MILITARY INSTITUTE OF SCIENCE & TECHNOLOGY

CSE 453 Data Mining Assignment # 2

Submitted by:

Group – 02

(Ayon Roy)

Group Members:

1. 201514178 – Capt Akib Zaman
2. 201714014 – Abdullah-al-Sheak Jaber
3. 201714018 – Ayon Roy
4. 201714024 – Md. Aqib Alfaz
5. 201714043 – Nafiz Imtiaz Khan

Submitted to:

Col Siddharth Malik, SM

Question

Set 2

Tutorial A

Web Search has its root in Information retrieval

- Briefly explain how is Information Retrieval carried out?
- Briefly explain the Vector space model for information retrieval?

Tutorial B

How do Meta search engines work?

"How do meta search engines work?"

→ 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 result. Sufficient data is gathered, ranked and presented to the users.

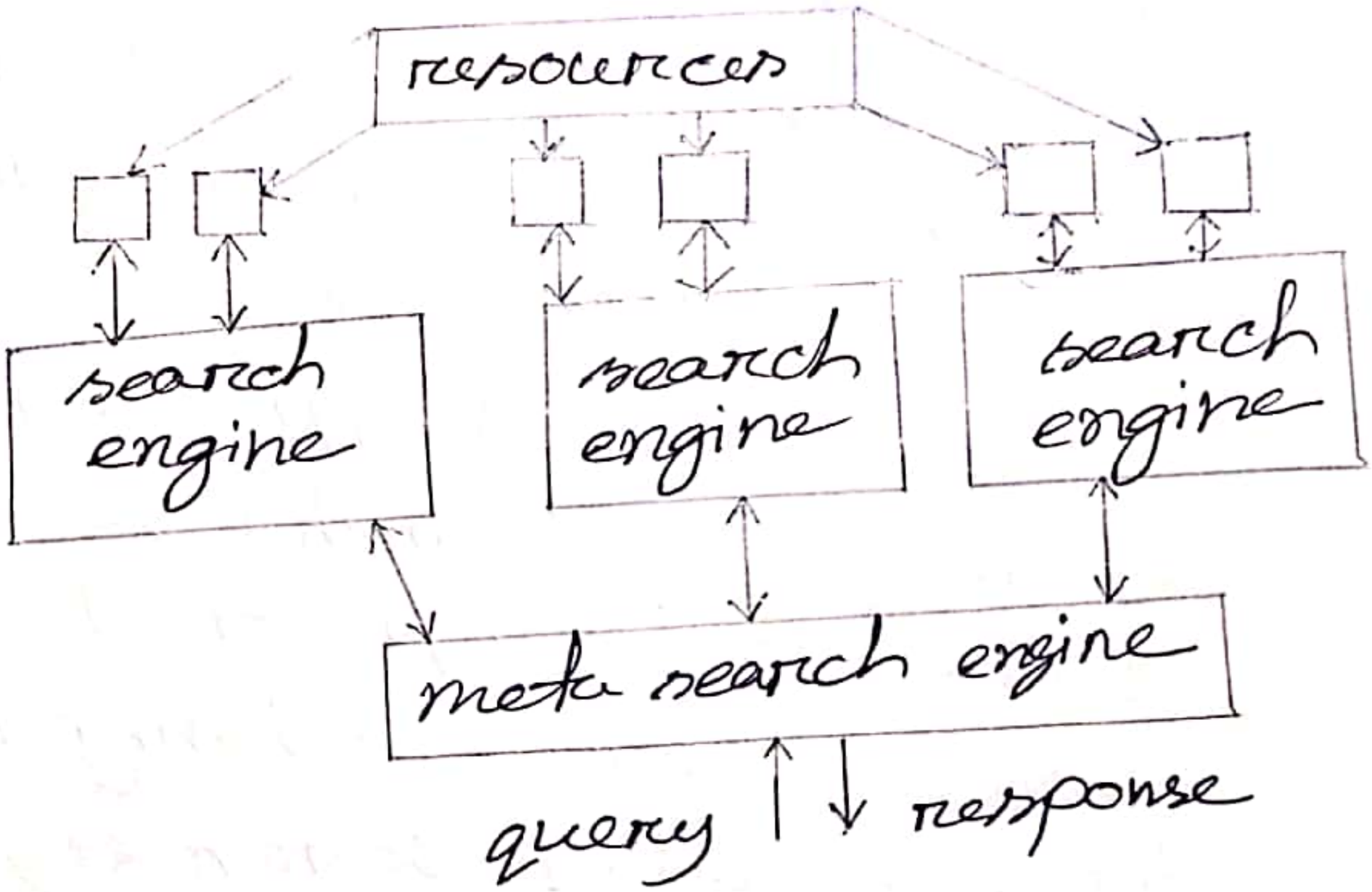

Fig: Architecture of a meta search engine

By sending multiple queries 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 ~~unig~~ 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 searching is also a useful approach in purpose 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 (Ix Quick Metacrawler, and Vivismo).

A meta search engine accepts a single search request from the user. This search request is then passed on to another search engine's database. A meta search engine does not 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 its data and applies its own algorithm. A revised list is produced as an output for the user.

Architecture of ranking:

Web pages that are highly ranked on many search engines are likely to be more relevant in providing useful information. However, all search engines have different ranking scores for each website and most of the time these scores are not the same. This is because search engine prioritise different criteria and methods for scoring. Hence, a website might appear highly ranked on one search engine and lowly ranked

on another. This is a problem because meta search engines rely heavily on the consistency of this data to generate reliable accounts.

Fusion: A meta search engine uses the process of Fusion to filter data for more efficient results. The two main fusion methods are:

(1) Collection Fusion: It's also known as distributed retrieval, deals specifically with search engines that index unrelated data. To determine how reliable, valuable these sources are collection fusion looks at the content and then ranks the data on how likely it is to provide relevant information in relation to the query. From what is generated collection fusion is able

to pick out the best resources from the rank. These chosen resources are then merged into a list.

(ii) Data Fusion: Deals with information retrieved from search engines that indexes common data sets. The process is very similar. The initial rank scores of data are merged into a single list, after which the original ranks of each of these documents are analysed Data with high scores indicate high level of relevancy to a particular query and are therefore selected. To produce a list, the scores must be normalized using algorithms such as CombSum.. This is because search engines adopt different politics policies of algorithm resulting in the score produced being incomparable.

There are two main classes of meta-search combination (or fusion) algorithms: ones that use similarity scores (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. It is worth nothing that the first class of algorithms can also be used to combine scores from different similarity functions in a single IR system or in a single search engine. Indeed, the algorithms below were originally proposed for this purpose. It is likely that search engines already use some such techniques (or their variations) within their ranking mechanisms because a ranking algorithm needs to consider multiple factors.

(1) Combination Using Similarity Scores:

Let the set of candidate document to be ranked be $D = (d_1, d_2, d_3 \ldots, d_N)$. 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}.

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

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

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

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

iii) CombMNZ : It is defined as.

$$CombMNZ(d_j) = CombSUM(d_j) \times n_j$$

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

(2) Combinations using Rank Positions :

We now discuss some popular rank combinations methods that use only rank positions for each search engines. The algorithms discussed below are based on voting in elections.

In 1770 Jeans Charles de Borda proposed "election by order of merit". Each voter announces a linear preference order on the candidates. For each voter the top candidate recieves n points (if there

are n candidates in the election), the second candidate recieves 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 among the unranked candidates. The candidate with the most points wins. This method is called the Borda ranking.

The condorcet ranking algorithm is a majoritarian method where the winner of the electon 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 losses to all other ranked candidates. All unranked candidates tie with one another.

Yet another simple method, called the 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 thier final total scores. This rank strategy gives much higher weight than Borda ranking to candidates that are near the top of a list.