Information Retrieval

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 - 2019

Scenes and shots

- Scenes in a video can be decomposed into a sequence of shots
- Shot is a sequence of thousands of images from a single camera
- The images can then be read and understood what the scene is about

Object detection - cars, text, people

- Requires a model that can identify objects
- This can be difficult due to lightning, the setting, video quality and angles

Key frames

- Randomly chosen frames for analysis
    - Random frame are hit and miss
    - First and last frames can be problematic to understand
    - Choosing somewhere in the middle is the easiest approach

Suggest how multiple audio and visual features extracted from a video could be combined 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 - 2019

Audio recognition - model built to recognize audio - wedding song during a movie, high noise levels when a goal is scored
Visual recognition - model that can recognize visual images e.g frames - number change in the score tab

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 - 2019

Relevance feedback (RF) methods are applied in information retrieval (IR) systems 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
Explicit feedback - user marked relevant documents
Implicit feedback - clicked docs are assumed to be relevant
Blind feedback - assumed top ranked documents 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.

Can either or both of precision and recall be calculated reliably when using an information retrieval test collection created using pooling? - 2019

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 - 2019
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.
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 2019

Effective RF can improve recall by guiding query expansion and improve precision by rewording the query.

What is the purpose of an information retrieval system? How does a standard information retrieval system attempt to achieve this purpose? - 2019

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.

- Easy to understand, searching process is very easy to understand
- Saves time of the reader when they search for new information
- Can serve multiple users at once
- Searching cost is less than manual search
- Most of the results returned are up to date

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? - 2019

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.

Question answering system is a system that is concerned with automatically answering questions posed by humans in a natural language.

- Employees who seek quick answers on company policy
- Casual questioners who ask simple factual questions
- Consumers who look for specific product features or prices

Give the standard definitions of precision and recall as used in the evaluation of information retrieval systems. Briefly explain what each of these metrics is designed to measure - 2018

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

Precision measures fraction of retrieved items that are relevant
Recall measures fraction of available relevant items that have been retrieved

What are the three components of an information retrieval test collection?
Explain how these should be chosen to evaluate the effectiveness of an information retrieval system for a specific task - 2018

A test collection consists of

1. A set of documents from which items are set to be retrieved
2. A set of search queries expressing the information needs
3. Relevance data telling us which documents are relevant to each request

Explain the principles of the tf and idf components in term weighting for information retrieval. In your answer, describe the individual key concepts underlying the use of tf and idf, and make clear why they are generally more effective when used in combination. - 2018

TF-IDF, which stands for term frequency — inverse document frequency, is a scoring measure widely used in information retrieval (IR) or summarization. TF-IDF is intended to reflect how relevant a term is in a given document.

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

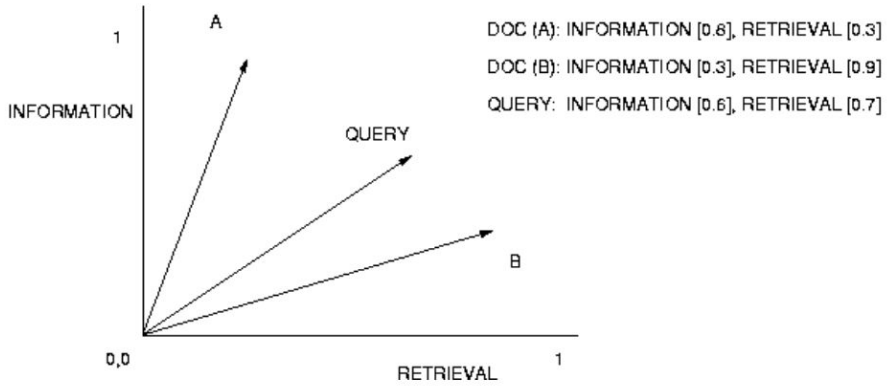
Inverse Document Frequency (IDF) is a measure of how rare or unique the term is.
Apple = 1000, Documents = 50000, 1000/50000=0.5

It's much more effective to use both TF and IDF together as we can take into the account term frequency and the length of the document.

Using a simple example explain how the vector space model creates a ranked list of retrieved documents to be returned to the searcher. - 2018

In vector space model, both the documents and queries are represented as vectors in a t-dimensional space, where t is the number of unique index terms in the collection.

The degree of similarity between document d(j) and query is calculated as the cosine of the angle between two vectors. The documents are then ranked in decreasing order of similarity.



Hypertext, Metadata and XML

What are the differences between HTML and XML document markup? Use examples to illustrate your answer - 2019

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> First paragraph. </body>
</html>

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 - 2019

Content annotation is used to add notes and descriptions about the topic e.g description of an image

For example XML can be used to define the attributes of an image.

- Time of capture
- GPS location
- Image quality and other features, size, resolution
- Automatic content analysis - shapes, places, people

Metadata can be used to annotate enterprise content with facets relating to the content items. Give three examples of typical facets in enterprise content - 2019
In an enterprise search collection, you can select facets to filter and narrow the search results to find particular documents of interest.
A facet provides a way to organize or classify content so that users can refine and navigate a collection along multiple paths, where each path presents a different view or perspective of the content

Three examples:

- A date facet is a four-level facet that is constructed from a date parametric field. The facet consists of year, month, day of the month, and hour values.
- With a flat facet, all possible paths are at the same level, such as a facet named State that includes values for New York, California, Virginia, and so on. By selecting facets, you can quickly narrow the results to documents that reference a particular state or states.
- A hierarchical facet allows you to explore nested levels of classification, such as a music category that lets you drill down by genre, artist name, song title, and so on..

What is the general reason for adding hypertext annotation to a collection of documents or other content items? - 2018

Hypertext annotations briefly describes the collection of documents

Explain the role of nodes, links and anchors in a hypertext - 2018

Hypertext is text displayed on a computer display or other electronic devices with references (hyperlinks) to other text that the reader can immediately access.

Nodes are chunks of information - basic unit of information in a hypertext system - nodes may be decomposed into smaller units - HTML page is a node.

Links connect two or more nodes - web link between HTML pages.

Anchors are persistent selection in documents and nodes - phrases, strings in text, icons

What does it mean to “jump into a hyperspace”, such as the WWW? - 2018

This means that the search engine suggest entry points into the hypertext which is judged to be the most likely to provide content relevant to the user’s information need

How do web search engines such as Google support users of the WWW to do this? - 2018

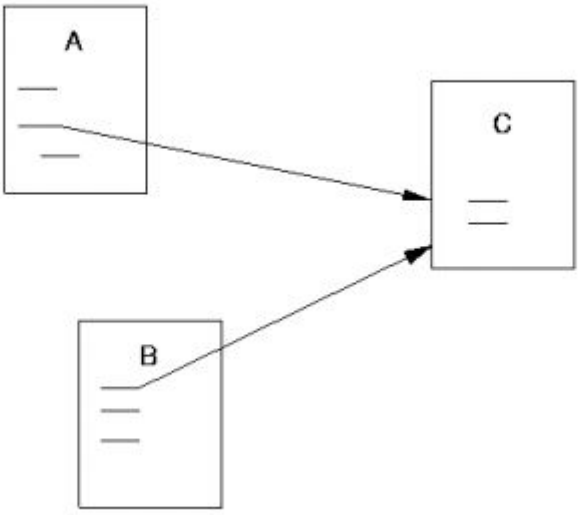
Google does this by providing the link the exact paragraph, rather than the homepage.

What is the PageRank algorithm as used in WWW search? - 2018

PageRank takes advantage of the link structure of the web to produce an approximate global importance score for each page based only on the link structure of the Web.

At any point in time, each page on the Web has PageRank scores associated with it.

PageRank score is independent of the actual content of the page.



If Page A links to Page C, then Page A is indicating that Page C is an important page with respect to Page A.

What does it mean to say that a hypertext containing multiple linked nodes has no beginning and no end? - 2018R

I think it's that they all link back to each other and hard to distinguish beginning and end

Text Retrieval

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? - 2019

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words. Stemming algorithms are designed to:

- Removing any suffix that is identified
- Matching ending of a word to a suffix dictionary
- Checking whether any context-sensitive rules apply

The algorithm requires manual or automatic creation of a dictionary that contains suffixes.

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

Give three examples of English stop words, and explain why they are stop words.
Why are stop words often removed in information retrieval systems? - 2019

as, and, is - the most common words in a language

Stop words usually never carry any information or relevance in regards to the document
Memory required for document representation lowered significantly (often more than 50%)
Improve search efficiency, no matching of stop words which user included in request.

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. - 2019

An inverted index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents

hello	**(1, 1)**
everyone	**(1, 2)**
this	**(2, 1)**
article	**(2, 2)**

Hashing is the transformation of a string of characters into a usually shorter fixed-length value or key that represents the original string. Hashing is used to index and retrieve items in a database because it is faster to find the item using the shorter hashed key than to find it using the original value.

Recording proximity of terms within documents in an information retrieval system enables it to take account of whether a pair of terms are close together or far apart within a document.

Why can taking term proximity into account be a useful factor in determining the potential relevance of a document to a search query containing such a pair of terms? - 2018R

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? - 2019

Collection frequency weighting is a concept which states that terms that occur in fewer documents are often more valuable than onces which occur in many documents.

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

Document length normalisation concept states that document relevance is independent of document length.

k1 factor determines the impact of term frequency in a document. A typical value for k1 would be 1.5

b determines the degree of document length normalisation. The value of b can vary from 0 to 1.

Summarisation

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. - 2019

The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources.

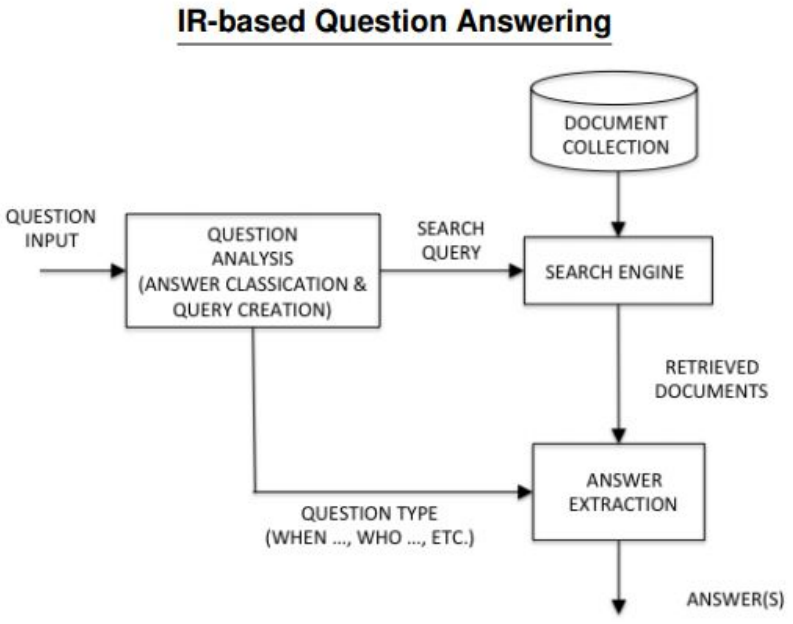
Knowledge graph are a powerful and useful way to capture interrelated information. The most popular form of knowledge graph in search focuses on:

- entities: persons, organisations, locations, products - usually real world objects
- relations: join entities
- facts: combination of entities and relations

- entities: `Lionel Messi, Argentina`
- relations: `<plays_for>, <was born_in>`
- facts: combine the previous two

  `<Messi, plays_for, Argentina>`

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. - 2019



Typical Question Answering workflow

Summarization is content reduction through selection or generalisation on

what is important in the source. Using an example, explain what is meant by the concepts of selection and generalisation in this definition. - 2019

Summarization is content reduction through selection or generalisation on what is important in the source. Snippet creation is a form of content summarisation.

A key challenge in snippet creation is to determine which content to include that is most likely to assist the user in determining the potential relevance of each document.

Selection is forming a summary focuses on a subset of the topical content of the source document in detail.

Generalisation is forming a summary which overviews the entire topical contents of the source document.

How could the suitability of a range of possible summarization methods for items shown in the output of a recommender system be evaluated using A/B testing? Assume that items are selected for inclusion in the output of the recommender system based on a collaborative filtering method - 2019

Web Retrieval

Explain the concept of “learning-to-rank” as used in Web search. - 2018

Learning to rank concept gives a composite score for each document to determine the ordering of ranked retrieval list. This concept used machine learning methods to automatically train the ranking mode.

Outline three features typically used in learning-to-rank for Web search. - 2018

1. **URL length -** shorter URL pages are usually associated with the root and likely to be more relevant
2. **No. of inlinks/outlinks** - from/to the page
3. **No. of matching query terms in the page**

Semantic Search

What is the “semantic gap” in search of visual media? Why does the semantic gap pose a challenge for multimedia search systems? - 2019

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

For what type of user query is a question answering system a suitable means of addressing a user’s information need? - 2018

Question answering system is a system that is concerned with automatically answering questions posed by humans in a natural language.

- Employees who seek quick answers on company policy
- Casual questioners who ask simple factual questions
- Consumers who look for specific product features or prices

Using an example, show how a knowledge graph encodes information from source texts. - 2018

The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources.

Knowledge graph are a powerful and useful way to capture interrelated information. The most popular form of knowledge graph in search focuses on:

- entities: persons, organisations, locations, products - usually real world objects
- relations: join entities
- facts: combination of entities and relations

- entities: `Lionel Messi, Argentina`
- relations: `<plays_for>, <was born_in>`
- facts: combine the previous two

  `<Messi, plays_for, Argentina>`

Knowledge graph encodes information from source texts by using intelligent machine learning algorithms. They provide a structure and common interface for all of your data and enables the creation of relations throughout your database.

When annotating the features or attributes of an entity, e.g. a named person or place, how can a knowledge graph capture all the important features for this entity if they are not found in an individual source text? - 2018

Knowledge graph can capture all of the important information, even if it's not found in individual source text by using hybrid combinations of document based question answering and using information retrieval models.

Explain how search engine companies such as Google use information summary "cards" created from knowledge graphs to provide information about common entities as part of a Search Engine Results Page (SERP) - 2018

Google use knowledge graphs to construct summary 'cards' about search entities. Cards are displayed when significant entity appears in the search query. Google includes facts for each entity which are most relevant to the object. Example - people also search for related people and places. These cards have reported a problem incase the data is incorrect.

Enterprise Search

What is enterprise search? Why is enterprise search of increasing importance? - 2019
Enterprise search refers to information retrieval within an organisation. Employers in many organisations can spend 10 hours per week searching for information.

Enterprise search is increasing in importance as it is practical and economical.

Compare and contrast enterprise search with Web search in terms of user requirements and system specifications. - 2018

Enterprise search refers to information retrieval within an organisation

Web search engine is a system that is designed to carry out web search - which means to search the WWW for particular way based on web search query.

Enterprise search is more optimised for protection not search. Enterprises have vital company data to protect. Enterprise search is usually only allowed to be performed on private or closed network.

How can facets be used to support search of partially remembered content in enterprise search in combination with suitably designed rich user interfaces, to facilitate effective enterprise search? - 2019

Faceted search is a technique for accessing information by filtering items based on facets of the information.

Each facet typically corresponds to the possible values of a property common to all objects, e.g. author, language, format, date, source, etc.

Faceted search can be useful since the searcher may remember one or more details of the item that they are looking for, even if they can’t remember enough details to create a meaningful search query, e.g. they may remember that they received the email that they are looking for from a particular person

Faceted search can combine text search choices in facet dimensions, e.g. first narrow by sender, then by date, then by document type.

Recommender Systems

Explain the following concepts as they apply to the goals of an operational recommender system: relevance, novelty, serendipity, diversity.
Why is a successful recommender system likely to incorporate all of these factors in determining its output? - 2019

Recommender systems are designed to predict items that may be of interest to the user. These systems can make their recommendations based on the profile of an individual user and their feedback on previously viewed items. One of the challenges of recommender systems is that reliable recommendations can require large amounts of feedback on previous items.

Concepts of operational recommender system:

- **Relevance** - recommended items relevant to the user
- **Novelty** - recommend new items relevant to the user which they have not seen before
- **Serendipity** - recommend unexpected or surprising items which the users finds relevant.
- **Diversity** - if a diverse list of items are recommended, there is a greater chance that one of them will be relevant to the user (and purchased!)

What is the cold start problem in recommender systems?
How does the cold start problem pose a challenge for new items introduced into the catalogue of an e-commerce website? - 2019

There are two main types of cold start problems in regard to recommender systems.

User-side - relates to the difficulty of making recommendations for new users who have so far provided very little rating information
Item-side - where new items haven't been rated enough to provide to make reliable recommendations

This poses a challenge for new items introduced into an e-commerce website as items haven't been rated enough or at all.

What are the features of a knowledge-based recommender system? For what tasks are knowledge-based recommender systems well suited? Why would a content-based recommender system or a collaborative filtering approach not be suitable for these tasks? - 2019

Knowledge based recommender systems are useful for items that are purchased frequently - home, cars , luxury goods

Sufficient ratings are usually not available for these items to make a reliable recommendation to a user using other methods

KBRS make recommendations based on customer requirements and items descriptions.
KBRS allow the user to specify what they want - price range, make, model.

What is A/B testing as applied to online web applications? - 2019

A/B testing is a randomized experiment with two variants, A and B. For example, using two slightly different online web applications and seeing which generates the most attention.

What is the purpose of a recommender system? - 2018

Recommender systems are designed to predict items that may be of interest to the user. These systems can make their recommendations based on the profile of an individual user and their feedback on previously viewed items.

What are the two main sources of information used by recommender systems? - 2018

User profile data, previous rating on different items

Describe in outline the operation of a recommender system based on collaborative filtering. - 2018

Collaborative filtering is not based on item content. CF systems provide recommendations based on ratings of items provided by other users who share common interests with the current user.

This makes CF systems potentially good and useful for recommendations on any type of item as it’s not based on item content.

Multimedia Search

What is meant by “human-in-the-loop” in image and video search? - 2019

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.

Why is the use of suitable data structures vital for the implementation of effective search systems. - 2019
Suitable data structures allow the system to process it easier. This will allow the system to return results in seconds. This will also minimise computational efforts and costs

Automatic image analysis for multimedia information retrieval is typically broken into three levels: image primitives, iconography and iconology. Explain these different levels of image processing. In your answer makes clear the relative complexity of using each level, and how it relates to the semantic gap and human interpretation of images - 2018

Three levels of image processing are

1. Image primitives - based on extracted features
2. Iconography - based on derived attributes
3. Iconology - inferred abstract attributes

Level 1 - image primitives

- **Colour** - the full RGB spectrum - image containing blue/green, yellow and orange could be a sunset
- **Texture** - a measure of properties such as smoothness, coarseness, regularity
- Natural textures are often random, wheres as artificial are often very similar
- **Shape** - geometric shapes, 2D - bart simpson detector, complex shapes very difficult to process
- **Spatial placement** - X and Y coordinates in a rectangle, find images containing object of interest in the top corner

Level 2 - iconography

- Derived attributes such as presence of specific objects - chairs around the table, Bill Gates
- Enables queries such as: Bill Gates meeting Gerry Adams

Level 3 - **Iconology**

- Describes pictures deeper artistic significance - if we have football players, a goal and a football = football match
- Queries like: images of football match > images of a footballer and football

Speech and video are temporal media. What does this mean? - 2018
Temporal media means that it's always changing constantly and is not static.

Locating relevant information within individual speech and video documents

is typically much more time consuming than locating relevant information in individual text documents. Explain why this is the case - 2018.

A lot more information needs to be analysed when looked at speech and video documents. It's much easier to analyse an individual text document than trying to analyze a video. Video is usually 30 frames per second, so for a minute long video you need to analyze 1800 frames which makes it a lot more time consuming.

Using simple sketches give examples of interactive tools and visualisations that have been developed to enable individual speech and video documents to be searched efficiently for relevant content. - 2018.

Collection Frequency

Key concept: Terms that occur in fewer documents are often more valuable than ones which occur in many documents.
Collection frequency weights (also known as inverse document frequency weights) are defined as follows for a search term t(i).
Given
n(i) = the number of documents term t(i) occurs in,
N = the total number of documents in the collection archive.
The cfw for term t(i) is
cfw(i) = log N
n(i)
The logarithm can be taken to any convenient base.

Term Frequency

Key concept: The more often a term occurs in a document, the more likely it is to be important for that document.
This refers to the term's within-document frequency.
Thus, while a term t(i)'s collection frequency is the same for any document, its document frequency varies.
The term frequency for term t(i) in document d(j) is:
tf(i, j) = the number of occurrences of term t(i) in document d(j).

Document Length

Key concept: Document relevance is independent of document length.
A term that occurs the same number of times in a short document and in a long document, is likely to be more valuable for the former.
Without compensating for document length, longer documents will tend to have higher matching scores merely because they are long, e.g. the tf(i, j) will tend to be higher for long documents.
Document length of document d(j) is
dl(j) = the total number of term occurrences in document d(j)
One length compensation method requires a normalised average document length defined as follows:
ndl(j) = dl(j)
average dl for all documents

“tf × idf”

One commonly used empirically derived weighting scheme is often referred to as “tf×idf”. (Note: idf(i) = cfw(i).)

$w(i, j) = f(tf(i, j)) * idf(i)$

where w(i, j) is the weight of term i in document j. Some tf(i, j) functions,

$f(tf(i, j)) = tf(i, j)$

$f(tf(i, j)) = 0.5 + 0.5$

$\frac{tf}{maxtf}$

$f(tf(i, j) = \log(tf(i, j) + 1)$

where maxtf is the maximum term frequency in document j.