Dublin City University

School of Computing

CA4009: Search Technologies

Section 7: Enterprise Search

Gareth Jones

October 2016

Background

Most people’s experience of information retrieval relates to Web search or possibly desktop search applications on their own computer.

A less public but increasingly important area of search relates to corporate or enterprise content.

Enterprise Search refers to the application of search technologies to information within an organisation, e.g. a business or public body.

Background

It is forecast that the biggest growth industry for search technology in the coming years will be for systems which enable search of structured and unstructured data in ways that allow users to make faster, better decisions.

Such content can derive from emails, printed and scanned documents, databases, project documentation, intranet content, and any other information sources within an enterprise.

Taking into account all these sources, the amount of information being created and stored within enterprises is staggering

Consider the amount of information created even in a small office, and then think about the situation for a multinational organisation.

Background

Interestingly something like 80% of enterprise content is in unstructured form, i.e. not in a structured database!

There is evidence that employees spend a significant amount of time searching for information needed in their work, and often failing to find it!

Employees in many organisations can spend 10 hours per week searching information and are not successful between 30% - 50% of the time!

According to one survey as much as 10% of a company's salary costs are wasted on ineffective search.

So enterprise search is practically and economically important, and currently often not very good!

Background

Once all the available information has been stored - a key technical challenge is finding relevant information within the records of the organisation(s) involved.

Electronic information is usually accompanied by metadata that is not found in paper documents and that can play an important part as evidence.

For example the date and time a document was written could be useful in a copyright case.

The preservation of metadata from electronic documents creates special challenges to prevent spoliation.

Background

Company information in databases:

- Vital company data protected in transactional systems.
- Optimized for protection not search.
- Skilled users with knowledge of the database formulate search queries.

Problems:

- These skilled users will not be able to formulate a database query if no one knows what they are actually looking for in the database!
- For a search relating to a legacy database possibly no one with knowledge of the content or structure of the database may still work for the enterprise!

How can we look for relevant information in such cases?

Introduction

Enterprise search systems are intended for use within an organisation by employees or other authorities seeking information held internally in a variety of formats and potentially at a number of locations.

And potentially to legal agencies in the case of legal actions involving the organisation.

Public search services, such as web search engines, give the perception that search is easy. Enterprise search however raises a number of challenges.

Introduction

Typically enterprise search does not make “all” of an organisation’s content available for search. Constraints on available content include:

- Considerations of security.
- Inability to index specialised content, e.g. Flash files, images.
- Difficulty integrating structured and non-structured content.
- Sheer cost, time and difficulty required to incorporate the diverse content repositories held within the organisation.

Challenges of Enterprise Search

The major challenge faced by enterprise search is the need to index content from a wide range of sources within an organisation, and then to search it effectively and efficiently.

The objective in enterprise search is generally to form an integrated ranked list in order of decreasing likelihood of relevance.

Why will this generally be difficult?

Challenges of Enterprise Search

Users of enterprise search systems:

- (Class 1) Members of an organisation who may be familiar with the information or documents that they are searching for, or
- (Class 2) who may be looking for information held within the organisation, but they do not know where it may be found.
- (Class 3) Third parties (e.g. legal investigators) who are looking for information relating to a topic of interest. They may have anything from extensive knowledge of the information held by the organisation to almost no knowledge.

Challenges of Enterprise Search

The different types of user will generally have differing abilities to generate suitable search queries.

Consider the differences between current company employees and legal investigators.

For example, what vocabulary should be used (e.g. there may be vocabulary specific to the enterprise concerned), or knowledge of the subject (as it relates to a specific enterprise).

Users will have differing abilities to recognise relevant content, e.g. can they recognise the item that they are looking for easily or do they need to study it in detail to discover whether it is or might be relevant?

Challenges of Enterprise Search

The metadata associated with information from different sources in an enterprise may have different formats (e.g using an XML markup standard) and/or have different metadata fields (e.g. email, meeting minutes, etc.).

This means that it is difficult to index it in a consistent way within the search index file.

The varied nature of content means that it can be difficult to produce meaningful document rankings.

Challenges of Enterprise Search

Similar to web search, it can be difficult to achieve and maintain coverage and freshness of the contents of an enterprise search system, i.e., it is difficult to locate and extract all content from within the enterprise, index this, and to keep the indexes of the search system up to date.

Some issues includes:

- the presence of multiple copies of the same content in different places,
- near duplicate detection,
- difficulty in determining content which has changed recently,
- network bandwidth issues (e.g. between offices in different parts of the world),
- difficulty of link extraction from JavaScript and Flash.

Challenges of Enterprise Search

Successfully crawling content within an enterprise can be very challenging, e.g. from records management systems, customer relation management systems, and content management systems.

Often there is no link-structure to assist with document ranking (i.e. we cannot use PageRank type algorithms to help determine which items are likely to be relevant to most users).

Metrics such as popularity, recency of update, spam score of email, document type, source repository (i.e. where within the organisation the document originates - some places may be known to be more reliable or useful) may be used in query-independent weighting of documents.

Since it is difficult to rank documents in a meaningful way, rich user interfaces can be used to enable exploration of collections.

Ranking Problems in Enterprise Search

In addition, queries may often fail due to differences between the language of the query and the language of some or all of the documents. Consider the previously identified user classes.

Enterprise search systems can include tools such as:

- stemming - as discussed in Section 2
- thesaurus expansion - as discussed in Section 2
- relevance feedback - as discussed in Section 2
- query suggestion - of queries used in previous searches

to help bridge this gap between terms occurring in queries and documents.

Database Search

- Queries to a database typically use a complex language form - such as SQL.
- Requirement to know how the database is structured.
- Difficult to pose similar queries to multiple very different databases.
- No ranking of output from each database.
- No meaningful way to integrate the output from multiple databases.
- Where should the user start looking among the retrieved content?

Database search is like Boolean search in information retrieval.

Modern Information Retrieval Search

Characteristics of best-match information retrieval as studied in this module:

- Queries can be posed in simple language - although complex structures can be accommodated.
- No requirement to know how database is structured.
- The same query can be posed to any indexed document collection.
- Meaningful ranking of retrieved documents can be produced.
- Ranked lists can be integrated into a single list - although reliable merged ranking is difficult.
- User can start looking at the top of the list.

Comparison of Database Access and Search Technologies

- Database technology users need training.

  No training required for users of best-match search engines.

- Search engines - subsecond response time.

- Search engines - index can be updated pretty much in real time now.

- Database systems can’t work with peak loads.

  Search engines don’t have this problem.

Database Offloading

Database offloading consists of the following process:

- Queries are applied to a database to probe what is contained in the database.
- Each line of results retrieved from the database in response to a query becomes a pseudo ”document” which is indexed in a search engine.
- Standard unstructured queries can then be applied to the indexed contents of the database to find out if it contains relevant information.

The database still exists, but the database itself is not the main access source to the information it contains.

Connection to Databases

Best-match search queries can be used to search databases in two ways:

- - Connect directly to the database and transform the search request into SQL or the appropriate native database query language.
  - Apply transformed query to the database.
  - Collect the output. Note: This will mean that retrieved documents will not be ranked.

Connection to Databases

- - Use a search engine spider to crawl the pages generated from the database using database offloading.
  - Index the generated pages into a standard text information retrieval system.
  - Apply best-match unstructured queries to the data indexed in the information retrieval system.
  - Collect the ranked output.

Access Control

A key consideration in enterprise search is the right of access to information.

Different people have the right to see different things.

Data may be restricted to certain users.

In these cases controls can be used to limit access to those with the authority to view certain information, e.g. only the human resources office will typically be authorised to access personnel files, and only the finance office will have access to all the financial data.

Access Control

“early binding security” - access control attributes are stored when the document is indexed.

“late binding security” - each entry in the results list is checked at display time.

There is a trade off here in performance:

How “fresh” the content is - if it’s very fresh, there may be no opportunity to set controls at index time

How quickly a system can respond to the query - filtering at retrieval time can be time consuming and inefficient.

Compliance, eDiscovery

Legislation in many countries now requires organisations to store much of this information in case it is required for future legal disputes, e.g. financial malpractice, illegal knowledge sharing, actions by or against employees.

- Email can be equivalent to paper-based documents as valid and admissible in a court of law. Typically 40,000 emails per year can be associated with each employee.
- Obligation to store all business communication

The world post-Enron!

Laws relate to businesses, and health and civil organisations.

Compliance, eDiscovery

eDiscovery (electronic discovery) is discovery in civil litigation which deals with the exchange of information in electronic format.

Usually (but not always) a digital forensics analysis is performed to recover evidence for use in legal cases.

Data is identified as relevant by lawyers. Evidence is then extracted and analysed using digital forensic procedures.

Only need to hand over information required by law - search can be used to cleanse content of other information which is not legally required.

Individuals working in the field of electronic discovery commonly refer to this field as *litigation support*.

User Interfaces

User interfaces for Web search engines are typically deliberately kept very simple.

Think of the *Google* interface.

It needs to be simple so that untrained users can use it, but also a simple interface means that users will be less likely to be distracted from clicking on the paid ads! (which is the main way that web search engines make money!)

Enterprise search systems typically give users much richer user interface functionality. This can occupy significant amounts of screen space - leaving no room for ads!

However, there is an expectation that finding information within an enterprise should be fast and efficient, and that it should be done through a single interface.

User Needs in Enterprise Search

Users of enterprise search systems within an enterprise (i.e. employees) are often looking for a single *known-item*, i.e. they already know what item they are looking for (since they either created it or they have seen it on a previous occasion), but they need the search system to help them find it.

This can often be better facilitated through *exploration* of the available content (sometimes referred to as *exploratory search*), rather than trying to create a single ranked list.

This can be supported through methods including:

- Document clustering
- Faceted search

Supporting User Needs in Enterprise Search

The more complex user interfaces used in enterprise search allow the user greater control of the output to support methods such as cluster-based browsing and faceted search.

These more complex interfaces may require the user to be trained in the use of the system and its interface.

Since search is part of their job, they will be invest time in learning to use the system (their employer may even send them on a course) and to spend time when performing search activities.

Document Clustering

Documents can be clustered based simply on metadata fields, e.g. from the same email sender, written on the same date;

or based on their contents where documents with similar content (words, phrases, etc) are placed into clusters as potentially related items (algorithms for content-based clustering are beyond the scope of this module).

Users can then browse clusters of documents rather than looking at documents one by one.

This is useful for enterprise search since users will typically remember certain features of the document(s) they are looking for or about the general subject matter of their target document(s). Or want to explore to see what they might be available, e.g. legal investigation.

Faceted Search

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.

Thus they could start their search by looking at items from this person.

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