1. Given a query that a user submits to an IR system and the top N documents that are returned as relevant by the system, devise an approach (high level algorithmic steps will suffice)

- to suggest query terms to add to the query. Typically, we wish to give a large range of suggestions to the users capturing potential intended query needs, i.e., high diversity of terms that may capture the intended query context/content.

ANSWER:

Given a query Q that a user U submits to an IR system and the top N documents that are returned as relevant by the system, the solution needs to deliver a list of very different but connected suggestions (S). To the query Q suggestions S can be included, which may then be used on the IR system to obtain documents.

The algorithm to suggest query terms to add to the query will be Select, Link and Rank Algorithm which has been cited in the paper [1].

Basically, it has three phase: select, link and rank. In select phase, relevant documents to the given query submitted by user are retrieved. Then, terms whose distribution inside documents differs from corpus of documents is chosen using Bo1 model (informativeness metric that quantifies divergence from randomness).

$$\mathrm{Bo1(t)} = \mathrm{f}\big(\mathrm{t},\ \mathrm{ResK}(\mathrm{Q},\ \mathrm{D})\big) \times \log_2 \frac{1 + \left(\frac{\mathrm{f(t,\ D)}}{|\mathrm{D}|}\right)}{\frac{\mathrm{f(t,\ D)}}{|\mathrm{D}|}} + \log_2\left(1 + \left(\frac{\mathrm{f(t,\ D)}}{|\mathrm{D}|}\right)\right)$$

Here, $\mathrm{f}\big(\mathrm{t},\ \mathrm{Res}_n(\mathrm{Q},\ \mathrm{D})\big)$ = frequency of term t in the document collection $\mathrm{Res}_n(\mathrm{Q},\ \mathrm{D})$

$\mathrm{Cand(Q,\ D)}$ = selected terms

In link phase, words in Cand(Q, D) are linked to Wikipedia pages for more details which are then added to the query. We receive relatedness score which is useful for ranking words for expansion. In rank phase, a graph is created of Wikipedia entity network. Based on the significance weight is assigned to each node in graph. For linking entities based on their diversity and relevance, a diversity-aware version of PageRank that performs a vertex reinforced random walk (VRRW) is used to score the nodes in the network.

Steps for algorithm are as follows -

Input: Documents D, Query Q

Output: List of expansion terms, E

SELECT PHASE

1. We obtain documents that matches search query Q

2. Select terms as Cand(Q, D)

LINK PHASE

3. Link each term t in Cand(Q, D) to wikipedia pages
4. Linked entities be t.E
5. Relatedness score be r(t, e)

RANK PHASE

6. Construct G(Q), a graph of linked entities and neighbours
7. Score each entity using relatedness to linked terms
8. Perform VRRW on G(Q), entity scores initialized using (7)
9. Collect the top-scored entities based on VRRW scores as E
10. Construct E, a diversified term ranking using entity scores and term-entity relatedness

2. Consider the following scenario: a company search engine is employed to allow people to search a large repository. All queries submitted to the system are recorded. A record that contains the id of the user and the terms in the query is stored. The order of the terms is not stored and neither is any timestamp. Each entry in this record is effectively an id and a set of terms. Any duplicate terms in a query are ignored.

The designers of the search engine, decide to use this information to develop an approach to make query term suggestions for users, i.e., at run time, once a user an entered their query terms, the system will suggest potential extra terms to add to the query.

Given the data available, outline an approach that could be adopted to generate these suggested terms. A brief outline is sufficient to capture the main ideas in your approach.

The designers of the system wish to take into evidence in previous queries and also any similarities between users. Identify the advantages and disadvantages of your approach (briefly).

ANSWER:

According to the given scenario, it is a company search engine, which means the data stored is domain specific. Additionally, all the queries submitted to the system are recorded. Therefore, based on this information the best approach for giving suggestions to the users at run time would be the combination of Contextual and Collaborative Filtering.

Firstly, we can use contextual filtering which takes into account the content and context of the queries and the documents available in the large repository. This includes analyzing the content of document using topic modelling or keyword extraction, and then creating user profiles based on previous queries they submitted. And then recommending query terms by matching user and items profile.

Secondly, we can use collaborative filtering as this considers the similarities between the users based on their previous search activity. If set of words used in queries frequently overlap between two users, then they are considered similar users. And then we can suggest more query terms based on what other similar users have search for and found useful in the queries.

ADVANTAGES

1. Using the past queries submitted by the user we can give highly customised recommendations by looking into specific topics or terms that users of the company has shown their interest into.
2. As our approach understands and also focuses on the content, using the company's large repository we can give more relevant recommendations.
3. Because we also consider user similarities, we get more information about user preferences, have expanded user search and ultimately leads to satisfaction with the query recommendations.

DISADVANTAGES

1. Because of the extensive examination of the content and the user queries, the efficiency of searching for recommendations will decrease.
2. We can be constrained with showing new search keywords if there is a limited search history performed by the users.

REFERENCES:

[1] Select, Link and Rank: Diversified Query Expansion and Entity Ranking Using Wikipedia (Lecture Notes in Computer Science book series (LNISA,volume 10041)