internal/external factors that seem to strongly influence how users judge relevance
- One of the most important document attributes is topical relevance

(Saracevic ’07)

Relevance Clues
the bad news

- A user's internal/external factors affect which document attributes are most important (there are interaction effects)
- A user's internal/external factors change over time, so relevance changes over time
- On Monday, we'll see an example of how the external context (time pressure) affects the document attributes that matter most

(Saracevic '07)

Batch Evaluation Assumptions

- Most IR test-collections are constructed and used under the following assumptions about relevance:

  1. Type of relevance?
  2. Discreet or continuous?
  3. The relevance of a document is impacted by the relevance of another document?
  4. Relevance judgement are consistent across judges?
  5. Relevance is stable over time?

(Saracevic '07)

Batch Evaluation Assumptions

- Most IR test-collections are constructed and used under the following assumptions about relevance:

  1. Topical: relevance is solely topical
  2. Binary: a document is either relevant or non-relevant
  3. Independent: the relevance of a document is not affected by the relevance of another document
  4. Consistent: relevant judgements are consistent across users/judges
  5. Stable: relevance is stable over time

(Saracevic '07)

Batch Evaluation Assumptions

relevance is topical

- As we've already discussed, relevance is not only topical
- Various other document attributes and internal/external factors affect a user's relevance judgements
- Topicality, however, is a key component

(Saracevic '07)

Batch Evaluation Assumptions

relevance is binary

- As you might expect, relevance is not binary (or even discreet)
- Users tend to judge relevance along a continuum
- However, relevance appears to be bimodal
- That is, most judgements fall within the two extremes (e.g., perfect/poor)

(Saracevic '07)

Batch Evaluation Assumptions

relevance is independent

- Relevance judgements are not independent
- Documents that are seen early have a higher probability of being relevant
- Suggests that novelty is important

(Saracevic '07)

Batch Evaluation Assumptions

relevance is stable

- As previously discussed, relevance is dynamic
- The user's internal/external factors are dynamic
- Therefore, the document attributes that influence relevance judgements are dynamic
- What internal factors change as the user searches?

(Saracevic '07)

Batch Evaluation Assumptions

relevance is consistent across users

- Relevance is in the eye of the beholder
- In general, overlap between assessors tends to be 30%
- The intersection divided by the union = 30%

(Saracevic ’07)

Batch Evaluation Assumptions

relevance is consistent across users

- Yes, relevance is in the eye of the beholder
- However, there are some regularities!
  - agreement is greater when assessors have a high level of expertise on the subject
  - overlap can be as high as 80%
  - using relevance grades, overlap is greater on the most relevant grade (arguably the most important findings?)
  - overlap goes way down as we introduce more than two assessors

(Saracevic '07)

Batch Evaluation Assumptions

relevance is consistent across users

- Because assessors disagree, it is very important that all judgements for a given query be made by the same assessor!

(Saracevic '07)

Batch Evaluation Assumptions

- Most IR test-collections have been build under the following assumptions about relevance:

1. ~~Topical: relevance is solely topical~~
2. ~~Binary: a document is either relevant or non-relevant~~
3. ~~Independent: the relevance of a document is not affected by the relevance of another document~~
4. ~~Consistent: relevant judgements are consistent across users/judges~~
5. ~~Stable: relevance is stable over time~~

(Saracevic '07)

Batch Evaluation Assumptions

- So, are decades of batch-evaluation results meaningless?
- What do you think?

Batch Evaluation Assumptions

- As it turns out, when we average across queries, the ranking of systems by performance stays the same when we use different assessors
- The best system remains the best system
- The second best system remains the second best system
- The worst system remains the worst system (and so on...)
- That is, <u>when we average across queries!</u>
- For individual queries, changes in the ranking of systems can occur

(Saracevic '07)

Interactive Information Retrieval

- How are relevance judgements affected by a user's many internal states (cognitive, affective, belief states)?
- How are relevance judgements affected by a user's many external/situational states?
- How can these internal and external states be communicated to the system?
- How can these internal and external states be predicted by the system?
- How do these states change as a task evolves and how does this affect changes in relevance judgements?