

CS 6890: Lecture 13

Vladimir Kulyukin

Department of Computer Science

Utah State University

Outline

- A Gentle Introduction to Information Retrieval
 - Relevance Feedback
 - Automatic Thesaurus Construction
 - Term Selection Models

Relevance Feedback

Motivation

- No IR system is likely to retrieve all documents relevant to the user's query.
- The basic reason is that the users frequently employ terms that are not contained in the documents.
- One way to alleviate this problem is to modify the original query to increase performance.

Relevance Feedback

- Query modification
 - Take the list of retrieved documents;
 - Have the user rank the documents as relevant and non-relevant;
 - Increase the weights of the terms that occur in relevant documents.
- Query expansion
 - Expand the original query with new terms that are deemed related to the query.

Standard Rocchio Formula

$$Q_{i+1} = Q_i + \frac{1}{n_1} \sum_{i=1}^{n_1} R_i - \frac{1}{n_2} \sum_{i=1}^{n_2} S_i, \text{ where}$$

Q_i = the vector for the query at step i ;

n_1 = the number of the relevant documents;

n_2 = the number of the non - relevant documents;

R_i = the vector for the i - th relevant document;

S_i = the vector for the i - th non - relevant document.

Generalized Standard Rocchio

$$Q_{i+1} = Q_i + \alpha \sum_{i=1}^{n_1} R_i - \beta \sum_{i=1}^{n_2} S_i, \text{ where}$$

Q_i = the vector for the query at step i ;

α, β are some (carefully chosen) constants;

R_i = the vector for the i - th relevant document;

S_i = the vector for the i - th non - relevant document.

Observations

- Standard Rocchio and its numerous derivatives do not perform query expansion.
- Standard Rocchio reweights the existing terms in the query.
- What happens if the query is retrieved no relevant documents?
- What happens if the query's terms do not match any terms in the relevant documents?

Query Expansion W/O Term Reweighting

- New terms are added to the query on the basis of a thesaurus: synonymy and co-occurrence.
- Automated construction of thesauri is fundamental.
- Manual thesauri are also important, e.g. WordNet.

Comments on Relevance Feedback

- Relevance feedback is a technique that is of marginal value to most users.
- Why? Most users are interested only in an answer to their query.
- Relevance feedback is a technique that is of great value to power users, i.e. users that are interested in completeness (recall) or accuracy (precision).

Automated Thesaurus Construction

Automated Thesaurus Construction

Bookstein, A., Kulyukin, V., Raita, T.,
Nicholson, J. Adapting measures of
clumping strength to assess term-term
similarity. *Journal of the American Society
for Information Science and Technology*,
54(7):611-620, 2003.

Automated Thesaurus Construction

- Automated IR relies heavily on statistical regularities that emerge as terms are deposited to produce text.
- Question 1: Are there any detectable statistical patterns that can be expected of terms that are semantically related to each other?
- Question 2: Can we derive measures of how tightly terms are semantically associated?

Term Clumping

Behavior of Content-Bearing Terms

- Subject content in a document shifts from topic to topic.
- Topics constitute islands of content.
- Terms that are relevant to a particular topic should have a tendency to gather in the islands of content relevant to that topic.

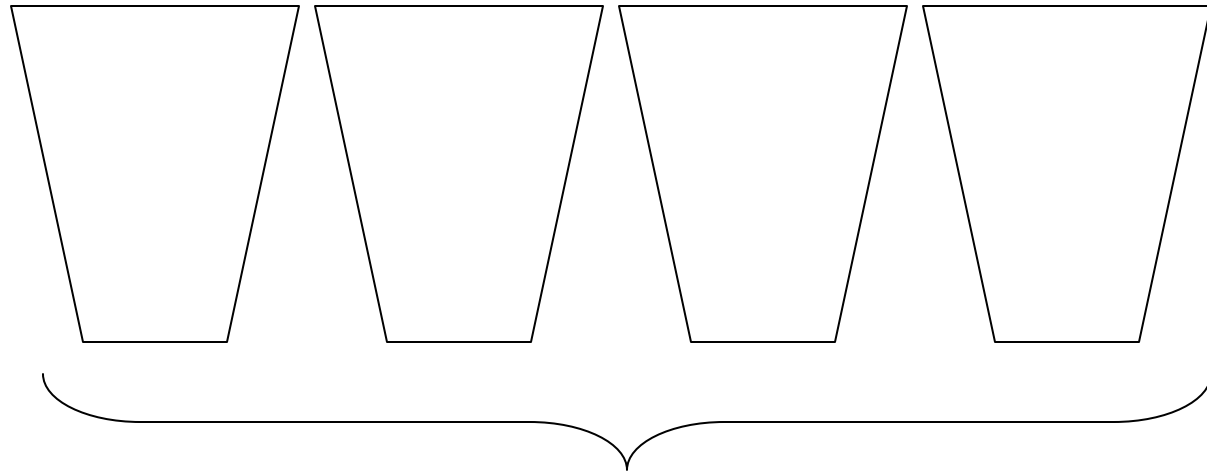
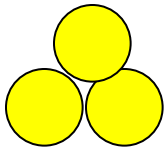
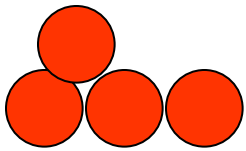
Behavior of Content-Bearing Terms

- If, instead of a single document, one looks at a collection of documents, then islands of content are individual documents.
- Content-bearing terms should have a tendency to clump in the documents to whose topics they are related.
- Measures of the degree to which the actual pattern of a term's occurrence deviates from randomness indicate its ability to carry content.

Text Production Model

- A text is a sequence of segments, i.e. paragraphs, sections, chapters.
- In the case of a document collection, we can conceptualize segments as individual documents.
- This model should be statistically similar to the standard ball/urn model of distribution.

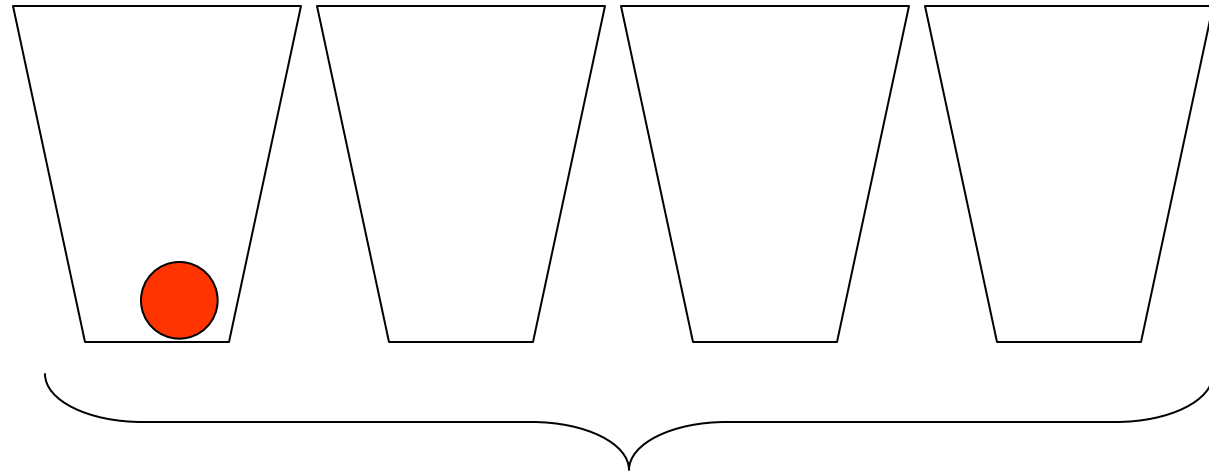
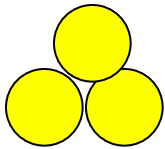
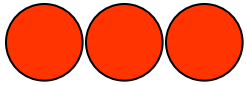
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

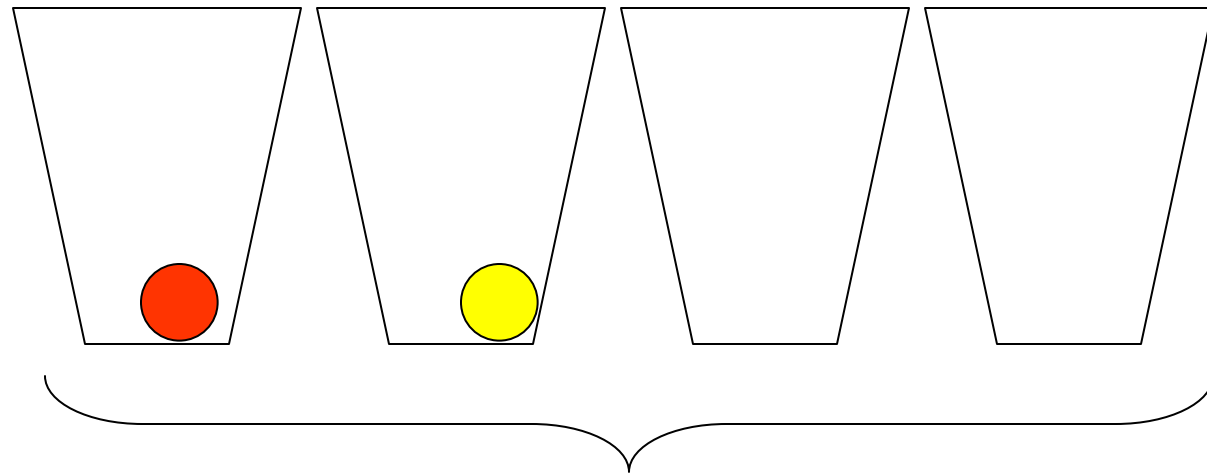
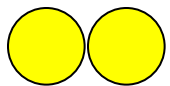
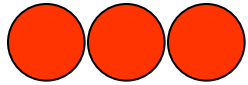
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

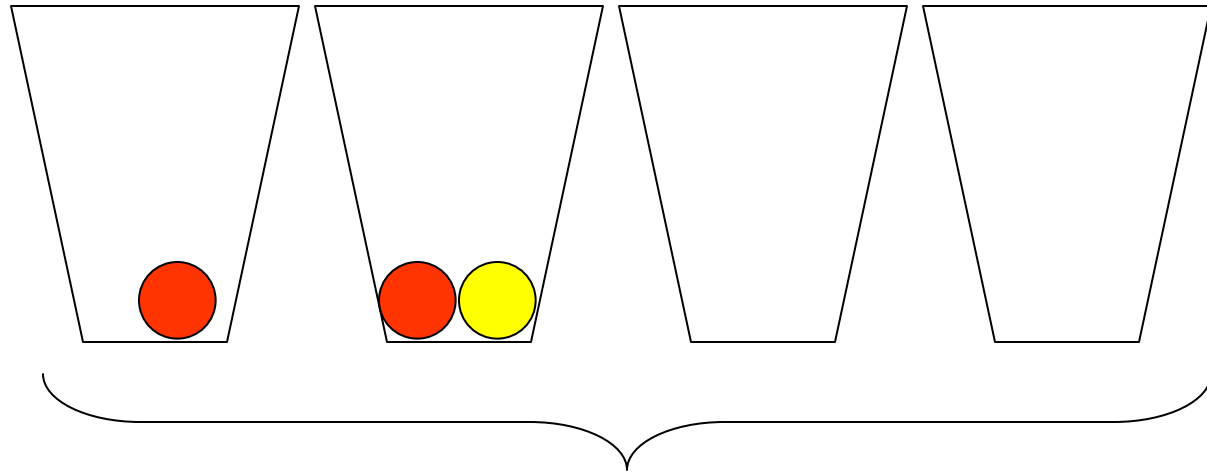
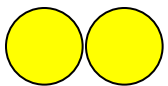
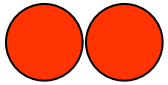
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

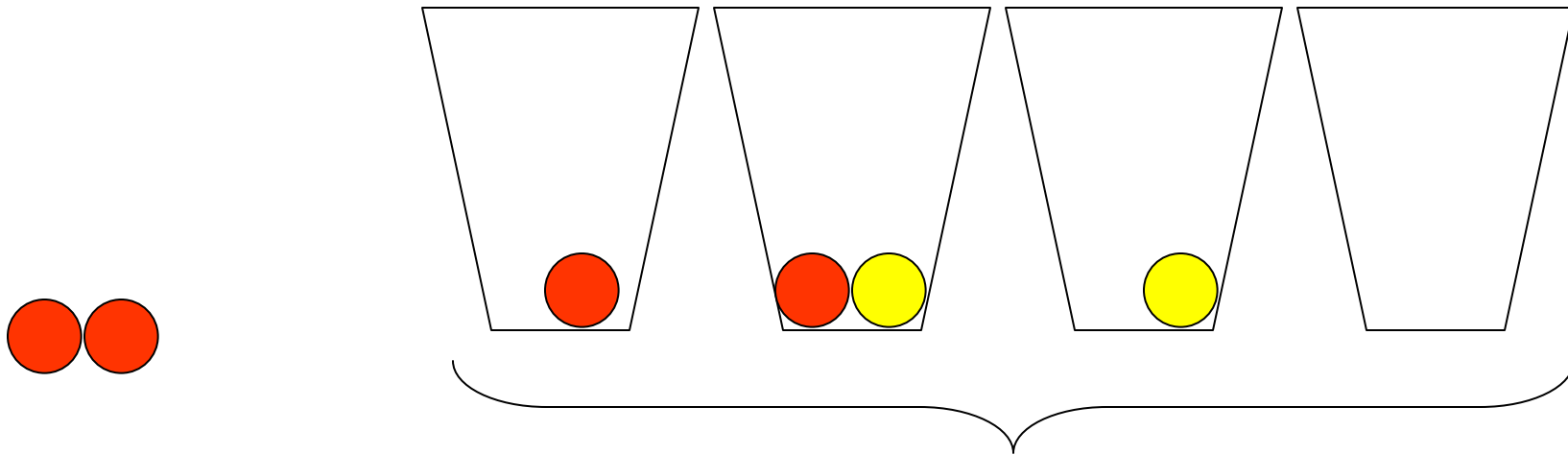
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

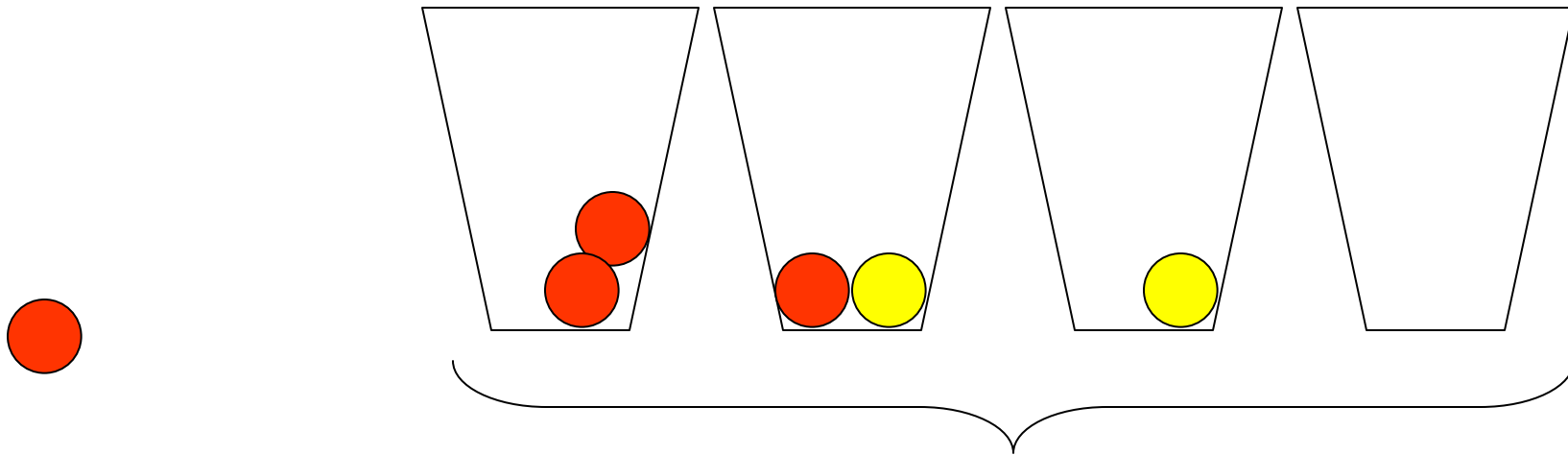
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

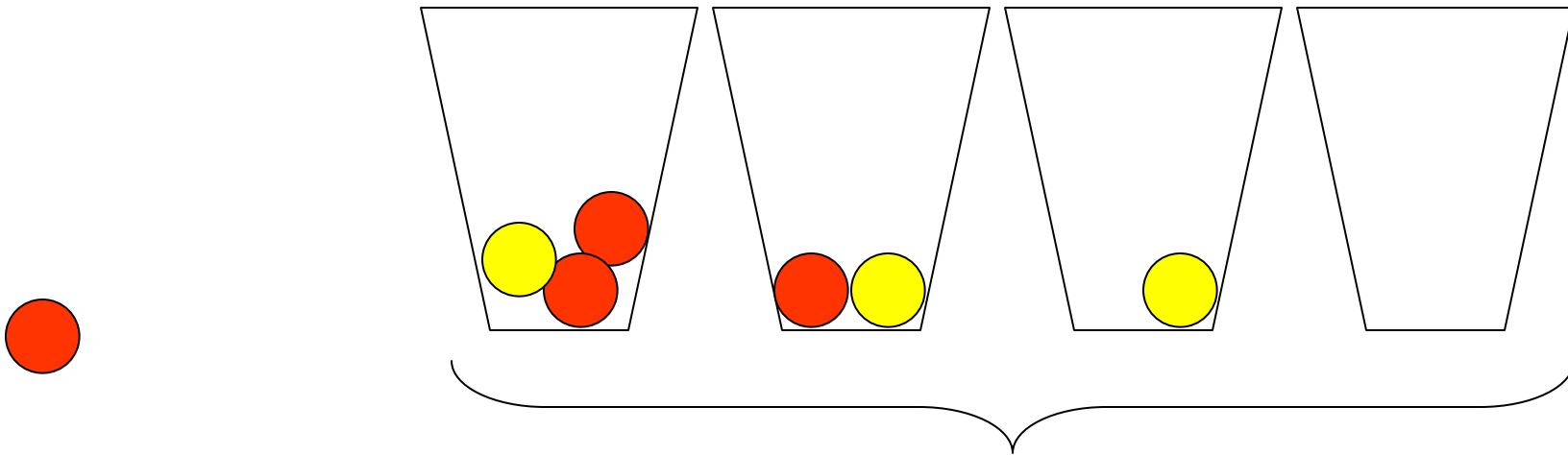
Ball/Urn Model



M URNS

N_1 red balls
 N_2 yellow balls
etc.

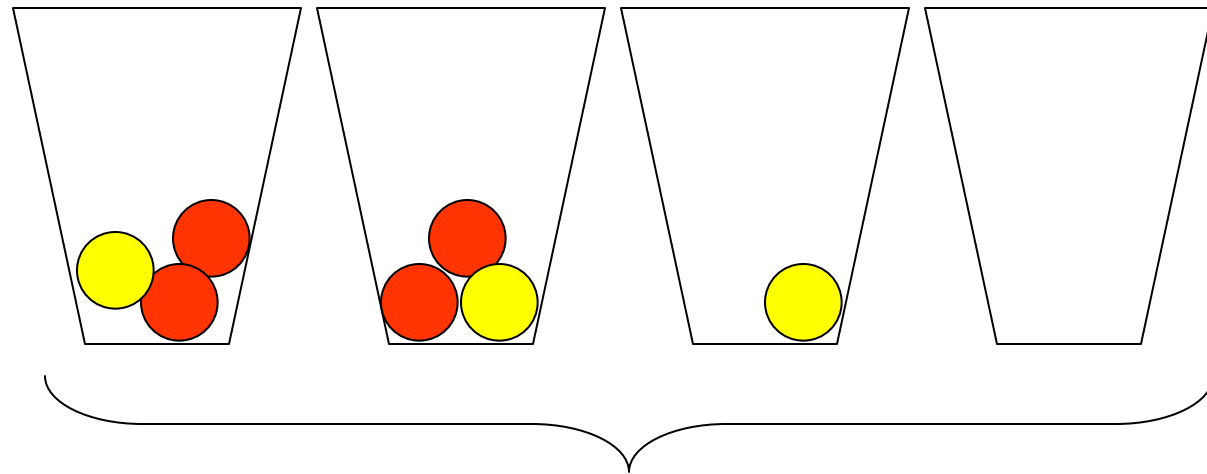
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

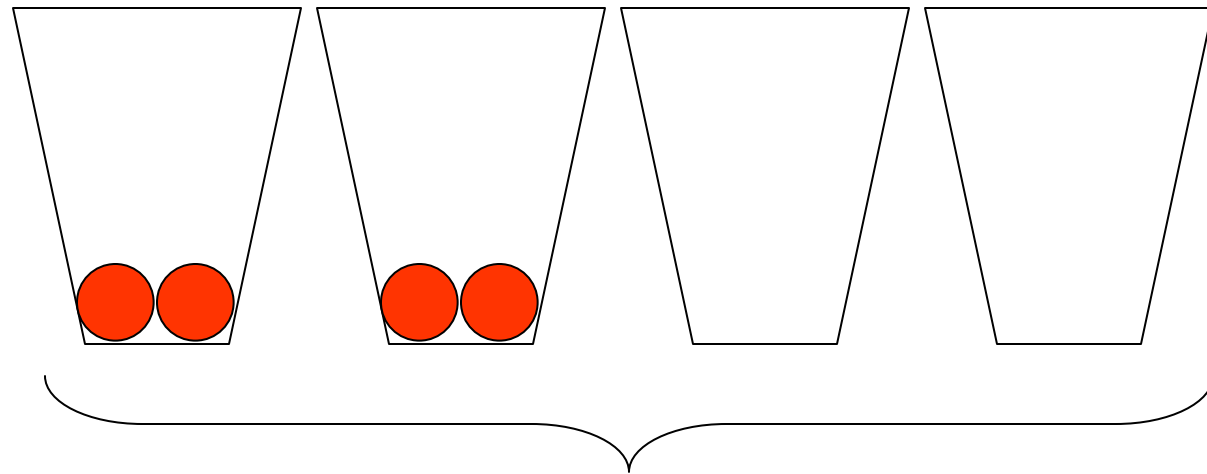
Ball/Urn Model



M URNS

N1 red balls
N2 yellow balls
etc.

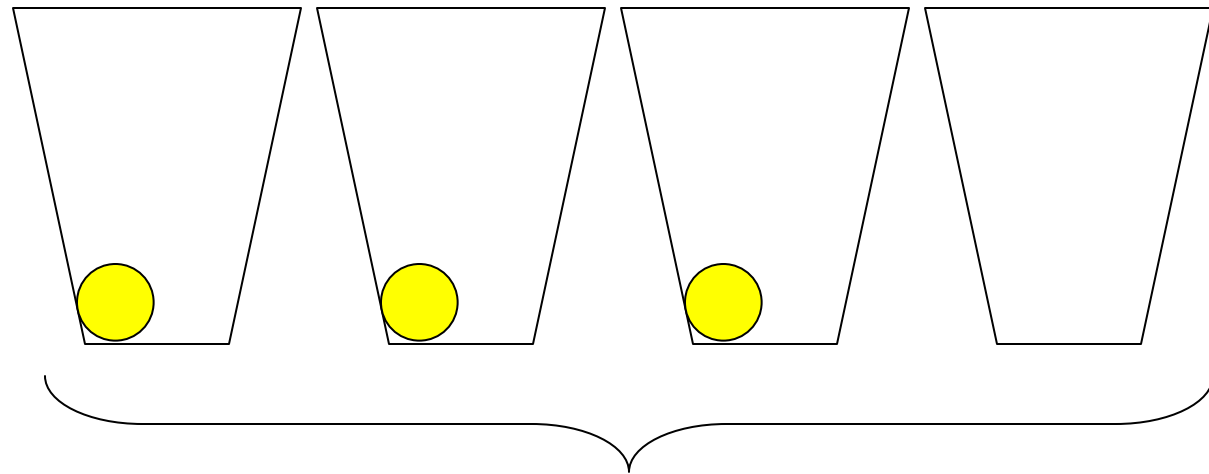
Red Ball Distribution Pattern



M URNS

N1 red balls
N2 yellow balls
etc.

Yellow Ball Distribution Pattern



M URNS

N1 red balls
N2 yellow balls
etc.

Term Distribution Model

- The urns can be taken to mean the segments of text.
- The balls can be taken to mean the actual occurrences of terms.
- We study the distribution of terms in text units in the same fashion that we study the distribution of balls in urns.

Single Term Distribution Model

- A text whose content causes it to attract more occurrences of a specific term can be likened to an urn with a wider mouth.
- We can compare the distribution of a given term to the random distribution where each urn is equally likely to attract a term.
- Such comparisons (measures) value terms whose distribution patterns deviate from random distribution.

Term Pair Distribution Model

- Suppose there are 2 terms: A and B .
- Suppose that there are the clumping measures for A occurrences and for B occurrences in the entire collection.
- If A and B are associated, it is likely that the clumping measure for A over the entire collection will differ from the clumping measure of A over the units that contain B .
- If A and B are associated, it is likely that the clumping measure for A over the entire collection will differ from the clumping measure of A over the units that do not contain B .
- Thus, there are three contexts of associativity: units with B , units without B , the entire collection.

Expected Number of Documents

If T_A occurrences of a term A are randomly distributed over D documents, then the expected number of documents that contain at least one occurrence of A is

$$E_A = D \left[1 - \left[1 - \frac{1}{D} \right]^{T_A} \right].$$

A Clumping Measure

Let N_A be the actual number of documents where A occurs at least once. Then

$$M_A = \frac{E_A}{N_A} = \frac{D}{N_A} \left[1 - \left[1 - \frac{1}{D} \right]^{T_A} \right] \text{ is a clumping measure.}$$

M_A is close to 1 if A is distributed randomly.

Distribution Equations

N_{AB} = number of documents that contain at least one occurrence of A and one occurrence of B .

$N_{A\bar{B}}$ = number of documents that contain at least one occurrence of A and no occurrence of B .

$$N_A = N_{AB} + N_{A\bar{B}}.$$

$T_{A|B}$ = number of occurrences of A in the documents that contain at least one occurrence of B .

$T_{A|\bar{B}}$ = number of occurrences of A in the documents that contain no occurrence of B .

$$T_A = T_{A|B} + T_{A|\bar{B}}.$$

Distribution Equations

$N_{\overline{A}B}$ = number of documents that contain
no occurrence of A and at least one occurrence of B .

$N_{\overline{A}\overline{B}}$ = number of documents that contain
no occurrence of A and no occurrence of B .

$$N_{\overline{A}} = N_{\overline{A}B} + N_{\overline{A}\overline{B}}.$$

$$N_A + N_{\overline{A}} = D.$$

$$N_{AB} + N_{\overline{A}B} = N_B.$$

$$T_{B|A} + T_{B|\overline{A}} = T_B.$$

Environment with B

$$M_{A|B} = \frac{N_B}{N_{AB}} \left[1 - \left[1 - \frac{1}{N_B} \right]^{T_{A|B}} \right]$$

Environment without B

$$M_{A|\bar{B}} = \frac{N_{\bar{B}}}{N_{A\bar{B}}} \left[1 - \left[1 - \frac{1}{N_{\bar{B}}} \right]^{T_{A|\bar{B}}} \right]$$

Environment with A

$$M_{B|A} = \frac{N_A}{N_{AB}} \left[1 - \left[1 - \frac{1}{N_A} \right]^{T_{B|A}} \right]$$

Environment without A

$$M_{B|\bar{A}} = \frac{N_{\bar{A}}}{N_{B\bar{A}}} \left[1 - \left[1 - \frac{1}{N_{\bar{A}}} \right]^{T_{B|\bar{A}}} \right]$$

Overall Measures

$$M_A = \frac{D}{N_A} \left[1 - \left[1 - \frac{1}{D} \right]^{T_A} \right]$$

$$M_B = \frac{D}{N_B} \left[1 - \left[1 - \frac{1}{D} \right]^{T_B} \right]$$

Association Measures

Association Measure 1

Contrast the clumping behavior of A over all segments with the clumping behavior of A over segments that contain B.

Association Measure 1

$$S_{A;B}^{(1)} = \frac{M_A}{M_{A|B}}$$

Association Measure 2

Contrast the behavior of A in the presence of B to the behavior of A in the absence of B.

Association Measure 2

$$S_{A;B}^{(2)} = \frac{M_{A|\bar{B}}}{M_{A|B}}$$

Association Measure 3

Combine the clumping tendency of A in contexts with B and without B.

Association Measure 3

$$S_{A;B}^{(3)} = \left(\frac{M_A}{M_{A|B}} \right) \left(\frac{M_A}{M_{A|\bar{B}}} \right)$$

Evaluation

- Association measure 1 is evaluated.
- The document collection is the online version of the Columbia Encyclopedia from www.bartleby.com.
- Each article is a separate document.
- Articles with fewer than 15 terms are eliminated from consideration.
- No stemming or stoplisting is done.
- Result collection includes 23,235 documents.
- Total number of terms is 103,178.

Evaluation

- 20 terms are randomly selected from the list of terms: ALGEBRA, BIRD, BUDDHISM, LEUKEMIA, MAMMAL, etc.
- The association scores are computed b/w each of the 20 terms and each other term in the collection.
- The top 100 pairs for each term are chosen and evaluated by a human judge.

Evaluation

- Each judge is instructed to rank each pair as + (definite yes); - (definite no); and ? (do not know).
- The overall success rates taken as percentage of +'s in four human judges indicated that the system was able to find meaningful associations.

Knowledge Discovery

- {Buddhism, Kalmykia}
 - Kalmykia is a region in Southern Russian where Buddhism has very strong roots.
- {Dog, Toshiro}
- {Dog, Mifune}
- {Dog, Kurosawa}
- {Dog, Rashomon}
 - Toshiro Mifune is a renowned Japanese actor who starred in many films by Akira Kurosawa, including “Stray Dog” and “Rashomon.”

Knowledge Discovery

- {Algebra, 1766}
- {Algebra, 1828}
- {Science, Pharmacy}
- {Fauna, Bear}
- {Music, Mozart}
- {Actor, Theater}