**Q.1.**

1. (a)  Describe the vector space model approach to information retrieval. Your answer should include a description of the query and document representations and also the comparison approach used. (8) **见2018-2019 Q2-a**
2. (b)  The accuracy of the vector space model depends on the quality of the weighting of the terms in both the query and documents. Discuss, with reference to well-known weighting schemes, the main components and properties of a good weighting scheme. (9) **见2018-2019 Q2-b**
3. (c)  The Extended Boolean model has been used to allow users express their queries in Boolean algebra and have documents ranked in order of relevancy by adopting a more expressive weighting scheme than afforded in the traditional Boolean model. Explain such an approach. Discuss the advantages and limitations of adopting such a model. (8) **见论文或去网上搜**

Extended Boolean model combines the control of the Boolean model and the ranking capability of the vector space model into a uniform framework. Users can use AND and OR in the queries (as in the Boolean model) but keywords are weighted (as in the vector space model). Furthermore, documents are ranked by a similarity function that is designed to exhibit intended behaviors.

In the following, we assume that the weight of a keyword is normalized to [0,1]. For example, the tfxidf weighting formula can be modified as:

Diagram, text

Description automatically generated

where wx,j is the weight of term x in document j, the term frequency tfx,j is normalized by the highest tf in document j, and the idf of term x is normalized by the highest idf among all keywords in the collection.

**Q.2.**

1. (a)  Empirical evaluation of information retrieval systems plays an important role in information retrieval research. Define and discuss the following metrics that can be used to measure the performance of an Information Retrieval system: precision, recall, novelty and coverage. (9) **见NOTES**

**Precision and recall** are the most commonly used metrics. Given a set D and a query Q, let R be the set of documents relevant to Q. Let A be the set of documents actually returned by the system.

Precision is defined as the percentage of documents returned to the user that are actually relevant to the user: 

Recall is defined as the percentage of relevant documents in the whole col- lection that are returned to the user: 

Generally, the IR systems return a ranked list to the user in descending order. We can obtain a more accurate representation of the quality of the ranking by **plotting precision against recall** for a number of points. Besides, if we are comparing several systems across many collections, we can use single value measures like MAP, E measure, etc.

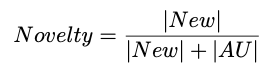
Another two metrics are **coverage and novelty** which involve user’s experience. Let U be a subset of the relevant documents in the collection previously known to the user (U ⊂ R). Let AU be the set of returned documents known to the user.

Coverage will be the percentage of the known relevant documents that are returned in the answer set to the user.

Text

Description automatically generated

A related concept is that of novelty. Let New refer to the set of relevant doc- uments returned to the user that were previously unknown to the user. Clearly, if coverage is high, then novelty will be low and vice versa. We can define novelty as:



1. (b)  Discuss potential uses of clustering algorithms in the domain of information retrieval. Outline any clustering algorithm and discuss any limitations associated with that algorithm. (8) **见2019-2020 Q4-C**

Clustering involves the task of grouping data points into homogeneous classes or clusters. So that items in the same cluster are as similar as possible and items in different classes are as dissimilar as possible. All applications of clustering in IR are based (directly or indirectly) on this cluster hypothesis. Van Rijsbergen’s original wording: “closely associated documents tend to be relevant to the same requests”.

Clustering can be used to improve search recall: Cluster docs in collection are a priori; When a query matches a doc “d”, also return other docs in the cluster containing “d”.

Table

Description automatically generated

1. (c)  Discuss suitable approaches to evaluating the usefulness of a clustering algorithm. (8) **见2018-2019 Q3-c**

**Q.3.**

1. (a)  Link analysis has been used in many modern web search engines. Describe, with a suitable example, how link analysis can be applied to the answer to obtain a new ranking over the returned set for a specific query. Outline any difficulties or limitations of this approach. (8) **见2019-2020 Q4-a**

**Answer:** A company has a large set of scientific articles. They wish to rank all papers that are relevant to a given query and to then re-order the papers in the answer set according to how authoritative or influential the papers are. We can use **HITs algorithm** to re-rank the returned articles.

* The returned answer set is the root R. We initialize base set S to R. And add to S all articles pointed to by any paper in R; Add to S all pages that point to any page in R.
* Assign to each article p∈ *S:*

an authority score: *ap* (vector***a***)

ahub score*: hp* (vector***h***)

Initialize all *ap = hp* to some constant value. (i.e.let this constant be 1)

Here, the authority score denotes the how authoritative or influential the article is. If an article appearing in a large number of good articles, then it normally has a high authority score. If an article points to lots of good authorities, it has a high hub score.

* Define M to be the adjacency matrix and make Mij =1 where .
* MTM can generate the authority vector ***a***. And MMT can return the hub vector ***h***.

Update the authority and hub scores:

***a←*** MTM ***a***

***h←*** MMT ***h***

* Iterate until convergence and the final authority and hub scores are generated.
* Re-rank the articles by the authorities scores in descending sequence.

**Difficulties/ Limitations**

**There exists problems with identifying authoritative pages:**

•authoritative pages do not necessarily refer to themselves as such

•many links are purely for navigational purposes

•advertising links

**Limitations of link only approach:**

1. on narrowly focussed query topics, there may not be many exact references and the hubs may provide links to more general pages
2. potential drift from main topic. All links are treated as being equally important. If there is a range of topics in a hub, the focus of the search may drift
3. timeliness of recommendation is hard to identify
4. sensitivity of malicious attack
5. edges with wrong semantics
6. (b)  Define what is meant by collaborative filtering. Describe, with a suitable example, the main stages involved in generating a recommendation for a user via collaborative filtering. (10) **见2018-2019 Q3-A**
7. (c)  Classical collaborative filtering considers ratings provided by users for items. In many domains, extra information is also available regarding users and items; discuss how this extra information could be used to increase the coverage of a collaborative filtering system. (7) **见2019-2020 Q3-b**

Collaborative filtering has many issues like data sparsity problem and cold start problem. For example, it is unable to provide accurate recommendations when users and items have few ratings, resulting in reduced coverage. However, we can extract information from items and users and finding regularities in the content. Incorporating the content filtering and collaborative filtering can improve IR performance.

We can use user profile information such as gender, postcode, occupation, and their tastes, preferences to more accurately cluster similar users and identify their preferred items; Or we can use knowledge about how a particular object satisfies the user needs from the item description or user judgements to more accurately group similar items or match the target users. These need some NLP techniques because we need to extract text information. The information is extracted and represented as keywords with weights. And then, similarity between the user profile vector and the item feature vector can be calculated based on cosine angle or some traditional machine learning models. Hence, we can generate user-item matrix and user-user matrix based on the above content information.

As we know, collaborative filtering creates a user-user matrix based on history ratings of users and calculate user similarity based on the vectors of rating representations. To add content-based filtering, when generating a user-user or user-item similarity, we can combine the similarities generated by content filtering and collaborative filtering together with suitable weights respectively. For example, assign equal weights to content and collaborative filtering at the beginning. When facing a new user or item, the weight of content filtering score can be increased and the weight of collaborative filtering can be decreased.

**Q.4.**

(a) User feedback is often used to modify a user’s query with the aim of improving retrieval performance. Outline such an approach for the vector space model. (8) **见20180-2019 Q4-b**

(b) Query augmentation can also take place without explicit user feedback. Outline an approach to automatically generate suggested keywords for a user to augment their query. (9) **见20180-2019 Q4-c**

(c) Describe the term query difficulty and outline measures to predict the difficulty of a query. (8) **见2019-2020 Q1-C**

**Difficulty**

Most IR systems exhibit large variance in performance in answering users’ queries. These can be caused by the query itself (ambiguous terms), vocabulary mismatch problem or missing content queries, as well as the robustness problem in IR. For example:

- failure to recognise all aspects in the query

- failure in pre-processing

- over-emphasis on a particular aspect/term

- query needs expansion

- need analysis to identify intended meaning of query (NLP)

- need better understanding of proximity relationship among terms.

**Measures to predict**

Approaches can be categorized as pre-retrieval approaches (estimate difficulty without running the system) and post-retrieval approaches (run the system against query and examine results).

* **linguistic approaches (Pre-retrieval)**

-Use some NLP approaches to analyse query

-Use external sources of information to identify ambiguity etc.

(Most linguistic features do not correlate well with performance.)

* **Statistical approaches (Pre-retrieval)**

-Take into account **the distribution of the query term frequencies** in the collection (e.g., consider idf and icf of terms)

-Take into account ***specificity* of terms** (Queries containing non-specific terms are considered difficult).

***- Term relatedness***

If query terms co-occur frequently in collection, we expect good performance.

Mutual information or Jaccard coefficient etc. can be used

***- Query scope***

what percentage of documents contain at least one query term, if a lot then this is probably a difficult query.

***- Simplified query scope***

Measures difference between language model of collection with language model of query.

* **Three main categories of post retrieval approaches:**

**-Clarity measures**

Attempts to measure the coherence in the result set. The language of the result set should be distinct from the rest of the collection. Compare language model induced from answer set and one induced from the corpus/collection. Related to the cluster hypothesis

**-Robustness**

Explores robustness of system in the face of perturbations to:

1. Query

Overlap between query and sub-queries. In difficult queries some terms have little or no influence

1. Documents

Compare system performance against collection C and some modified version of C

1. Retrieval performance

Submit same query to many systems over same collection; divergence in results tells us something about difficulty of query

**-Score analysis**

Analyse score distributions in returned ranked list:

- difficulty can be measured based on distribution of values; is cluster hypothesis supported?

- can look at distribution of scores in answer set and document set and attempt to gauge difficulty

- relatively simple measures shown to be effective