Multimedia oral examination -Text retrieval

1.What is Information Retrieval? Why is it challenging?

Information retrieval (IR) obtain (relevant) information from a collection of resources.It is used to reduce information overloading.

The challenges come from the semantic gap and big scale of data.

Semantic gap means that words,pictures,videos are informative and depends on its context.So the information acquired by the computer will be different with human's understanding ,resulting in semantic gap.

Large-scale data makes it challenging to store, organize, and retrieve data in order to achive instance response.

2.Search Engines

(1)What is a Search Engines and its framework.

Search engine search for information on the Web under a given query.

It can maintain information by a web crawler keeps them in a centralized manner by inverted file. In online search,for a given query,search engine analyze the query and do searching and ranking,then return top-k ranked information back to user.

(2)Structure of a crawler?

Crawler is used for updating web content by Web search engines.It starts with a list queue of URL,repeat visiting,saving and adding hyperlinks to the list.

3.Document Retrieval- Preprocessing

(1)Preprocessing steps

For a given text or document, first tokenize them into meaningful tokens,then remove stop words,do case-folding, normalization and spelling correction,group nouns and do stemming.This is the whole pre-processing procedure.

We just group nouns because verbs,adverbs and adjectives are not helpful to understand the meaning of the text.

(2)Tokenization and its challenge

Tokenization **breaks a stream of text up into words**, phrases or other meaningful elements,which are called tokens.

challenge 1:We need maintain a **corpus** to identify tokens with special meaning, and need to update from time to time. challenge 2:**Numbers** are usually ignored but they are sometimes meaningful,such pi,911 and post code challenge 3:The difficulty of tokenization in different languages is different.In most of the **reflecting languages**, for example English,words are separated by spaces.But for Chinese Japanese , it require split **compound word**.

(3)stop word and its challenge

The most common words such as “the”, “a”, “there”, “be” are meaningless. We usually remove them according to a stop words list. But there are some exceptions like “to be, or not to be” which is meaningful but consists of stop words.

(4)Case-folding, Normalization and spelling correction

Case folding reduces all the letters to lower case. Normalization normalize different writings of forms into a general form. Spelling correction corrects spelling mistakes by searching for closest words (Hamming distance)

(5)Stemming

Stemming reduces words into their roots .For example “automatic”, “automatically” to “automat”.

4.Information Retrieval - Model

(1)KMP

(1.1)What is KMP

KMP is a linear-time algorithm for string matching(search for occurrences of word in a string).

It determines where the next match could begin when mismatch occurs to bypass re-examination of previously matched characters.

(1.2)how to do IR based KMP

IR can be modelled as a string matching problem.Given a query string,KMP matches it against the whole corpus and returns all the locations that the query string occurs.

(1.3)Pros and cons

KMP is a linear-time algorithm for string matching.

But First, the size of corpus could be very large ,KMP need to match through the entire corpus in the worst case,which is time consuming but users require instant response. Second, It is not error tolerant

(2)Boolean model

(2.1)What is Boolean model

Boolean model represent document as a binary vector based on a vocabulary.The binary vector share the same length with the vocabulary, each element indicates whether a term appears.

(2.2)how to do IR based on Boolean model

Given a query string,Boolean model represent it as a binary vector,then use some distance function such as Jaccard distance to compute similarty,then return top-K ranked documents back to user.

(2.3)Pros and cons

Advantages: Boolean model is very intuitive and can express complicated retrieval request by **disjunctive normal form**. Disadvantages: It does not support partial matching. Some specific terms should have higher weight.But it cannot give different words different weight. Not all query can be expressed as boolean expression smoothly

(3)Vector model

(3.1)What is Vector model

Some specific terms could have higher weight,so in Vector model ,each term is associated with a **weight** wich is caculated using the combination of Inverse Document Frequency and term frequency,and thus **suppresses the weights of highly frequent terms**.

F refers to Frequency,means total number of occurrences of term in the corpus. DF refers to Document frequency,means the number of documents that term occurs.Apparently,DF is lesser than F. IDF refers to Inverse Document Frequency,computed by the log of reciprocal of DF,and thus suppresses the weights of highly frequent terms.

(3.2)how to do IR based on Vector model

Given vector representation of query and document with TF-IDF weighting,then use some distance function such as Euclidean distance and Cosine distance to compute similarty between them,then return top-K ranked documents back to user.

(3.3)Pros and cons

Vector model overcome most of the pitfalls of Boolean model:

term is weighted according to its importance;

supports partial Matching

weight is usually sparse so inverted file can be used which is very efficient.

But Vector to vector comparison is not very efficient and vector model views each terms are independent from each other,which is not the fact and will reduce accuracy.

5.Information Retrieval - Searching

(1) Why is it challanging?

Searching is challanging because the complexity grows exponentially with the number of dimensions,whose complexity is O(D*N). So approaches such as Brute-force are apparently not efficient because users require instant response. But if data is sparse,we can use inverted file which is efficient.

(2) Inverted file

(2.1) Inverted file construction and the information it keeps.

In inverted file ,terms point towards documents where the term occurs. For one cell in the inverted list, many informations will be kept,including document ID,URL,the position of the term,term frequency (TF), Click-in frequency,Pagerank and so on.

(2.2) pros and cons

Inverted file can optimize the speed of search if the data is sparse.If we use forward index ,it requires iterating through each document and each word. But with inverted index, only terms appears in query need to be considered ,a lot of computation cost can be reduced by precomputing. The disadvantage is that processing cost when a document is added to the inverted file are increased .

(3) Retrieval on Inverted files

We need to compute distance between query and documents based on inverted files. Elements of term vectors are partitioned into two groups and calculated differently.

In Euclidean distance, for terms appear in the query,the Euclidean distance between them can be caculated, because we can access the weight of the words in each document.

For terms do not appear, we cannot access their weight but we know that the weight of the query is zero.So we can precompute the the square sum of the vector,then distance can be calculated by subtracting the square sum of the appeared elements from the sum of the squares of the entire vector.

In the cosine distance, we pre-calculate the l2 norm of each document to obtain the denominator. The numerator only needs to consider the terms appears in the query because the weight of non-appeared terms are zero.

6.The whole Information Retrieval procedure

First,Documents are pre-processed offline,by tokenization,removing stop words,do case-folding, normalization and spelling correction,group nouns and do stemming.

After pre-procession, we can user vector model to represent documents and use inverted file to organize them.

In online searching procedure,we do the same pre-procession and vectorization for a given query,and measure similarity by some distance functions such as l2 and cosine.Then return top-K ranked documents back to user.

7.Word Embedding

(1) What is Bag-of-Word?

Bag-of-Word embed each word by the frequency they occur in documents

(2) What is Co-occurrence Matrix?

The idea behind it is that if two terms co-occurres frequently,they may play similar roles in sentencesand may be semantically similar.

It builds co-occurrence Matrix by counting the co-occurrence frequent of words appear in a fixed size window.Then obtain the embeded word vocabulary by reducing the dimension of the matrix by PCA or SVD(Singular value decomposition).

8.Others

(1)Entropy

Quantity of information kept in one document can be estimated by entropy,defined using the probability of words. The lower of the entropy, the lower of the uncertainty(more words but less popular words).