Modern Information Retrieval

Chapter 15

Enterprise Search

Introduction
Enterprise Search Tasks
Architecture of Enterprise Search Systems
Enterprise Search Evaluation
Potential Reasons for Dissatisfaction
Context and Personalization

Introduction

- **Enterprise Search**: the application of IR technology to information finding within organizations
- Enterprise search may be interpreted as search of digital textual materials owned by an organization, including
  - search of their external Web site
  - company intranet, and
  - any other electronic text that they hold

Characteristics and Applications

- Many characteristics of enterprise search represent a significant challenge for IR system designers
  - Information in the enterprise may be structured or unstructured
  - Documents are produced by a variety of sources, generally without formatting standards
  - Metadata may be created according to a number of different schemes, or may not be added at all
  - Not all users have the same access rights to all information
  - Some information are highly confidential

Characteristics and Applications

- There is evidence that employees spend a significant amount of their time searching for information needs
  - According to IDC, a company with 1,000 information workers can expect more than $5M in annual wasted salary costs because of poor search
    - people spend 9 – 10 hours per week searching
  - According to Butler Group, as much as 10% of a company's salary costs are wasted through ineffective search
  - A 2007 Accenture survey of 1,000 middle managers found they spend as long as two hours a day searching for information
    - more than half the information found by searching is useless

Characteristics and Applications

- Another area of high financial impact for enterprise search is the area of **e-discovery**
- Search tools capable of auditably searching all the sources of information within an organization are in increasing demand
- **E-commerce** sites are often driven by a search tool whose functions include
  - supporting search for product information and reviews
  - location of the actual product purchase page
  - search-driven advertising
  - intelligent recommendations

Characteristics and Applications

- Data produced by enterprise search tools can provide
  - information about trends
  - information on sudden spikes in customer interest
  - extensive reporting capabilities
- Search tools are used within organizations to locate expertise when putting together project teams
- Automatically generated internal reports may include summaries derived from search results
- Benefits of Web publishing and search are taken advantage of on almost all corporate intranets

Characteristics and Applications

- Enterprise search almost inevitably includes a small-scale Web search dimension
- Upstill *et al* showed that anchor text and PageRank variants were very effective in small-Web contexts
- Hawking *et al* investigated whether links from outside an organization could be used to improve the quality of Web site search for that organization
  - most external links to the organizations under study tended to reference only the site entry page
- Hawking and Zobel studied the value of topic metadata for enterprise search
  - topic metadata was of extremely low value in answering queries

Enterprise Search Software

- Companies such as FAST Search & Transfer and Autonomy are well-known for their enterprise search
- Major companies like IBM, Oracle, and Google have also developed products intended for this market
  - Google's Search Appliance is popular due to its ease of use and the familiarity of the Google name
- Smaller companies offering enterprise search products may create a niche through special features
  - Vivisimo, Endeca and Funnelback

Workplace Search

- There is a distinction between enterprise search and other types of search conducted by employees
- **Workplace search:** all searches conducted by employees
  - also covers search of information sources held external to the organization
  - examples: the Web, patent databases, legal resources, subscription information services
  - pilot survey reported by Paul Thomas illustrates diversity of sources accessed by different employees

Enterprise Search Tasks

Search-Supported Tasks

- Many tasks carried out by employees are either made possible or made more efficient by search tools
- Following, we present some examples of tasks which may be supported either
  - by application-specific, or
  - by integrated information retrieval tools

Approving a Travel Request

- In order to decide whether to approve a employee travel request, a manager requires a variety of information:
  - At what level of seniority is the employee?
  - How beneficial is the event likely to be to the employee and the company?
  - How much has the employee spent on travel in the past year?
  - What is company policy on this type of travel?
  - Is the employee performing well?
  - Would their proposed absence of work cause a loss of production or the failure to meet a deadline?
- A newly appointed manager in this circumstance would need to search a variety of information sources

Responding to Calls in a Call Centr

- Many call centres rely on efficient search tools operating over carefully prepared documentation
- If the search tool always finds the right answer page and reduces time wasted looking for answers, then
  - a less skilled and less trained workforce can do the job for lower wages
  - support calls can be shorter and more calls can be handled with the same number of operators

Course of Dispute

- When projects fail a company may need to follow the trail of communication leading up to the adverse event
- Effective search for critical emails and project documents may be critical to making the right response

Writing a Proposal

- For a private company, responding to a large “Request for Proposal” opportunity can be a costly business
- Many such RFPs require responses to hundreds of questions, and result in a large document
- The cost of responding can be significantly reduced if a search tool can quickly and accurately
  - locate the best response paragraphs from previous RFPs
  - locate the best images from previous RFPs
  - locate other useful and current company documentation

Obtaining and Defending Patents

- Industrial companies typically subscribe to commercial patent database and use specialist patent search tools
- Patent search poses many challenges, including
  - obscure language of patent attorneys
  - the need to search patents in all languages
  - the need to search for diagrams and chemical structures as well as text
  - the need to recognize variants of chemical and biological names
  - the need to impose relational constraints over factors such as reaction temperatures

Obtaining and Defending Patents

- Information about searches is of course highly confidential information
- Intellectual property (IP) searching takes several forms:
  - **Patent landscaping**: Identifying patent gaps in a particular field in order to target the company's research into fruitful areas
  - **Freedom to operate**: Does a technology created by the company violate any patents held by others?
  - **Novelty search**: Is a new discovery potentially patentable?
  - **Patent invalidity search**: Can we discover prior art in a field which would enable us to strike down a patent held by a competitor which is impeding our business?

Selling to an Existing Customer

- The probability of making a successful pitch to a customer can be substantially increased if
  - pitch is targeted at solving problems the customer actually has
  - vendor can present themselves as competent, professional and attentive to the customer's needs
  - vendor can identify who are the most useful contacts within the customer organization and what roles they play
- Successful customer relationship management (CRM) relies on the ability to
  - effectively search
  - analyse and present all the data relating to that customer

Expertise Finding

- Expertise finding is a particular problem in large organizations
  - In some cases, a dedicated application maintains a register of expertise which can be queried in normal database fashion
  - In other cases information created and published for other purposes can be mined for expertise
- In the latter type of system, identifying the set of candidate experts is a significant problem

Expertise Finding

- An expertise finding developed in CSIRO intersected the crawl of Web pages with a employee database
- Passages of text including the name of a employee were extracted and added to a surrogate document
- When a query was processed against the experts collection, documents representing people were ranked
- The contact details for the top-ranked people were returned as the search results
- Subsequent research within the TREC Enterprise Track has showcased improved methods

Operating an E-commerce Site

- Some businesses rely on e-commerce Web sites for some or all of their revenue
- A typical e-commerce site provides
  - product search, coupled with query suggestion
  - faceted navigation
  - automatically generated cross-sell recommendations
- Ranking algorithms for e-commerce sites must take into account a variety of non-traditional factors
  - stock levels, use-by-dates, and profit margins on different products; and whether items are “on-sale” or part of some promotional campaign

Search Types

- Broder identified three distinct types of Web search: navigational, transactional and informational
- Queries of all three types may be submitted to enterprise search engines, *e.g.*
  - **navigational:** ‘library’, ‘HR’, ‘plastics division’
  - **transactional:** ‘buy parking permit’, ‘renew library card’, ‘claim expenses’.
  - **informational:** ‘IP policy’, ‘customers in Spain’, ‘product xyz - error 57’

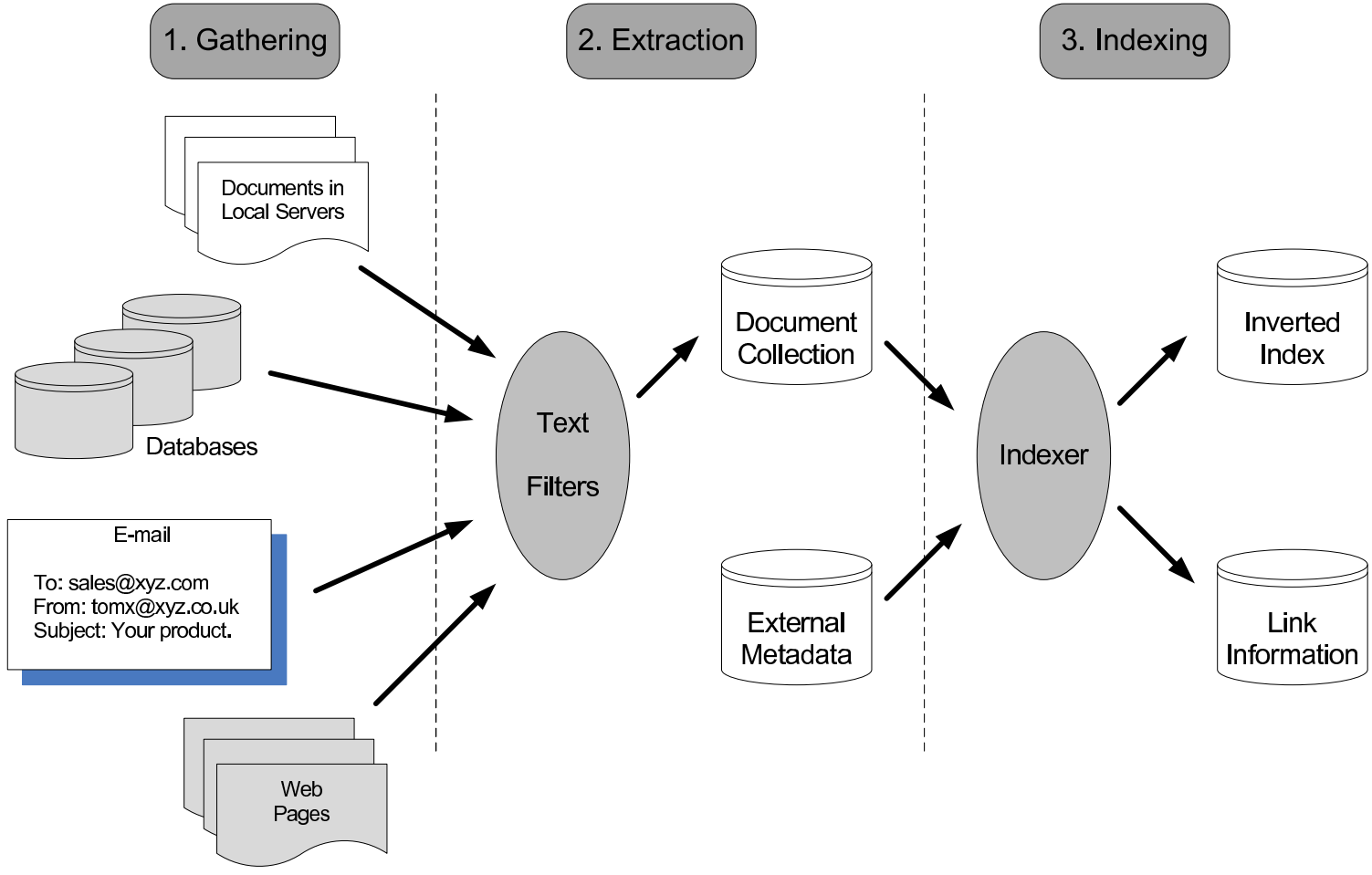
Studying Enterprise Search

- It is very difficult to study search behaviour within an organization of which you are not an employee
  - query logs are unlikely to be made available for publication
- For similar reasons, it is unlikely that experimenters will be permitted to follow employees around
- However, the TREC Enterprise track in 2007/8 provided information need statements and relevance judgements for two real tasks
  - provided by Science Communicators from Australia's government research organization CSIRO

Architecture of Enterprise Search Systems

Architecture

- Phases involved in building a unified index of heterogeneous enterprise data



Gathering

- Crawling of search engines may be very complex
- First, coverage and freshness can be subject to many of the challenges faced on the external Web
  - redirections
  - publication of multiple copies of the same content
  - difficulty of identifying content which has changed recently
  - near-duplicate detection
  - network bandwidth issues
  - difficulty of link extraction from JavaScript and Flash
- However, techniques allowing servers to supply lists of changed content may be deployed without risk

Gathering

- In addition, **scanning of file systems** is relatively straightforward
- Further, crafting of SQL queries may allow the extraction of all and only the information needed in **database gathering**
- Second, successful gathering depends upon the availability of appropriate APIs or adapter software
  - Records management, customer relationship management (CRM) systems, and content management (CMS) systems are some of the generic classes of such systems

Gathering

- Particular issues arise with systems such as Lotus Notes
  - In this system, objects of various types are accessible both through a native API and through standard Web publishing
  - A single document may be synthesised from multiple content fragments while
  - On the other hand, multiple views of the same basic content may be published at multiple URLs

Gathering

- To illustrate, all of the following URLs from an anonymized organization represent the same document
  - `.../d/xyz%40.nsf/mf/3240.1?OpenDocument`
  - `.../D/xyz@.nsf/b06660592430724fca2568b5007b8619/1c87d9876bc11ee8ca256fd5007722a8!OpenDocument`
  - `.../D/xyz@.nsf/5087e58f30c6bb25ca2568b60010b303/1c87d9876bc11ee8ca256fd5007722a8!OpenDocument`
  - `.../d/xyz@.nsf/w2.2.2/1c87d9876bc11ee8ca256fd5007722a8!OpenDocument`
  - `.../d/xyz@.nsf/w2.2.1/1c87d9876bc11ee8ca256fd5007722a8!OpenDocument`

Gathering

- Third, in many applications it is necessary to gather access control lists (ACLs) and external metadata
- Fourth, whether **email** can be made searchable then either:
  - a corporate decision that email sent to an organizational address is not private to an employee, or
  - the implementation of a system for segregating organizational and private email
- Fifth, some organizations have embraced and adapted so-called “Web 2.0” approaches

Gathering

- Sixth, the gathering process may take a long time and generate telecommunications charges
- Very large efficiency gains can usually be achieved by taking an *incremental* approach
  - In a large intranet or database, it is unlikely that even 1% of content will change in the course of a day
- Note that gains from incremental gathering may flow through to filtering and indexing

Extracting

- Extracting (or filtering) text from binary documents seems as though it should be a simple task
- In practice, filtering issues may be a major cause of user dissatisfaction with enterprise search results
  - Failures in filtering may result in meaningless titles, poor quality cached copies, and garbled summaries
- Further, critical documents on a topic may not even be recognised as matching the query

Extracting

- Why is filtering harder than it seems?
- First, the use of proprietary document formats
- Second, the loss of text semantics when encoding a document in a presentation-oriented format
- Third, the representation of metadata
  - many well-known formats such as MSWord, PDF, OpenDocument, and JPEG are capable of storing metadata such as title, author, subject, and date
  - in practice, however, these metadata are usually missing

Extracting

- In any case, many commonly used document formats are limited in the types of metadata they can record
  - leads to reliance on external metadata repositories for recording details about the document
  - leads to the absence of certain metadata which might be useful in retrieval
- Fourth, textual information may be present in a document only in scanned form

Extracting

- Fifth, the accessibility of content within a document
  - Documents may be compressed in a variety of schemes and may be encrypted
  - PDF documents may be flagged internally to prohibit the extraction of text
- Finally, important structure within a document may be represented in many types of document by typesetting conventions
- Filtering of large collections of binary format documents may be very time consuming
- However, the time taken can usually be dramatically reduced by incremental filtering

Indexing

- There is no particular reason for index used by enterprise tools to differ from those used on the Web
- In enterprises, however, there is a particular need to index fielded data
- Indexing systems vary in:
  - how well they support the different types of data to be indexed
  - the rate at which they can index data
  - their ability to efficiently support phrase and proximity operations
  - the compactness of the indexes they produce

Indexing

- A challenge in designing an indexer lies in how best to deal with incrementally updated content
- An incremental indexer deals with updates as follows
  - new documents are dealt with by appending a new entry to the document table
  - new entries at the end of the postings lists corresponding to each of the terms the new documents contain
- This potentially requires a lot of random-access I/O
- Further, this may be in conflict with the document ordering necessary

Indexing

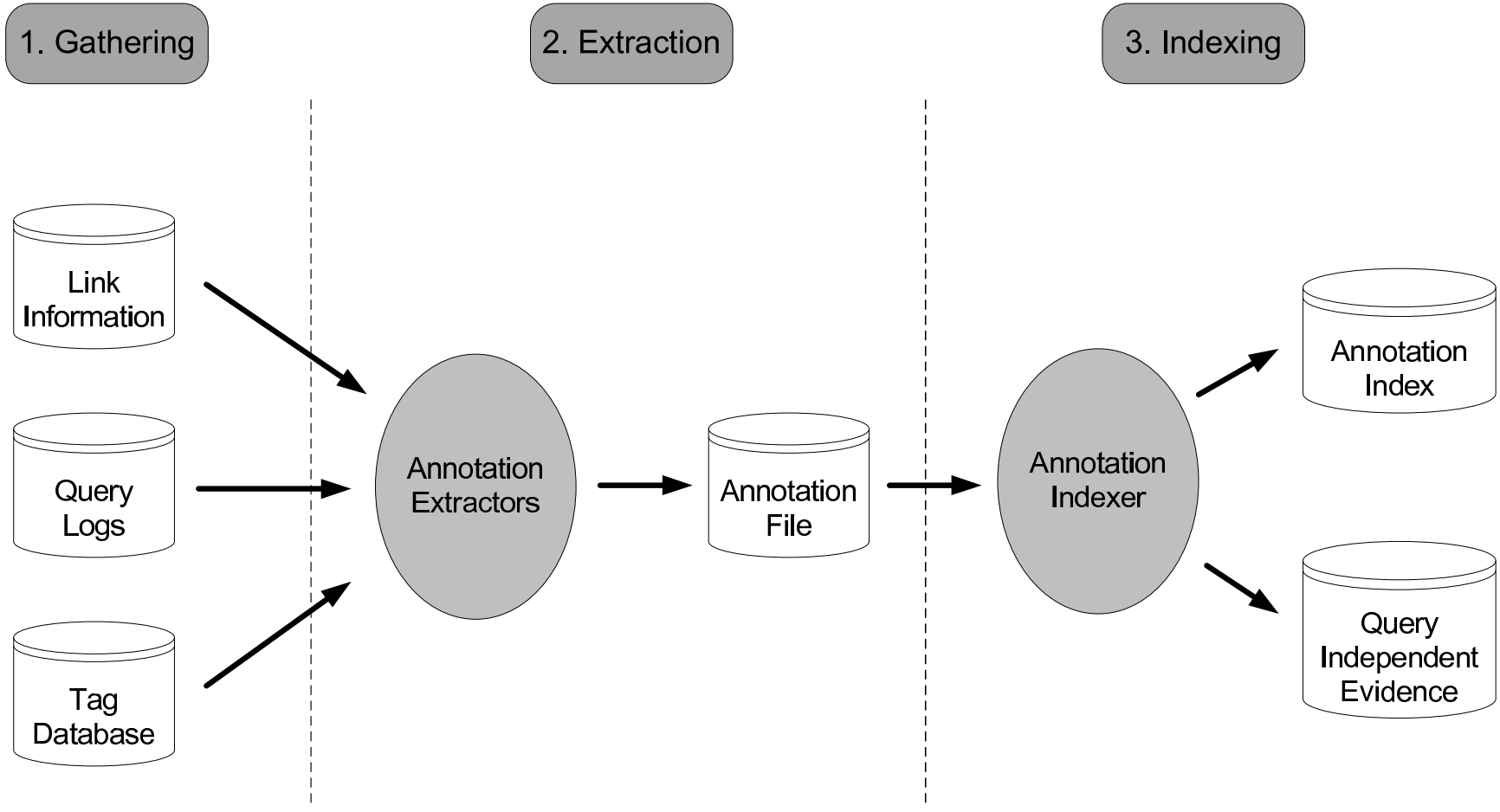
- Deletion of documents from an incremental index poses even more problems
  - This is particularly true when the postings lists are compressed
- A typical solution is to leave the postings where they are but to mark the document as deleted
  - However, the index can grow significantly in size while access speeds decline due to fragmentation and loss of locality
- An alternative to using an incremental index is to maintain a combination of baseline and update indexes and search them in parallel

Indexing Textual Annotations

- As on the Web, enterprise documents may be annotated using various mechanisms
  - anchortext from links
  - officially applied metadata
  - click-associated queries
  - folksonomy tags

Indexing Textual Annotations

- Processing of annotations to support effective enterprise search



Indexing Textual Annotations

- Annotations can be used to provide query-dependent and query-independent scoring components
- Aggregated annotations may be:
  - scored separately and combined with text scores, or
  - treated as fields of the original document

Query Processing

- Queries often fail in enterprise search because of a difference between
  - the language of the query, and
  - the language of the documents
- To help bridge this gap, enterprise search systems may provide tools such as:
  - thesaurus expansion
  - query suggestion
  - stemming, and
  - relevance feedback

Query Processing

- Enterprise search may combine text-derived scores with a static score
  - The static score formulation may need to be tuned to the characteristics of the particular publication environment
- For example, in some organizations, link counts and/or URL lengths may not be useful for ranking
- Further, there may be no inter-linking and the static score may take into account new factors
  - frequency of access to a resource
  - recency of publication
  - spam score of emails
  - document type or genre
  - repository

Query Processing

- In a e-commerce site, static scores may consider the number of times an item has been borrowed or bought
- In addition, benefit to the publisher may be considered in determining scores
  - profit margin on an item, the perishability of goods or the availability of stock

Query Processing

- The value of query-independent contributors to the static score may vary from organization to organization
  - Publication dates may be reliable in some organizations and not in others
  - The value of URL features is grossly diminished in an organization where URLs are generated in fixed formats by CMSs
  - Links are not useful if they arise from a single navigation template, created by a single person

Query Processing

- Search rankings must suppress near-duplicate results and encourage diversity, in order to overcome
  - the likely presence of multiple drafts and versions of documents
  - the presentation of sales materials in multiple forms
  - the presence of the same documents on file shares and in email attachments
  - the universal problem of publishing the same material at multiple URLs

Query Processing

- Scoring the organization's Web documents may potentially use factors such as
  - URL structure and length
  - link counts and anchortext that are not present in email
  - the staff database
  - the CRM system or the records management system (RMS)

Presentation of Search Results

- In many cases, the type of a search result may be a signal of how useful that result is likely to be
- Results in a unified ranking may include
  - icons, representing source or document type
  - thumbnails, images for products or staff profiles
- Alternatively, the search result list may be segmented by source or type
  - results from the staff directory may be presented above results from the intranet

Presentation of Search Results

- It is also very common within organizations to provide searches which are **scoped**
  - the search box on a Human Resources intranet Web site may restrict search results to URLs from that site only
- Other search interfaces may be provided to search:
  - only the staff directory
  - only policies and procedures
  - only the staff bulletin
- In some cases, sorting options should be provided
  - E-mail search results may be sorted, most recent first
  - Publications may be sorted in alphabetic order of title

Presentation of Search Results

- Enterprise search tools may provide a range of extra facilities, including
  - clustering
  - metadata facet counts
  - multi-document summaries
  - spelling suggestions
  - related queries

Presentation of Search Results

Oxfam Australia *Intranet*

Welcome to the Oxfam Australia Search Engine

SEARCH POWERED BY funnelback | swanson | search

1 - 10 of **174** search results

Didn't find what you are looking for? Ask the Help Desk or provide feedback on the search results.

Location
- Intranet (99)
- Help Desk (63)
- Program Database (10)
- Contracts Database (1)
- Phone Book (1)

Year Published
- 2008 (84)
- 2007 (32)
- 2006 (26)
- 2009 (22)

Format
- Online Resource (173)
- PDF (77)
- Word (5)
- Excel (2)
- PowerPoint (1)

Mike **Swanson**

Position: Knowledge & Information Services Team Leader
Section: International Programs
Unit: Program Development
Office: Melbourne, Australia
Email: [redacted]
Office Phone: [redacted]
Workstation: 0.24

Countries where we work
Last reviewed: 8 April 2009. This document will be reviewed periodically. For queries or comments relating to this information please contact Mike **Swanson**.
intranet.oxfam.org.au/programs/regions/

[PDF] The modified Program Database
If you have any questions on using the modified Program Database, don t hesitate to contact Mike **Swanson** on x.3416 or [redacted]
intranet.oxfam.org.au/programs/ips/pdu/modified-program-database.pdf

Program Database
and for any questions please contact Mike **Swanson**.. Browse Projects. All. By Geography. By Managing Unit. By Partner. (Alphabetical.). By Sector. (DAC.). By Strategic Objective. Status. Active Projects Only.
ips.oxfam.org.au/program_db/

Currently browsing...

Search Terms
swanson

Have you tried...

Swanson By Type
Mike...
Michael...

Swanson By Topic
...Knowledge

Presentation of Search Results

- Enterprise search systems may also include tools for
  - analysing a deep results set
  - showing results in the context of a map
  - featuring thumbnail images for image/product results
  - including key frame thumbnails for video segments
  - activating automatic alerts via RSS or email, based on user interest profiles

Presentation of Search Results

- Search can be integrated into enterprise application
  - email search results may be presented as a virtual email folder, supporting all the usual folder operations

Presentation of Search Results

- Enterprise search tools can play key roles in the customisable employee interfaces
- A customised portal page can provide
  - alerts
  - targeted summaries of the things that a particular employee needs to know about
  - action links
  - personalized search facilities
- Manber *et al* provide a number of lessons learned from *My Yahoo!* which have applicability within the enterprise

Security Models

- Enterprise search tool must be given omniscient acces to privileged users
- Tool must therefore enforce the security of that information
- Rights to view documents or search results depend upon the user login

Security Models

- Threats which may arise from the use of an enterprise search tool include
  - Unintended actions carried out when an internal crawler accesses an active server page or CGI script
  - Allowing someone to see content via the search engine to which direct access would be forbidden
  - Preventing someone from accessing content via the search engine to which they would be granted direct access
  - Providing means by which a malicious unprivileged user may deduce the existence of a sensitive document
  - Externally accessible enterprise website search is potentially vulnerable to cross-site scripting, JavaScript and other types of injection and to buffer overflow vulnerabilities

Security Models

- Access to an organization's documents is generally controlled by Access Control Lists (ACLs)
- These may specify which of a range of actions are available to particular individuals or to groups
- Folders or directories may also be subject to ACLs
- Organizations may use a network authentication protocol such as Kerberos

Security Models

- There are interesting differences between email messages and other documents from the point of view of security
- In most systems, the sender cannot specify an access control list for the message
  - Instead, copies of the message may be stored either in folders in the recipients' own file systems or in a central mail database
- Access control to the message is thus determined by the recipients or by administrators of recipients' mail databases

Collection-level Security

- Ideally, information can be simply divided into collections with uniform access rights
  - For example: a general-access collection, a finance collection, a senior-management collection, and an HR collection
- Unfortunately, in most cases, the applicable security model is much more complex than this

Document-level Security

- Access controls can be applied at the level of individual documents
  - Behind the scenes, the documents of a ranking results must then be filtered, result by result, to exclude all and only the documents not accessible to the user
- Different organizations impose different requirements on the security restrictions which should be applied when searching

Federation/Metasearch

- Sometimes it is not feasible for a enterprise search engine to index all the information of a organization
  - For example, the gathering processing from a particular source may be too slow, or the network traffic too expensive or slow
  - Alternatively, the data may be locked into a proprietary application which provides no export facility (yes, really!)
- Consider, however, that the problematic source provides its own search facility
- In this case it is possible to provide a unified search by taking an approach known **search federation**

Federation/Metasearch

- **Metasearch**
  - query is received by a broker and forwarded to the search of the federated sources
  - broker combines the separate result sets into a single set to be returned to the user
  - ranks and scores returned by the different search interfaces may be quite incompatible
    - top document from one source may be a weaker match than the 50th rank from another that is more oriented to the topic
  - in the simplest case this arises because IDF values for some terms may be subject to large variation
  - in more complex cases, score variations may be due to different static weightings or annotation scores

Federation/Metasearch

- Search federation may pose particular problems for maintaining document level security
- User credentials must be forwarded by the broker to the individual search services
- A reliable single-sign-on mechanism across all the sources to be integrated seems almost indispensable

Five Sub-problems of Metasearch

- Five problems to be addressed in a metasearch application
  - At service definition time, **identifying** and choosing the sources to be federated
  - At service definition time, and as often as necessary during the operation of the service, **characterising** the sources
  - At query time, **selecting** the subset of available sources to be included in the search
  - **Translating** the query into the query language accepted by each of the federated sources
  - At query time, **merging** the result sets returned from the search facilities at each of the sources

Five Sub-problems of Metasearch

- The majority of work in metasearch area has been evaluated by partitioning TREC Ad Hoc data
  - These partitions show much less variation in documents than would be expected across federated enterprise repositories
- Sources to be federated may cooperate with the broker in various ways
  - For example, they may supply accurate statistics about collection size and document frequencies
- In the uncooperative case, it is necessary to sample documents in order to characterise the server
- Thomas and Hawking proposed a efficient *multiple queries* sampler which attained representative accuracy

Five Sub-problems of Metasearch

- Estimating the size of a collection generally relies on methods developed for estimating animal populations
  - The number of documents in common between two independent, unbiased samples can be used to estimate the population size
- The number of distinct published methods for source selection now exceeds forty!
  - Some rely on information obtained from the sources using probes
  - Others assume the availability of term frequency data, while yet others assume no cooperation

Five Sub-problems of Metasearch

- The CORI method treats each collection as a document and the set of collections as the collection
- Then, it uses a standard relevance calculation as the basis for selection
- In the uncooperative case, its analogues of TF and IDF must be estimated from samples
- Selection methods may take into account estimates of the effectiveness of the retrieval systems operated at each source

Five Sub-problems of Metasearch

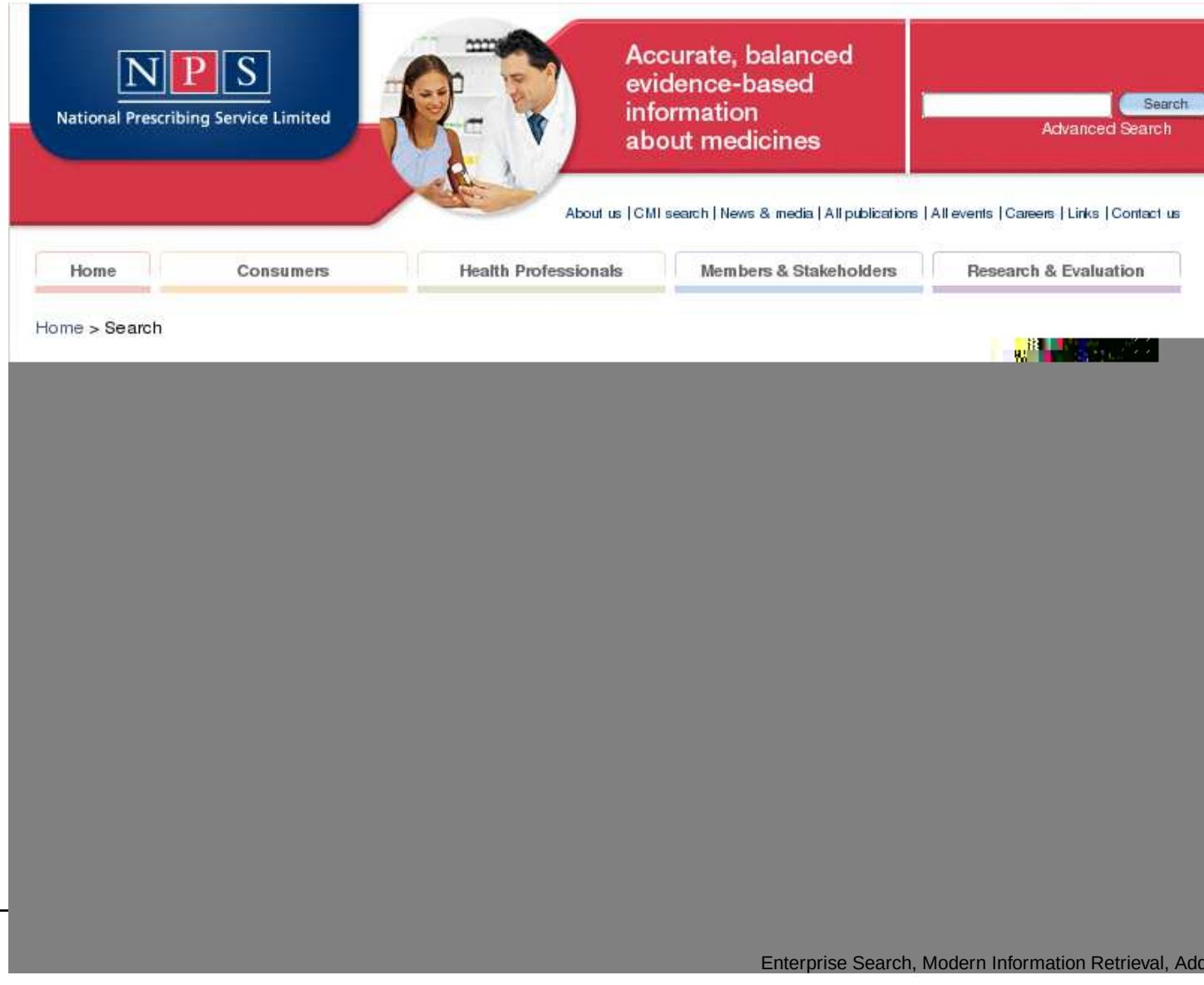
- Lawrence and Giles proposed a merging method
  - all the documents from the primary result lists are downloaded
  - the documents are locally ranked for relevance
- Rasolofo *et al* propose and evaluate strategies for merging results in the case of a current news metasearcher

Five Sub-problems of Metasearch

- Presentation of the results of a metasearch is important to get right in a heterogeneous environment
  - Different types of result may need to be presented differently
- One way to avoid merging problems is to to present result lists for different sources in separate columns

Five Sub-problems of Metasearch

- Segmented result list presentation



Enterprise Search Evaluation

Published Test Collections

- Some difficulties in constructing test collections for enterprise search
  - big differences in information holdings from one organization to another
  - most enterprise information is company confidential
  - differences in the types of searches that are conducted during the normal operation of each company

The TREC Enterprise Track

- The only publicly available collections are those created by the TREC Enterprise Track
- The Enterprise Track Corpora comprise only material published on external Web sites:
  - crawl of `w3c.org`
  - mailing lists from `w3c.org` (converted to Web pages), and
  - crawl of `csiro.au`

Enterprise Search Evaluations

- Reasons why evaluation of effectiveness of enterprise search tools is carried out in practice
  - R&D carried out by a search company to improve algorithms
  - Product comparisons leading to purchasing decisions
  - Tuning of an existing system to make it perform better, within the context of a particular implementation
  - Such tuning may
    1. cut costs by increasing the proportion of public enquiries which can be handled through the Web
    2. increase employee productivity and company competitiveness
    3. increase sales
    4. improve the quality of decision making
    5. reduce complaints

Enterprise Search Evaluations

- Evaluation of enterprise search is no different in principle to evaluation of other types of search
- Comparing two search facilities by presenting their results side-by-side has advantages in enterprise environment

Enterprise Search Evaluations

- Because the comparison tool replaces the search tool normally used by the people studied
  - validity of inferences about the group studied may be subject only to unbiased sampling error
  - there is no need for experimenters to understand what tasks are being performed by the searchers
  - results sets are evaluated in their entirety
  - a person evaluate the result sets obtained according to how well that set meets the need behind their query
  - side-by-side evaluations are conducted in real rather than simulated contexts

Enterprise Search Evaluations

- Concern is sometimes expressed at the lack of sensitivity in side-by-side comparisons
  - no significant preference between systems A and B may be found even after a reasonable numbers of queries
- Making n-way comparisons using the side-by-side tool is optimal for $n = 2$
  - $n > 2$ can interfere with the work of the searcher
- An even bigger limitation of the side-by-side method is the inability to use it for tuning purposes

Enterprise Search Tuning

- For tuning an enterprise search system one can
  - use a conventional but private-to-the-company test collection, or
  - take a machine learning approach and gather large quantities of data of the form
- Judgments for a test collection
  - may be made manually by employees of the organization, or
  - could rely on user click data
- When using a test collection for tuning, the collection should accurately represent the real situation
  - otherwise, optimal parameter settings determined from the test collection might be far from optimal in actual use

Enterprise Search Tuning

- Workload of an enterprise search is faithfully recorded in the query logs
- An approach to unbiased evaluation is to use a uniform random sample from the query log
  - the most useful answers to each information needs can be identified
- Level of performance predicted for a search engine can vary substantially, depending upon how the query set is chosen
- A limitation of the workload sampling approach is the need to infer information needs from queries in the log

Enterprise Search Tuning

- Another issue is that an enterprise search engine is operated primarily for the benefit of the publisher
  - A publisher may want to focus evaluation on the queries which are business critical to the organization
- A desire to bias search engine performance in order to achieve business or political goals is found in many organizations

Enterprise Search Tuning

- The open-source C-TEST Toolkit provides a formal way of representing evaluation test files
- It is capable of modelling many of the factors which are necessary for meaningful enterprise search evaluation
  - weights can be associated with each query in the test file, reflecting importance
  - multiple interpretations for an individual query can be represented
  - the fact that multiple documents in the collection are of equivalent value can be represented
  - relevant answers can be assigned weights reflecting their contribution to meeting the information need
  - the test file entry for a query can specify the appropriate judging depth given the need behind the query

What is it Reasonable to Expect?

- The performance of an enterprise search lies somewhere on the continuum between the following extremes:
  - the best possible answer at rank one
  - a set of partial answers sprinkled among a lot of irrelevant results

The Answer at Rank One

- Consider a query for the name of a company
- In this case, it is expected that the best answer page for the company will appear as the very first search result
- This success relies on the richness of the Web search environment
  - link graph, anchortext, URL length and structure, user-behaviour data, etc
- It also relies on the availability of information published specifically to answer these needs
  - company Web sites and Wikipedia or other sites created to provide high quality definitions and explanations

Set of Partial Answers

- The task modeled by the TREC **ad hoc** evaluation campaign was essentially intelligence gathering from newspaper archives
- Some queries achieve no satisfactory results
  - economic impact of recycling tyres
  - dangers posed by the spread of fissionable materials from the states of the former Soviet Union
- There is no link structure, no anchortext, no site structure and, in TREC ad hoc, no user behaviour data
- No single document exists which is designed to meet the information needs behind either of these queries

Set of Partial Answers

- Furthermore, the query may not use the same words as documents it should match
  - for instance, states of the former Soviet Union should match Russia, Ukraine, etc
- In this case, good search engines perform text-related operations
  - query expansion (stemming, US-UK conflation, pseudo-relevance feedback, thesauri)
  - document length normalization
  - relevant passage upweighting

Where Enterprise Search Lie?

- In a well-organised intranet which looks like a microcosm of the Web, the experience of searching can be at the happy end of the continuum
- On the other hand, baseline performance will be lower and users will be unhappy when
  - the information to be searched consists of blocks of plain text in a database
  - office documents are dumped into an unstructured fileshare without metadata or a naming convention

Potential Reasons for Dissatisfaction

Reasons for Dissatisfaction

- Employee satisfaction with enterprise search and visitor satisfaction with Web site search are often low
- Satisfaction depends upon “search and searchability”
  - depends on both the effectiveness of the search technology, and how effectively information and services are published
- It is sometimes the case that the best answer to a query does not match that query
- In many cases, it seems better to improve the way the information is published, rather than to attempt to modify the search technology

Reasons for Dissatisfaction

- All the current search technologies of which we are aware are statistically based
- The rank of the document naturally depends upon scores achieved by other documents
- Tuning a ranking algorithm can make a great difference to effectiveness

Reasons for Dissatisfaction

- If a key document is in the index, why isn't it top-ranked for this query? Primary diagnostic:
  - Does the document actually match the query?
  - Does the document match the query text less well than other documents?
  - Do shorter documents match the query as well as this one?
  - Do higher ranked documents receive more links or textual annotations which match the query?
  - Do higher ranked documents have features which may indicate that they are more popular or important?
  - Do higher ranked documents come from different repositories to the desired result?
  - Is the target document very similar to another document appearing in the ranking?

Reasons for Dissatisfaction

- Why doesn't a wanted document seem to be in the search tool's index? Diagnostic:
  - Does the document exist?
  - Is it within scope?
  - Am I allowed to see it?
  - Is it reachable by the gatherer?
  - If this document is in a binary or proprietary format?
  - Did the document exist when the index was last updated?
  - Is the document flagged to prevent its display?

Reasons for Dissatisfaction

- There are a number of matching issues which can cause ranking problems within organizational search
  - Often there is a set of matching pages with large numbers of incoming links and strongly matching anchortext which are not good answers to this query

Context and Personalization

Context and Personalization

- Except when indexes are updated, a simple search engine delivers the same results for a query
- But in reality, search performance may be improved if answers to the following questions can be obtained
  - Who is searching?
  - What role are they playing?
  - What are they interested in?
  - Why are they searching?
  - Where are they located?
- Personalized information retrieval represents only one aspect of a broad field of Personalization research

Context and Personalization

- Teevan *et al* quantified the potential for gain from personalizing search
  - For queries supplied by the experimenters, they found a diversity of imputed intents
  - Even when the imputed intents were the same, subjects disagreed substantially in the ratings they gave to results

Context and Personalization

- Pitkow *et al* studied the value of using a client-side personalization system termed as *Outride*
  - Outride builds up a model of the user-based on their searching, browsing, demographic and application use profile
- Search results are reranked with reference to a vector-space representation of the user's profile
- The authors observed dramatic reductions in both:
  - the time taken to complete a search task
  - the number of user actions such as mouse clicks or keyboard entries

Context and Personalization

- Search tools can be customised according to characteristics of groups, individuals, or tasks
- There is also a particular potential for contextualizing search by employees within an organization

Controls for Contextualization

- By what means can that context affect the behaviour of the search system?
- In general there are five categories of search engine controls
  - scoping, static ranking, query manipulation, dynamic ranking, and presentation
- The settings of these controls may be recorded in a **search profile**

Scoping

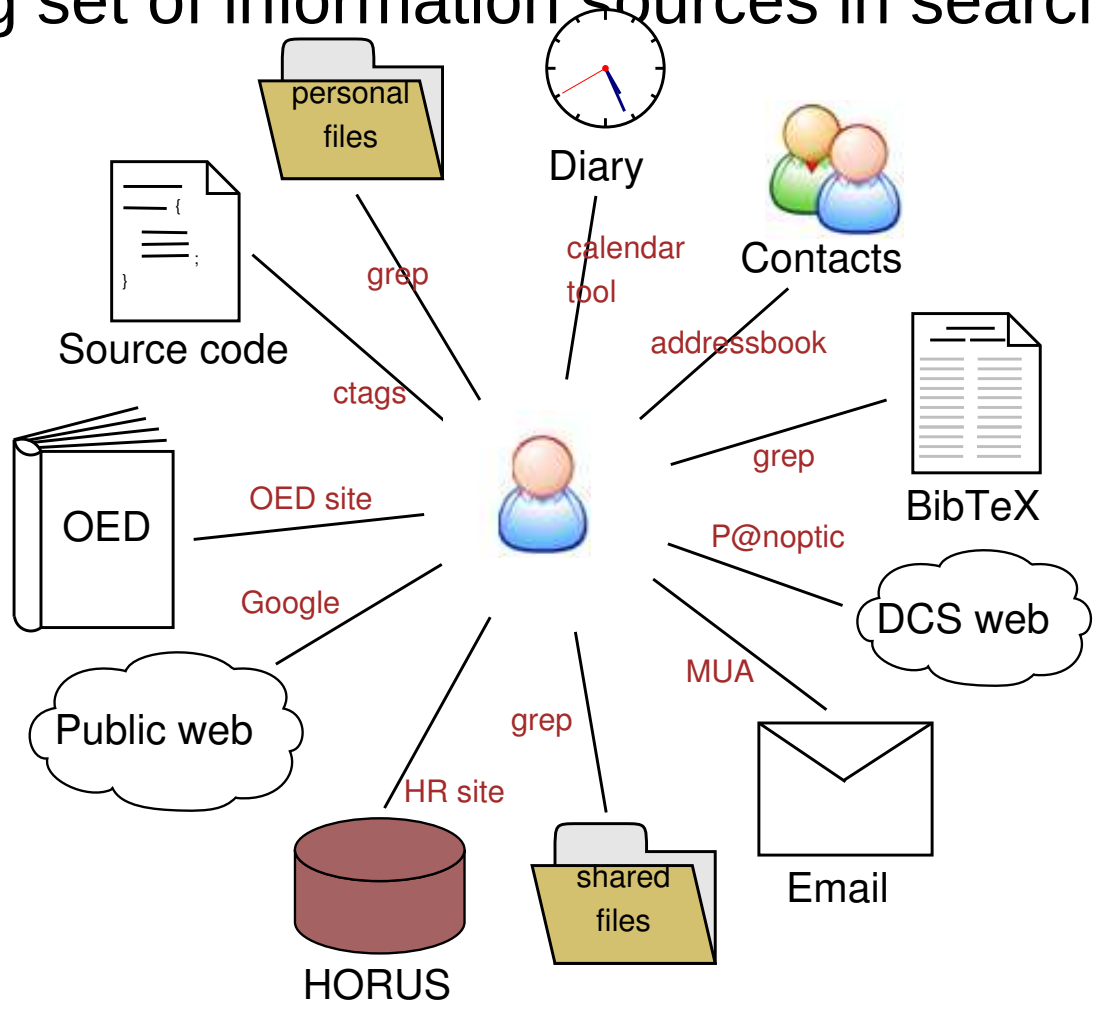
- The scope of a search is the complete set of documents which may be matched against the query
- Scope is controlled by the repositories which are included in the search, by
  - any exclusion filters which are applied to matching documents within those repositories
  - access restrictions applying to this particular search
- It is easy to see that scope is a powerful tool for contextualizing search
  - For example, a technician in the R&D department may be more satisfied with search results if the CRM and finance repositories are not included in the search

Scoping

- Exclusion filters can exclude documents from included repositories on the basis of file type
- Documents can be filtered from a results list because the user is not authorised to access them
- Another familiar example from the Web is adult content filtering
- Personal metasearch is a particular example of scoping
  - Users can choose the set of sources likely to be important to them over time

Scoping

- A particular personal metasearch configuration, showing set of information sources in search



Biasing

- Rather than totally excluding a particular category of content by scoping, it may be more effective to bias the ranking algorithm against it
- Modern enterprise search engines combines many query-independent features such as:
  - author popularity scores, user popularity scores, document recency, etc
- Features can be used to bias search results for or against the feature
  - For example, instead of excluding technical manuals, the ranking function may be biased against them

Manipulating Queries

- The query submitted by a searcher may be manipulated in various ways in order to contextualise results
- In a simple case, group-specific thesauri may be used to interpret ambiguous
- For example, the query *CRM* may be augmented with:
  - **customer relationship management** for the sales department
  - **carbon reinforced mouldings** for the production department
- An extra degree of sophistication is achieved by using a variant of the technique of pseudo-relevance feedback
  - The search results are reranked by locally re-scoring the documents using the expanded query

Manipulating the Ranking Function

- Almost all functions for scoring the relevance of document content to a query are parametrized
- There is therefore potential for adjusting parameters to improve result quality for particular users or groups
- Particular linguistic features or transformations may potentially suit some users better than others
  - For example, people may prefer heavy stemming, light stemming, no stemming, or stemming targeted at a particular language

Customization of Human Interfaces

- The presentation of search results may be customised to the needs of individuals or groups
  - Do you like to see a ranked list of results or a grid of thumbnails?
  - Do you have preferences regarding colours or fonts?
  - What language do you prefer?
  - Do you like to see facet counts, query suggestions, and other add-ons?
- If an employee is unable to use the organization's search facilities, they may struggle to be productive
- Finally, employees may access enterprise search facilities via mobile devices with limited screen area

Contextualization

- It is common for individual employees playing knowledge in an organization to be allocated a PC
  - Sometimes that PC is accessed from centralized enterprise servers, and files and email are stored there
  - However, in many cases the PC must operate in a standalone fashion
- Whatever the detailed arrangements, PCs are able to collect vast quantities of personal interaction data
  - A small but increasing exception is that a great deal of information interaction now tends to occur on mobile devices

Contextualization

- An important element of contextualization is the nature of the task being performed
  - There is great potential to exploit this, because the task may be accurately inferred from the application currently being run
- That inference would be even more reliable if the search function were embedded in the application

Client or Server?

- A client PC is clearly the best place to maintain personal and task profiles for search
- If profiles are kept only on a secure PC, the privacy risks are controlled
- Unfortunately, part of the task of personalizing searches conducted on external search engines is done at the search engine
  - This is where modifications to scope, static scores, and ranking can be fully effective
- Web search engines keep profile information for individuals, maintained from one query to the next
  - However, the profile so maintained is incomplete

Privacy of Profiles

- There is good reason to maintain the privacy of interaction histories and personal search profiles
  - The material could provide very valuable information to a competitor, very useful data for targeting advertisements
  - In less likely scenarios, this is a compelling material for divorce lawyers, blackmailers, police, and foreign intelligence services

Creating and Maintaining a Profile

- There are some researchers who are interested in the total context of search behaviour
  - The objective is to understand humans and their online behaviour
- However, it is very difficult to specify precisely what information should be recorded
- The interaction information can be recorded by:
  - software run at the local operating system level
  - add-ons or plugins to browsers
  - other personal or enterprise applications
- Other less complete records can be made by proxy servers or search engines

Creating and Maintaining a Profile

- Another question to be answered is how to determine which profile should be applied to a search
  - The same person may prefer different profiles depending upon the activity they are engaged in at a particular time
- Can we build an automatic system for determining the right profile which achieves 100% accuracy?
  - Likely not
- But otherwise how do we explain to a user what profiles are available and how they differ
- The best approach may sometimes be to have the user select among a small set of obviously named group profiles

User Modeling

- The process of automatically generating profiles may be described as **user modeling**
- A number of different types of user models have been described

Ontology Vectors

- Pretschner and Gauch describe a system in which
  - user's profile consists of a vector of weights for each of 4,400 hierarchical categories in a published ontology
  - each category is represented by a document vector representing the amalgam of ten exemplar documents for the category
  - when the user visits a Web page, that page's similarity to each of the categories is computed
    - Then, their profile weights are updated
  - research found that
    - profiles converged over time
    - profiles could be used to rerank
  - modest but worthwhile gains on 11-point average precision were obtained through reranking

Ontology Vectors

- Pitkow *et al*
  - achieved substantially greater gains, albeit with a quite different evaluation methodology
  - used a similar ontology vector profile though they do not give precise details
  - their method used query augmentation as well as reranking

Relevance Feedback Methods

- Teevan *et al* describe methods for reranking top-50 results based on a range of different types of profiles
  - A method producing substantial gains is to promote in the ranking URLs from domains recently visited by the user
- A much more sophisticated model represents the user by the index of their desktop search tool
- Unfortunately, both the original Web and URL reranking outperformed this personalized reranking
  - However, a mixed method was found to improve on the raw Web ranking slightly but significantly

Relevance Feedback Methods

- User profiles studied by Waern consisted of long lists of terms
  - The terms were either manually constructed by the user or automatically derived
- The study found that users were generally unable to improve on machine learned profiles
- However, it pointed out that user involvement in profile maintenance was essential to enable correction of errors made by an automatic profile generator

Characterizing Users by their Clicks

- Dou *et al* performed a large scale evaluation of five personalized search strategies
- Two of them were based on clicks and the others based on automatically derived profiles
- They found that the benefit of personalization varied considerably from query to query
  - They found that queries with high click entropy are the ones which benefit most from personalization
- The method based entirely on a user's past clicks is only capable of improving queries which this user has previously submitted

Language Models

- Tan *et al* describe the extension of the language model framework of information retrieval to include both:
  - short-term (current session)
  - longer-term historical language models derived from click behaviour
- Such historical models can be considered another form of profile

Biasing PageRank

- A quite different form of user profile is the personalized PageRank vector model proposed by Jeh and Widom

Implicit Measures

- Kelly and Teevan provide an overview of the use of implicit measures derived from user behaviours
  - Their paper lists five classes of user behaviour – Examine, Retain, Reference, Annotate and Create
- Researchers at Microsoft have also studied the use of implicit measures
  - Fox *et al* established that a probabilistic combination of implicit measures could predict explicit judgements made by users
  - Agichtein *et al* extended this work to include query-dependent measures and proposed a distributional model robust to noise
  - Agichtein *et al* showed that implicit measures, when combined in a large scale machine learning framework, can be used to improve Web search performance

Implicit Measures

- White *et al* point out that explicit relevance feedback can impose a load on searchers
  - They analyse an implicit version of relevance feedback (IRF) in which user actions are used to infer relevance judgements
- Although IRF is less likely to be beneficial, they report that users, particularly novices, preferred it

Information Filtering

- Personalizing search results can be seen as a bringing together of tools from Information Retrieval (IR) and from Information Filtering (IF)
- Hanani *et al* provide a detailed conceptual framework for IR and IF and compare and contrast the two
- Search results are filtered to remove items in which the particular user is unlikely to be interested
  - Personalization aims to achieve better results from searches by combining
    - the specification of an immediate need (the query)
    - term profile

Information Filtering

- In the case of routing and alerting systems there is no immediate query
- Instead, a long-term profile is registered with a search service
  - Newly created or discovered documents are matched against this profile
- For decades, organizations like Lexis-Nexis have offered selective dissemination of information services
  - In these systems, users register a profile consisting of a Boolean query and are sent all documents which match the filter
- In this model the user takes on the responsibility for creating the filter query

Information Filtering

- More recently, Google has provided a similar alerts facility in their Web search engine
  - Documents which newly arrive among the top ranked results for a user-registered query are candidates for being sent to the user
- Google researchers Yang and Jeh discuss perceived problems with this alerting service
  - They describe and evaluate methods for automatically extracting alert profiles from a user's search history
- The challenges are to identify long standing interests in the query log

Information Filtering

- Another approach is to associate an individual with a group and to use a group profile to customise results
  - This is called Collaborative Filtering and (CF) systems which use it are sometimes known as Social Recommender Systems

Social Recommender Systems

- Modern search engines perform a type of generic collaborative filtering in their base ranking methods
  - Documents which are linked to by many authors, or which are tagged or clicked on by many readers, tend to appear higher in result rankings
- Individual searchers thus benefit from the wisdom of the populations of authors, browsers and searchers
- To achieve personalization within a CF framework, one can identify groups within the overall populations and associate individuals with appropriate groups

Social Recommender Systems

- Heer and Chi studied methods for categorising user sessions on the `xerox.com` Web site
  - The authors performed a user study in which users were asked to perform realistic information finding tasks
- By using a combination of features they were able to achieve very high clustering accuracy
- It is not clear how quickly a new visitor or a new browsing session could be classified

Social Recommender Systems

- An online shopping can draw a customer's attention to items that they are likely to be interested in
- People who bought item X also bought item Y:
  - This problem can be addressed by treating an item selected by a customer as a query and retrieving related items
- However, Amazon found that search-based methods failed when users make large numbers of purchases
  - Instead they use a related item method, with the vast item-item similarity matrix computed offline
  - Two items are considered closely related if they both tend to be purchased by the same customers