

MECT / MECB

CA557

INFORMATION ACCESS

2004-2005

These are the course notes for the course **CA557: Information Access** for 2004/2005. These notes, available on the web in PDF format, contain the basis for the material which will be presented in class. My lecturing style is to work directly from these notes along with OHP slides for illustration.

This module aims to familiarise the student with aspects of information management which impact the eCommerce area ...

- databases,
- access to text documents, primarily web search engines because this is so important for enabling customers to reach businesses;
- access to multimedia information, because it will become so much more important;
- the whole area of markup / metadata covering XML and how that impacts searching, and presentation of information through web browser;

We introduce and develop issues related to the nature of information including:

- information sources – who is producing info, and for whom, and why;
- information content – what's there;
- structure – how information is structured and organised;
- representation – how information is represented – particularly important in the context of markup/metadata (HTML, XML, and a whole litany of other standards) and also of encoding/compression formats for text (HTML, XML), and for other media (WAV/MP3, GIF/JPG/SVG, MPEG family);
- access – how can people/organisations find out where to look for information, and when they've found where to look, how can they find or access information;
- presentation – how is/can information be presented;
- what are the resources needed to empower efficient and effective information access, in this digital, ecommerce-driven world, where are we lacking, where are the "holes/gaps";

The course syllabus is structured into 6 "chapters" of unequal size as follows:

1. **Introduction to Information Access & Information Retrieval**, covering the nature of information and its management, the nature of media and the need to manage information... in the context of eCommerce... leading to an introduction to Information Retrieval
2. **Databases** – relational, basic organisation of information as tables with foreign/primary keys and constraints - take a complex worked example and show it as a R.DBMS, just for illustration;
3. **XML and XML databases** – all the history, the terminology, the status, the trends – how XML is/can be used in B2B, CEC (consumer-oriented eCommerce) – what are the pitfalls and the potential.
4. **Information Access now (content retrieval of text)**, types of information access people are doing - principally **web searching**. Hypertext searching vs DB searching... a discussion of web searching and web search engines. Examining text-based IR, boolean, weighted

terms, ranking, Hal Varian's economics of search, using links information in web searching, question-answering, TREC... and after all that, a discussion of how to evaluate retrieval systems.

5. **Content retrieval of non-text information** ... why bother? How to retrieve directly from audio, images and digital video, using online (WWW) examples, and how to retrieve using metadata... ?) and current technologies for capturing, storing, presenting and accessing such media – and why ?
6. **Data mining** from web usage information – what information can we extract from web usage, and how, and what can we use it for ... strictly speaking this isn't information access by people, but it is a spin-off advantage from logging information access... also collaborative filtering & personalisation;

HOW IS CA557 TO BE EXAMINED?

We have 70:30 ratio for exam/continuous assessment...

Exam is 2h, usual format; continuous assessment will be discussed in week 5/6 and will be due at the end of week 12:

SUPPORTING MATERIALS FOR THE COURSE

There are a variety of sources of information for the course, and each of you will use your own... I list references in these notes and will add others during class, plus the online material and pointers.

I especially recommend the WWW as a source of information... but be careful... it's very easy to publish...

- Textbook ... there is none, sorry ! But I will reference a few as the course progresses... they will be available in the library.
- Journal/conference articles ... loads, and most of them will be online.
- Websites ...!
- Systems and demos – available from the web.

To bring these all together I've a single "Resources" page from the course home page, which I'll update throughout the course:

<http://www.computing.dcu.ie/~cgurrin/CA557/>

These pages will be password protected to protect intellectual property, the single userid and password will be handed out during class... now!

These notes I'll copy (x40 or so) and distribute in class (PDF version also on web page).

CHAPTER 1

INTRODUCTION TO INFORMATION ACCESS & INFORMATION RETRIEVAL

Information is essential to society ... c.70% of all work in USA and in Europe requires processing information rather than processing materials.

A consequence of connectivity and replicability of information is its increased availability, which is nice, but there is a glut and it can suffocate... e.g. too much information the night before an exam...

Technology has been and remains great at generating and replicating and distributing and delivering information in different media, but not in managing that same information; this isn't just electronic documents ... look at your own information sources... your college notes... cd collections... videos... photos...

These days, information finds us wherever we hide ... it is difficult to switch it off, to get away from voicemail and answering services, fax, courier and express mail, paging, bulletin boards, news, e-mail, mobile phones, SMS, MMS, active badges and so on.

... all these technologies find us wherever we are and constantly provide information, bombarding us, and it is getting worse ...

- PDAs, XDAs, Blackberrys and even *wearable computing* for mobile communications.
- Airlines aimed at the business market are the best (worst ?) for facilitating constant communications ... on-board telephones, modem jacks in seat armrests, on-board fax, seat-to-seat, seat-to-ground, etc...

As individuals, we are bombarded with information from all angles, which we cannot refuse or ignore, and the technology has, to date, been used only to generate and deliver the information, not so much to manage it... for example, how do you manage your digital photographs?

That electronic management of information and access to this information is currently far short of ideal... It is this that we will examine in detail over the next twelve weeks.

1.1 INFORMATION ACCESS IN THE CONTEXT OF ECOMMERCE

Information access implies the technologies, infrastructure, resources required to create, store, transmit, render/display, index, access and retire, all kinds of information.

Information Access is essential for eCommerce to be successful. The internet allows commerce to take place much quicker and much cheaper, as long as the customer can access the information.

In other courses you will have examined the nature of eCommerce, the major stakeholders (businesses and consumers) and how much of this has come about as a result of disintermediation. The middle man now makes less and less, if indeed there is one at all... e.g. AerLingus.com, dell.com. Some examples of intermediaries:

- Amazon.com... creating a "virtual" catalogue or portal which is the sum of the catalogues of all its suppliers. Amazon.com is an example of a symbiotic¹ relationship – Amazon provide the new service (electronic book buying) for a fee and the suppliers/publishers get to sell their books to a new market, i.e. increasing their sales volumes. In addition Amazon.com make extensive use of collaborative filtering in an effort to 'help' users to find products that they are likely not looking for, but may find themselves interested in. The old market of Barnes and Noble still remains, but it's the NEW market that Amazon taps into.
- Online security brokerages to facilitate online day traders led to large cost savings for little people like us who wanted to play the stock market. That's because there is so little overhead in online trading and there has been an explosion in this type of activity with shrinking margins as they undercut each other.
- 123.ie provides a single location from which to get insurance quotations for cars, households, etc., in Ireland. 123 is the middleman while Hibernian, Eagle Star, etc., are the suppliers... try it, it works!

The problem with these early adopters and first movers, and there are many in almost every sphere of commerce/e-commerce, is that the value they add to the service – economy and independence of suppliers – is easily replicable. Anybody can set up a website, buy and sell things.

It then follows that if it is so easily replicable then it is difficult to preserve a competitive advantage for the first-to-market (Amazon, for example).

In fact, a successful first-to-market in a domain encourages those who are higher up the value chain, to get into the game. So Amazon.com led to BarnesandNoble.com and that's what has happened. Amazon.com, under pressure, had to save themselves by diversifying into other product ranges, other

¹ A Symbiotic relationship – one in which both parties gain from partaking in the relationship.

media – two reasons for such diversification are profit or survival – and they've had to do this to survive. They've done this and have started to turn a profit.

A lesson to be learned from this is that in order for such ecommerce disintermediation to be sustained and stable, the value added must be

- proprietary,
- unique,
- legally protected
- difficult to replicate.

What Amazon.com tried to do to make themselves unique was to make extensive use of Collaborative Filtering and include online reviews from customers, unmoderated reviews, but they're not great, in fact they can be totally misleading!

But, not to go too far off the point, in order for eCommerce to be successful, whatever form it takes, the lack of a human intermediary requires that Information Access methodologies in place operate successfully, whether it be a simple R.DBMS or a personalisation engine behind an on-line store or even a full search engine providing advanced text (or other media) search over an archive/library of information...

I have often spent hours trying to find the 'best' flights and trains to a particular location. Many people argue this point and claim that this is a bad thing... but I'm not convinced... If I don't have the time to spend I will not do so! The cars I rent online, the books I buy, the flights I book, the software/hardware I purchase... all worthwhile transactions from my point of view – I do get better choice, I may get better prices² or I may not, and I may even get better service, but I feel that I'm winning, and I get convenience, so I'm happy... e-tickets for flights, convenient hotel bookings and car rentals... without even leaving my desk!

Looking into airlines as a good example... here we see predatory discrimination. Airlines now have their own online reservation web sites to sell seats and they have an enormous advantage when selling their own product, offering perks and fares only to their own online customers, blocking cheaper seats for their own online customers, making their own rules on tickets and penalties, controlling airmiles/points, basically favouring their own. Aerlingus.com started to do this and their cheapest fares are available on their website.

So what about travel agents? they are neutral³ w.r.t. airlines... will ecommerce force them out of business. And where does expedia.com fit in? The only fares they can get offered by the airlines will naturally not be their cheapest fares, so expedia.com is for the higher market...

² Recent Purchase (Dec 04) – USB-phone data cable for €20 from English WebSite v.s. €80 in Vodafone store, OMNI.

³ In fact they are not wholly neutral since their interest is in selling customers the most expensive ticket since they make 7% of the purchase price, no matter what cost the ticket ... interesting paradox ... they make more money selling more expensive tickets, yet customers want the cheapest.

The online air reservations case is slightly different to Amazon.com vs. Barnes and Noble, since people do and will always like to browse bookstores, but the attraction of visiting/calling a travel agent is somewhat less. With online stock brokers the position is somewhere in between ... people like the security of an experienced brokerage, but the convenience of online trading is undeniable.

Given that ecommerce is both good and bad (but mostly good) and information access is a crucial component to successfully launching an ecommerce venture (along with infrastructure, standards, cultural adoption, security, interfaces, and others), what type of information access have we got at present, and what is rolling down the tubes towards us? *That's what this module is about.*

The major components of the course and their relationship to each other, are:

- (a) Info access to structured / databases
- (b) Info access to text / web search
- (c) Representation of information – data markup
- (d) Data mining and indirect applications of info access
- (e) Info access to non-text

Given that the web is (currently) the infrastructure supporting ecommerce, **(a)** is crucial to helping people find information in the first place ... structured data in databases is ideal for this.

(b) is set to become important for interaction as the interface moves from being a web browser on a computer with text in, image out to being a visual interaction on a domestic TV or mobile device;

(c) is part of the underlying support structure for almost all ecommerce enterprises, migrated from legacy environments to become "web enabled" this is where XML fits in;

(d) is something really special and unique to ecommerce. A never-advertised advantage (for the business) of eCommerce is the ability to do detailed, personalised tracking of usage and purchases. This is great for stock management keeping warehousing costs down and reducing product spoilage, is great for strategic decisions on what products to put marketing drives on, and what products to withdraw, and for personalising the interaction between consumer and business. Look at Amazon.com recommendations... Because I buy a CD by Radiohead I will be recommended music by Queen, or books about... Even, aside from eCommerce, when I shop in Supermarkets using loyalty cards (especially Superquinn Swords) the chains are constantly building databases of my purchases, and deciding what personalised special offers to tailor and send out in the monthly newsletters.

(e) is the topic which we have not yet got quite right, but we may be almost there. If mark-up of all kinds of information we use was accurate

and reliable and detailed and widespread, then web searching would be easier and faster (but less fun) and content access to non-text information would not cause the problems that it does.

SO, INFORMATION ACCESS IS IMPORTANT AND WE NEED TO UNDERSTAND THE TECHNOLOGY AND THE APPROACHES, IN ORDER TO APPRECIATE THE DIFFICULTIES, POSSIBILITIES, AND THE LIMITATIONS.

Using technology, there are multiple ways to organise information, some techniques being inherited from pre-digital days which electronically replicate the manual information management we had/have, while others provide functions like content search which can not be achieved with analogue information.

So this is good, we can do (electronically) what we could have done in analogue, plus we can do more ... what can we do ?

- **Flat collection of homogeneous objects** ... library card catalogues, files in a large directory, database records, Search Engines indexes... objects are sorted and keyed by some attribute or may be content-searched.
- **Hierarchical organisation** ... used extensively in manual systems ... filing cabinets, drawers, files, documents with chapters, sections, sub-sections, Dewey decimal classification of books, Hierarchic filing system on most operating systems ... information seeking is satisfied by traversing the hierarchy, may be combined with electronic search.
- **Cross-reference** ... as an adjunct to either of the above, footnotes, see-also notes, encyclopaedia references ... easily replicated electronically.
- **Hypertextual** ... an extended form of cross-references, multiple information links and no superimposed hierarchy, navigation by following links, maybe use search to find starting point in the network / web, other peoples recommendations or bookmarks...

In an electronic world, information management means this also but additionally it means stemming the flow or isolating or retrieving or filtering only relevant information...

Users' information needs can vary:

- verificative vs. explorative
- precise vs. vague information needs
- shifting vs. static topics

As an example, information systems can retrieve with

- precise (DB) or fuzzy matching (Search Engine)

Query languages can be:

- precise query language (SQL) or ambiguous query language (natural language)

Putting together some of these combinations we have popular information systems such as:

- Databases
- Expert Systems
- Bibliographic databases
- Web search engines

1.2 INTRODUCTION TO INFORMATION RETRIEVAL

Information Retrieval (IR) has been receiving increasing levels of attention since the end of the Second World War...the US government began to pump huge amounts of money into research and development with a corresponding increase in the volume of scientific literature being produced... the need to have search and retrieval facilities provided over this literature, along with a growing dissatisfaction with the then current manual processes and the hope that automation might hold the answers led to the development of a new field of research, which we now call Information Retrieval.

In principle, the problem of information storage and retrieval is simple. If a person has an information need that can be fulfilled from reading or examining each document in a given set of documents, retaining documents containing relevant material and discarding all others, this is called **manual information retrieval** (or brute-force information retrieval). Note that these documents could be text, images, audio files, video clips, entire movies, etc...

Manual information retrieval is clearly impractical in the majority of cases⁴... the amount of effort involved is directly proportional to the size of the collection of data... In a library scenario, an individual may be seeking information contained between the covers of only a handful of books contained within a vast library. A person has neither the time nor the inclination to read a whole document collection to fulfil an information requirement.

Therefore, it follows that the advent of computer technology from the mid-40s onwards posed possibilities for **automatic information retrieval** as opposed to manual information retrieval.

Automatic information retrieval removes the human from the relevance ranking process, leaving the human only to compose a query and examine the results automatically produced by the retrieval system.

⁴ Manual Retrieval will not be discussed in this course and we are henceforth assuming that IR refers to automatic information retrieval.

WHAT IS INFORMATION RETRIEVAL?

Information Retrieval is the name given to a process that stores, retrieves and provides maintenance functions over some body of information. Information in this context can be composed of text, images, audio, video and other objects. Information retrieval can be either data retrieval (structured data) or information retrieval (unstructured data). Both retrieve information and will be examined in this course, but what is the difference?

Data retrieval is concerned with looking for an exact match between queries and documents while **unstructured information retrieval** mostly seeks a partial match and then from this partial match, a small (manageable) number of the best documents are selected (best match).

DATA RETRIEVAL

Data retrieval is best exemplified by a user's interaction with a DBMS. Here there is no ambiguity in the query, which will generally be expressed using some artificial query language such as SQL, and each query will only generate one possible (complete and exact-matching) set of results based on the underlying data. Ranking of this set of results is not possible (unless defined using 'sort by' or 'order by' commands within the query) as all results are equally valid. For example, the following query to a DBMS:

```
SELECT author FROM books WHERE title = 'Information Retrieval'
```

can only have one possible answer, which is a listing of all the authors who have written books called 'Information Retrieval'.

We will look at this in chapter 2, but now turn our attention to introducing Information Retrieval in the unstructured domain.

UNSTRUCTURED INFORMATION RETRIEVAL

Unstructured Information retrieval, on the other hand, is best exemplified by a user's search engine query. A further distinction can be made with respect to unstructured information retrieval... this distinction is between information retrieval and document retrieval.

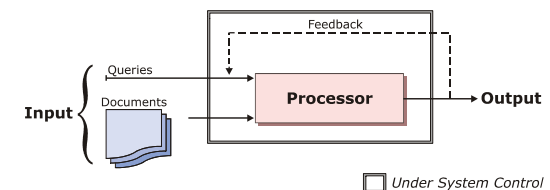
In the strictest sense, information retrieval could refer to the retrieval of minimal information (such as Question Answering systems), which satisfies a user's query. In most information retrieval systems, however, the unit of retrieval is the document (in a ranked list) and not some smaller unit of information, resulting in the need for a user to browse a document in order to locate the required information. Techniques such as highlighting query terms in the result document to aid the user in locating the relevant information are commonplace, but the underlying unit of retrieval is still the document and it is this document retrieval aspect of IR that we will focus on in later chapters of this course.

IR systems exist that will attempt to provide the user with precisely the information (answers) that the user is requesting. A good example of such a system is the IONAUT Question Answering system which attempts to answer a user's query with specific content, extracted automatically from a collection of web documents... this is the URL, try it out:

<http://www.ionaut.com:8400/>

1.3 OPERATION OF AN INFORMATION RETRIEVAL SYSTEM

The following diagram illustrates, in a very simplistic manner, the overall construction of a typical IR system. As can be seen, the system consists of three components: input, processor and output.



Inputs

Looking at the inputs into an IR system, the primary task is to convert each input (both queries and documents) into an internal representation to support fast search and retrieval... The vast majority of IR systems will only store a representation of their inputs, as opposed to the full documents and queries... This is a one-way process, in that it is not possible to convert the internal representation of a document back into the original document.

A second phase of input could allow a user to modify the articulation of the query (information need) in light of previous output of the system. This process is called '(relevance) feedback' and may be done both automatically (without the user even knowing that the feedback process is taking place) and manually... We will see more of this later... A good example for the present though from an eCommerce domain is, given a book that I like, find me "more books like this".

Processor

The next component of a typical IR system is the processor. The processor is concerned with generating the data structures to support the provision of speedy, efficient and effective results in response to a user's query input.

The processor will accept the internal representation of the query and calculate the documents that best match the user's information need as articulated by the query (input).

Output

Finally, we examine the output component, which is primarily composed of a set of documents that are returned (from the processor) in fulfilment of a user's information requirement. This output may consist of a set of unranked document identifiers or the identifiers may be ranked in decreasing order of relevance.

We are now in a position to outline the steps an IR system must carry out in order to operate effectively.

STEPS IN PERFORMING INFORMATION RETRIEVAL

We can identify four distinct steps that a typical IR system must follow in order to be able to fulfil its task. These are:

1, Document Gathering

This is the process of gathering the documents that are to form the core content of the IR system... once again these documents could be text, images, audio files, video clips, entire movies, etc... If working with a fixed and readily available set of documents, then this is simply a process of knowing the location of each file on disk and gathering them before converting them into a searchable internal representation (*document indexing*)... but it may be necessary to actively seek out content for the indexing stage, as is the case with search engines... they send out crawlers to seek and gather content for indexing.

In this stage also, some parsing of unnecessary content may take place (especially if the documents are textual in nature). For example:

- Unnecessary mark-up of text may be removed.
- Many frequently occurring words that are of no benefit to the automatic retrieval process may be removed. These words are called stopwords and we will discuss stopwords in a later chapter.
- Terms within documents may be truncated to term stems (stemming).

2, Document Indexing

The documents gathered in the document gathering phase are converted into a fast searchable internal representation. This will usually be implemented using some programming language dependent data structures which provide fast searching facilities such as arraylists, vectors, sets, multi-sets, maps or multi-maps.

Non-text documents such as images, audio files, video clips will be indexed using some features which support user searching... any examples of how we could index an image...?

-
-

3, Searching Support

This process involves accepting a query, processing it, finding possibly relevant documents, calculating the degree of similarity between each document and the query for each (possibly relevant⁵) document, sorting the set of highly ranked documents and returning these to the user in groups (usually) of 10. All this has to be done as efficiently and quickly as possible. For example, the IR system that operates as the Google search engine accepts and processes (as of July 02):

- 150 million queries per day.
- 6.25 million per hour.
- 105,000 per minute.
- 1,700 per second.

4, Document Management

In the previous three steps, we have gathered documents, indexed them and are now allowing users to search their content. However, in many scenarios such as web searching, the documents that have been indexed will be unstable and constantly changing. Consequently, we must validate that:

- The documents that comprise the internal representation of the document collection are as up-to-date as possible.
- The documents included in the internal representation are actually still in existence.

This will involve re-gathering documents, at periodic intervals, or even completely rebuilding the internal representation of the document collection, which necessitates returning to the document gathering phase and starting again...

We will return to unstructured information retrieval in Chapter 4 – 6.

⁵ In a **simple** IR system, given a query "cat dog" it is clearly futile to calculate the degree of relevance of documents that contain neither term... hence possibly relevant documents.