**Q.1.**

(a)  Learning mechanisms has been used successfully to search for suitable means to combine sources of evidence in information retrieval. Discuss such an approach applied to a problem of your choice in information retrieval. Your answer should also identify the strengths and weaknesses of this approach. (10)

**Answer:**

**Genetic Algorithms**

Inspired by Darwinian theory of evolution.

At each step of algorithm, the best solutions are selected while the weaker solutions are discarded.

Uses operators based on crossover and mutation as the basis of the algorithm to sample space of solutions.

**Genetic Algorithms**

Inspired by Darwinian theory of evolution.

At each step of algorithm, the best solutions are selected while the weaker solutions are discarded.

Uses operators based on crossover and mutation as the basis of the algorithm to sample space of solutions.

Diagram

Description automatically generated

**Fitness**

Need an an evaluation function which will discriminate between better and worse solutions.

**Crossover**

Example of one-point crossover: **11001011** and 11011***111*** gives **11001*111*** and 11011**011**

**Mutation**

Occurs in the GA at a much lower rate than the crossover.

Important in order to add some diversity to the population in the hope that new better solutions are discovered and therefore it aids in the evolution of the population.

***Tournament Selection****:*

A number individuals are selected at random with replacement from the population

The individual with the best score is selected

This is repeated n times.

**Case study 1: GA approach**

Each genotype can be a vector of length N (the size of the lexicon).

Set all rates randomly initially.

Run system with a set of queries to obtain fitness; select good chromosomes; crossover; mutate.

Effectively searching landscape for weights to give good ranking

Several examples: Etzioni (one of the first).

**Case study 1: GA approach**

Each genotype can be a vector of length N (the size of the lexicon).

Set all rates randomly initially.

Run system with a set of queries to obtain fitness; select good chromosomes; crossover; mutate.

Effectively searching landscape for weights to give good ranking

Several examples: Etzioni (one of the first).

**Case study 2: application of genetic programming to IR**

Evolutionary computing approaches:

- evolutionary strategies

- genetic algorithms

- genetic programming

**Advantages**:

* Produces a symbolic representation of a solution which is useful for further analysis.
* Using training data, MAP can be directly optimised (i.e. used as the fitness function)
* Solutions produced are often generalisable as solution length (size) can be controlled

**Issues**:

* **Choice of representation for encoding individuals**
* **Definition of fitness function**
* Definition of selection scheme
* Definition of suitable genetic operators
* Setting of parameters:

size of population

number of generations

probability of crossover

probability of mutation

etc.

(b)  Many modern web-based search engines attempt to take into account the web link structure in addition to the content of the pages. Describe the Page Rank algorithm that uses information embedded in the web link structure to return relevant documents to a user. (10)

**A Markov chain has two components**:

* A graph/network structure; each node is called a state
* A transition probability of traversing a link given that the chain is in a state.

(The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.)

A sequence of steps through the chain is called a random walk.

**Random Sufer Model**:

* Assume the web is a Markov Chain.
* Sufers randomly click on links, where the probability of an out-link from page A is 1/n, where there are “n” out-links from A.
* The sufer occasionally gets bored and is moved to another web page (teleported 传送), say B, where B is equally likely to be any page.
* The PageRank of a web page is the probability that the sufer will visit that page.

**Formula**:

Where W is a web page

Wi is the web page that has a link to W

is the number of outlinks from Wi

T is the teleportation probability

N is the size of the web

**Advantages**:

* Efficiency: early experiments on Google showed convergence in 52 iterations on a collection with 322 million links. Number of iterations required for convergence is empirically O(log n) where n is the number of links. This is quite efficient.

**Weakness**:

* Issue of Dangling pages. A page with no out-going links; cannot pass on rank.

(Solution: assume page has links to all pages with equal probability)

* Rank Sink: Pages in a loop accumulate rank but do not distribute it.

(Solution: “Teleportation” (隐形传送), i.e. with a certain probability the sufer can jump to any other web page to get out of the loop)

**Difference between HITS and PageRank**

* PageRank is independent on the query. We have ranked the pages in advance and just pay attention to the positions of pages on the graph.
* HITS  algorithm is in the same spirit as  PageRank . They both make use of the link structure of the Weg graph in order to decide the relevance of the pages. However, unlike the  PageRank  algorithm,  HITS  only operates on a small subgraph (the seed  S Q ) from the web graph. This subgraph is query dependent; whenever we search with a different query phrase, the seed changes as well.  HITS  ranks the seed nodes according to their authority and hub weights. The highest ranking pages are displayed to the user by the query engine.

(c)  Illustrate, with a suitable example, how the rank of the pages may be calculated according to the PageRank algorithm. (5)

**Answer**

Assume a small universe of four web pages: A, B, C and D. Links from a page to itself, or multiple outbound links from one single page to another single page, are ignored. The link structure is as follows:

PageRank is initialized to the same value for all pages, let’s assume 1/4.

The original formula is:

(Where W is a web page, Wi is the web page that has a link to W, is the number of outlinks from Wi, T is the teleportation probability, N is the size of the web)

Here, we ignore T/N and (1-T) and just use:

The calculation is as follows:

Text, letter

Description automatically generated

The algorithm works until convergence. Here, we simply assume it convergences after two iterations. And then we rank the pages.

**Q.2**

(a)  The vector space model for Information Retrieval is one of the most commonly adopted models. Outline the model explaining both the representation of queries and documents and a means to calculate similarity. Discuss the advantages and disadvantages of such a model. (9)

**Answer:**

Representation and comparison technique depends on the information retrieval model chosen. The choice of feedback techniques is also dependent on the model chosen.

**(Representation)**

We can view any IR model as comprising:

* D is the set of logical representations of the documents.
* Q is the the set of logical representations of the user information needs (queries).
* F is a framework for modelling these representations (D and Q) and the relationship between D and Q.
* R is a ranking function which defines an ordering among the documents with regard to any query qi .

We have a set of index terms: t1 ...tn  
A weight wi,j is assigned to each term ti occurring in document dj

We can view a document or query as a vector of weights:

= (w1,w2,w3 ...)

= (w1,w2,w3 ...)

(**Similarity**)

The vector space model attempts to improve upon the Boolean model by removing the limitation of binary weights for index terms. In the vector space model, terms can have a non-binary value for both queries and documents.

There terms weights are used in the comparison between documents and queries. We can then sort the documents based on their degree of similarity and return a ranked list to the user. The similarity is calculated between the vector representations of the document and the query. We can measure the cosine of the angle between the two vectors:

Sim(q,d)=cos() = =

**(Advantages)**

* Simplicity;
* Term weights are not binary;
* Improved performance over the Boolean model due to weighting schemes Partial matching allowed which gives a natural ranking;
* Allows computing a continues degree of similarity between queries and documents.

**(Disadvantages)**

* Terms are considered to be mutually independent.
* The order in which the terms appear in the document is lost in the vector space representation.
* Semantics sensibility: Suffers from synonym (同义词)and polysery (多义词); needs NLP techniques.

(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 of a good weighting scheme. (10) 同19年Q2(c)

**Answer:**

A good weighting scheme comprises local factors, global factors (collection wide) and query related features:

sim(q,d)=

Where TF(D) is the normalized term frequency in a document; gWt(C) is the global weight of a term across a collection; qWt(Q) is the query weight of a term in Q. For example, BM25/Okapi weighting scheme penalizes the long document, high term frequency and the words in a large number of documents, providing high retrieval accuracy

Text

Description automatically generated



An axiomatic approach to IR (Fang and Zhai 2005) has been developed which refines a number of constraints (axioms) (Fang et al. 2004) to which all good weighting functions should adhere.

* Constraint 1: adding a new query term to the document must always increase the score of that document.
* Constraint 2 : adding a non-query term to a document must always decrease the score of that document.
* Constraint 3: adding successive query terms to a document should increase the score of the document less with each successive addition. For example, encountering a term 30 times does not increase the likelihood of relevance by a factor of thirty.
* Constraint 4: Ensuring that the document length factor is used in a sub-linear function will ensure that repeated appearances of non-query terms are weighted less.

(c)  Prior to calculating the similarity between a query and a document, documents and queries are often pre-processed using stemming and stopword removal prior to assigning weights to terms. Explain what is meant by pre-processing and the possible effects it may have on the retrieval process. (6)

**Answer:**

Application of a set of well-known techniques to the documents and queries prior to any comparison. The document pre-processing phase involves applying a well known set of techniques to the document collection to convert it to a format more suitable to the task at hand. There are many potential pre-processing approaches with certain approaches developed for specific requirements of particular domains or particular user needs.

* Stemming: Stemming algorithms attempt to remove common suffixes from terms occurring in the documents. The overall goal is to reduce sim- ilar words to a common root form by identifying morphological derivations of words. There are many approaches in the literature and in commercial systems. Lovin’s algorithm and Porter’s stemmer are two of the well- known classic stemmers
* Stop word removal: This involves the removal of highly frequent words/terms from documents. These words add little semantic meaning to the docu- ment and usually include articles and conjunctions.
* Thesaurus (词库) construction: Thesauri used to try to identify synonyms within the documents. Manually or automatically created.

**Effects:**

* The 100 most frequent words add-up to about 50% of the words in a document. Hence, stopword elimination improves the size of the indexing structures.
* As stemming replaces all the variants of a word with the single stem of the word, it improves the storage and search efficiency. For example, less terms are stored. Also, it improves recall. For example, we want to match the term “connection” between documents and the query. After stemming, we can match addition documents which have “connected”, “connecting”, “connects” as well. However, on the other hand this decreases the precision because users can no long target just a particular form.
* Thesaurus (词库) construction standards the index terms that were selected which helps users locate proper query terms.
* Preprocessing can also reduce document length, save space and processing time. In a vector space model, it can also reduce vector dimensions and increase computation efficiency.

**Q.3.**

(a)  Explain what is meant by collaborative filtering. Explain briefly the main stages involved in generating a prediction/recommendation for users. (5)

**Answer:**

Collaborative filtering collects human judgments and match people who share same information needs and tastes. Users share their judgments and opinions; echoes “word of mouth" principle.

It is based on the principle that users who have similar interests normally make the similar ratings.

**Steps:**

* Create a user-item matrix based on the data sets of user identifiers, item identifiers, and ratings by users of items which are numeric values in some predefined range.
* Calculate user correlation which indicates how similar each user is to every other user. (i.e. Pearson correlation)
* Form groups or neighborhoods of users who are similar. (i.e. using correlation thresholding)
* In each group, make recommendations based on what other users in the group have rated. (i.e. compute the weighted mean of all neighbors’ ratings)

(b)  Discuss what you consider to be the main limitations of a collaborative filtering approach and suggest, with suitable examples, approaches to overcome these limitations. (10)

* **Issue of sparsity of matrix**:

there would be many users and many items, but any user would only have rated a small percentage of all items in the dataset.

Approaches: We can use a technique such as singular value decomposition (SVD) to reduce the data space. Thus, due to this reduction a correlation may be found between similar users who do not overlap ratings in the original matrix of ratings. Also, principle component analysis can also be used to reduce the dimensionality.

* **the cold start problem:**

when a new user or item has just entered the system, it is difficult to find similar ones because there is not enough information. New items cannot be recommended until some users rate it.

Approaches: Perhaps use weighted average of global mean and user’s (or item’s).

* **Issue of size of matrix**:

In general, the size of matrix is very large. It affects the computational efficiency.

Also, SVD can be used to improve scalability by dimensionality reduction.

* **Noise in matrix**:

Increased noise (or sabotage破坏活动) is another challenge, as the user population becomes more diverse. Ensembles of maximum margin matrix factorizations and instance selection techniques are found useful to address the noise problems of CF tasks

* **the ratings would change for items over time**

Thus, making the ratings not reliable.

Approaches: make the model time dependencies. i.e. add annotations to the document which maintains a set of values regarding the temporal aspects such as creation time, modification times, last modification time, etc.

* **Shilling Attacks (先令攻击)**

The ratings may be not reliable. For example, fake accounts may lead to batch or malicious scoring.

Approaches: It is desirable for CF systems to introduce precautions that discourages this kind of phenomenon, and they can use specific algorithm or model to evaluate and detect shilling attacks. For example, we can remove global effects in the data normalization stage of the neighbor-based CF and working with residual of global effects to select neighbors.

* **The size of neighborhood is hard to choose**.

This affects predictions.

Approaches: Visualization of neighborhood and summarization of main themes/features of neighborhood will help.

* **Gray Sheep:**

User do not consistently agree or disagree with any group of people and thus do not benefit from collaborative filtering.

Approaches: Combine content filtering and collaborative filtering by basing a prediction on a weighted average of the content filtering and collaborative filtering.

Overall, it is good to combine collaborative filtering and other approaches. Also, machine learning techniques can be applied. Besides, there is a trend that giving a recommendation but also attempting to explain the recommendation to the user.

(c)  With reference to a clustering algorithm of your choice, describe suitable approaches to measuring the quality of the clustering algorithm. Your answer should distinguish between internal and external criteria. (10)

**Reference**: https://nlp.stanford.edu/IR-book/html/htmledition/k-means-1.html

**Answer:**

1. **Internal criteria**

Clustering is compared only with the result itself. Typical objective functions in clustering formalize the goal of attaining high intra-cluster similarity (documents within a cluster are similar) and low inter-cluster similarity (documents from different clusters are dissimilar). This is an internal criterion for the quality of a clustering.

K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance of documents from their cluster centers where a cluster center is defined as the mean or centroid  of the documents in a cluster  :



A measure of how well the centroids represent the members of their clusters is the residual sum of squares or RSS , the squared distance of each vector from its centroid summed over all vectors:



RSS is the objective function in K-means and our goal is to minimize it. Since N is fixed, minimizing RSS is equivalent to minimizing the average squared distance, a measure of how well centroids represent their documents.

1. **External Criteria**

However, an internal criterion often does not evaluate the utility of a clustering in the application. Alternatively, we can use external criteria which implies that we evaluate the results of a clustering algorithm based on a pre-specified structure (gold standards). The gold standard is ideally produced by human judges with a good level of inter-judge agreement. We can then compute an external criterion that evaluates how well the clustering matches the gold standard classes.

There are some measures: **purity** is a simple and transparent evaluation measure. Normalized mutual information can be information-theoretically interpreted. The **Rand index** measures the percentage of decisions that are correct and penalizes both false positive and false negative decisions during clustering. The F measure in addition supports differential weighting of these two types of errors.

* Formula of purity:

Text, letter

Description automatically generated

* Formula of index:

Text

Description automatically generated

**Q.4.**

(a)  Explain the terms precision and recall and discuss their suitability as a means of measuring the performance of information retrieval system. (6)

**Answer:**

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: 

**Suitability**

* Precision recall measure (and variations) are useful metrics for a number of reasons. They are in widespread use. Precision-recall are definable quantifiable measures and they provide a useful summary of a system’s behaviour.
* However, there are certain limitations and criticisms that can be levelled at these measures. In many domains we do not have available test collections and even in domains where we have large collections, meaningfully calculating recall is not possible as the collection is too large. Moreover, these metrics are useful in capturing the usefulness of the return set for a given query; they are useful for one off adhoc queries but are inadequate in capturing the complexity of an interactive session.
* Another important limitation is the underlying assumption that the set of relevant documents for a given query is the same for every user. We need to distinguish between the notion of a document being ’on-topic’ for a query and the notion of a document set being relevant to the user’s information need. In- deed, there has been some research showing that the measure of precision and recall do not reflect user satisfaction.

(b)  Query modification is often used by systems to attempt to improve precision and recall for a given information need. Discuss an approach, given user feedback on the returned answer set, to improve the performance of the query. (10)

**Answer:**

The Rocchio method can be applied here to modify the query to improve performance. It is based on the observation (and assumption) that relevant documents have similarly weighted term vectors.

Its formula is:

A picture containing text

Description automatically generated

where Dr denote the set of relevant documents returned by the system; the set of relevant documents returned by the system; α, β and γ are constants that are used to determine the importance of feedback and the relative importance of positive feedback over negative feedback.

Given the formula, it modifies a query to distinguish relevant documents from non-relevant ones. The use of these feedback mechanisms has shown that it makes marked improvement in precision and recall of system.

Text, letter

Description automatically generated

(c)  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)

**Answer:**

We can use local analysis to automatically expand query without explicit user feedback based on the returned answer set.

1) In local analysis, a term-term correlation matrix Mi,j is created to quantify the connection between term i and term j. There are many different means to develop the correlations.

* Association Clusters

Calculate the number of co-occurrence of term i and term j in the returned document collection using:

* Metric clusters

take into account position within documents. Let dis(ti,tj) bet he distance between two terms ti and tj in the same document. Can define term-term correlation matrix by:

A picture containing shape

Description automatically generated

* Scalar clusters

If two terms have similar neighbourhoods there is a high correlation between terms. Similarity can be based on comparing the two vectors representing the neighbourhoods.

2) Then, we can develop an association cluster for each term ti. Given the ith row from the matrix and select the top N values from the row. These are the values which correspond to the top N correlates for term ti. After the cluster for each query term has been created, we can get |q| clusters. Note that N is usually quite small to prevent the query from becoming too large and potentially “drifting” in terms to topic of the query.

3) Add these new terms to expand the original query. Also, may take all terms, or those with the highest summed correlation end itemize (结束项目).