**2019**

**QUESTION 2 [Total marks: 25]**

**2(a) [4 Marks]**

**What are the differences between HTML and XML document markup? Use examples to illustrate your answer.**

XML is Extensible Markup Language, HTML is Hypertext Markup Language.

XML is a framework for specifying markup languages and HTML is predefined markup language.

XML is designed to store data, and HTML is used to display data.

HTML has predefined tags and XML doesn’t.

XML is .xml and HTML is .html.

XML is content driven and HTML is format driven.

**XML**

<?xml version="1.0>

<address>

<name> Krishna Rungta</name>

<contact>9898613050</contact>

<email>krishnaguru99@gmail.com

</email>

<birthdate>1985-09-27</birthdate>

</address>

**HTML**

<!DOCTYPE html>

<html>

<head>

<title> Page title </title> </head>

<body>

<hl> First Heading</hl> <p> First paragraph.</p> </body>

</html>

**2(b) [6 Marks]**

**i. What is meant by “human-in-the-loop” in image and video search?**

Human in the loop describes a process when the machine or computer system is unable to offer an answer to a problem, needing human interaction. .

**ii. What is the “semantic gap” in search of visual media? Why does the se-**

**mantic gap pose a challenge for multimedia search systems?**

The difference between machine and human description of visual media is referred to as the semantic gap

The biggest challenge is that visual media can be interpreted differently by humans and machines. How we view objects, depends on what our task is and what we are looking for exactly

**2(c) [4 Marks]**

**How can XML be used for content annotation in multimedia information retrieval for**

**items such as images and video? Use examples to illustrate your answer.**

Content annotation is used to add notes or more information about the topic, in this case about images or videos. Videos can contain subtitles which can explain the topic more or add more value.

**2(d) [8 Marks]**

**Before it can be searched, a collection of video recordings must be analysed to identify its visual and audio features. These extracted features for the video are then entered into a multimedia search system.**

**Outline a range of analysis techniques that can be applied to the visual and audio**

**content of a video, such as a television news broadcast or sports match, to prepare it**

**for indexing by an interactive video search system.**

Multimedia content is usually accompanied by text metadata which can be used for search for relevant items without requiring the analysis of the multimedia content. E.g. for a movie, title, actors, subtitles

Metadata can be created manually the content provider e.g. manual transcription of audio content, manual annotation of images (people, places, events)

**2(e) [3 Marks]**

**Suggest how multiple audio and visual features extracted from a video could be com-**

**bined to identify an event in the video, e.g. a goal in a soccer match, a wedding in a movie, or the key points in a scientific lecture.**

Wedding in a movie - wedding songs starts playing - model built to recognize a wedding dress

Goal in a soccer match - number recognition in the score tab - loud screams

**QUESTION 3 [Total marks: 25]**

**3(a) Relevance feedback (RF) methods are applied in information retrieval (IR) sys-**

**tems following an initial retrieval pass or run. RF is designed to improve the search effectiveness of an IR system by adjusting the parameters of the IR system and/or**

**expanding the user’s search query to better express their information need.**

**i. [5 Marks]**

**Application of RF requires the identification of relevant documents retrieved in a first**

**retrieval pass.What 3 sources of relevance information for use in RF are potentially available in an IR system? Comment on the likely reliability of each of these sources of relevance information.**

elevance feedback (RF) methods are applied in information retrieval (IR) sys-

tems following an initial retrieval pass or run. RF is designed to improve the search effectiveness of an IR system by adjusting the parameters of the IR system and/or

expanding the user’s search query to better express their information need.

3 possible sources of relevance information

1. Explicit feedback - user marked relevant documents
2. Implicit feedback - clicked docs are assumed to be relevant
3. Blind feedback - assumed top ranked documenRts are relevant

Most reliable source of relevance would explicit feedback, as the documents in question have been marked relevant by users. Implicit feedback would be somewhat relevant, as clicked documents by the users are marked relevant. Blind feedback can be the least relevant as document returned are assumed to be relevant.

**3(b) [4 Marks]**

**i. Give the standard definitions of precision and recall as used in information**

**Retrieval.**

Precision and recall are the measures in an information retrieval system to measure how well information system retrieves the relevant documents requested by the user.

Precision = total number of documents retrieved that are relevant/ total number of documents that are retrieved

Recall = total number of documents retrieved that are relevant/ total number of relevant documents in the database

**ii. [7 Marks]**

**Pooling is a popular method used to identify a set of relevant documents when constructing an information retrieval test collection. Describe the pooling procedure as it is used to identify relevant documents for an information retrieval test collection. In your answer, identify the assumptions made in the pooling procedure.**

Pooling is when every document is retrieved a judged manually. The judgement should be independent, that us the documents is judged either relevant or not relevant irrespective of the relevance of other documents. Documents are also presented in a random sequence to avoid sequential bias.

This is not a practical way to determine the relevance of each document as there would be way too many manually to judge.

With pooling we can test multiple top ranked documents retrieved using multiple IR systems and use this as a sample for the whole dataset.

We assume that we can never judge all of the retrieved documents as we may miss some from the IR retrieval.

**iii. [2 Marks]**

**Can either or both of precision and recall be calculated reliably when using an in-**

**formation retrieval test collection created using pooling?**

**iv. [3 Marks]**

**Effective application of RF improves the rank of relevant documents in retrieval runs**

**carried out following the application of RF.**

**Will effective RF improve precision, recall or both? Explain the reasoning underlying your answer.**

**4(a) [4 Marks]**

**Give three examples of English stop words, and explain why they are stop words.**

**Why are stop words often removed in information retrieval systems?**

as, and, is

1. Memory required for document representation lowered significantly (often more than 50%)
2. Improve search efficiency, no matching of stop words which user included in request.

**4(b) [5 Marks]**

**What are stemming algorithms as used in automatic indexing for information retrieval? Explain what is meant by under-stemming and over-stemming.**

**For stemming of English language text, why do we generally want to stem suffixes, but not prefixes?**

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 the meaning of it. E.g. Cycling to cyc

We generally don’t want to stem prefixes, as this can completely change the meaning of the word. E.g pre-war to war

**4(c) [7 Marks]**

**i. Why is the use of suitable data structures vital for the implementation of**

**effective search systems.**

**ii. Using an example, explain the use of inverted files in text search systems.**

**Your answer should illustrate how hashing is used for efficient processing of search terms.**

**4(d)**

**i. [4 Marks]**

**What is enterprise search? Why is enterprise search of increasing importance?**

Enterprise search refers to information retrieval within an organisation. Employers in many organisations can spend 10 hours per week searching for information. So enterprise search is practically and economically important.

**ii. [5 Marks]**

**Metadata can be used to annotate enterprise content with facets relating to the content items. Give three examples of typical facets in enterprise content.**

Faceted search is a technique for accessing information by filtering items based on facets of the information. Each facet usually corresponds to the possible values of a property common to all objects. E.g author, language, format, data, source etc

**5(a) [3 Marks]**

**What is the purpose of an information retrieval system? How does a standard in-**

**formation retrieval system attempt to achieve this purpose?**

Information retrieval system is a system for tracing and recovering specific information from stored data.

Standard information retrieval attempts to achieve its purpose by satisfying users' information needs by locating documents relevant to this information need.

**5(b) [8 Marks]**

**The Okapi BM25 term weighting function for best-match information retrieval is given**

**by the following equation:**

**cw(i, j) = cfw(i) ×**

**tf(i, j) × (k1 + 1)**

**k1 × ((1 − b) + (b × ndl(j))) + tf(i, j)**

**where**

**i = the current search term**

**j = the current document**

**cw(i, j) = the overall BM25 combined weight of search term i in document j**

**cfw(i) = the collection frequency weight of search term i**

**tf(i, j) = the within document term frequency of term i in document j**

**ndl(j) = the normalised length of document j**

**k1 = an experimentally determined constant**

**b = an experimentally determined constant**

**With reference to the Okapi BM25 model as described by the equation above, explain**

**the concepts of:**

**• collection frequency weighting,**

**• term frequency weighting,**

**• document length normalisation**

**How do the k1 and b factors operate in the equation for the Okapi BM25 model?**

**5(c) [5 Marks]**

**Knowledge graphs encode information extracted from source texts. A knowledge**

**graph typically describes the relationships between entities and the attributes of the**

**entities.**

**Give a simple example of illustrate these features of a knowledge graph.**

**5(d)**

**i. [4 Marks]**

**What is the difference between a conventional information retrieval system and a question answering system?**

**Even if high quality question answering systems were available commercially, why**

**would there still be a need for information retrieval systems?**

**ii. [5 Marks]**

**Sketch the standard workflow of a question answering system based on document**

**retrieval.**

**Suggest how a knowledge graph could be used for question answering instead of**

**retrieving documents.**