CSE-453

Data Mining

Assignment-02

Submitted by:

Group-02

Written by:

Ayon Roy

(201714018)

Sec-B

CSE-17

Ayon Roy 201714018

Question

Tutorial B:

→ How do Meta Search engines work?

Answer:

A meta search engine is an online Information retrieval tool that uses the data of a web search engine to produce its own results. Meta search engines take input from a user and immediately query search engines for results. Sufficient data is gathered, ranked and presented to the users.

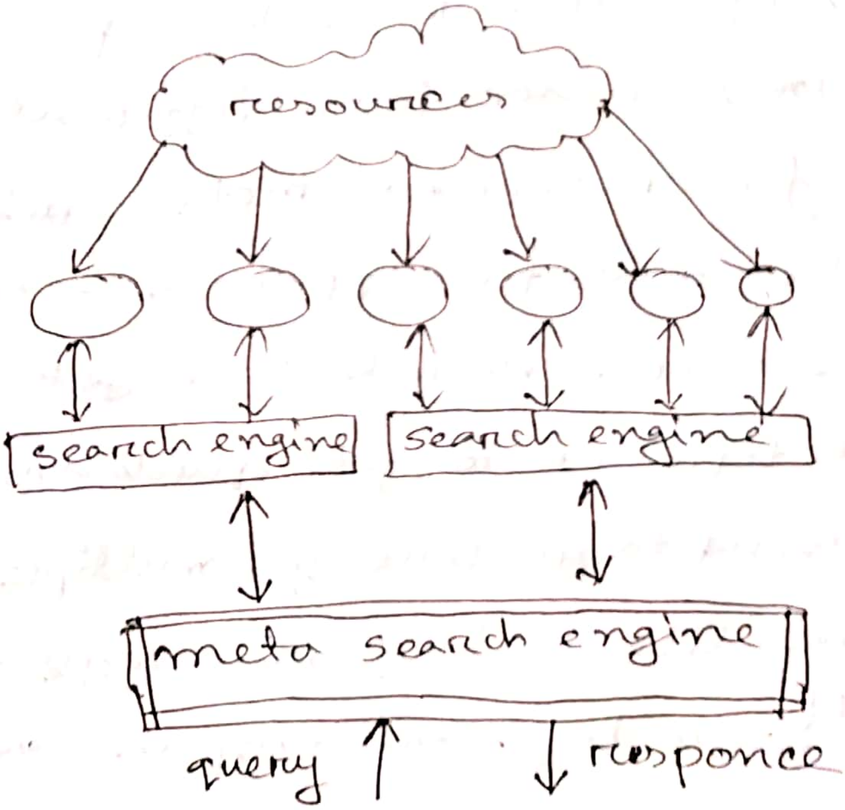

Fig: Architecture of a meta search engine

By sending multiple quercies to several other search engines this extends the coverage data of the topic and allows more information to be found. They use the indexes built by other search engines, aggregating and often post-processing results in unique ways. A meta search engine has an advantage over a single search engine because more results can be retrieved with the same amount of exertion. It also reduces the work of users from having to individually type in searches from different engines to look for resources. meta ~~seach~~ searching is also a useful approach if the purpase of the user's search is to get an overview of the topic or to get quick answers. Instead of having to go through multiple search engines like Yahoo! or Google and comparing results, meta search engines are able to quickly compile and combine results. They can do it either by listing results from each engine queried with no additional post-processing (Dogpile) or by analyzing the results and ranking them by their own rules (IxQuick, Metacrawler and Vivismo).

A meta search engine doesnot create a database of web pages but generates a Federated database system of data integration from multiple sources.

Since every search engine is unique and has different algorithms for generating ranked data, duplicates will therefore also be generated. To remove duplicates, a meta search engine processes this data and applies its own algorithm. A revised list is produced as an output for the user.

There are two main classes of meta-search combination (or fusion) algorithms: ones that use similarity scores returned by each component system and ones that do not. Some search engines return a similarity score (with the query) for each returned page, which can be used to produce a better combined ranking. We discuss these two classes of algorithms below.

(1) Combination Using Similarity Scores:

Let the set of candidate documents to be ranked be $D = \{d_1, d_2, \ldots d_N\}$. There are 'k' underlying systems (component search engines or ranking techniques). The ranking from system or technique i gives document 'd_j' the similarity score, s_{ij}.

⊞ CombMIN:

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

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

⊞ CombMAX:

The combined similarity score for each document d_j, is the maximum of the similarities from all underlying search engines:

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

⊞ CombMNZ:

It is defined as:

$$\text{CombMNZ}(d_j) = \text{CombSUM}(d_j) \times r_j$$

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

(2) Combination Using Rank Positions:

The algorithms discussed below are based on voting in elections.

Election By Order of Merit: 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. The points from all voters are summed up to give the final points for each candidate. If there are candidates left unranked by a voter, the remaining points are divided evenly among the unranked candidates. The candidate with the most points wins. This method is called "Borda Ranking".

Condorcet Ranking: The condorcet ranking algorithm is a majoritarian omethod where 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: Reciprocal ranking sums one over the rank of each candidate across all voters, For each voler, 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. This rank strategy gives much higher weight than Borda ranking to candidates that are near the top of a list.

——— × ———