Dublin City University

School of Computing

CA4009: Search Technologies

Section 1: Introduction

Gareth Jones

September 2016

Rationale

We are surrounded by many rapidly growing archives of digital media

Content can be from many sources (e.g. publishers, web, personal, ...) and be in one or more different media forms or languages.

- text: books, professional publications, the web, social media (Facebook, Twitter, Linked, ... etc.), enterprise content
- speech: radio, TV, lectures, meetings
- music: recordings, scores
- video: movies, TV, survillance
- images: scanned images, trademarks, photo archives (Flickr, Picasa, ... etc.)

Rationale

The ability to store and retrieve digital content online gives us the possibility of instant access to potentially unlimited amounts of information.

This information can be used for many purposes, e.g. education, business, research, entertainment.

Storing and distributing all this information is in itself a significant achievement, but it only has real value if useful content can be located reliably when it is of value to users.

Scope of CA4009: Search Technologies

The fundamental question addressed by this module is:

How can we efficiently identify and access useful information from within collections of various media types for different application areas?

This module is about: the challenges, technologies and evaluation of information retrieval or search as applied to multiple content media types.

i.e. how can we automatically search for things in collections of documents or other digital items? - why this is difficult, current commercial and research technologies which seek to do this, and how we can test how well they work.

This module is *NOT* about: how to use the latest facilities of *Google*, how to stream music and video from the internet, search engine optimisation (as least not directly!).

Definition of Information Retrieval

The purpose of an **information retrieval (IR)** system is *to satisfy a user's information need*.

The IR system seeks to locate *relevant* content.

Relevant content is that which *the user* deems to completely or partially fulfil their information need, e.g. a web page, image, video or audio clip.

- IR is concerned principally with unstructured information, e.g. as found in natural language texts or videos.
- the lack of structure precludes use of traditional database technologies.

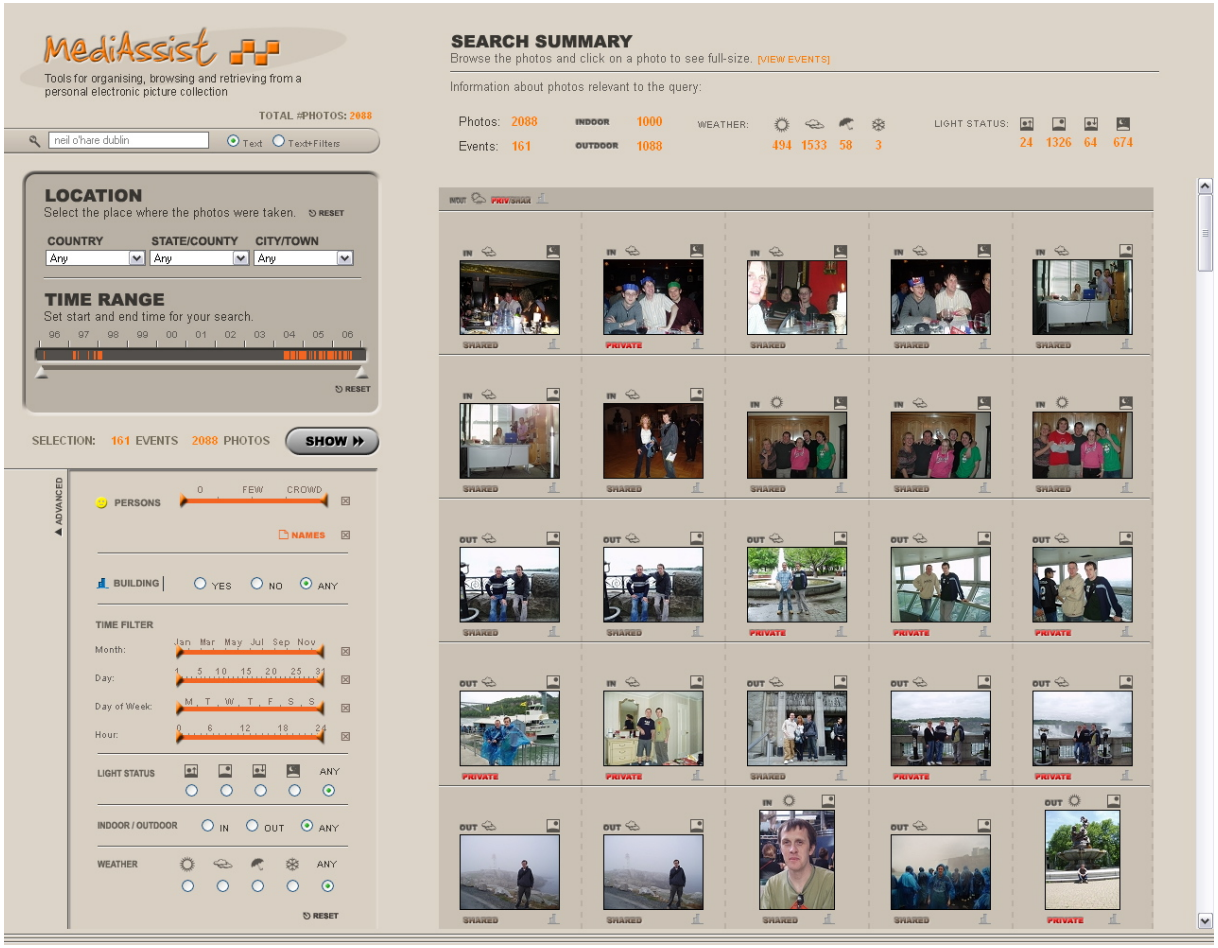
Given the visibility of databases in computing courses, it is perhaps surprising that the majority of digital information is actually stored in unstructured form.

Challenges of Information Retrieval

How would we search for this image?

Challenges of Information Retrieval

MediAssist digital photo search with location and time.



Challenges of Information Retrieval

The MediAssist system enabled users to search personal photo archives based on *time* and *location* at which photo was taken.

But also using features such as *light status*, and *likely weather conditions*.

The usability of the system was enhanced by ’“smart” design ideas such as mapping the date to time periods, e.g. weekend, middle of the week, summer, winter, etc.

We can think of these ides as ”smart” rather than ”clever” since they are based on insights into what might be useful, rather than being technically sophisticated.

Such ”smart thinking” is a common feature of successful IR systems.

Introduction to Information Retrieval

Research in manual indexing to enable location of relevant information has been undertaken by library scientists since the mid-1800's, e.g development of the Dewey decimal system for relative location of books in a library.

Early work in *automatic* indexing and document abstracting began in the 1950's. As such information retrieval and summarisation are one of the oldest applications of computers.

Computers were used to search for manually labeled items and also for items labeled automatically from the title, abstract, etc.

Introduction to Information Retrieval

Tests at Cranfield, U.K. in 1966 showed that automatic indexing produced equivalent results to manual indexing. This was a major surprise to those developing the systems. Why might automatic systems be as good as or better than human indexes?

IR systems were once used almost exclusively by expert users such as librarians.

Largely due to the WWW, they are now used by a wide variety of users.

- What are the implications of this change for interaction with IR systems?

Introduction to Information Retrieval

Archives may contain billions of documents, e.g from the web, usually *only a small number are relevant* to the *user's information need*.

Finding the the ones which are relevant is very difficult.

This really is a needle in a haystack problem!

As a user, what is your current mental model of how a web search engine works?

What do you think the mental model of the average users of a web search engine is?

Introduction to Information Retrieval

It is important to realise (and I think easily forgotten!) that the only information available to the IR system describing the user's *information need* is their *information search request*.

The user may have a clear picture of what they are looking for, but if their query does not describe this adequately the IR system may perform poorly.

Requests entered to web search engines typically consist of about 2.5 words.

What does this generally mean in terms of the description of the information need?

Introduction to Information Retrieval

Users have various forms of information need. For example:

- verificative *vs* explorative: “is this correct?” *vs* “tell me about”
- precise *vs* vague: “Who directed *Gravity*?” *vs* “tell me about the movie *Gravity*”.
- shifting *vs* static topics: “Keep me informed on a breaking news stories from the Rugby world cup’ *vs* “What is the Rugby 6 nations championship?”

These different types of information need may be addressed by a single fact, single relevant document, multiple relevant documents, or by a regular feed of updated relevant documents.

All of which require the IR system to function differently.

Introduction to Information Retrieval

Interaction: Users interact with systems.

Relevance:

- Stated requests are NOT the same as information needs.
  You can try to describe your information need in natural language, but unless your request is very simple, your description will fall short of expressing your often complex need. What you are looking for can depend on context, e.g. time, location, personal experience, interest, subject knowledge.
  These factors are generally not included in the search request.
- Relevance needs to be judged in relation to *information need* NOT the *stated search request*. This means that it needs to be judged manually, ideally by the person with the information need.

The Searcher and the IR System: A Cognitive View

- An information need arises from an anomalous state of knowledge (ASK).
- The process of resolving an ASK is essentially a cognitive process on the part of the user.
- Information seeking is part of the process.
- Users' models of information seeking are strongly influenced by systems.
- Conceptually trying to resolve an ASK makes IR hard, since it implies that the searcher needs to create a search request to look for information that you don't know. This can be stated more formally as the idea that there is a "non-specifiability of need" problem.

The Searcher and the IR System: A Cognitive View

The ASK means that the user may not know or use language correctly to form a search request which properly describes their information need.

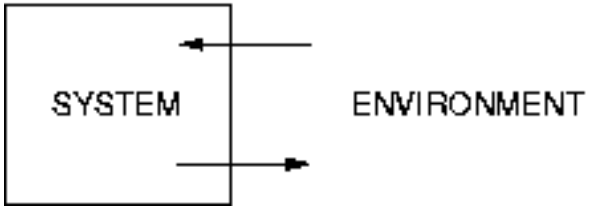
This presents a fundamental challenge to IR systems:

If users are not able to accurately describe their information need, then how can an IR system which is designed to returned documents which match the query, reliably resolve an ASK?

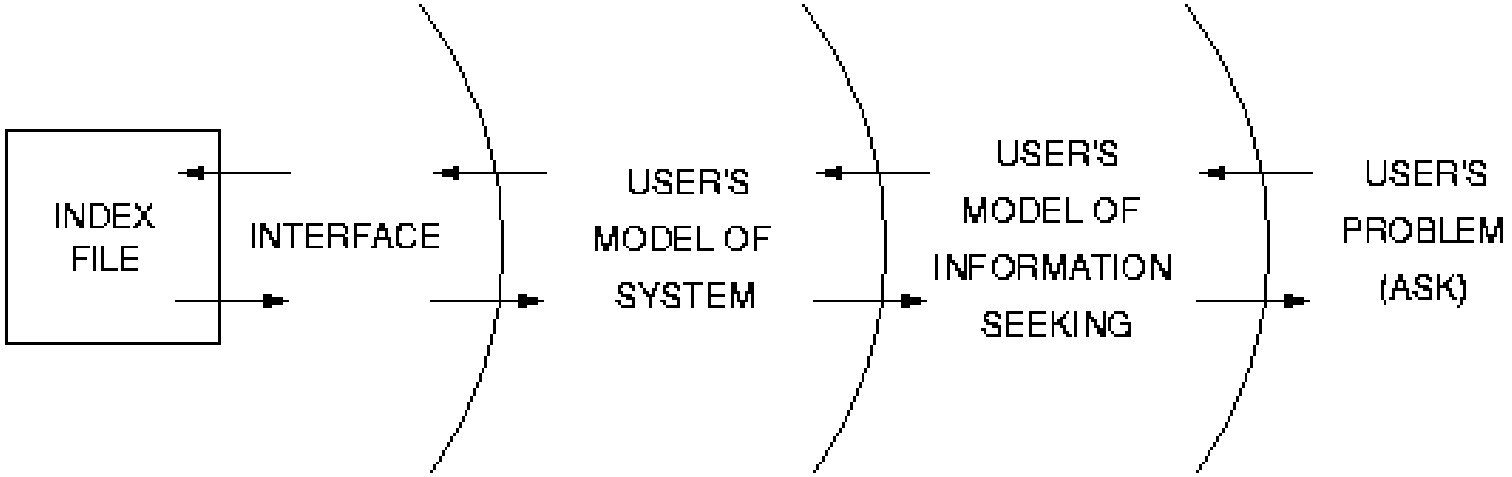
This is a challenge for all search engines.

- How do you deal with this problem when using a search engine?

The Searcher and the IR System: Boundaries



What is the “system” and where is the user?



It is very difficult to design experiments to assess existence and placement of these boundaries.

The Searcher and the IR System: A Cognitive View

The figure shows the process of resolving an ASK using an IR system.

- The user recognises that they have an need for information.
- The user creates a description of their information need as a query and enters it into the IR system.
- The IR system returns potentially relevant items.
- The user then evaluates the relevance of the items returned by the system, and determines if their information need has been satisfied.
- User typically accesses items until their information need has been satisfied, or they revise the query and try again, or they give up!

User Queries

Searchers generally enter short, ambiguous and often inaccurate queries.

Example requests from a web query log:

- ecko
- football graduate assistanships
- heliocopter
- quick drive
- shoes

Some people try too hard by using complex logical constructs:

- Horses AND OR AND Equine AND or AND Horse AND AND AND Anatomy

System Issues

In practice, most users are only concerned with the results of their search.

They rarely understand or consider the underlying system.

They even more rarely make use of sophisticated tools provided by a search engine.

- 1% use advanced search facilities.
- 10% use query syntax - usually incorrectly!

Average length of search queries is around 2.5 words!

- Perhaps surprisingly this is true for both general interest and professional users carrying out search as part of their work, e.g. medical doctors.

System Issues

IR systems are generally concerned with performance in terms of two key concepts:

- *Precision* (fraction of retrieved items that are relevant) is generally more important than
- *recall* (fraction of relevant items retrieved).

Although there are special applications such as patent search where high recall is very important.

Precision and recall will be formally defined in a later lecture.

Module Objectives

- Gain an appreciation of the diverse technologies that can be useful in accessing information contained in unstructured content from different sources, for different user needs, in different languages and/or in different media.
- Gain a basic technical understanding of several of these technologies.
- Understand the importance and difficulty of meaningful evaluation in the development of IR systems.

Module Outline

The following topics will be covered in this module:

- Hypertext, the WWW & XML
- Text Retrieval
- Text Retrieval for the WWW
- Information Summarisation
- Question Answering
- Enterprise Search
- Recommender Systems
- Multimedia Information Retrieval: audio - speech and music; images; video

Module Assessment

70% – written examination

- details of format will be made available in advance.

30% – Continuous Assessment

- 15% - Laboratory exercises - week 7 onwards)
- 15% - Design project - details and deadline will be published shortly.

Please note: Much of this module is based on material derived from the discontinued module *CA437: Multimedia Information Retrieval*. However, CA4009 is a new module, and includes new material and excludes some old material from CA437. The module will be assessed based on the material covered in CA4009 in the current academic year.

Module Texts and Study Resources

Main texts:

Introduction to Information Retrieval, C. D. Manning, P. Raghaven and H.Schutze, Cambridge, 2008.

Supplementary texts:

Modern Information Retrieval: the concepts and technology behind search (Second edition), R. Baeza-Yates and B. Ribeiro-Neto, Addison Wesley, 2010.

Natural Language Processing for Online Applications Text Retrieval, Extraction and Categorization, P. Jackson and I. Moulinier, JB, 2002.

Study resources for this module will be made available via Loop.

Lecture notes, suggested reading, past examination papers, etc.

Introduction to Information Retrieval

Approaches to Information Management

IR systems seek to enable us to search for information more efficiently and more effectively.

Without IR facilities to do this, we potentially need to read all the available books, articles, etc to find the information that we need.

This is a completely brute force search approach, and obvious entirely impractical for anything more than a very small collection of information items.

With IR, we can be directed to potentially relevant materials which may contain the information necessary to satisfy our information need.

Introduction to Information Retrieval

Approaches to Information Management

Information can be formally organised to assist manual searching. For example:

Flat collection of homogeneous objects:

- library card catalogues, files in a large directory, database records, ...

Hierarchical organisation: used extensively in manual information structuring:

- filing cabinets, files, documents with chapters, sections, subsections, Dewey decimal classification of books,...

Whle this physical organisation can be efficient, if it is clear how to organise the information, often however, it is not clear how the information should be organised or where individual items should be placed.

Introduction to Information Retrieval

Approaches to Information Management

One way to associate information with multiple places is *cross-referencing*: links to related information, possibly of interest to the reader:

- These can be implemented using: footnotes, “see also” notes, post-its, encyclopedia references,...

A generalisation of this idea is *hypertextual* linking:

- an extended form of cross-references, multiple information links and no superimposed hierarchy, navigation by following links, ...

links between WWW pages and within *wikipedia* are classic examples of hypertextual information structuring.

But this method of information linking, seeking and navigation does not scale.

Introduction to Information Retrieval

Approaches to Information Management

Links and structures are limited to the specific intentions of the creator of the links.

Often people will not have the time or resources to place all the links that might be included , or it will not be worth the investment of their time to do so.

As structures become larger and more complex, the absence of all possible useful links means that it will not be possible to use them to locate relevant information efficiently.

In this context search provides a means to locate relevant information within such structures.

Multimedia Information Retrieval

Developments in computer hardware and networks mean that archives increasingly contain non-text content such as speech, image video and music.

Spoken content can be transcribed using automatic speech recognition or manually (but this will be expensive), we can then apply text search methods.

For image and video search, we can search textual annotations or use multimedia content features by directly matching on the visual contents.

Also, there is much legacy material in hardcopy printed form which cannot be searched automatically. This content must first be digitised to enable it to be searched online.

Multilingual Information Retrieval

We will begin by considering English language text information retrieval.

However, while English remains the dominant language of the web, but the proportion contained in other languages continues to increase:

– it is generally agreed that English is now significantly $<$ 50% of total web content.

Thus it makes sense to look beyond English language documents.

There are various aspects of this topic.

We need IR systems which enable search in different languages.

This enables users to search online for information in any language with which they are familiar.

Multilingual Information Retrieval

But, also users may want to search for information only available in languages with which they are not familiar.

To support this, *machine translation* can be used to translate search requests, from one language to search for documents in one or more other languages.

Machine translation is currently far from perfect, but automated translations can enable users to gist the subject matter of retrieved material.

Retrieved documents can then be translated into the searchers native language to enable them to access the information.

Note: Many people can read material in a second language much better than they can write in it! For these users, document translation may not be needed.

Enterprise Search

While people often think in terms of search relating only to web search engines, another important area is enterprise search.

Enterprise search is the practice of identifying and enabling content across an enterprise to be indexed, searched, and displayed to authorized users.

A major challenge is the need to index data and documents from a variety of sources such as: file systems, intranets, document management systems, e-mail, and databases; then to present an integrated list of ranked resources from these sources.

Access controls are often vital if user access is to be restricted only to data and documents to which they have been granted access to within an enterprise.

Information Omnipresence

The availability of information is now becoming ubiquitous via increasingly widespread internet connectivity.

Internet connected devices can search for and download information in a wide variety of locations, or more generally *contexts*.

This means that users often now seek to exploit or rely on the web and other information repositories as ever present source of informations to address their information needs instantly whenever and wherever they arise.

Information Omnipresence

The selection and delivery of information can potentially take into account the user's *context*.

- location, time, weather, current associates, diary, potentially their current activities - walking, driving, etc.
- Information can be supplied without interactive query input, e.g. while driving - travel conditions - wet roads, traffic congestion, etc. can be **pushed** to the user.

  The same information may be pushed to all users matching a specific context, or delivery may be personalised based on a model of the user's interests.

Personalization and Context

Search engines behave the same for all users for any search request.

But!! : *different* users have *different* information needs at *different* times.

Should a search engine respond differently for different users?

Should a search engine respond differently for the same user in different contexts?

Should a search engine respond differently depending on the platform to which the information is being delivered?

Personalization and exploitation of context (time, location, biometric state, ... etc) offer the possibility of adaptation IR to deliver different results to different users in different settings.

An Emerging Topic – Lifelogging

Lifelogging relates to capture, storage and exploitation of as many sources of information relating to an individual's life experiences as possible.

This can include all computer activities, smartphone use, GPS tracking, but also images captured using a device such as the Microsoft SenseCam, and biometric sensor data capturing features such as heart rate, skin conductivity and skin temperature, which can vary with engagement with an activity or emotional response to a situation.

Lifelogs enable an individual to search for information from their past, reminisce or reflect on past experiences, share experiences (such as a holiday), etc.

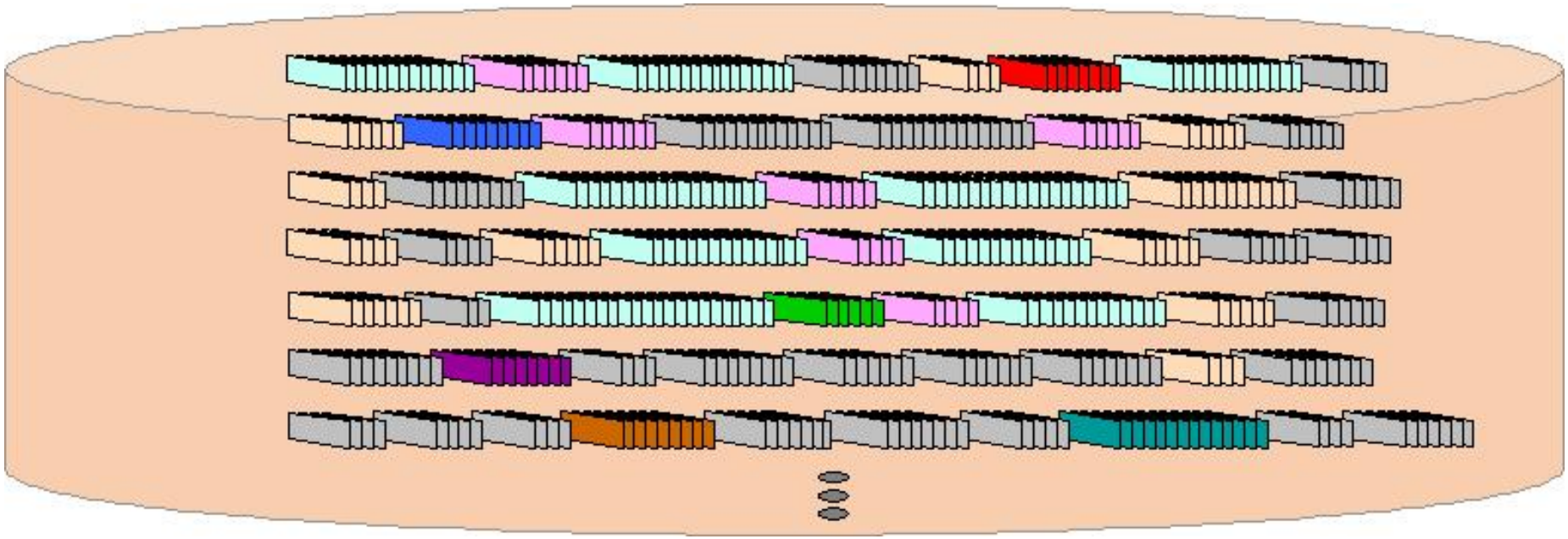
An Emerging Topic – Lifelogging

Ubiquitous image capture - Microsoft SenseCam[a]

[a] You can now buy these as Vicon Revue `http://www.viconrevue.com/` proactive camera.

An Emerging Topic – Lifelogging

How can we organise the many thousands of images captured by a SenseCam in a meaningful way to allow search and browsing to find useful and interesting things?

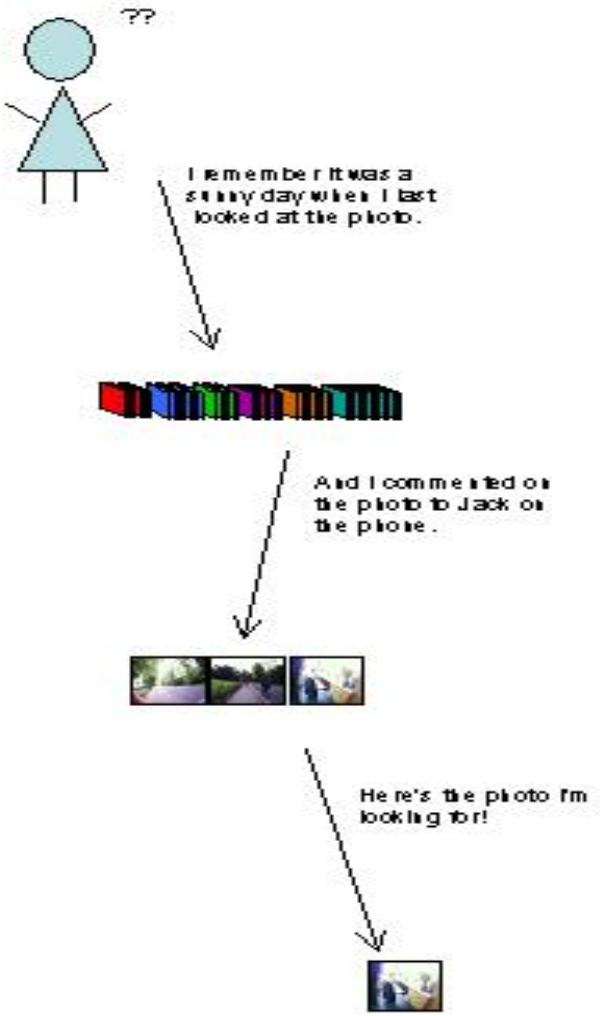


An Emerging Topic – Lifelogging

How can we find things based only on context?

I remember it was a sunny day and I was talking with Jack on the phone when I last saw the photo.

Linking content and context for searching a “human digital memory.”



Web Search Engines

What are the features of a good search engine?

Components of an IR System

The increasing challenges of search arising from the complexity of and volume of material available, and the advances in technologies such as personalisation and exploitation of data indicating the context of search, are increasing the complexity of search applications.

Nevertheless, a core component of a search application is standard text information retrieval.

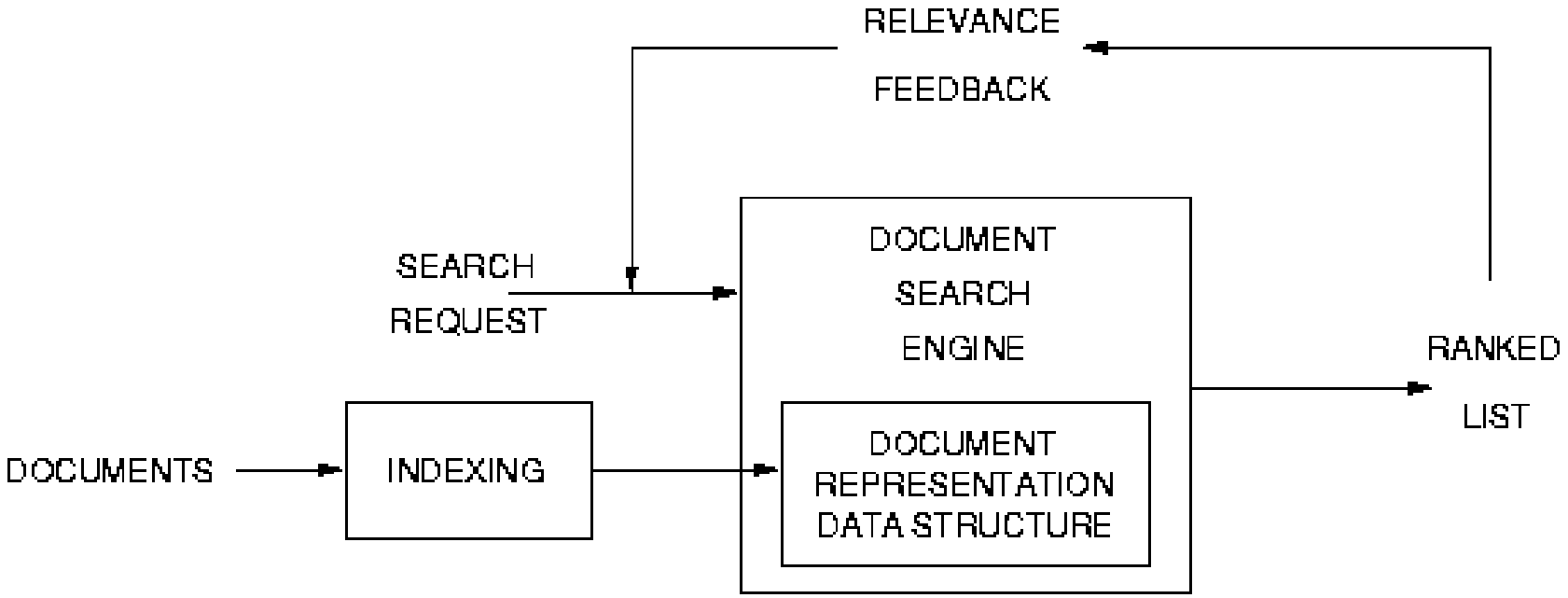
Components of an IR System

Such text IR systems typically consist of a standard set of components, the exact specifications of the components for a specific IR system are subject to design decisions taken to develop a system suitable for the specific application for which it is intended.

The standard components are shown in the following diagram and then described in overview.

We will look at these in detail in later lectures of this module.

Components of an IR System



Components of a standard information retrieval system.

Components of an IR System

- Document Collection:
  - existing document set, e.g. business reports, newspaper articles.
  - web documents, collected using a web spider, crawler, robot, bot.

  Documents usually need to be preprocessed to standard format, e.g. case conversion, removal of HTML, stop word removal, stemming.
- Document Indexing:
  - convert documents into a fast searchable format,.e.g. inverted file.

Components of an IR System

- Search Request:
  - enter search request expressing user information need.
- Document Searching:
  - calculate set of potentially relevant documents, and return to user, usually ranked by matching score. (which is intended to rank documents in order of likelihood of relevance)
- Relevance Feedback:
  - modify search, e.g. expand query using extra search items, based on relevance data; and run the search again.

Web Search

Speed response is crucial.

Also, even with the resources of a Web search company, efficiency is vital to minimise the computational cost of providing a search service.

Potentially pre-compute results for popular queries - perhaps daily - then provide these results rather than re-computing the results each time this query is entered.

The scale of the Web means that the thousands of documents may match each user search request.

Text IR is an important component of Web search. However, it cannot reliably differentiate between these documents which match the query.

Web Search

How should the search engine determines which ones are most likely to be of interest to the user?

In addition to text IR, Web search engines use many other features to help determine document rank including:

- Web link structure,
- frequency with which users click on each document,
- subjective “quality” of each webpage or website.

If you can think of a method to improve efficiency or effectiveness of an operational Web search engine, the chances are that it is already being used or has been tried and withdrawn because it was not useful for some reason.

Historical Perspectives on Information Availability

How did we arrive at the current situation of vast online information resources?

Before the development of written scripts most information was passed by word of mouth.

Then before the development of the printing press documents were handwritten.

Handwritten copies of books were usually held in libraries; often in monasteries with no right of public access, but how many people could read anyway?

This means that individual copies of documents were rare and precious.

Historical Perspectives on Information Availability

With the advent of printing and greater literacy, books became more widely available and more widely read. But,

- you still needed a copy of the book,
- manual typesetting meant that the printed book itself is the only source of the information.

It has been argued that history will view the development of the Internet as the third great development in knowledge dissemination.

The internet makes information is available from anywhere at any time - assuming that you have access to a networked computing device and permission to access the information.

Historical Perspectives on Information Availability

What are the implications of this easy availability of information for literacy in society as a whole?

What might be the impact on those without access to the required technologies and/or access permission to the information?

Historical Perspectives on Information Availability

Curiously early electronic preparation of documents was seen as a replacement for mechanical preparation, not as the creation of accessible digital content.

This was driven both by the mental model of document preparation and purpose - people expected to prepare and receive physical documents, and had no expectation of the existence of an electronic version.

Also, storage space was limited, and people were anxious to free up space on early PCs and macs for the production of their next document by deleting documents with which they were now finished.

The introduction of the internet, expectations of electronic versions, reduced storage cost and the practical advantages of making electronic content available online, have changed perspectives of this completely.

Historical Perspectives on Information Availability

In the area of multimedia, early radio and television broadcasts were transmitted live with no copy being made.

Recording was regarded as a means of time shifting between performance and broadcast, and not as a means of making a permanent copy.

The master tapes from recorded radio and TV shows were often reused since they were not regarded as potentially valuable, and the physical master tape was comparatively expensive.

The world has moved on: recording and broadcasting agencies now have large archives and home recording and archiving is now common.

The means for people to make their own copies means that it is difficult to revise history!!

Historical Perspectives on Information Availability

There are now significant efforts to digitise the achieves of national broadcasters and other agencies.

Similar issues with photography. Traditionally photographs were produced as prints or slides with the negatives often stored, Many people are now digitising archives of physical photographs.

Digitized copies of documents and other content can be made available online for research purposes. This means that people have access to content that would either be too remote for them to visit or too precious or fragile for them to have permission to access. For example, rare handwritten manuscripts written hundreds of years ago.

Moving from the Industrial Age to the Information Age

The industrial revolution saw the dawn of the **Industrial Age**. This encapsulated the concept of the means of production with the concept of economies of scale in production. The age of atoms.

We now see the development of the information revolution opening up the **Information Age**: Economies of scale with less regard for time and space. Manufacturing of information can take place anywhere at anytime.

The cloud for services and data storage is enabling the arrival of the **invisible computer**.

Moving from the Industrial Age to the Information Age

In the information age broadcast of media reaches ever larger audiences.

The proliferation of broadcast channels gives access to huge amounts of information with increased possibilities for *narrowcasting*, where the audience is a small group wishing to receive specialist programming.

We are now entering a form of *post-information age* (Negroponte) where we can have an audience of size one, and the information provision is personalised potentially for each individual.

Note that this is quite different to narrowcasting.

Moving from the Industrial Age to the Information Age

Information provided to the individual can be augmented to include additional information and links related to the broadcast sources.

This additional information may be in different media or languages.

Video-on-demand or catch up TV services can be uniquely configured for the individual.

Post-Information Age

Information retrieval and access is related to both narrowcasting,

- how do we find the content related to a topical area that we are interested in?

and, post-information age personalisation,

- how do we provide interesting and useful material to the individual?

In these applications search can help to provide this unique personal experience.

How should we best represent the user in this setting?

How should we best represent the user's context?

Information Overload

With 4 or 5 TV channels you could “channel surf”- - with 4000 channels you can’t – at least not efficiently.

You need a personal assistant or a personal agent to:

- choose your TV programmes for you,
- organise news (newspaper, TV) for you,
- schedule your diary and answer the telephone.

More generally:

- How can we filter out irrelevant information?
- How can we find other relevant material?
- Can we discover information within information? (information inference)

Information Overload

Search is very important in addressing the problem of information overload.

The diversity of media and languages of online information means that we need to be interested in multimedia IR, multilingual IR and personalised IR.

Conceptually, we can think of search as providing a dynamic narrowcast to a personalised audience of one.

Issues in Information

It is perhaps easy to assume that the intention of the information society is that everyone will have access to all information, once everyone has access to computing devices connected to the internet.

But even a little thought reveals that this is obviously not true.

So we can begin to ask some important questions:

Who has the information?

Who has access to it?

Can we depend on the timing of its delivery?

Issues in Information

Security of Information

Security of information is important to businesses, governments (personal and national security issues), individuals.

We see regular breaches via hacking, etc.

The effects on individuals are rarely fully reported - identity theft issues, etc.

How can unauthorised access be prevented?

Issues in Information

Reliability of Information

Information can be wrong.

Individual written records may be wrong:

- people make mistakes,
- software contains errors,
- sensor values may be wrong - faulty sensors.

There is so much information that we cannot hope to manually check everything - and even if we could - the correct values may not be knowable (missing sensor data) or there may be further human mistakes.

Issues in Information

Reliability of Information

Attempt to correct automatically:

- spelling correction,
- smoothing to look for outliers in values from sensors,
- use of knowledge sources to check values, e.g. facts in documents.

Issues in Information

Reliability of Information

Distribution of copies mean that information can be verified. If the copies are different, which one should be believed?

Digital information can be distorted.

- every copy is a perfect replica of the original. No carbon dating!

Centrally held digital libraries can be modified.

- online newspapers, images , video footage, “paintings”

Will we be able to trust the legacy of the digital age?