CA4009 2021/2022

Student Name: Vincent Achukwu
Student ID: 17393546
Exam Number: 111562

Q1

Scenario 3

Analysis of the search requirements of the end-users of the system

It will be required that students understand the process of searching for content the same way they would research topics beyond the lecture notes. The exception is that LearnOnline will be the place where the students can do this. Students will be able to search for lecture notes, terms which may be present within the lecture notes themselves, and search queries related to what was mentioned in the video lecturers themselves. The users should also be able to do keyword searches so as to not specify the entire topic name (e.g instead of "Software Engineering: Processes, Principles, and Methods", the user could search "Software Engineering" and relevant results should follow). It will be expected that students are familiar with the basics of searching with this system as with how they may search via browsers. For videos, users would be expected to understand how to click on the timestamps that are relevant to their search query and view the video from that point onwards. This will incorporate video summarization so as to save the users time from watching the entire video or most of the video just to get to a certain part of the video.

Analysis of the domain and the search expertise of the end-users

Given that the search queries should only relate to the course the students are involved with, the search system should only retrieve relevant information based on the topics that the students are learning about (i.e it should retrieve relevant documents related to the material students are learning about and not random content on the web). The search tool should also allow searching for terms included in the notes, hence, summarization would have to be incorporated into the system so the snippets display the search query that also appears in the notes in bold. They can then click on the retrieved link and view the PDF of the notes or the videos related to their search query, as well as the other related material on the web which provides more details about the topic. It would also be useful for the system to highlight the section related to their search query. This would be more effective if the user uses longer search queries in order for the system to highlight text that is related to the search query. Along with this, the system would automatically scroll to the relevant section of the notes while highlighting the portion of relevant text. As for videos, timestamps could be shown to indicate which part of the video is related to the search query. (e.g if a user searches "matrix multiplication linear algebra", the system would show the PDF notes as well as the timestamp(s) of where the lecturer demonstrates this mathematical concept.) Lectures themselves could also use the timestamps/video chapters (similar to YouTube's feature) to help students know which topics are discussed in the video. As for linking the lecture content to other relevant sources, this would require the use of machine learning and OCR (Optical Character Recognition) in order to determine which external source is related to the lecture material.

Consideration of the types of queries that might be entered by the users

Users may use short or long search queries. Short search queries may result in more irrelevant search results since the system can also provide results based on terms specified in the lecture notes or external resources (which may be in PDF/Word Doc/Slides, etc.). For instance, if a user search query is along the lines of “algorithm” and the user is expecting to see only “Heap Sort Algorithm” notes, the user will end up seeing many results regarding the term “algorithm” and will have to take time to find exactly what they’re looking for. Hence, short search queries may not be helpful and lead to unreliable top results. It would be best for the user to search via longer search queries. This would also be an issue with video summarization since the machine learning algorithm will also use speech recognition, hence, using short search queries will make the system retrieve videos and video timestamps based on the times those terms may have been mentioned without full context.
Query expansion may also be incorporated into the system so as to allow for the retrieval of more relevant documents using relevance-feedback, derived by Robertson.

Available search technologies that could be used in a new search application to address this problem

There is an existing plugin called “VidIQ” that allows users to view statistical information about YouTube videos such as tags, likes-to-dislikes ratios, top keywords searched for relevant videos, SEO (Search Engine Optimization) features, etc. This can help lecturers and students to find videos accordingly by knowing what terms to search. Given that LearnOnline will be used for searching lecture notes and external resources, it would be useful to know when and what was said in the notes or videos, hence, including keywords or tags in the notes and videos can optimise search results to retrieve more relevant results.

Selection of a set of required components for your new search application and how these would be combined or used within the new system

The system will require a machine learning algorithm in order to process the videos for speech recognition. This will accommodate for the processing of words used in the videos to be translated into text (captions) as well as for comparison of the search queries with the terms used in the videos. OCR (Optical Character Recognition) will also be a key feature in this system as some video materials often contain text, hence, words can be extracted from the video keyframes to allow for timestamps to be generated. Video summarization would provide accurate and effective results for LearnOnline.

How the new system could be evaluated, including the features of a suitable test collection and choice of evaluation metrics

The system should go through response time testing given that there can be a lot of data processing and machine learning taking place. The accuracy would be considered the most important aspect of the system given that this system should be able to provide relevant results to the students and lecturers search queries. A suitable test collection would consist of a set of documents from which items are set to be retrieved, a set of search queries expressing the information needs, and relevant data telling us which documents are relevant to each request. Precision and recall would contribute to the system for effective information retrieval.

Q2

a) The non-specifiability-of-need problem derives from ASK (Anomalous State of Knowledge), which means the user may not use or know the language well enough

to form a search request which accurately represents their information need. Users encounter this problem because IR systems assume that users know what they want. However, the issue is that people have an inexpressible need for the information (some information need cannot be explicitly specified). Users have an anomaly with respect to the problem and will need to seek help by specifying their search request to the IR system. For example, users who are not familiar with cryptocurrency may simply search “Ethereum” expecting to see results about “Ethereum smart contracts”, but results only show the current price of Ethereum and the latest news about it. In essence, the user is not familiar with the correct terminology to use in order to satisfy their information need. Language barriers may also cause similar issues.

b)

i) It can only be determined by a human relevance assessor because assessors determine which of a number of categories a document should be assigned to.

ii) Using the 4 classes, we can determine how a document can be relevant to addressing a user’s information need:

1) Class 1 - does the document fully meet the initial information need?
2) Class 2 - does the document answer part of the information need?
3) Class 3 - does the document merely help the user to extend their knowledge of the topic of their information need without having to answer the problem, in which case it might be shown to be at least somewhat relevant to the information need?
4) Class 4 - is the document completely non-relevant to the information need?

iii) Query expansion can be used to bring more relevant information to satisfy the user’s information needs. The use of Boolean IR systems can help too by having the user specify a query using boolean conjunction.

c)

i) Precision and recall are the measures in an information retrieval system to measure how well the information system retrieves the relevant documents requested by the user. Precision is known to be the total number of documents retrieved that are relevant/total number of documents that are retrieved. The recall is the total number of documents retrieved that are relevant/total number of relevant documents in the database.

Precision measures the fraction of retrieved items that are relevant, whereas recall measures the fraction of available relevant items that have been retrieved.

ii) Pooling is the process where every document is retrieved and judged manually. The judgement should be independent, that is, the document is judged and considered to be either relevant or not relevant irrespective of the relevance of other documents. Documents are also presented in a random sequence to avoid sequential bias. Hence, this is not a practical way to determine the relevance of each document as there would be way too many manually to judge. Therefore, query expansion can improve recall by managing query expansion at the expense of precision. This is because larger recall implicitly produces a decrease in precision, given that parts of the recall are part of the denominator. It is also assumed that a bigger recall

negatively affects the overall search result quality since many users do not want more results to brush through, regardless of the precision.

d) A/B testing is a randomized experiment with two variants (A and B). An example of A/B testing would be using two slightly different versions of an online web app and seeing which generates the most attention by slightly varying some features.

A/B testing of an operational setting allows developers to evaluate new ideas by getting very fast feedback on their effectiveness. For instance, if we were to try a new idea to improve the accuracy of the retrieved relevant documents, the search engine provider can produce a live online analysis to assess the user response to the new relevance-retrieval method.

A/B testing helps you avoid additional risks by allowing you to target your resources for the best effect and efficiency. The downfall is that it takes time to see results for A/B testing.

Q3)

a) The learning-to-rank concept gives a composite score per document to manage the ordering of the ranked retrieval lists. This concept makes use of machine learning methods to automatically train the ranking model.

It is used in the development of web search engines with these 3 features:

1) URL length - shorter URL pages are usually correlated with the root and are likely to be more relevant
2) The number of inlinks or outlinks - i.e from or to the page
3) The number of matching query terms in the page

b) PageRank takes advantage of the link structure of the web to create an approximate global importance score per page based only on the link structure of the web. At any point in time, each page on the Web has PageRank scores correlated with it, but the PageRank score is independent of the actual content of the page.

The importance of the PageRank algorithm in the operation of an effective web search engine is that if a page has relevant links pointing to it, it has a high PageRank value, hence, its links to other pages also become important.

A simple example calculating the PageRank value of a webpage is as follows:
The formula for calculating the PageRank is **PR(A) = (1-d) + d (PR(T1)/C(T1) + … + PR(Tn)/C(Tn)).** If we start with 40 we get:

PR(A) = 40
PR(B) = 40

After the first calculation we get:

PR(A) = 0.15 + 0.85 * 40 = 34.25
PR(B) = 0.15 + 0.85 * 0.385875 = 29.1775

Doing this again, we get:

PR(A) = 0.15 + 0.85 * 29.1775 = 24.950875
PR(B) = 0.15 + 0.85 * 24.950875 = 21.35824375

And so on. Hence, we can see the values are working their way down to 1.0 and stop.

c)

i) Webspam is considered to be content that will never be considered relevant to a query.

Content that is considered to be spam include:

1) All content that the user does not wish to see
2) Content that is intended to affect the behaviour of a search engine in a detrimental way

ii) Link farms are believable-looking web pages that attempt to alter PageRank-type measures of the pages that they link to by increasing link counts of their destination pages.

As soon as a link farm has been identified by the search engine providers and doesn't seem to be effective anymore, the operator of the link farm has the choice to change it in some way (e.g perhaps by changing the page or site structure, location, content, etc. to evade exposure again). Then it might go undetected again until the search engines operators alter their link farm detection methods. This back and forth between search engine providers and the manipulators of the search engine behaviour in an unfavourable way is also known as adversarial search.

d) Click models try to model user search behaviour based on web search logs and to determine future activities.

It can be useful because the model can be used to estimate document relevance based on search behaviour.

Some disadvantages of this model are that it can be inaccurate and misleading. If many users click on the same link based on the same query, it is more likely to be relevant than a link clicked by a smaller number of users. This can be done by the same one person or multiple bot-users to alter the relevance of the result page (i.e it can make use of web spam)

Q4)

a) The semantic gap is referred to as the distinction between machine and human description of visual media. It is important because one of the biggest challenges is that visual media can be understood differently by humans and machines. The way we view objects depends on what our task is and what exactly we are looking for. Examples include the way we perceive colours in comparison to how machines do, or sounds, facial recognition, and speech recognition.

b) Selection means forming a summary by focusing on a subset of the topical content of the source document in detail. Generalization means forming a summary that overviews the entire topical contents of the source document.

c)

i) We come across a shot boundary when a new camera is used or a recording instance ends and a new one begins (known as shot cuts). A simple process of shot boundary detection is by examining every frame and looking for

shot-cuts that can be based on a change in colour, texture or intensity/brightness of the adjacent frame.

Problems that can typically be encountered are:

1) Dropped shot boundaries such as fade-in and fade-out, dissolving, morphing, or wipes.
2) False shot boundaries such as zooming and panning, tilting, booming and tracking, or events in the content (e.g camera flashes)

ii) Automatic Speech Recognition (ASR) systems can be used to generate imperfect index information for spoken content. ASRs are typically composed of 3 components - lexicon, acoustic model, and the language model. These components help decode an audio signal and provide the most appropriate transcription. Lexicons account for all of a word's possible phonetic variants. Acoustic modelling involves separating an audio signal into small time frames. Acoustic models analyze each frame and provide the probability of using different phonemes in that section of audio. The language model makes use of Natural Language Processing (NLP) via machine learning. It operates in a similar way to acoustic models by using deep neural networks trained on text data to estimate the probability of which word comes next in a phrase. This can help with video summaries. However, problems encountered with this include audio out of sync (i.e inaccurate audio to video synchronisation for video summarization), background noises affecting audio quality, and difficulty in understanding accents.

d) Reducing the semantic gap can help construct better video summaries because human perception is deemed to be more accurate in comparison to machine perceptions of visual media. Machine learning can have low accuracy and can have a bad effect on video summaries. For instance, audio and video may be out of sync in a video clip. Machines can have trouble detecting this when producing timestamps and captions via speech recognition for video summarization, but humans will notice instantly when the video/audio is out of sync. If a system is producing timestamps to identify a key part of the video based on a search query, but the video and audio are out of sync, this may or may not satisfy the user's information need.

Q5)

a)

i) Stemming is the process of removing suffixes via matching the ending of a word against a suffix dictionary and removing any suffix. An example is as follows: shockingly -> shocking -> shock. They are often used because they are designed to remove any suffix that is identified, match the ending of a word to a suffix dictionary, and check whether any context-sensitive rules apply.

We generally don't want to stem prefixes because this can completely change the meaning of the word. For example - fungus to fun. This can lead to irrelevant results.

ii) Conflation is the process of bringing together words that are related to each other in some way. There are two main classes of conflation algorithms -

Stemming and String similarity measures. Examples of the need for conflation are as follows:

1) Incorrect spelling: memory and memorie
2) Alternative spelling: colour and color
3) Multi-word concepts: cryptocurrency and crypto currency
4) Affixes: single and singular

An example of String similarity measures is fuzzy string matching also known as approximate string matching (via levinstien distance, n-grams, and suffix-tree).

iii) It can degrade the search effectiveness because of under-stemming and over-stemming. Under-stemming is not removing enough of the suffix so the word is incorrect (e.g. Playing to playi). Over-stemming is when we stem too much of the word and lose its meaning (e.g. Cycling to cyc)

b) Collection frequency weighting is a concept that says that terms that occur in fewer documents are usually more important than terms that occur in many documents.

Term frequency weighting means that the often a term occurs in a document, the more likely it will be more important for that document. This is usually calculated by the number of times a term appears in a document by the total amount of words. For instance: Apple = 5 , Words = 100, 5/100 = 0.5.

Document length normalisation means that document relevance is independent of document length.

The k1 factor determines the impact of term frequency in a document. They are experimentally determined constants that control the effect of tf(i,j) and the degree of length normalisation respectively. The typical value for k1 would be 1.2, and b determines the degree of document length normalisation. The value of b can vary in the range of 0 to 1.

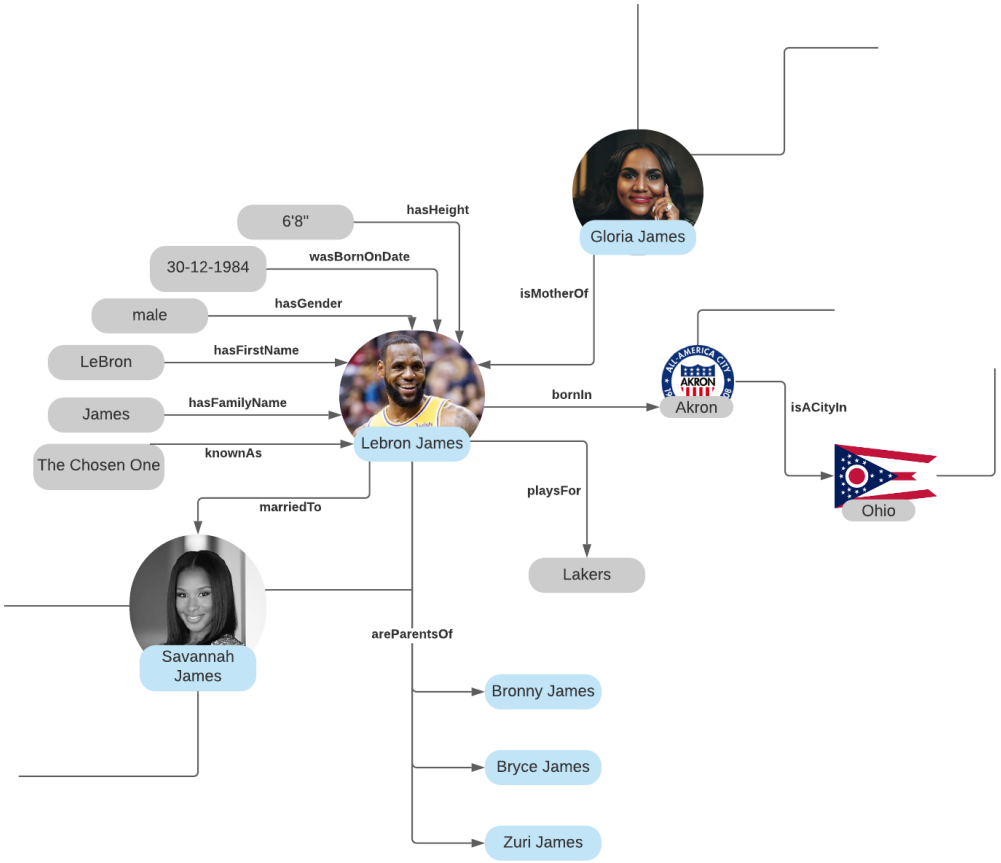
c) It gives the means to rank these terms by a predicted utility as expansion terms for the query. By getting the product of the 2, we can score terms by their selectivity and potential to occur in relevant documents.

Higher r(i) and rw(i) are likely to be the best overall expansion terms because since the original query terms were entered by the user, they can be regarded as more reliable than automatically selected expansion terms.

Q6)

a)

i) The below knowledge graph describes LeBron James with his entities and attributes.



ii) It can add missing information to a knowledge graph since it can provide the potential for improved efficiency of answering “easy” questions using the knowledge graph, and it also offers more professional questions where the knowledge graph does not contain the answer using documents.

b) It can be used by applying formal natural language processing and linguistic resources. The question can go into a semantic-parsing model which is trained to predict queries. Then, the question goes into an entity resolution model which links parts of the sentence to IDs in the knowledge graph.

c) It is a learned representation for text where terms that have the same meaning have a similar representation. Through embedding, it allows words and phrases to be expressed in terms of vectors denoting their shared properties. For example, “dad” = [0.1548, 0.4848, …, 1.864], and “mom” = [0.8785, 0.8974, …, 2.794].

It enables word mismatch to be overcome because similar words in a semantic sense have a smaller distance (Euclidean, cosine, or some other form) between them than words that have no semantic relevance. For instance, words such as “mom” and “dad” should be closer together than words like “mom” and “grass”, or “dad” and “metal”.