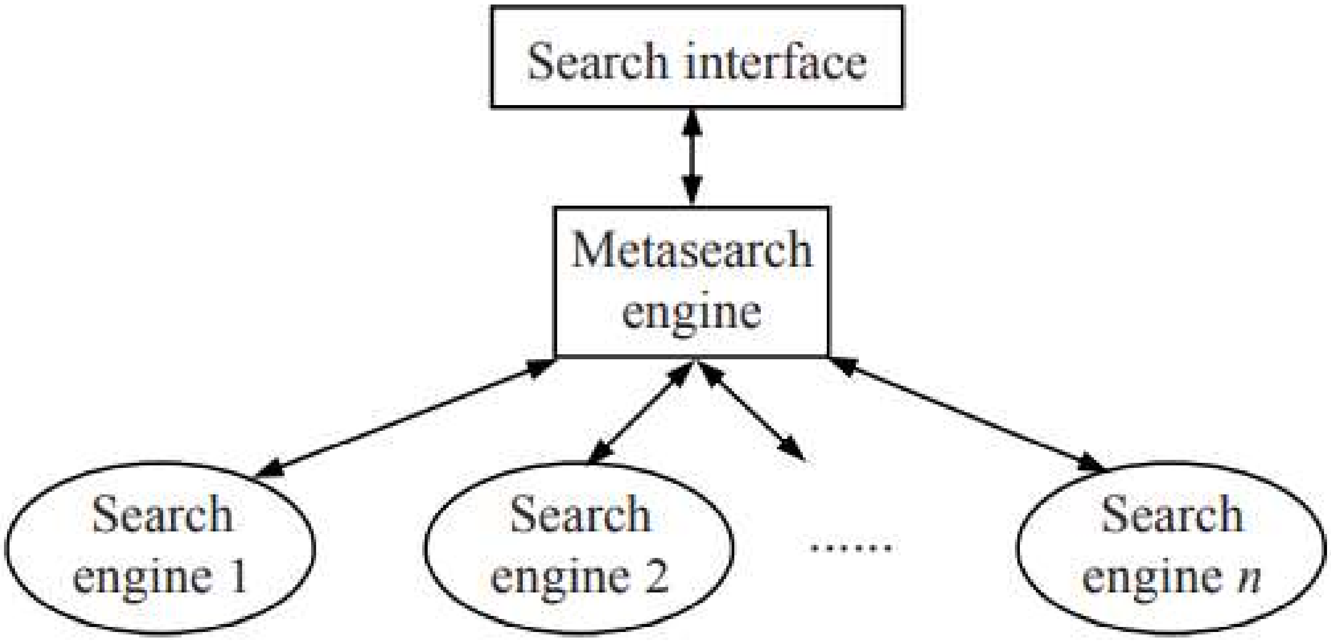
Web Search: Work of Search Engine

- Crawls pages in the Web
- Crawled pages are parsed, indexed and stored
  - Parser parses HTML pages to produce a stream of tokens (terms to be indexed)
  - Inverted index is created using multiple schemes
- Index used at query time for efficient retrieval

Meta-Search and Combining Multiple Ranking

- Meta Search Engine
  - A search system that does not have its own database of Web pages. Instead, it answers the user query by combining the results of some other search engines which normally have their databases of Web pages
  - After receiving a query from the user through the search interface, the meta-search engine submits the query to the underlying search engines (called component search engine). The returned results from all these search engines are then combined (fused or merged) and sent to the user.

Meta-Search


Search interface
Metasearch engine
Search engine 1
Search engine 2
......
Search engine *n*

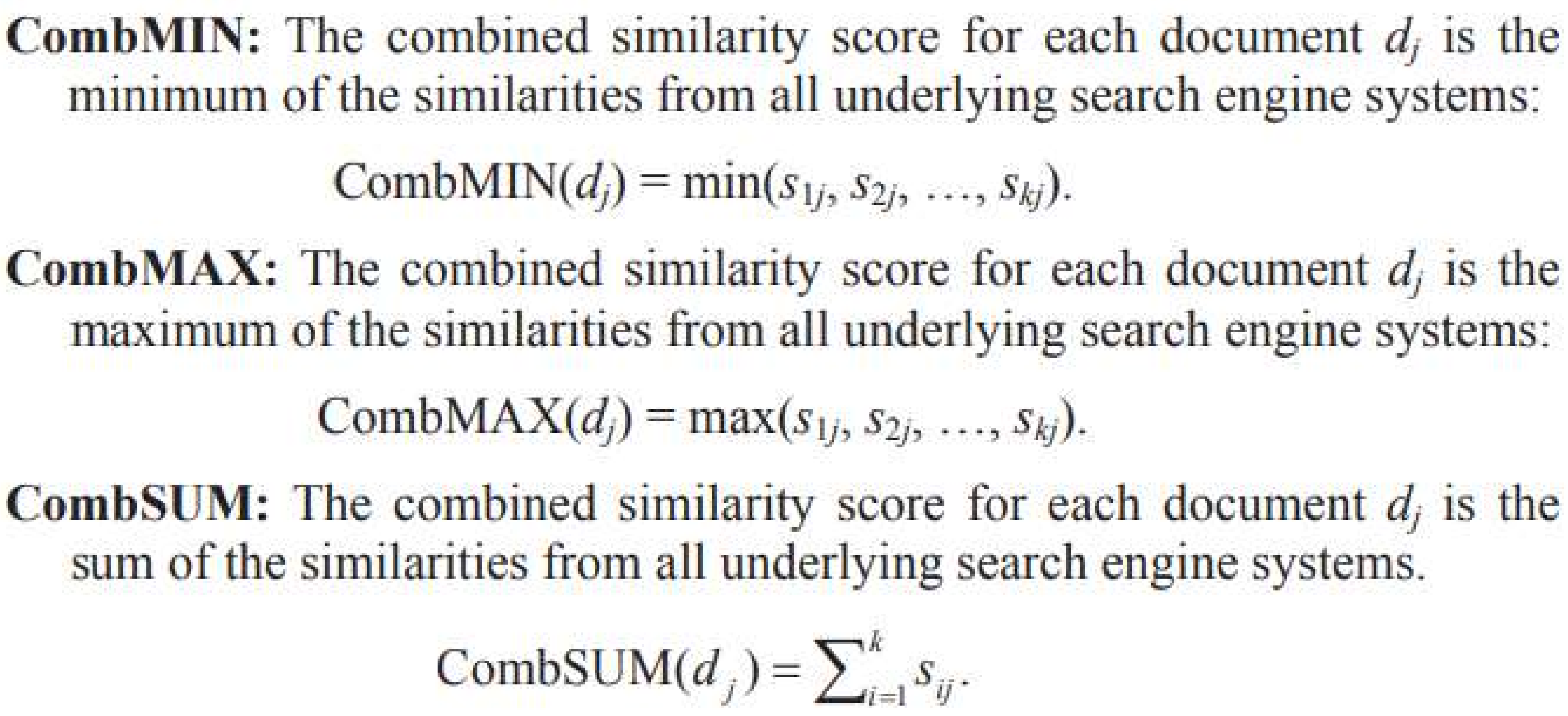
Advantages of Meta-Search

- It increases the search coverage of the Web. The Web is a huge information source, and each individual search engine may only cover a small portion of it.
- Makes search more effective. Each component search engine has its ranking algorithm to rank relevant pages, which is often biased. By combining the results from multiple search engines, their biases can be reduced.

Combining Using Similarity Scored

- Let the set of candidate documents to be ranked be D = {d1, d2, …, dN}.
- There are k underlying systems (component search engines or ranking techniques).
- The ranking from system or technique i gives document dj the similarity score, sij.

Combining Using Similarity Scored



CombMIN: The combined similarity score for each document d_j is the minimum of the similarities from all underlying search engine systems:

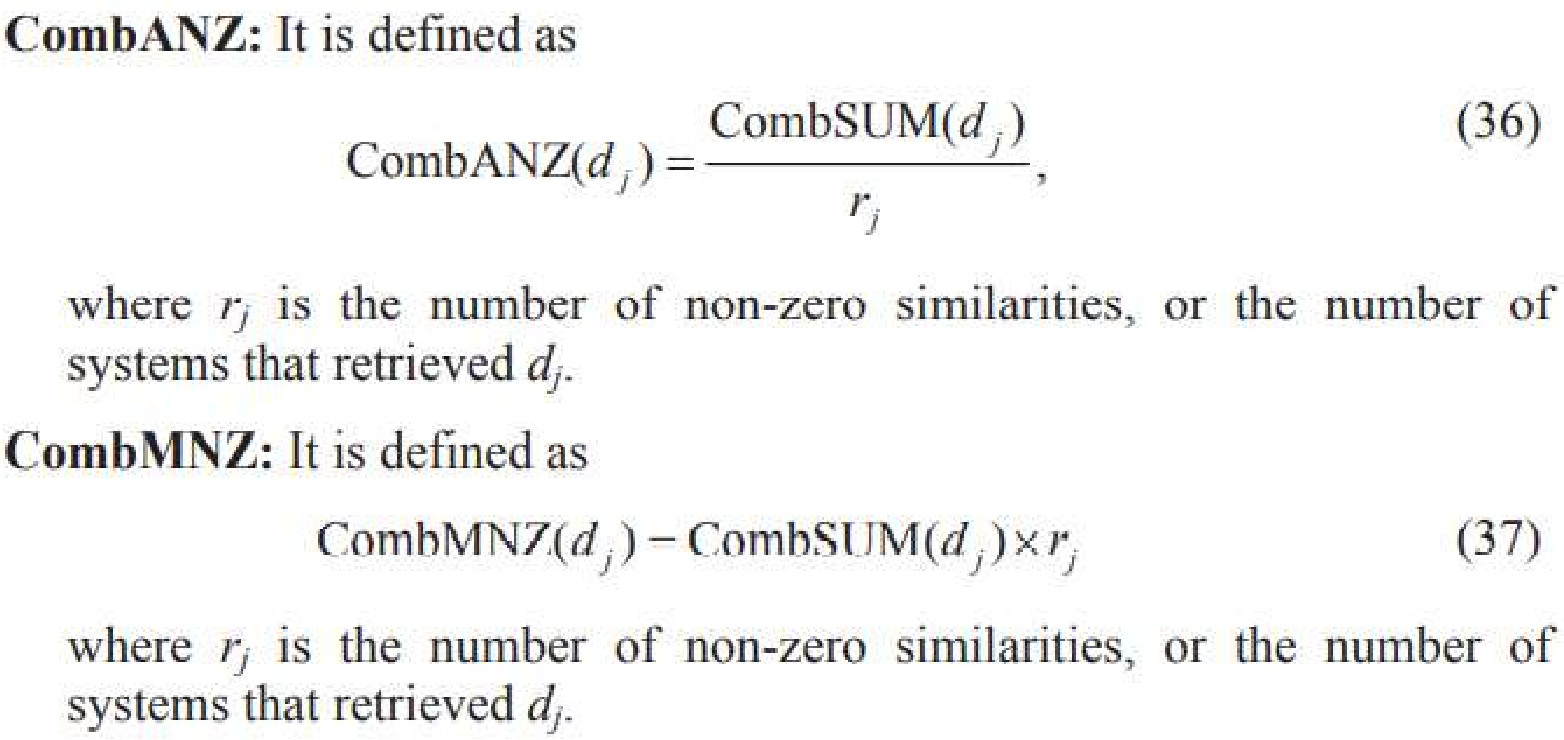
$$\text{CombMIN}(d_j) = \min(s_{1j}, s_{2j}, \ldots, s_{kj}).$$

CombMAX: The combined similarity score for each document d_j is the maximum of the similarities from all underlying search engine systems:

$$\text{CombMAX}(d_j) = \max(s_{1j}, s_{2j}, \ldots, s_{kj}).$$

CombSUM: The combined similarity score for each document d_j is the sum of the similarities from all underlying search engine systems.

$$\text{CombSUM}(d_j) = \sum_{i=1}^{k} s_{ij}.$$



CombANZ: It is defined as

$$\text{CombANZ}(d_j) = \frac{\text{CombSUM}(d_j)}{r_j}, \qquad (36)$$

where r_j is the number of non-zero similarities, or the number of systems that retrieved d_j.

CombMNZ: It is defined as

$$\text{CombMNZ}(d_j) - \text{CombSUM}(d_j) \times r_j \qquad (37)$$

where r_j is the number of non-zero similarities, or the number of systems that retrieved d_j.

- It is a common practice to normalize the similarity scores from each ranking using the maximum score before combination. Researchers have shown that, in general, CombSUM and CombMNZ perform better. CombMNZ outperforms CombSUM slightly in most cases.

Combining Using Rank Positions

- The social choice theory studies voting algorithms as techniques to make group or social decisions.

Borda ranking

- Each voter announces a (linear) preference order on the candidates.
- For each voter, the top candidate receives n points (if there are n candidates in the election), the second candidate receives n-1 points, and so on.
- If there are candidates left unranked by a voter, the remaining points are divided evenly among the unranked candidates.
- The points from all voters are summed up to give the final points for each candidate.
- The candidate with the most points wins.

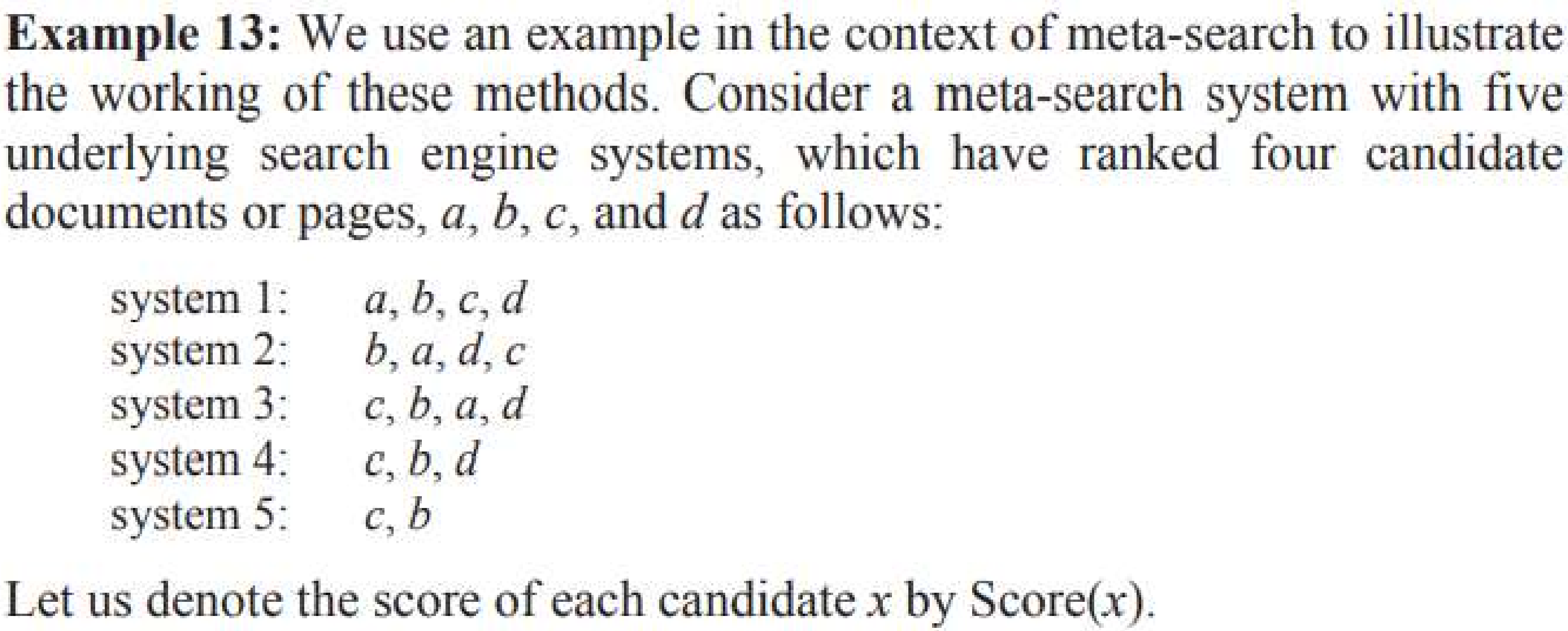
Condorcet ranking

- The winner of the election is the candidate(s) that beats each of the other candidates in a pair-wise comparison.
- If a candidate is not ranked by a voter, the candidate loses to all other ranked candidates.
- All unranked candidates tie with one another

Reciprocal ranking

- Sums one over the rank of each candidate across all voters.
- For each voter, the top candidate has the score of 1, the second ranked candidate has the score of 1/2, and the third ranked candidate has the score of 1/3 and so on.
- If a candidate is not ranked by a voter, it is skipped in the computation for this voter.
- The candidates are then ranked according to their final total scores.

Example

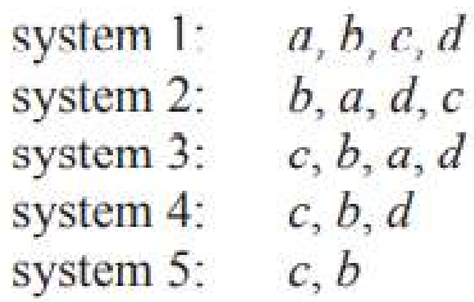


Example 13: We use an example in the context of meta-search to illustrate the working of these methods. Consider a meta-search system with five underlying search engine systems, which have ranked four candidate documents or pages, *a*, *b*, *c*, and *d* as follows:

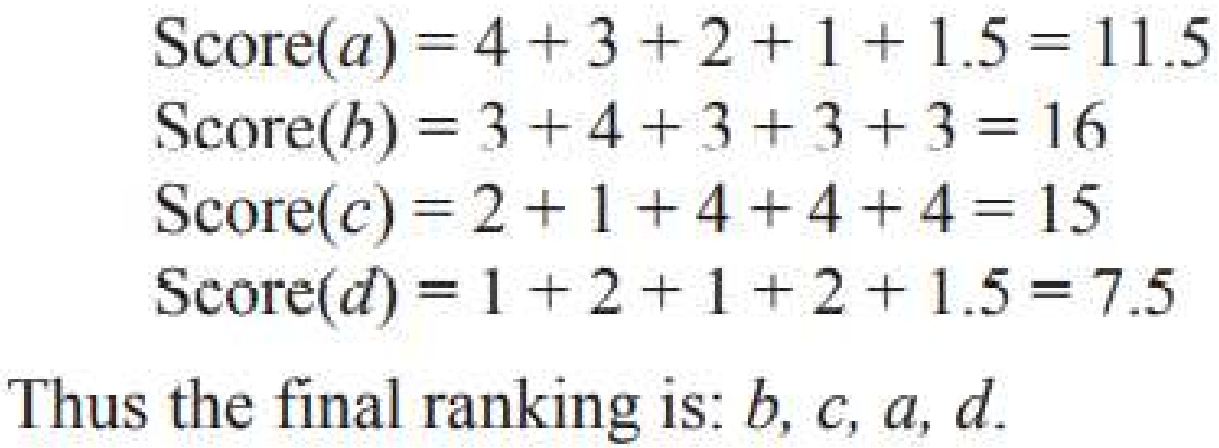
system 1:	*a*, *b*, *c*, *d*
system 2:	*b*, *a*, *d*, *c*
system 3:	*c*, *b*, *a*, *d*
system 4:	*c*, *b*, *d*
system 5:	*c*, *b*

Let us denote the score of each candidate *x* by Score(*x*).

Borda Ranking



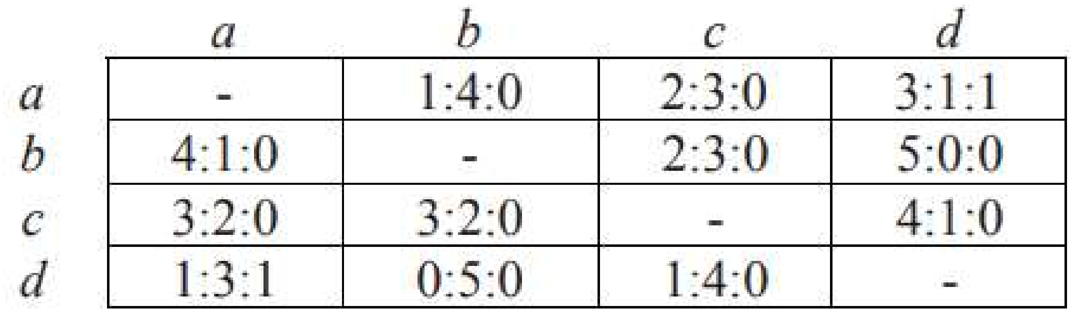
system 1:	a, b, c, d
system 2:	b, a, d, c
system 3:	c, b, a, d
system 4:	c, b, d
system 5:	c, b



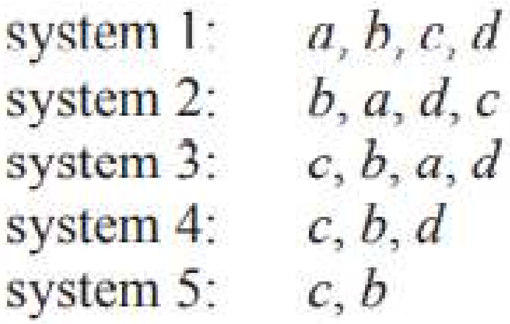
$$\text{Score}(a) = 4 + 3 + 2 + 1 + 1.5 = 11.5$$
$$\text{Score}(b) = 3 + 4 + 3 + 3 + 3 = 16$$
$$\text{Score}(c) = 2 + 1 + 4 + 4 + 4 = 15$$
$$\text{Score}(d) = 1 + 2 + 1 + 2 + 1.5 = 7.5$$

Thus the final ranking is: *b*, *c*, *a*, *d*.

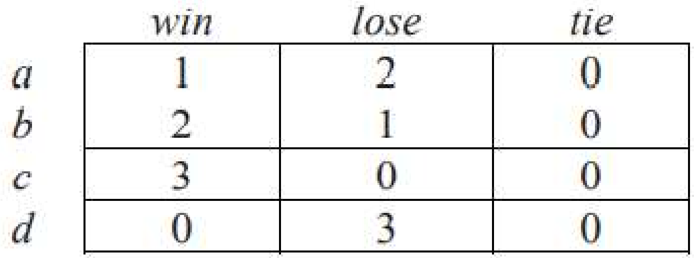
Condorcet Ranking



	a	*b*	*c*	*d*
a	-	1:4:0	2:3:0	3:1:1
b	4:1:0	-	2:3:0	5:0:0
c	3:2:0	3:2:0	-	4:1:0
d	1:3:1	0:5:0	1:4:0	-



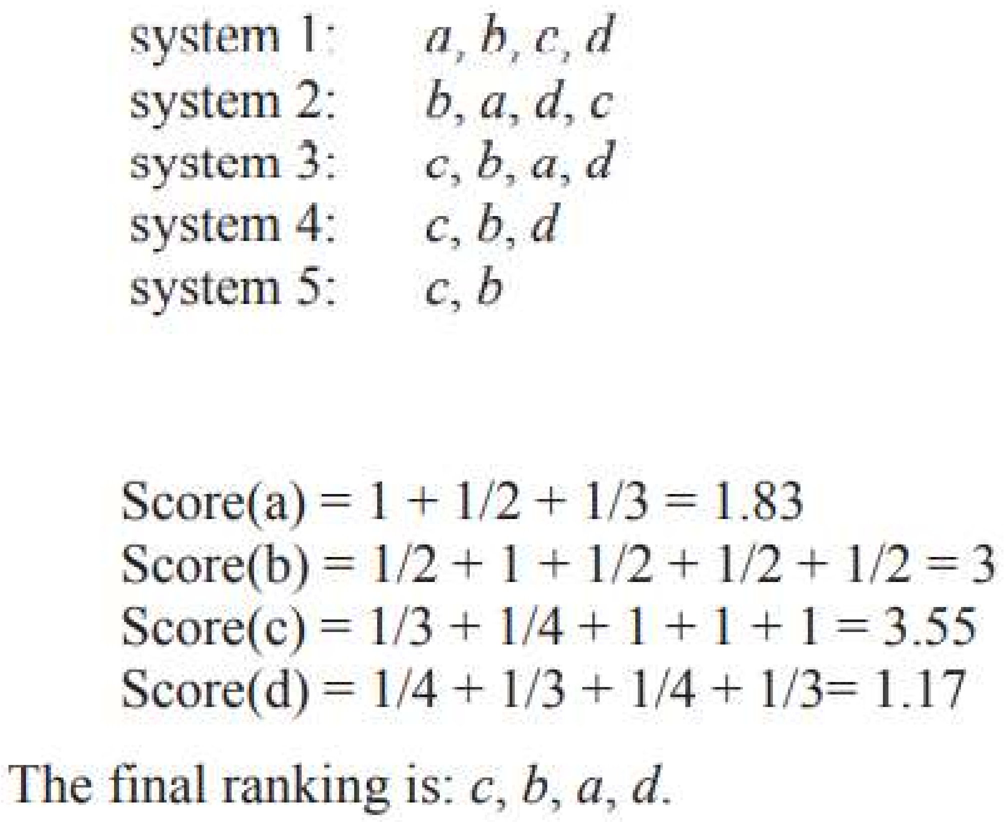
system 1:	*a*, *b*, *c*, *d*
system 2:	*b*, *a*, *d*, *c*
system 3:	*c*, *b*, *a*, *d*
system 4:	*c*, *b*, *d*
system 5:	*c*, *b*



	win	*lose*	*tie*
a	1	2	0
b	2	1	0
c	3	0	0
d	0	3	0

The final ranking is: c, b, a, d.

Reciprocal Ranking



system 1:	*a*, *b*, *c*, *d*
system 2:	*b*, *a*, *d*, *c*
system 3:	*c*, *b*, *a*, *d*
system 4:	*c*, *b*, *d*
system 5:	*c*, *b*

Score(a) = 1 + 1/2 + 1/3 = 1.83
Score(b) = 1/2 + 1 + 1/2 + 1/2 + 1/2 = 3
Score(c) = 1/3 + 1/4 + 1 + 1 + 1 = 3.55
Score(d) = 1/4 + 1/3 + 1/4 + 1/3= 1.17

The final ranking is: *c*, *b*, *a*, *d*.

Web Spamming

- Human activities that deliberately mislead search engines to rank some pages higher than they deserve.
- Example
  - Assume that, given a user query, each page on the Web can be assigned an information value.
  - All the pages are then ranked according to their information values.
  - Spamming refers to actions that do not increase the information value of a page, but dramatically increase its rank position by misleading search algorithms to rank it high.

Content Spamming

- Important words/phases may be added at
  - Title
  - Meta-Tags
  - Body
  - Anchor Text
  - URL

  For example, a URL may be http://www.xxx.com/cheap-MP3- player-case-battery.html

Main Techniques Used

- Repeating some important terms: This method increases the TF scores of the repeated terms in a document and thus increases the relevance of the document to these terms.
- Dumping of many unrelated terms: This method is used to make the page relevant to a large number of queries.

Link Spamming

- Creating a honey pot: If a page wants to have a high reputation/quality score, it needs quality pages pointing to it.
- Posting links to the user-generated content (reviews, forum discussions, blogs, etc): There are numerous sites on the Web that allow the user to freely post messages, which are called the user-generated content.

Hiding Technique

- Content Hiding: Spam items are made invisible. One simple method is to make the spam terms the same color as the background color.

```
<body background = white>
<font color = white> spam items</font>
 …
</body>
```

Hiding Technique

- Cloaking: Spam Web servers return a HTML document to the user and a different document to a Web crawler. In this way, the spammer can present the Web user with the intended content and send a spam page to the search engine for indexing.
- Redirection: Spammers can also hide a spammed page by automatically redirecting the browser to another URL as soon as the page is loaded.