Foundations and Trends® in Information Retrieval
Vol. 11, No. 1 (2017) 1–142

DOI: 10.1561/1500000053

the essence of knowledge

Searching the Enterprise

Udo Kruschwitz
University of Essex, UK
udo@essex.ac.uk

Charlie Hull
Flax, UK
charlie@flax.co.uk

Contents

Abstract

Search has become ubiquitous but that does not mean that search has been solved. Enterprise search, which is broadly speaking the use of information retrieval technology to find information within organisations, is a good example to illustrate this. It is an area that is of huge importance for businesses, yet has attracted relatively little academic interest. This monograph will explore the main issues involved in enterprise search both from a research as well as a practical point of view. We will first plot the landscape of enterprise search and its links to related areas. This will allow us to identify key features before we survey the field in more detail. Throughout the monograph we will discuss the topic as part of the wider information retrieval research field, and we use Web search as a common reference point as this is likely the search application area that the average reader is most familiar with.

U. Kruschwitz and C. Hull. *Searching the Enterprise*. Foundations and Trends®
in Information Retrieval, vol. 11, no. 1, pp. 1–142, 2017.
DOI: 10.1561/1500000053.

1

Introduction

“Enterprise Search doesn’t work”. Does that sound familiar? Well, it is a view commonly held by employees trying to find information within their organisation. On the other hand, an information retrieval (IR) researcher or student might never have heard this or even thought about it.

Given the wide-spread sentiment among users and search practitioners that enterprise search does not deliver on its promises, the question arises as to what is it that causes these perceptions and do they properly reflect the reality? One of the fundamental issues underlying the overall discussion is the question of how does enterprise search relate to search in general and Web search in particular. This monograph will provide a thorough discussion of the topic and outline implications and guidance resulting from this. We will focus on both theoretical and practical issues as well as their interplay.

We expect our main audience to be researchers and PhD students with some background in information retrieval who want to learn something about enterprise search. Apart from that, we do hope that practitioners facing the challenge of having to implement an enterprise search system and those that need to understand the technical and user interface problems associated with enterprise search will also find the monograph valuable.

There are at least two ways of approaching this topic academically. One approach might be to review the refereed literature and provide a contextualisation of enterprise search by comparing and contrasting it with related work. A second approach could do this by highlighting research gaps in enterprise search and providing a research agenda in the spirit of the *Strategic Workshops on Information Retrieval (SWIRL)* in Lorne.[1] Given the relatively limited academic interest that enterprise search has attracted, in particular a lack of detailed comparison with other search areas, we opted for the first approach as the first step. We do however need to note that due to the applied nature of the field the refereed literature will only be able to paint a partial picture and not necessarily represent where the technology is at the present time. We therefore complement the analysis with appropriate references to studies and surveys from the practitioners' community. In concluding the review we will also provide a number of interesting future research directions.

1.1 Overview

Web search is a relatively recent development that has attracted much attention and for many it has become a synonym for 'search' in general. But Web search is just one – perhaps the most prominent – search context. There are many other application areas, enterprise search being one of them, which require fundamentally different solutions. Note that we do not simply want to reduce our discussion to a comparison of searching the Web with searching within an organisation – other types of search include Web site search, database search, and desktop search among others. Identifying the key features of each search type allows a systematic comparison and these features can be characterised by investigating a number of dimensions that reflect, for example:

- Generic properties characterising the document collection (e.g., scale of the collection, document formats)

[1] http://www.cs.rmit.edu.au/swirl12/

- Specific structural and organisational differences (e.g., link structure, internal document structure, distribution of documents across data silos)
- Individual document properties (e.g., time stamp, version number, metadata)
- Different types of information needs (e.g., navigational search *versus* attempting to find an expert in a particular subject)
- Differences between the target users (e.g., a heterogeneous set of Web searchers on the one hand and on the other a clearly defined user population whose members have different information needs, interests and access rights according to their role within the enterprise)
- The level of support needed (e.g., working out of the box *versus* requiring continuous support).

Unpacking these differences results in a fairly complex picture with equally complex implications. To choose just one way of contrasting enterprise search with related areas, we will see that while Web search is aimed at high *precision*, in enterprise search *recall* is often at least as important as precision.

Comparing and contrasting enterprise search with other search applications allows us to work out exactly what the fundamental features of enterprise search are and, following from that, what needs to be done in order to make enterprise search work. In short, we will identify the importance of 'putting the user in control', and present customisation and continous tuning as *essential* requirements for those wishing to maximise the value of their investment in a practical search solution – in other words to avoid failure.

The survey aims to be a thoroughly compiled resource and a primer for a range of interested readers as outlined earlier. References that offer a more general entry point into enterprise search include [Mukherjee and Mao, 2004] and [Hawking, 2010]. White approaches the problem from the perspective of a manager who has been put in charge of enterprise search [White, 2007, 2015b]. Any publication on enterprise search exemplifies a particular characteristic of the area, the fact that it is difficult to separate the fundamental technical challenges from organisational and pragmatic considerations.

1.2 Examples

Enterprise search tools must provide support for many more functions than simply indexing and query processing so that a basic search tool will not be enough [Hawking, 2010]. To illustrate this we present some introductory examples, all drawn from real applications in industry and academia. Each of these highlights a number of core issues that enterprise search has to deal with, including different types of data structures, access to data silos, the need for manual customisation, the application of domain-specific taxonomies, the varying user needs that need to be catered for etc. One could also argue that none of the examples might be seen as *traditional* enterprise search implementations which just shows what variety of problems need to be considered in any specific use case.

Some of the key concepts that we will encounter again are *emphasized.*

1.2.1 Reed

Reed Specialist Recruitment is part of Reed Global, which also includes Europe's biggest jobs Web site[2] receiving more than 1.5 million job applications per month.[3] Founded in 1960, Reed is a specialist provider of permanent, contract, temporary and outsourced recruitment solutions, and IT and HR consulting, with more than 3,000 permanent employees working out of 350 offices worldwide. As part of a major IT development programme the company moved to a new search framework, accessible to staff in local offices and centrally, that had to deal with a mix of document *repositories* of varying structure including a complex database consisting of millions of records (i.e. *structured* data)

[2] `http://www.reed.co.uk`

[3] `http://www.flax.co.uk/wp-content/uploads/2015/07/reed_case_study_oct2011.pdf`

as well as a database of CVs represented as flat files in a multiplicity of different formats such as Microsoft Word and PDF (i.e. *unstructured / semi-structured* content). Most of the files are in English, but other languages include Polish, Arabic and Chinese. The *open-source* Apache Lucene/Solr platform was chosen as the new search framework which provides *faceted search* and geospatial filtering and ranking based on complex business rules as well as custom *boost* options. A custom-built performance tester was built for *continuous monitoring* of the system's performance.

1.2.2 Australian National University

An example from academia is reported by Li et al. [2013] who investigate methods for *federated search* at the Australian National University. Information at this university is available in a variety of formats including *structured* databases such as a telephone directory, a course catalogue, and a library catalogue. Furthermore, *semi-structured* documents are sourced from more than 500 different Web servers. *Email lists*, *file shares*, local Web servers and other resources add further internal repositories. The university also makes use of external services such as Twitter, Facebook and YouTube. The individual sources *vary in size* ranging 'from quite small to more than a million documents', in subject matter, and in language among other characteristics. In this setting the creation of a central index is impossible due to the range of sources and *restrictions on access*. Instead, different repositories need to be accessed individually and then results *aggregated*. The authors conclude that this setting is not just realistic for the chosen institution but also for many others.

1.2.3 IBM

A number of studies looking at different aspects of search within IBM[4] have been published illustrating yet again the specific problems aris-

[4] Obviously, it needs to be appreciated that many of the studies being published by the research teams of large companies such as Microsoft and IBM might represent experimental applications and might never be the core engines underlying the organisation's enterprise search application.

ing in enterprise search. We will look into a number of these studies throughout the survey but here we only focus on one aspect which is people-focussed searches.[5]

Guy and colleagues investigate *expertise finding* as a central information need within an enterprise, i.e. finding people knowledgeable in a given topic [Guy et al., 2013]. They explore enterprise social media applications and what makes this another typical enterprise setting is the multitude of sources including blogs, wikis, forums, bookmarks, microblogs, communities, shared files, and people tags. Each of the different data sources turns out to cover a different fraction of the 400,000 employees within the organisation, ranging from around 20,000 to about 290,000 with the overlap among the individuals retrieved based on each application being very low so that each social media application tends to identify different people.

Another *people-searching* study using the internal tool 'Faces' demonstrates that enterprise people search should be considered a very important tool for the workforce in a large enterprise [Guy et al., 2012]. Faces goes beyond expertise search and offers searching the name, organisation unit, management chain, phone number, email, office location etc. A rapid adoption was reported within the organisation gaining tens of thousands of users per month.

1.2.4 GOV.UK

As a last example, we would like to introduce GOV.UK[6], the Web site of the UK government which offers a single access point to information and services for citizens and businesses, guidance for professionals as well as information on government and policy. This is an example of *site search* rather than *enterprise search* and it illustrates the point that

[5] We would like to refer to a concern raised by Treem and Leonardi who review the use of social media in organisations and who observed that a disproportionate number of studies referenced in their review are the result of research conducted at IBM and involving that organisation's employees simply because they are among the most active in publishing work on social media use in organisations [Treem and Leonardi, 2012]. We note the same is true for research published on enterprise search developments.

[6] `http://www.gov.uk`

different search areas, such as site search, Web search and enterprise search, share some properties but differ in others. GOV.UK indexes about 300,000 items of content and about 250,000 downloadable files[7] all driven by Elasticsearch[8]. These documents are reported to originate from 870 different organisations[9] covering 140 different *formats*[10]. Issues to be tackled include *duplicated* pages which might be identical or older *versions*[11]. The difficulty in finding the right information has led to the conclusion that a *taxonomy* covering the entire content of GOV.UK needs to be developed[12] and that *tagging*[13] content needs to be an integral part of the publishing process. The use of *'best bets'* makes sure that some fixed results will always be at the top of the result list for certain queries.[14]

This example demonstrates clear differences to Web search (e.g., the size of the collection, the use of hard-coded matching, control over the publishing process), and close similarity with many of the features observed in the three enterprise search examples. Nevertheless, in contrast to enterprise search, it is also worth pointing out that there was no mention of *email* search, or of *access control* issues. In addition to that, all the content on GOV.UK is actually intended for publishing rather than just being *deposited.*

[7] `https://insidegovuk.blog.gov.uk/2016/12/05/gov-uks-content-operating-model-whats-next-after-discovery/`

[8] `https://insidegovuk.blog.gov.uk/2014/06/13/how-gov-uk-site-search-works/`

[9] `https://insidegovuk.blog.gov.uk/2014/05/12/new-search-results-page-design-unified-search/`

[10] `https://insidegovuk.blog.gov.uk/2017/01/09/formats-and-templates-whats-the-difference/`

[11] `https://insidegovuk.blog.gov.uk/2013/06/12/duplicate-titles-in-site-search/`

[12] `https://insidegovuk.blog.gov.uk/2017/03/21 /presenting-our-new-taxonomy-beta/`

[13] `https://insidegovuk.blog.gov.uk/2017/04/18/making-tagging-part-of-publishing/`

[14] `https://insidegovuk.blog.gov.uk/2014/06/13/how-gov-uk-site-search-works/`

1.3 Perception and Reality

The heterogeneous structure and variety of formats of underlying data sources turns out to be a particularly prominent feature of enterprise search but there are other such features that make searching in an enterprise stand out. For example, a 'typical' non-enterprise search scenario might be characterised by a user trying to find a document that contains some relevant information, but a more common use case in an enterprise is the search for people who have the right expertise and a simple reason for that might be to avoid spending time and resources on work that has already been conducted within the organisation [Hertzum and Pejtersen, 2000].

With these motivating examples in mind let us step back a bit and look at the extent to which search and findability actually affect an everyday worker within an organisation. According to the most recent 'Enterprise Search and Findability Survey'[15], two thirds of responding organisations state that more than half of their employees depend upon good findability of information in their *daily work* [Findwise, 2016]. We conclude that enterprise search is not a *nice-to-have* but an essential requirement to work effectively within an enterprise context. Note that this need is in contrast to the perception of actual enterprise search users, as in the same survey almost half of the respondents expressed they are *dissatisfied* or *very dissatisfied* with existing search applications within their organisation. This discrepancy is also highlighted by another major enterprise search survey conducted by the Association for Information and Image Management (AIIM)[16] which found that while almost three quarters of organisations polled expressed that search is vital or essential, hardly more than ten percent actually have an enterprise search capability in place that allows search across the organisation, a number that is consistent across different sizes of organ-

[15] The *Enterprise Search and Findability Survey* is an annual survey of enterprises conducted by Findwise focussing on the state of play of search and findability within enterprises. While the overall objective is to observe trends across years the questions asked are not identical every year. This is also the reason why we reference three different surveys as they each provide insight into different aspects in addition to the overall picture.

[16] `http://www.aiim.org/`

isations [Miles, 2014]. Obviously, not much has changed then in more than 15 years [Feldman and Sherman, 2001].

1.4 Recent Developments

Despite this monograph approaching the topic from an academic perspective, we do want to offer a glimpse into the enterprise search market. What is remarkable is the rapid change in the enterprise search landscape in recent years. To illustrate the point, David Hawking's milestone publication [Hawking, 2010] lists a broad range of enterprise search software systems but hardly any of them are still available, most prominently Google's Search Appliance (GSA) is now being retired[17], FAST Search & Transfer has disappeared once acquired by Microsoft, Autonomy was taken over by HP, Vivisimo was acquired by IBM and so on. Companies like Funnelback[18] on the other hand have become more prominent providers of enterprise search solutions, and a number of new vendors such as Sinequa[19], Coveo[20] and Mindbreeze[21] have appeared. The biggest shift has however been the rise in open source solutions. Elasticsearch[22] and Apache Lucene/Solr[23], both based on the Apache Lucene[24] library, have developed into powerful tools that are widely applied. Bloomberg, for example, does not just deploy Apache Lucene/Solr in over 100 of its applications but the company also actively engages in the community by committing code.[25] Enterprise search is a core part of these applications.[26] More broadly speaking, the deployment of open source code has become mainstream. For example, in an attempt to create a level playing field between pro-

[17] http://fortune.com/2016/02/04/google-ends-search-appliance/

[18] https://www.funnelback.com

[19] https://www.sinequa.com

[20] http://www.coveo.com

[21] https://www.mindbreeze.com

[22] https://www.elastic.co/products/elasticsearch

[23] http://lucene.apache.org/solr/

[24] http://lucene.apache.org

[25] http://www.bloomberg.com/company/announcements/open-source-at-bloomberg-expanding-our-engagement-with-solr/

[26] http://www.bloomberg.com/company/announcements/open-source-bloomberg-solr-work-enhance-enterprise-search/

prietary and open source software, the UK Government IT strategy[27] explicitly states that government will procure open source solutions where appropriate given that “open source presents significant opportunities for the design and delivery of interoperable solutions” [Cabinet Office, 2011].

1.5 Outline

As part of plotting the landscape we will first look at the changing face of search in Chapter 2 before defining enterprise search and then contextualising it with many other common search applications, such as general Web search and more specialised applications like patent search. This analysis should offer useful insights into the different types of search areas and goes beyond enterprise search. As such the mapping of the search landscape into some form of ‘feature vector of search applications’ should be a self-contained chapter which can be used as an easy reference and overview of where enterprise search fits within the bigger picture.

The second part will be dedicated entirely to enterprise search. We decided to split the discussion into four main chapters.

Chapter 3 starts by providing a systematic overview of what defines enterprise search. We drill down into the actual characteristics by adopting a topical structure that should allow the reader to easily refer back to the discussion that contextualises enterprise search within the broader world of other search applications in Chapter 2. This chapter is meant to be a survey of the academic literature providing a solid account of the state of the art in the field while also highlighting issues around enterprise search that need to be addressed properly in order to make it work successfully.

Chapter 4 is devoted to the discussion of how to evaluate enterprise search. We present commonly applied metrics and evaluation approaches and contrast them with other areas of search. In line with the previous chapter, we will again highlight the fundamental difficulties emerging from evaluating enterprise search applications.

[27] https://www.gov.uk/service-manual/making-software/open-source.html

Chapter 5 then picks up the issues and difficulties identified in Chapters 3 and 4 in a less theoretical and more practical discussion of what can and needs to be done to fully exploit the potential of enterprise search. This is the chapter that we expect to be of most interest to the practitioners among our readers while at the same time offering our academic readership an understanding of why enterprise search so often fails to perform.

Chapter 6 will look at current and future developments in this exciting application area and also identify a number of research directions that are emerging from the survey. We conclude in Chapter 7.

2

Plotting the Landscape

2.1 The Changing Face of Search

It seems like a long, long time ago that Vannevar Bush developed his ideas of how to make collective knowledge automatically accessible based around a device he called a 'memex' in which "an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility", yet when he talks about how one might use this framework to explore, for example, the differences between the short Turkish bow and the English long bow one gets a feel that he is actually talking about the construction of knowledge graphs as they are now applied in any modern Web search engine [Bush, 1945]. It is fascinating how the field of 'search' has changed so dramatically over what is really a fairly short period of time compared to other scientific disciplines or research areas.

For decades search applications followed the Cranfield approach of single-shot queries resulting in a list of documents that could then be assessed for their relevance to the original query [Cleverdon, 1997][1]. Pushed by the emergence of the World Wide Web the field has only

[1] This is a more readily available reprint of the original paper published in 1967.

fairly recently diversified, almost beyond recognition, with more sophisticated approaches to search emerging so quickly that we take many features for granted which have not actually been around for long, e.g. even knowledge graphs have only found their way into mainstream search engines a few years ago, speech-driven search and autocomplete are other such examples.

Enterprise search is just one of many search areas that reflects these changes. However, despite rapid developments in search engine technology there is a mismatch with the actual application of this technology in an enterprise context on the ground with a large proportion of end users still being dissatisfied with existing search applications, most notably in larger organisations of 1,000 or more employees [Findwise, 2014]. To understand this discrepancy we will need to drill down to the underlying issues.

2.2 Defining Enterprise Search

Enterprise search technology goes back a long time – relative to Web search – all the way to the 1960s when it became necessary to search large databases of scientific, commercial and legal information and to provide support for legal teams involved in some large anti-trust suits [White and Nikolov, 2013],[Bourne and Hahn, 2003, p.101-139]. Despite this relatively long history it remains hard to properly define enterprise search. The different factors that play a role make it difficult to come up with a solid, comprehensive definition. Hawking keeps the scope quite broad by referring to enterprise search as the "application of information retrieval technology to information finding within organizations" [Hawking, 2010] and then elaborates that this may be interpreted as search over digital textual content owned by an organisation such as the external Web site, the company intranet, as well as emails, database records and shared documents. The AIIM defines it as the "practice of identifying and enabling specific content across the enterprise to be indexed, searched, and displayed to authorized users".[2] Here is yet another definition provided by White [2015b]:

[2] http://www.aiim.org/What-is-Enterprise-Search

> Enterprise search is a managed search environment that enables employees to find information they can rely on in making decisions that will achieve organizational and personal objectives.

While it is by no means intended to be capturing all aspects of enterprise search it does highlight a number of key concepts that appear to be core to enterprise search including *employees*, *making decisions*, *organizational objectives*, and *personal objectives*. The picture that emerges is that of an area that is deeply rooted in business processes and requirements which excludes, for example, typical Web search scenarios such as searching/browsing for entertainment.[3]

However, we do go a step further and broaden the scope of what we consider to be included in enterprise search. First of all, we do want to *include* search over an organisation's *external Web site* since this may be critical to the business of the enterprise. A lot of information for employees might only be published there. Examples of external-facing information useful internally include course and scholarship information in a university, job vacancy information for a government agency, and published terms, conditions and rates for a bank. As an implication of widening the scope we also include all the users of the enterprise search system and not just employees, e.g. students in a university even though they may only have access to limited subsets of the university's internal repositories, and customers that search for documentation on an organisation's Web site, e.g. for product manuals. We furthermore, include intents that are not directly related to making decisions, e.g. to collect background information.

Figure 2.1 sketches a typical enterprise search architecture that highlights some specific characteristics including the heterogeneity of data sources, the central role that security (via access control) plays, the need for customisation and the overall fairly complex setup. The architecture we adopt here makes the assumption that enterprise search is a single application, and some argue that the aim is indeed to have a single interface that allows users to access all available repositories resulting in an aggregated list of results. e.g. [van der Lans, 2013, Grefen-

[3] Even in a social media platform embedded in an enterprise it was found that information needs exceed entertainment and social factors [Guy et al., 2016].

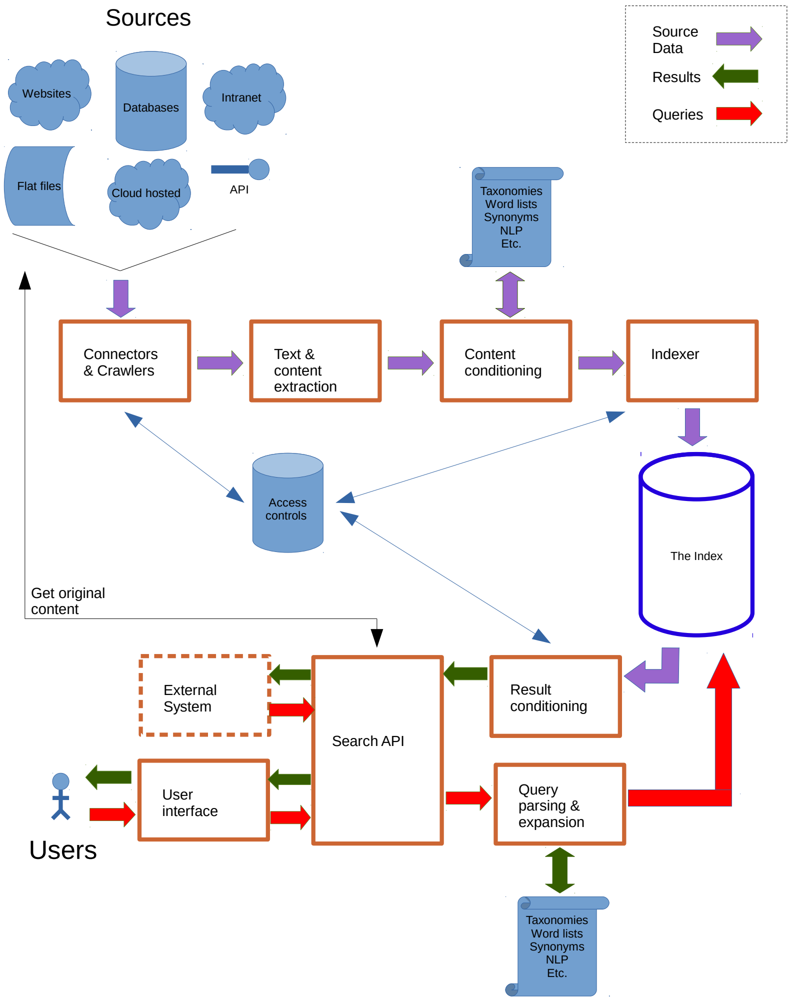

Figure 2.1: Typical enterprise search architecture

stette and Wilber, 2010], but this architecture also runs some high risks which is partly the reason why White stresses the flexibility of enterprise search when defining it as a *managed search environment* that can easily contain a multitude of individual search applications as long as it is guaranteed that the employee does find the information he or she needs [White, 2015a].[4]

If we go beyond the purely technical architecture and identify those parts where the system needs to be managed and maintained and where content is being created or needs to be managed (see Figure 2.2), the picture gets even more complex. One of the observations to pick out here is the fact that there are *many* different roles and the interplay between these different people is critical to the success of an enterprise search application. We will not discuss these roles here but will get back to three of them – *admin team*, *domain experts* and *search developers* – in Section 2.5.6 when comparing what support is needed in enterprise search compared with other search application areas.

We will now compare and contrast enterprise search with other types of search by shortly introducing the different areas with the view of contextualising them before providing some concluding remarks.

2.3 Related Search Areas and Applications

It will have become quite clear by now that enterprise search is neither a niche area nor does it come with a handful of specific features that allows us to clearly separate such applications from other types of search. It is much more an area that picks and chooses approaches and then brings them together in a rather heterogeneous framework. In addition to that, there is a lot of overlap and fuzzy boundaries with other areas. In order to define the overlaps and draw the boundaries (no matter how fuzzy they are) and to uncover the distinct characteristics of enterprise search we will first provide a broad review of related search areas. We will look at the academic literature but also keep in mind a practitioner's angle.

[4] Obviously, success can never really be 'guaranteed' as if there are twenty different search interfaces, an employee might never search all of them and may often fail to search the repository that actually has the answer.

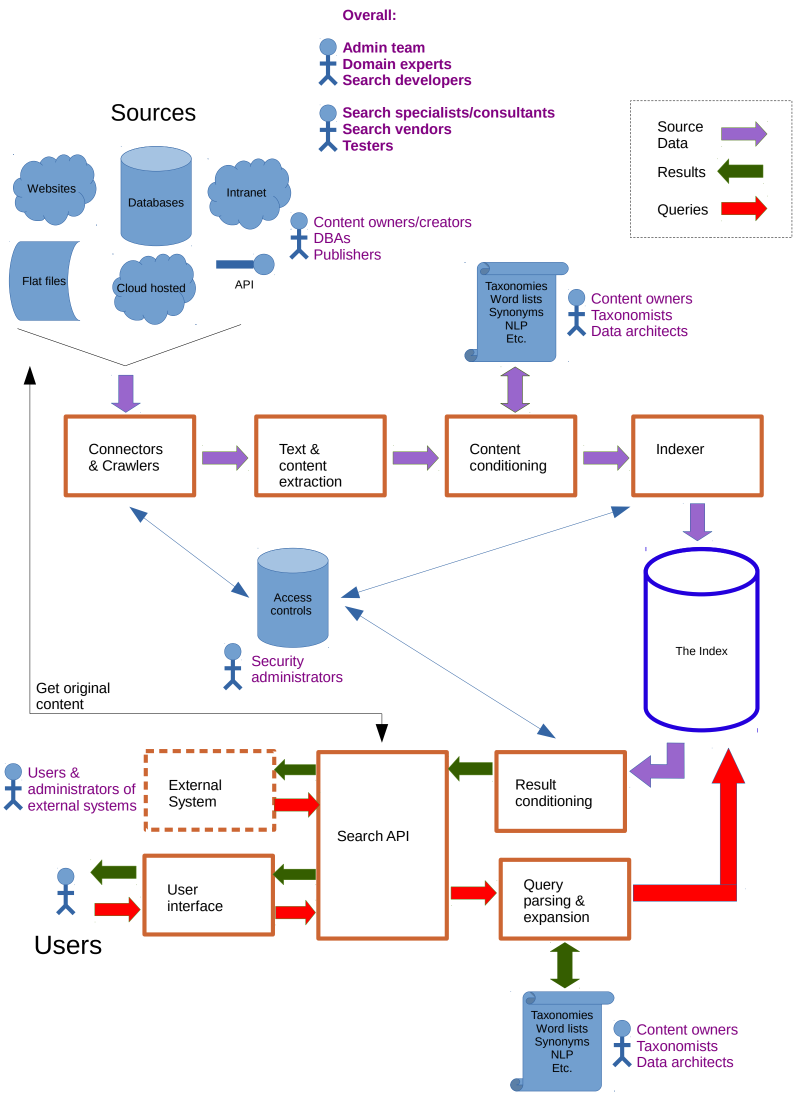

Figure 2.2: Adding people to a typical enterprise search architecture

There is obviously more than one way of classifying the related work. Ideally one would pick a specific dimension, such as the type of data or specific use cases at hand, and compare and contrast the research areas as defined by different instances of this dimension, but given the heterogeneity of the field we will be slightly less formal. We will first review any search areas and applications that are closely related to enterprise search be it that they are defined by *what* is being searched, e.g. Web search, desktop search and database search, or what the actual *applications* are, such as expertise search and e-discovery. We will then look at two specific search *techniques* that are essential building blocks of any modern enterprise search architecture but are also commonly applied elsewhere – faceted search and federated search.

Following on from this discussion we will extract a broad range of features along different dimensions that will allow us to more formally work out what enterprise search shares with other fields and how it differs. We will use tabular representations to do that.

In order to keep the monograph focussed we decided to exclude altogether a detailed discussion of any research areas that are not primarily concerned with search. This includes some fundamental textual processing steps that search applications are commonly based on. Much progress has recently been made in natural language processing, topic modelling and content analysis, for example, and text mining techniques have found their way into mainstream enterprise applications [Upshall, 2014]. Zhai and Massung actually see the two fields of information retrieval and text mining as the two core ingredients of a *text information system* which includes not just data access but also categorisation, clustering, topic modelling, summarisation and other types of analysis [Zhai and Massung, 2016].

2.3.1 Web Search

Information retrieval has been around for more than 50 years but was, despite its maturity as a research field, for decades a narrow area of interest restricted primarily to librarians and information experts, a situation that changed almost overnight with the invention of the Web [Berners-Lee, 1989]. Now searching the Web is by far the most common application involving information retrieval [Croft et al., 2010, p. 3]. The main trigger for the Web's popularity was however not just the provision of the fundamental infrastructure but the *freedom to publish*, i.e. the fact that suddenly anybody could publish their ideas and reach millions. This was a paradigm shift and can be seen as the birth of a new era in information processing and retrieval – the *e-publishing era* [Baeza-Yates and Ribeiro-Neto, 2010, p. 3].

In sharp contrast to this characterisation of the Web the 'freedom to publish' does not typically apply in an enterprise search context where data is at least in parts curated, vetted and controlled (though contents of file shares and emails are two common cases for which this is not true). On the other hand, Web search faces some problems that are similar to the problems enterprise search applications have to deal with. Examples include distributed data, data duplication (up to 30% of Web pages are reported to be duplicates or near duplicates), unstructured data, and heterogeneous data; while volume of data and data quality point at more Web-specific problems [Baeza-Yates and Ribeiro-Neto, 2010, p. 449]. We will discuss these underlying problems in more detail later once we have plotted the search landscape that will allow us to contextualise enterprise search in respect to related areas.

Web search can be a fallback strategy to find content not found on Web sites, in intranets or in the enterprise if users are faced with a poor information architecture, e.g. [Lund and Ørnager, 2016]. Many who have tried to find content on an intranet or enterprise system that has a public-facing component will have tried to find the *local* content through a Web search engine and often this might even have been quicker and more accurate in finding the information.

2.3.2 Site Search and Intranet Search

Moving from Web search to Web site or intranet search might appear like a simple scaling issue but it is a much more fundamental shift and we are actually moving slowly into enterprise search territory.[5]

[5] We consider Web site search and intranet search as closely related, in fact intranet search can be seen as a natural extension of site search in that intranet search tends to access more than just the publicly available content that represents

Going back a few years, Fagin and colleagues at IBM summarised the key differences between an *intranet* and the *internet* as a set of axioms, namely (1) intranet documents are often created for pure information dissemination as opposed to attract the attention of users;[6] (2) many queries have a small set of correct answers – often only a single one; (3) intranets are spam-free; and (4) large portions of intranets are not search-engine-friendly [Fagin et al., 2003]. This is still largely true, and these differences have implications on how users search but also – perhaps more importantly – what algorithms that are commonly employed in Web search work well and which ones do not.

The boundaries between site search, intranet search and enterprise search are often fuzzy, but as a rule of thumb one could assume that site search usually indexes *HTML* documents whereas intranet and enterprise search usually index a wider range of diverse document formats and repositories.[7] This means that algorithms that work well for site search are based on similar techniques as applied in *Web* search such as anchor text mining for query refinement [Kraft and Zien, 2004] while at the same time relying on manual customisation such as 'best bets' [Morville and Callander, 2010, p.89] or 'suggested matches' [Bao et al., 2012a], all of which are hard-coded links to map common/important queries to specific matches. We have already come across a site search example that matches this description, the UK government site GOV.UK as discussed in Section 1.2.4.

Kraft and Zien, experimenting on the IBM intranet, refer to their setup as a "large corporate intranet comprising a document corpus of 4 million unique HTML documents" [Kraft and Zien, 2004].[8] Now, this

a Web site. However, we do note that intranet repositories are typically larger, less organised and more varied than many corporate sites [Pernice et al., 2007, p.11]

[6] We will expand on this in Section 3.2.4 where we will argue that often new documents might just be created and deposited without any intention to disseminate.

[7] Do note however that there are also plenty of examples of site search that involve much more than HTML documents, for example amazon.com (searching a product catalogue) or GitHub (searching code, user profiles etc). The first example could in fact be treated as a specific type of search in its own right – e-commerce search.

[8] Fagin and colleagues identified about 50 million unique URLs when crawling the IBM intranet but reduced this to 4.6 million by removing database queries, applying de-duplication etc. [Fagin et al., 2003].

has certainly changed in recent years and intranets are, unlike Web sites, no longer a collection of HTML documents. Instead, a typical intranet contains technical reports, white papers, spreadsheets, presentations, marketing material, online forms etc [Morville and Callander, 2010, p.37]. Hence, PDF and Office documents tend to represent a much larger share of the content than static HTML documents. If we now also consider structured content, e.g. by adding search in an employee directory, we will encounter all the typical enterprise search dimensions such as search in data silos, research aggregation, presentation issues, domain-specific rules, etc. [Vaithyanathan, 2011].

In terms of general setup and challenges, we might conclude that site search and intranet search are somehow specific instantiations of enterprise search in that they share similar problems like sparsity at various levels, and which require similar solutions to work well (like 'best bets'). Enterprise search however adds a few more challenges as we will elaborate on a bit later.

A specific aspect that makes site, intranet and enterprise search different from Web search is the fact that queries tend to be even shorter overall. This is true for people search, e.g. a median of only seven characters for each query submitted to the IBM enterprise people search engine *Faces* [Guy et al., 2012]. It can also be observed in general site search query logs. Web queries tend to be be growing in length from around 2.35 words long on average, e.g. [Silverstein et al., 1998, Jansen et al., 1998, Beitzel et al., 2004, 2007] to more than 3 words [Taghavi et al., 2012], a trend that is likely to continue given the growing share of speech-driven input Web search engines receive. Queries submitted on intranets and to site search engines on the other hand are shorter, e.g. 1.8 words on average for queries submitted to the intranet of the UN's World Food Programme [Lund and Ørnager, 2016], and 1.4 terms according to the query logs of the corporate intranet of SwedCorp, a large multinational manufacturing company with over 60,000 employees [Stenmark, 2005b]. This figure has been shown to be consistent across different years [Stenmark and Jadaan, 2006]. A study of site search of a university Web site observed an average length of 1.81 query terms [Kruschwitz et al., 2013], and here as well we can see that this result is consistent with a study conducted about ten years earlier on the same site (average query length of 1.72) [Kruschwitz, 2003].[9]

The same appears to be true for sessions. The logs analyzed by Kruschwitz et al. [2013] contain on average 1.53 queries per session, compared to 2.02 queries per session on average on an *AltaVista* log [Silverstein et al., 1998], 2.8 for an *Excite* log [Jansen et al., 1998] and 2.31 as the average number of queries per session submitted to the meta-search engine *Dogpile* (using a comparable session definition) [Jansen et al., 2007]. Interestingly, another study of a local Web site has also come to the conclusion that sessions are shorter than general Web search sessions: on average 1.73 queries were submitted per session to the Utah government Web site [Chau et al., 2005]. Shorter sessions than what is typical for Web search have also been observed when analysing the logs of a people search engine with 78% of all sessions consisting of a single query and an overall average of 1.64 queries per session [Weerkamp et al., 2011].

In line with Fagin *et al.*'s axioms mentioned earlier, there is agreement that Web sites and intranets can be difficult to navigate, partly due to their static and possibly idiosyncratic organisation, e.g. [Karim et al., 2009, Berendt and Spiliopoulou, 2000]. Supporting this finding, a longitudinal study of retrieval technology on a fair number of German and Swiss *external* Web sites assessed according to more than 70 individual tests addressing result quality, index quality, user interface and user interaction found that many Web sites do not meet many of the requirements set out. On top of that no substantial improvement was found over the years [Mandl et al., 2015].

Stepping back a little bit from the academic and technical side and looking at the practical issues, White concludes that the skills needed to support Web site search are the same as those for internal enterprise search [White, 2015b].

[9] Unlike Web search queries that tend to get longer, enterprise search queries can be expected to get even shorter with the growing search support via filters and facets, but only if we do *not* treat filters and facet specifications as part of the actual query. We should also note that the (now) wide-spead use of auto-completion suggestions will almost certainly have an effect on the length of submitted queries.

2.3.3 Desktop Search / Personal Information Search

A further step to scale down the document collection is to move from Web search via Web site search to desktop search (also referred to as personal information access) which is understood to be the process of searching within a personal space, e.g. a desktop, which includes search in word processing documents, emails, visited Web pages etc [Elsweiler et al., 2010]. Another view is to treat desktop search as the "personal version of enterprise search" [Croft et al., 2010, p. 3]. There are indeed a lot of similarities between the two, e.g. access to heterogeneous data collections, and recall being possibly more important than precision in a typical search scenario. However, it just offers a simplified view as key enterprise search concepts like user roles and document-level security are largely absent from desktop search.

An interesting observation related to desktop search is that the user context is readily available, represented, for example, by location, activity, time, season, emotional state etc. [Dmitriev et al., 2010]. The fact that contextual information is readily available together with the client-side approach of desktop search allows a much richer and more cohesive search experience as the application context can be used to model the current search situation in some detail [White, 2016, p. 271].

Algorithms applied to personal information search will naturally focus on individual users and therefore address search patterns expected from such users, e.g., refinding documents the user is already aware of as proposed by *Stuff I've Seen* (SIS) [Dumais et al., 2003]. The approach provides a unified index to documents visited on the Web, files on the desktop, in email folders, etc. A study of 234 employees within a large company (Microsoft) resulted in users accessing alternative search tools less frequently after installing SIS. It also demonstrated that filters such as date and person name provide important cues. Users indicated in the post-installation questionnaire that SIS-like search services should be an essential functionality on any computer.

An extension of desktop search that also incorporates external sources commonly accessed by a searcher is 'personal metasearch' [Thomas and Hawking, 2008] or, framed within an enterprise context, 'workplace search' [Hawking, 2010].

Note that important business information as well as email communications are very often hidden away on personal or shared drives rather than added to business content management systems, e.g. when files stored locally are not indexed by the enterprise search system [Feldman and Sherman, 2001] or when users store information in their own private region of a common repository [Massey et al., 2014]. For *personal* information access this might not be a major problem given that users appear to still prefer navigation over search when accessing personal information management systems with search only being used as a last resort when users cannot remember the file location [Bergman et al., 2008]. However, as soon as documents need to be found by other employees this strategy is likely to fail.

Looking at the big picture though, desktop search *within an enterprise* is becoming less relevant as personal drives disappear, but it certainly remains an interesting field for personal information management.

2.3.4 Database Search

The related search areas discussed so far are to a large extent concerned with data collections that contain textual content in a largely unstructured format which need to be gathered/crawled and pre-processed before they can be searched. As a result of this, effective index structures have been developed for Web search which are very different from the record and table structures in relational databases [Baeza-Yates and Ribeiro-Neto, 2010]. Enterprise content however comes in a variety of formats, much of it fully structured such as employee databases, customer databases or databases holding product information. For such structured records relational database management (RDBM) technology has for decades been the backbone of any information system in an enterprise.[10]

[10] We will not dive into database technology as there are many textbooks that have covered this field extensively. We should also acknowledge that without a precise definition of ‘database search’ we are opening a can of worms. For simplicity reasons let us assume that issuing SQL queries is what we broadly understand here, but there are obviously many alternatives such as full-text search on all records, searching entities or searching metadata.

Given that enterprise search applications need to access both unstructured and structured content, RDBM systems form a very usual and important source of data in such a setting. We are particularly interested in how the gap between the two approaches is being bridged.

Two noticeable developments away from the strict RDBM framework have started to make a significant impact in database technology. One of them is the rise of various flavours of NoSQL[11] databases – scalable, typically distributed database systems that do not strictly follow the relational model. They emerged with the need to scale beyond the limitations that a relational database imposes, they work well with unstructured data and it turns out that databases do not always need all the features that a relational database offers [Leavitt, 2010]. While principles like ACID (Atomicity, Consistency, Isolation, Durability) will have to remain a core feature in applications that work with transactions (e.g., banks and stock markets), there are many use cases which work well without imposing these constraints [Pokorný, 2013]. Search (in cases where it is 'non-mission-critical') is certainly a candidate for the latter type.

The second development is that of *search-based applications* (SBAs) which are software applications that build on a search engine backbone to access the content of traditional databases [Grefenstette and Wilber, 2010]. The motivation is very similar to NoSQL databases which includes the need to scale, the ease of access, the rapid response time, and the observation that much of the complexity of a traditional RDBM system might not be needed for many applications. In SBAs database content is *offloaded* into a format that can be indexed and processed like search engine content[12] with the additional advantage of the presence of semantic structure derived from the database(s). Unlike NoSQL systems that might simply serve as the backend to a search engine, SBAs aim at domain-oriented tasks like decision intelligence and discovery.

Database technology is also commonly being applied to help manage the search architecture. Fagin and colleagues, for example, devel-

[11] Meaning "Not only SQL" or "Not Relational" [Cattell, 2010].

[12] Including basic natural language processing steps such as language detection, tokenisation and part-of-speech tagging but also more sophisticated semantic analysis like named entity recognition, sentiment analysis and event extraction.

oped the idea of a *search database system* (SDBS) to assist in interpreting and manipulating a user query [Fagin et al., 2010]. A keyword-based query is mapped into a database of structured content based around concepts and schemas that connect these concepts, database instances over a schema then link concepts to actual resources such as Web pages, documents, email messages etc. One way of embedding a SDBS in the search framework is to map the query to the most specific parse of that input [Fagin et al., 2010]. Alternatively, this could be employed to provide search administrators and domain experts with a way to specify and customise rewrite rules that transform user queries into revised interpretations [Fagin et al., 2011]. Fagin *et al.* found this to be a powerful and effective mechanism in the hand of administrators in an intranet search context, and it maps directly into the typical record structure of a RDMS.

We should also flag up research that has been going on for decades aimed at natural-language interfaces to databases [Popescu et al., 2003]. This is less commonly applied in an enterprise search setting.

2.3.5 Digital Library Search

Digital libraries are more than just an electronic version of a library providing access to books, journals and multimedia content. They represent very complex information systems having to address search in distributed, heterogeneous collections, accessing documents of varying structure and adhering to access security constraints [Fox and Sornil, 2003]. All this hints at some striking similarities to enterprise search although the typical content of a digital library catalogue lends itself much more naturally to some *domain-independent* classification and annotation such as the Dublin CORE elements describing an electronic resource by properties like the title, the author, the publisher, the date associated with the resource etc. [Weibel et al., 1998].[13] At first sight, a digital library therefore appears to offer a more consis-

[13] This does not mean that a generic standard like Dublin CORE is not applicable to enterprise settings. It has for example been used as a basis of a new metadata standard for NASA's Jet Propulsion Laboratory (JPL) that is content-neutral and application-neutral and which can be consistently applied to JPL's hundreds of active repositories [Mathieu, 2017]

tent structure similar to records in a RDBM system when compared to enterprise search repositories which tend to be less homogeneous and more domain-specific. However, going beyond the structured metadata of items in the digital library, the content itself might well be unstructured (just a PDF document, for example).

Access control is important for both application areas but in the case of digital libraries this appears to be largely driven by copyright issues whereas in an organisation access control is defined according to an employee's role.

An interesting problem that digital libraries and enterprise search share is the need of multi-disciplinary expertise. Designers of digital libraries are often library technical staff with no formal training in software engineering, or computer scientists without a background in information retrieval which ultimately results in implementations that do not appropriately acknowledge the state of the art in the different disciplines [Gonçalves et al., 2004]. Similarly, in enterprise search we observe that search is typically managed by administrators who are domain experts but not search experts and the question arises as to how to incorporate both types of expertise appropriately [Bao et al., 2012a,b].

2.3.6 E-discovery

E-discovery is closely linked to enterprise search and is an area of high financial impact [Hawking, 2010]. It describes "the process by which one party (for example, the plaintiff) is entitled to 'discover' evidence in the form of 'electronically stored information' that is held by another party (for example, the defendant), and that is relevant to some matter that is the subject of civil litigation (that is, what is commonly called a 'lawsuit')" [Oard and Webber, 2013]. The scale of such processes becomes apparent when considering that even early systems from the late 1960s that dealt with litigation support activity already had access to millions of documents as was the case for IBM's TEXT-PAC system [Bourne and Hahn, 2003, p.126-130].

Such processes may also be carried out internally (auditing) to preempt any such discovery or to be assured that certain regulations or laws are being adhered to. Internal auditing might include the identification of non-compliant behaviour, confidentiality breaches or fraud. E-discovery and auditing can be summarised as *discovery* which forms a major component of an enterprise search system according to the AIIM Enterprise Search survey with 50% of respondents reporting that they deal with internal compliance audits and 44% with pre-trial legal discovery [Miles, 2014].

This is a growing field and some of the characteristics are:

- The importance of both precision (to reduce unnecessary review costs) *and* recall (all evidence needs to be discovered), on balance however recall is the more important among the two metrics[14]
- A human in the loop who assesses the documents
- Potentially a large number of relevant documents, e.g. in the TREC 2008 Legal track topics were developed to approximate realistic e-discovery cases and in some such topics the number of relevant document in the collection was higher than 100,000 and even the number of documents judged as 'highly relevant' was 11,542 *on average* per topic [Oard et al., 2009].

E-discovery is therefore particularly interesting as it can first of all not really be separated from enterprise search and secondly, it further accentuates some of the problems of enterprise search. It does however apply mainly to larger organisations and within this scope primarily to law firms and large legal departments.

The concept of *provenance* is another example of discovery. Provenance is concerned with tracking the history of an artefact or the process by which it was created, all of which can also be applied to digital data. In fact, provenance systems were historically the focus of research in the database community [Carata et al., 2014]. Now, consider that in the construction industry, for a large number of building projects in

[14] Total recall is more like an idealised view as typically there is a trade-off between recall and effort [Grossman et al., 2016], and an acceptable recall level for e-discovery and other technology-assisted review applications might be in the region of 70-75% [Blair and Maron, 1985, Cormack and Grossman, 2016].

the United Kingdom the construction company has to hand over documents to the client once the project is completed enabling search and reuse of the digital assets [Khan et al., 2016]. Given that these documents could easily be in the region of tens or hundreds of thousands, the close similarity with e-discovery scenarios becomes apparent.

2.3.7 Patent Search

Patent search [Lupu and Hanbury, 2013] is another classical example of recall-focussed search and with more than one million patent applications submitted each year [Alberts et al., 2011] it is of huge economic importance. It is one particular type of 'professional search', i.e. search that is conducted on a professional, paid basis [Tait, 2014]. Patent search also features prominently in *some* enterprise search settings [Hawking, 2010]. Patent search comes in different flavours but typically follows a very rigid, structured approach making use of hierarchical classification structures like the International Patent Classification (IPC) scheme [Gomez and Moens, 2014]. It shares this reliance on structured knowledge with a typical enterprise search setting. In addition, it can be seen as an extreme case of recall-oriented IR (comparable to e-discovery) and hence provides a close link to enterprise search in general.

2.3.8 Expert and Expertise Search

Trying to find an expert on a particular topic is an everyday task and of particular importance within an enterprise context. However, standard search engines are not ideal as they return documents and not people and existing enterprise search solutions are reported to be very poor at helping with expert and expertise search, complicated by the fact that 'expertise' is a loosely defined concept [Balog et al., 2012, p. 2-3].

Expertise seeking in an organisation setting as the activity of selecting people as sources for consultation about an information need is a very common activity, and people are commonly ranked higher than other information sources, with workgroup colleagues and other nearby co-workers being the most frequent sources [Hertzum, 2014].

We should note that what seems to be a very narrowly defined task offers quite a broad range of approaches. To pick a more extreme example, the search of recruitment professionals (to find people with the expertise required to fit a specific job description) is characterised by a strong reliance on Boolean operators, lengthy search sessions, different notions of relevance compared to Web search and the use of specific domain knowledge [Russell-Rose and Chamberlain, 2016].

It turns out that expert and expertise search requests cover a fair proportion of information needs found within an enterprise and this warrants a more in-depth discussion later on.

2.3.9 Social Media Search

The massive growth of user-generated content and the formation of communities that share such content has triggered the need for new search algorithms which address social-media-specific concerns such as vulnerability to spam[15], short lifespan of articles and locality of interest [Santos et al., 2012], issues that make social media search interesting but also a challenge. Obviously, social media search can mean different things. For example, searching public tweets for news is different to organisations searching for customer feedback on public tweets, employees searching internal forums or Slack[16] channels for information etc. All this would be considered types of social media. Despite the potential of exploiting social media within enterprises there is still relatively little uptake of it which is why we only provide a short discussion of this rather multi-faceted field here.

User-assigned tags and search within communities are central to social media search [Croft et al., 2010, p. 401-412]. Obviously, there is a major difference to enterprise search where user-generated content is typically limited and vocabularies to assign any metadata are controlled centrally within an enterprise. A further point to note is the fact that social media communities tend to be joined by choice but

[15] Due to the significance and scale of various types of spam which are constantly being adapted to recent trends in search technology an entire research area referred to as adversarial information retrieval has emerged [Castillo and Davison, 2011].

[16] `https://slack.com/`

the place within an enterprise 'community' is defined by the role(s) the user is assigned to. With this background mass participation cannot simply be taken for granted [Han et al., 2015]. However, if social media is actually adopted by a large user base, as reported for IBM's *Dogear* social bookmarking service, then significant improvements in terms of precision can be obtained over a standard enterprise search engine by incorporating user annotations and page popularity [Amitay et al., 2009].

There is certainly potential in adopting ideas of social media search for the enterprise considering the difficulty of finding the right information and the wide-spead tacit knowledge in organisations [Treem and Leonardi, 2012, Mukherjee and Mao, 2004, Baumard, 1999] that can potentially be utilised in collaborative search efforts, and social media is gaining popularity within the enterprise [Guy et al., 2013]. In fact, social media adoption within organisations, via blogs, wikis, social tagging and microblogging, is occurring at a rapid pace [Treem and Leonardi, 2012] and can attract wide use among employees, e.g., [Mark et al., 2014]. However, it has also been argued that 'social intranets' will only work if there is a need for collaboration and that ultimately most enterprise social networking platforms fail [Mergel, 2016]. Management support and training are identified as crucial factors to avoid such problems.

Despite some rapid changes we still have to conclude that so far *search* in an enterprise still remains largely a solitary exercise [White, 2015b].

2.3.10 Other Areas of Search

There are many other areas of search that aim to address one or more core problems also found in enterprise search. We only provide a glimpse into some of these areas.

Personal lifelogging data presents an interesting mix of structured and unstructured information. On the one hand there is a lot of metadata such as time and location for any recorded item (such as an image) but on the other hand this generates huge archives of personal data (e.g. images taken in frequent intervals in the order of minutes or seconds over long periods of time), data with no manual annotations, no semantic descriptions, often raw sensor data [Gurrin et al., 2014]. For example, at the time the article appeared, one of the authors (Gurrin) had accumulated 14 million automatically-captured images of life-experience, along with time-aligned sensor data. All this shows some resemblance with large portions of enterprise data, e.g. scanned images and drawings which have no explicit internal structure but come with some metadata annotations. The challenges of turning unstructured data into semantically interpretable data are similar.

Trying to link a *medical record* to the relevant academic biomedical literature or to past cases has long been identified as an interesting search application [Frisse, 1988]. Looking at it through our enterprise search eyes this use case has similarities with trying to find the right expertise on a particular project an employee is working on [Hertzum and Pejtersen, 2000]. The TREC Medical Records Track framed this problem as trying to enable semantic access to the free-text fields of electronic health records [Voorhees, 2013]. These records do have free text but also have annotations such as discharge diagnosis codes, i.e. not that dissimilar to domain-specific taxonomies in enterprise settings. There is more similarity: the track organisers conclude that the language used within health records is sufficiently *different* from general use to warrant domain-specific processing.

The move away from returning documents to returning answers is the idea of *question-answering* (QA) systems and while QA systems have a fairly long history [Prager, 2007] they have only recently attracted much wider attention not least through the development and publicity of IBM Watson [Ferrucci et al., 2010]. In addition to that, state-of-the-art search engines now include QA features to answer large fractions of queries. For example, given the query "How many member states are there in the EU", both Google and Bing return structured output containing the correct result ahead of any documents *containing* the answer. The application of QA system within an enterprise beyond the identification of employees' contact details could offer real benefit.

Recent research has also looked at trying to predict search intent of queries submitted to a site search engine (internal search) based on referral information coming from a Web search engine (external search) [Ortiz-Cordova et al., 2015]. The authors investigate *external-to-internal search* but also acknowledge that users might leave the internal site, then go to an external search engine to locate specific content on the internal site and then return, a scenario that is very applicable to enterprise search if this includes search over publically available site content.

2.4 Search Techniques

Having reviewed different search areas and applications we also want to introduce two search *techniques* that have emerged as dominant search paradigms in enterprise search as well as in Web search and elsewhere. These techniques exploit document properties (faceted search) or the distributed nature of data sources (federated search), respectively.

2.4.1 Faceted Search

Faceted search (or guided navigation) relies on data that is partially structured [Tunkelang, 2009]. This structure is realised by document *facets* – orthogonal sets of categories – examples of which could be the document type, the date of creation, the language used in the document, other possible facets are derived from classifications like domain-specific taxonomies. Faceted search was first proposed as a way to search and explore large collections of images as a way to overcome limitations of the then dominant keywords-based and similarity-based interface types [Yee et al., 2003]. Faceted navigation has (in 2010) been described as the most significant search innovation of the past decade [Morville and Callander, 2010, p.95].

Given that enterprise search tends to access domain-specific data from heterogeneous data silos with a range of inherent structure faceted search appears to be a particularly good fit to access the data. While faceted search has become the main tool to find known items within an enterprise [Muchemi and Grefenstette, 2016], its uptake varies, e.g. comprising only 2.7% of all searches conducted among engineers in an oil and gas enterprise context [Cleverley and Burnett, 2015b].

Faceted search offers a way of rapidly conveying quantitative information such as the number of documents published in a particular month or the number of documents by a particular author, but it does not naturally represent more complex and more ‘qualitative’ relationships as might be important in an enterprise context. For example, when exploring contracts that a company has in other countries it might be more important to identify the country that has the largest average contract *value* than the one in which the largest *number* of contracts exist [Ben-Yitzhak et al., 2008]. Ben-Yitzhak *et al.* extend traditional faceted search to allow the exploration of more complex data models and to support correlated facets. Both these extensions aim at providing business intelligence (BI) capabilities, more specifically online analytical processing (OLAP), to the world of textual queries over semi-structured or metadata-rich data. Such analysis is traditionally only supported by databases over structured data.

The taxonomies underlying faceted search do not necessarily need to be hand-curated and they might well be automatically constructed for a specific document collection at hand [Muchemi and Grefenstette, 2016], and a hybrid approach that combines automatic techniques with manually created resources has the potential of outperforming each of the individual techniques, e.g. when identifying synonyms in a domain-specific document collection [Cleverley and Burnett, 2015a].

Faceted search may be even more useful if the source data is enhanced in some way, e.g. by named entity recognition, the search can then be faceted by entity such as person name, place name, organisation, email address, phone number etc. – a processing step that is finding its way into enterprise search [Hawking, 2010]. Faceted search, i.e. the provision of facets and metadata for the user to navigate and filter results, is seen as an essential criterion to support exploratory search [White and Roth, 2009].

2.4.2 Federated Search

Federated search has emerged as an important search paradigm whenever there are several different textual collections that need to be searched independently with results being merged and returned to the user. In a Web context this would allow a search engine to tap into the content of a collection that cannot be crawled directly (e.g. as it might form part of the *deep Web*) and it also allows the aggregation of results coming from different locations [Shokouhi and Si, 2011].

In enterprise search and site search data is often held in different repositories which lend themselves to federated search, or may simply be impossible to index in a central index [Li et al., 2013]. Government Web sites that provide access to material from different agencies are an example [Thomas et al., 2010]. Note that in an enterprise setting federated search techniques do not just provide a suitable framework for parallel search in multiple collections but it can be the only way to provide a single access point to internal and external resources simply because different departments index their data silos in different software systems [Mukherjee and Mao, 2004]. In addition to the number of different applications that are typically searched in an enterprise there tends to be the need that users have to be authenticated to access these applications, which makes federated search very appealing as authentication can happen centrally via an identification token that is passed back and forth when accessing each individual application [Delgado et al., 2005]. However, human-resource-intensive maintenance and crawling issues are among the potential drawbacks.

No matter what data collections are being indexed, there are generally three main challenges to be addressed by a federated search architecture: the issue of selecting the right collections for a query at hand (*collection selection*), the issue of keeping knowledge about each individual collection (*collection representation*) and finally the aggregation of the results (*result merging*) [Shokouhi and Si, 2011]. Result merging is perhaps the most difficult as each system searched (even if based on the same technology stack) will have different and possibly incomparable measures of relevance. Hawking adds another challenge, that of *translating* the query into the query language accepted by each federated service [Hawking, 2010].

2.5 Contextualisation

By having surveyed related work in the wider area of search we are now in a position to more clearly describe enterprise search as one particular region in a multi-dimensional space of search applications. For the sake of simplicity and generalisability we pick dimensions that are in our view core characteristics of the different fields and hence good discriminators when contrasting different areas. They are represented as features whose values we pick to present a *typical* application in that area. This sort of conceptualisation lends itself to some straightforward tabular representation, and as a result these features allow us to easily pinpoint each region in the broader space of search applications.

The tables are not meant to be definitive as it should be possible for almost all dimensions/features to create some niche examples that would contradict the abstraction we provide. Hence, we suggest to treat the tables as those describing a *typical* scenario for the chosen search context. There is of course a risk that these comparisons seek to generalise the categories beyond what one would consider a sensible extent. This is a fair concern. Nevertheless, we attempt to equip the reader with some tangible conceptualisations and hope that (taken with a pinch of salt) these tables are a useful rough guide, or perhaps the starting point for some heated discussions.

What are these features? We will look at data structures, information needs, types of users, evaluation metrics, security considerations, support and customisation, and a few others. Once we have set the scene we can dissect enterprise search in more detail in the next chapter. In the following we will narrow down the scope and limit our discussion to those search areas that we find provide the best way of contextualising and contrasting enterprise search. We leave it as an exercise for the interested reader to complete the picture based on the more detailed discussion earlier in this chapter.

2.5.1 Data

Data is a rather broad dimension as it includes, for example, data structures, data types, and data gathering. Let us start with data structure which can refer to the internal document structure, the organisational structure of these documents, the structure of document repositories etc. It would therefore not be unreasonable to represent them as several orthogonal dimensions. Here we are only concerned with the internal document structure but will dive into more detail in the following chapter to look at other data structures of interest.

We will take a simple three-way classification that distinguishes *unstructured* and *structured* data as the two extreme ends of a spectrum with *semi-structured* data sitting in-between the two. We should of course note that all data has some structure, whether it is explicit or implicit and often the real challenge is that underlying structure may be hidden in the data or even in the representation [Allan et al., 2012].

The type of structures in enterprise search tend to cover the full spectrum from *unstructured* to *structured* data. To be more precise, we refer to *structured* data as anything that comes with an explicitly imposed structure such as records in a relational database (e.g., customer records, phone book entries, spreadsheets); we refer to anything that comes with implicit structure which could be used to automatically infer some 'usable' structure[17] (e.g., Web pages, emails, Word documents) as a *semi-structured* document and we consider *unstructured* data as such that does not lend itself to being mapped into a structured format (e.g. a scanned document or an image without metadata). We leave this deliberately under-specified as we should also point out that there is no commonly agreed consensus about these types, and Web pages, word processing files and emails are quite commonly described as *unstructured*, e.g. [Mukherjee and Mao, 2004, Leavitt, 2010, Grefenstette and Wilber, 2010]. This is a fair point and as long as the internal structure is *not* being exploited documents like emails and Web pages should really be treated as unstructured, *but* structural information can and has been used offering insights directly derived from that structure, e.g. in Web search, for example by interpreting *schema.org* annotations, and in email search, e.g. [Graus et al., 2014]. Hence, we consider both unstructured and semi-structured content as being typical for Web search.

[17] This could for example be the title of a document, the author or date which can be mapped to attributes in a database structure [Croft et al., 2010, p. 2].

Table 2.1 puts all this in context, and we see that enterprise search is unusual in that all three types of structures represent a defining feature of this search application area.

Table 2.1: Typical source data

	Unstructured	Semi-structured	Structured
Web Search	x	x	
Site Search		x	x
Desktop Search		x	
Database Search			x
Enterprise Search	x	x	x

The level to which the data repositories are managed is reflected in Table 2.2. Large parts of enterprise search repositories can be assumed to be curated but there are aspects of them that are not. For example, many enterprise search applications might simply index 'everything on the shared drive' and this is certainly not curated; people might then assume the search engine itself will be able to figure out structure and replace the function of a curator. We have already noted that intranets and enterprise search applications are essentially spam-free which is a result of the way the data is managed. Desktops are at least partially curated as people organise their work in folders etc. The introductory example of GOV.UK is a site search example that we consider fully curated.

Table 2.2: Typical level of data curation

	Largely/fully	Partly	Not
	Curated	Curated	Curated
Web Search			x
Site Search	x		
Desktop Search		x	
Database Search	x		
Enterprise Search		x	

Another way of looking at the data is to assess the extent to which the collection changes over time. This can vary a lot even within one type of search application. Also, this reflects how quickly the data needs to be reindexed which should not normally be the biggest problem to solve. Having said this, we will get back to this issue when looking more closely at data gathering in the next chapter.

More interestingly perhaps when comparing different search areas is the question as to how frequently the underlying data characteristics can be expected to change, such as the main data structures, the language(s) used in the documents of the repository, the types of documents, the data sources, etc. Here we find that Web search faces a constantly changing document collection with new characteristics being introduced all the time, while in enterprise search this happens less frequently, it does however happen with the introduction of a new document management system or a merger with another company [Manning et al., 2008, p.134]. Once such change happens it is likely to require substantial customisation effort which is different to the Web search case in which changes will have be be dealt with largely automatically. See Table 2.3 that puts this in context with other search applications.[18]

Table 2.3: Frequency of change in data *structures*

	Never	Infrequent	Frequent
Web Search			x
Site Search		x	
Desktop Search			x
Database Search	x		
Enterprise Search		x	

2.5.2 Information Needs

Enterprise search environments feature a huge diversity between organisations not just in respect to quantity of information, number of

[18] Just as a reminder, the tables are meant to represent typical cases, hence we specify that database structures *typically* do not change once in production.

repositories, number of document types, but also the nature of searches conducted [Hawking, 2010]. This makes is difficult to capture and classify *typical* information needs.

One way of capturing what a user is searching for is to adopt Broder's classification of Web searches which is defined by three categories: navigational, transactional and informational [Broder, 2002]. Navigational requests aim at specific pages (e.g. 'bbc homepage'), transactional queries are aimed at conducting a transaction (e.g. 'tickets for the Colchester Beer Festival') and informational requests are more exploratory (e.g. 'what to do in Wivenhoe'). However, all three types can also be found in enterprise search making it difficult to differentiate what distinguishes one search area from another.

An alternative approach is to classify typical information needs according to their cognitive complexity, e.g. by mapping them into Anderson and Krathwohl's *Taxonomy of Learning* which distinguishes, for example, simple fact retrieval from analyzing and evaluating information [Krathwohl, 2002]. This has been applied to devise search tasks of varying complexity, e.g. [Wu et al., 2012], but given the abovementioned variability of information needs across organisations makes it difficult to generalise.

The approach we take here is a rather pragmatic one – in line with the applied nature of enterprise search. We adopt ten different information need categories that have been identified as representing ten common use cases with labels that should be familiar to most organisations [White, 2015b, p.139-141].

Here is a short summary of each of them (listed in alphabetical order):

- *Analysis*: looking for trends in performance requiring for example a defined set of reports with potentially a substantial element of numerical data; related to business intelligence

- *Compliance*: high recall tasks aiming at finding all critical information on a particular topic

- *Expertise*: finding people with specific expertise

- *Induction*: finding information relevant for a new member of staff or an employee starting in a new role
- *Item*: finding a specific document
- *Learning*: finding information on a topic that could be covered by a broad range of terms (essentially an exploration scenario)
- *Mobile*: queries submitted from a mobile device[19]
- *Monitor*: a typical monitoring task where the search requirements do not change much over time and the user needs to be informed when new information becomes available
- *Product*: finding information on a specific product with a near miss essentially being a failure
- *Task*: supporting the user in performing a standard task such as setting up a project team.

One observation worth pointing out here is that about half of these examples represent the higher, i.e. more complex, levels in the *Taxonomy of Learning*. This indicates another inherent reason why enterprise search is so difficult. Table 2.4 depicts how common these use cases are across other types of search. The results suggest that what can be considered common information need categories are overall much less well represented in the other search applications we use for comparison.

2.5.3 Users

Users in an enterprise search setting are defined by their role, their access rights and their position within the employee network across the enterprise. Roles define job functions within an organisation and users can then be assigned to these roles [Hawking et al., 2005, Sandhu et al., 1996]. This enterprise-specific setting provides a contrast to most other applications at hand here.

[19] This appears to be a bit orthogonal to the other use cases as it is defined by the type of device rather than the information needed but we include it for completeness.

Table 2.4: Typical information need categories in enterprise search (an 'x' means that this category also represents a typical type of information need in the application at hand)

	Analysis	Compliance	Expertise	Induction	Item	Learning	Mobile	Monitor	Product	Task
Web Search					x	x	x	x		
Site Search			x	x	x	x	x		x	
Desktop Search					x					
Database Search					x				x	
Enterprise Search	x	x	x	x	x	x	x	x	x	x

Looking at it slightly more generally, one way of approaching a comparison might be by defining different user types or use cases. While this is an interesting route to explore (and we will look at this in part in the next chapter), we adopt a different classification here representing the typical size of the user base.

Table 2.5: Typical size of user base

	Small	Medium	Large
Web Search			x
Site Search		x	
Desktop Search	x		
Database Search		x	
Enterprise Search		x	

Enterprise search takes on the middle ground in this respect with desktop search and Web search representing the extreme ends. Table 2.5 puts this in context with related areas. We should of course point out that enterprises might have 100,000s of users which would hardly count as a medium-sized user base but most do not have that large a number and we are more interested in making the case as a relative comparison with other fields.

2.5.4 Security

Security is paramount in enterprise search. At an application level this mainly concerns confidentiality, i.e. the provision of access to documents the user has a right to access according to his or her role in the enterprise, which means that forbidden access to a document should be prevented but it also means that preventing access to a document a user should have access to needs to be avoided [Hawking, 2010, van der Lans, 2013]. Such access is commonly controlled via Access Control Lists (ACL) granted at document-level which could be done via late binding (checking access levels at query time) or early binding (capturing the access level at indexing time) via caching. It would be desirable to control access at a collection level by simply granting access according to a user's role but the applicable security models typically turn out to be more complex than that [Hawking, 2010]. Figure 2.1 identifies access control as a central component of the overall enterprise search framework. If type-ahead search suggestions are provided when the user enters a query, the suggestions provided should also be restricted according to the user's access level.

Table 2.6: Essential security requirements (at application level): *Confidentiality* refers to the concealment of information or resources, i.e. making sure information is not disclosed to unauthorised users. *Integrity* refers to the trustworthiness of data or resources, i.e. users should not be allowed to make unauthorized changes. *Availability* refers to the ability to use the information desired, i.e. the information should be available whenever required by the user.

	Confidentiality	Integrity	Availability
Web Search			x
Site Search	x		x
Desktop Search			x
Database Search	x	x	x
Enterprise Search	x		x

In Table 2.6 we compare and contrast which of the three commonly used (basic) security requirements of confidentiality, integrity and availability [Bishop, 2003, p.3] are key features at the (search) application level.

2.5.5 Evaluation Metrics

Information retrieval research has always had a strong experimental element and the adoption of commonly accepted evaluation paradigms (including evaluation metrics) has been integral to progress in the field [Baeza-Yates and Ribeiro-Neto, 2010, p. 131-176].

A plethora of metrics have emerged and been adopted over the years, primarily measuring 'retrieval quality'. The two metrics that have dominated at least the academic IR world are *precision* and *recall* – representing the quality of returned results and the proportion of matching documents that have been returned, respectively. Many other measures are based on one or the other or a combination of both (such as F-Measure).

Deciding for each search area which one of the two basic metrics is the most 'important' one risks simplifying the comparison too much. However, given they are so fundamental to many of the standard evaluation measures and given they provide a suitable way of contrasting different areas using a simple check-list we do just that. Table 2.7 illustrates that Web search with access to a large document collection is typically precision-oriented while smaller collections (corresponding to smaller user bases according to Table 2.5) appear to be more concerned with recall – this includes enterprise search. This is not surprising given that not finding a result when searching the Web might not be noticed or not be an issue but not returning a set of known records or to not find *all* the matching content might be considered a failure in an enterprise setting– especially for discovery purposes [Miles, 2014].

We would like to reiterate that the table is only an approximation. There are Web search tasks that are clearly aimed at high recall, e.g., searching for medical information. There are also aspects of desktop search that are precision-oriented, e.g., known-item finding of an old file that the user knows exists. In fact, for every one of these, one could

Table 2.7: Dominant standard IR metric

	Precision	Recall
Web Search	x	
Site Search		x
Desktop Search		x
Database Search		
Enterprise Search		x

create scenarios where it is either precision or recall focused. What we have presented is a simple rule of thumb, and we refer the reader to Chapter 4 which will discuss enterprise search evaluation at the level of detail it deserves.[20]

2.5.6 Support

The type of different actors and corresponding support involved in a search application varies substantially. While a Web search engine like Google relies on 20,000+ engineers to look after the system, in enterprise search there is "rarely more than one lonely person with the responsibility for supporting the search application and making sure that it is tuned to meet user requirements" [White, 2015b], a statement further supported by the most recent Findwise Enterprise Search survey: less than half the organisations surveyed have more than one full-time person in charge of search and findability [Findwise, 2016]. The problem with this is that without (continuous) search support enterprise search does not perform.[21]

Support responsibilities might be distributed, in fact enterprise search is typically managed by administrators who are domain experts

[20] Database search is more difficult to generalise. If search is over free text fields, then it depends very much on the specific application context whether precision or recall is more important. If the search request resembles a Boolean-style SQL statement, for example, then it is more difficult to trade off one metric against the other. Hence, we left the relevant row unticked.

[21] For anyone who wants to take home just one message from this review, then this is it.

but not search experts which means that translating the domain knowledge into tuning an underlying retrieval model is non-trivial if not impossible [Bao et al., 2012a,b]. The problem is even more pronounced once text mining gets incorporated as it is rare to find a single individual who would bring in the necessary domain knowledge, knowledge of text mining software and how to configure it as well as natural language processing expertise [Upshall, 2014].

Table 2.8 provides a simplified overview of what groups of experts are involved in running the system. The roles we identified here are:

- *Search Admin*: the person/team who runs the search engine day-to-day, making sure it is available to users, logging and investigating any faults with the infrastructure or code (they may pass these to the search developer to fix), they are logging useful metrics (e.g. query logs, no-results pages, query response times).

- *Domain Expert(s)*: the person/team who understands the content indexed by the search engine and the likely queries that might be run on it, and what 'relevance' actually means in the context of the business e.g. in a law firm, someone with enough legal knowledge to understand the difference between 'clients' and 'matters'. Domain experts can judge how well the engine is performing with respect to relevance and can advise on how content might be restructured if necessary without changing the meaning/usefulness of it.

- *Search Developer*: the person/team who can install the search engine, develop indexers and user interface code that connect to it and change its configuration to improve relevance, performance etc. as necessary. Search developers are typically not domain experts so they will need the advice of one to help tune relevance or improve content indexing.

Due to the domain-specific and heterogeneous nature of the document collections it is not surprising to find that enterprise search applications need to be coordinated among a larger number of different actors than most other search applications. Of the chosen search

Table 2.8: The human (expert) in the loop

	Search Admin	Domain Expert(s)	Search Developer
Web Search	x		x
Site Search	x	x	x
Desktop Search			
Database Search	x		x
Enterprise Search	x	x	x

application types, desktop search appears to be the only one that does not require any admin support – the application gets installed once and then runs out of the box.

Note that there are more people identified in Figure 2.2 than are in the table as we only focus on the most important roles here, but it should be pointed out that there are others that might be involved at different stages. For example, the search engine will likely be bought from a vendor, but once it is installed the search developer makes any further changes.

2.5.7 Relevance Tuning

So far we have left the more pragmatic issues of getting a search system to perform well aside. Here we introduce customisation and relevance tuning as essential ingredients to make enterprise search work and contrast this with tuning methods that are core to other applications. We also provide a sneak preview on what will be discussed in more detail in Chapter 5.

Query and access log analysis offers a way of tracing the users' search behaviour and is a good starting point to improve the search system, e.g. [Hawking, 2010]. In intranet and enterprise search this is not sufficient and hard-coded suggestions linked to specific queries referred to as *quick links*, *suggested matches* or *best bets* are an important element of making enterprise search work [Wu et al., 2014, Morville and Callander, 2010, Bao et al., 2012a, Rowlands et al., 2007]. Through this mechanism the majority of top queries might actually be available as quick links [Lund and Ørnager, 2016].

Table 2.9 introduces these tuning elements as quality control mechanisms. We refer to 'analytics tools' as those tools that allow the ongoing analysis of how well the system is performing according to the requirements set out. For enterprise search this means that one typically has an internal test collection based on a sample of queries for which the ranking function needs to be tuned [Hawking, 2010].

Table 2.9: Quality control mechanisms

	Best Bets	Log Analysis	Analytics Tools
Web Search		x	x
Site Search	x	x	x
Desktop Search			
Database Search		x	
Enterprise Search	x	x	x

Table 2.10 provides a coarse-grained classification of what level of tuning is (typically) needed to keep a particular search application running satisfactorily. Unlike the continuous tuning necessary in enterprise search as discussed above, the 'Search Quality' groups of Web search companies work very differently in that matching and ranking *algorithms* are continuously improved while hard-coded links are not the centre of attention.

2.6 Concluding Remarks

Rather than attempting to provide yet another definition of enterprise search arising from the discussion and contextualisation we simply conclude that enterprise search is a multi-faceted area as well as an evolving field that shares a lot of features with other areas of information retrieval but differs in other respects.

We will now explore the characteristics and challenges of enterprise search in some more detail. The structure of the discussion and the

Table 2.10: Relevance tuning

	Little Needed	Continuous
Web Search		x
Site Search		x
Desktop Search	x	
Database Search		
Enterprise Search		x

checkbox approach in the last section, in which we compared and contrasted enterprise search with related areas, provides a suitable road map to do this. From this discussion we conclude that there are three broad areas that are worth dissecting in more detail – each one in its own chapter.

We will first provide a systematic overview of what defines enterprise search, this will be the focus of the next Chapter, entitled *Enterprise Search Basics*. We will drill down into the actual characteristics outlined in Tables 2.1 – 2.6 describing data, users, information needs and other core features of enterprise search. This chapter will provide a solid account of the state of the art in the field while also highlighting issues around enterprise search that need to be addressed properly in order to make it work successfully.

Following this we discuss how to evaluate enterprise search, again dedicating an entire chapter due to its overall importance. Table 2.7 provides a reference point to start the discussion. We will more broadly present commonly applied metrics and evaluation approaches and contrast them with other areas of search. In line with the previous chapter, we will highlight the fundamental difficulties emerging from evaluating enterprise search applications.

Finally, we pick up the identified issues and difficulties and work out in a less theoretical and more practical discussion what can and needs to be done to fully exploit the potential of enterprise search. That chapter, called *Making Enterprise Search Work*, aims at offering an understanding why enterprise search so often fails to perform.

We also expect that this chapter will be of particular interest to the practitioners among our readership as it will provide some important guidelines for making enterprise search work – backed up by findings reported in the refereed academic literature as well as in core readings from the practitioners' community. Tables 2.8 – 2.10 are key reference points in this discussion.

3

Enterprise Search Basics

This chapter will expand on the main characteristics that have emerged from the discussion of how enterprise search compares with other search application areas.

We will focus our analysis on the more technical aspects of enterprise search which on their own already provide enough material for this study. It should however be noted that technical solutions do not work if they are not supported by appropriate management structures and if users do not perceive the information management system at the enterprise as able to provide answers to day-to-day information needs, e.g. [Stenmark, 2006]. We will touch on some of this in Chapter 5 when we discuss how to put the user in control.

We will start with infrastructural issues – data sources, collection gathering and enterprise search architectures. We then move on to discuss typical information needs with an extended discussion of issues around common people-finding needs. Search context, user modelling issues and temporal factors will be examined before finishing with a brief look at existing tools, frameworks and resources.

3.1 Structure of Data

We already discussed in Section 2.5.1 that there is a multitude of data structures present within a typical enterprise search context and that section also compared and contrasted some of these structures with alternative search application areas. Here we will try to convey a more detailed picture by expanding on three types of data structures, namely the structure that characterises the documents/items in the repositories, the metadata and taxonomical structure that is linked to these documents, and the organisational structure of the repositories. How these data structures are to be exploited in a search framework will be discussed in the following sections that look at data gathering and architectural considerations.

3.1.1 Document Structure

A defining characteristic of enterprise search is the fact that data comes in a variety of different formats, ranging from largely unstructured data (e.g., scanned documents) to semistructured (emails, memos) and fully structured (internal phone directory) – a property clearly reflected by the unique position enterprise search takes in Table 2.1 which compares the typical data sources of different search areas. To make this point clearer, asked which document types are important for employees to search, the AIIM Enterprise Search survey found the most obvious types like Office documents and PDF files to be top of the list – as well as emails, but it also uncovered that about 80% of respondents reported scanned/OCR documents, 51% drawings and maps, and 46% photo images [Miles, 2014]. In that respect enterprise search fits perfectly one of the key research questions identified at the SWIRL 2012 strategic workshop on IR, namely how to move beyond simple document retrieval by better integrating structured and unstructured information which was identified as a promising but underdeveloped area of exploration [Allan et al., 2012].

Generally speaking, there is a lot of inherent structure in the documents which is somehow the reverse to the idea of automatically trying to extract structure from unstructured text. e.g., [Shin et al., 2015].

The presence of different types of document structure in enterprise search does however not just present a challenge but also offers potential benefits, for example, in query result diversification subtopics derived from structured data might offer high-quality terms but vocabulary gaps whereas subtopics extracted from unstructured sources might have a broader coverage but more noise, so that a combination of both types of sources will outperform each one of the two [Zheng et al., 2011]. Similarly, unstructured enterprise data might be used to first classify the type of information need of a query which then guides access to the semi-structured or structured data sources (relational databases) to improve the overall search accuracy [Liu et al., 2011], or for entity-centric query expansion [Liu et al., 2014].

3.1.2 Metadata and Taxonomies

Metadata and taxonomies are closely related concepts as metadata such as classifications assigned to a document might be driven by business-specific taxonomies. In addition to that there is of course more generic metadata like the type of document, the date of document creation or last access etc.

There are different questions we can raise. Does document metadata play an important role in making enterprise search perform well? Does taxonomical knowledge help in this process? Do documents carry sufficient metadata in the first place? The short answer to all of these questions is ‘yes, but’. Let us look at this more closely.

Metadata does indeed play an important role. Supporting evidence includes the fact that organisations that adopt metadata standards have generally more satisfied users [Findwise, 2014]. However, there are caveats. First of all, while content creation will typically be centralised to a small number of people and thus comply to certain company policies, consistency is still not guaranteed as there might be multiple organisational units with differing policies [Doane, 2010, Mukherjee and Mao, 2004] – a not at all new phenomenon as illustrated by some of the inconsistencies in the Domesday Book, published more than 900 years ago, in which measurements of pastures vary across different regions from acreages to (different) linear estimates (e.g. “17 furlongs by 17 furlongs" versus "one league in length and breadth"), and meadows are measured by acreages, or number of plough teams they could sustain or linear measurements, all making comparisons very difficult [White, 2012].

A second problem emerges when content is created by simply copying metadata from another document leading to incorrect metadata [Stocker et al., 2014, Li et al., 2005]. Similar findings and the implications of this on search performance are reported by Hawking and Zobel [2007]. The problem of low-quality metadata is by no means restricted to plain documents but applies to databases as well. An analysis of hundreds of tables with thousands of columns sampled from 29 databases within Microsoft's IT organisation uncovered that many frequently used *column* names are very generic such as 'name', 'id', 'description', 'field', 'code', and 'column' and they made up 28% of all sampled columns [Cortez et al., 2015] concluding that these names are useless in helping a user find tables.

A different way of arguing for metadata in enterprise search is by looking at some commonly observed user needs, and this is where a strong case can be made. The importance of not just finding any matching document but the *latest version* is one such example [Stocker et al., 2014, 2015], or in fact the need to identify all content published *after* a specific date [White and Nikolov, 2013] as it is common that users need to find the *earliest* version of a policy or a client pitch as this provides reassurance that there is no relevant information before a certain date narrowing down the search space [White, 2015a]. Note however, that the same caveat applies here in that publication dates may be reliable in some organisations and not in others [Hawking, 2010].

Versioning, e.g. knowing what is the most current and applicable version of a document, is not just important but of critical importance in an organisation [Cherkasova et al., 2009]. This can be hard for a search application to know but it can be essential information, first of all to identify the most up-to-date version of a document but also for compliance reasons.

Addressing the second question, a taxonomy can generally be seen as helping an organisation understand and get access to the information it holds but also to identify missing information [Lund and Ørnager, 2016]. Taxonomies are considered the base of a good information structure in an enterprise [Findwise, 2014], and a well-defined taxonomy alongside standardised metadata and a consistent classification scheme are key to findability [Miles, 2014]. Note that taxonomies are often used to provide a 'browse' function which can be used *instead* of search, common in organisations where the search function is, or is perceived to be ineffective. Other users simply want to browse through a well-structured hierarchy [White, 2015a], but it is not unreasonable to assume that users might not understand the taxonomy in which case this search could become frustrating.

In any case, much work remains to be done to bring experts in taxonomy and search together as neither really understand each others' worlds – an observation we already reported earlier to be true in the digital libraries space, e.g. [Gonçalves et al., 2004]. A successful example of this is the construction of a taxonomy to assist in searching the education content of GOV.UK, a collaboration between developers, information architects and data scientists who used topic models in combination with human intervention to identify the topics' meaning and to label them.[1]

Finally, document metadata, e.g. document title and semantically meaningful document category information, is commonly present at least in some of the repositories accessed by enterprise search applications. This might be a fairly obvious statement given that data management within an organisation is at least partly curated and controlled as reflected by Table 2.2 – but note here as well the caveats related to the quality of this information as discussed above. Apart from answering the earlier example queries which make use of some date stamp, there are other types of metadata that can then be exploited in a structured index as for example in IBM's *Gumshoe* intranet search framework [Bao et al., 2012b]. In the Gumshoe example the categories reflect for example whether a document resides in the employee directory or is a software page or wiki page.

[1] `https://gdsdata.blog.gov.uk/2017/01/12/using-data-science-to-build-a-taxonomy-for-gov-uk/`

To conclude, metadada and taxonomies are very valuable in enterprise search *if* handled properly. This is a big ‘if’ though which cannot easily be aligned with the fact that organisations often do not have the resources to build and maintain metadata properly, that people searching do not have the same taxonomic model as the custodians of the data and that they cannot use it to search effectively, that navigating through an enterprise taxonomy to find the answer to a specific question might be impossibly clunky compared to effective search etc. This is all part of a much more general and long-running discussion as it touches on some of the fundamental assumptions of the idea of a Semantic Web [Berners-Lee et al., 2001] and the more pessimistic take on this is that a world of exhaustive, reliable metadata would be a utopia that will never come into being [Doctorow, 2001]. On a more positive side, some metadata is useful and reliable such as automatically assigned labels like geo-location data and time; from/to/cc/date metadata on emails etc. Also, metadata used in e-commerce search can generally be assumed to be fairly accurate as it forms the backbone of any search application. Hence, in enterprise search topical and taxonomic metadata can be very useful but is virtually useless in the worst case.

3.1.3 Repository Structure

We established that unlike in Web search the data structures in enterprise search tend to change infrequently (see Table 2.3), a characteristic that should assist in the management of search applications. Nevertheless, the repository structures that hold the data present a fair number of serious challenges. To start with, large enterprises tend to have thousands of relational databases each having tens to hundreds of tables [Cortez et al., 2015]. But also beyond relational databases, data silos are typically found in an enterprise setting and often not even joined up in a single search application [Findwise, 2015]. In fact, we should say, *very commonly* not joined up, e.g. only 11% of organisations have a fully joined up enterprise-wide approach and another 18% across departmental content according to the AIIM Enterprise Search survey [Miles, 2014]. Complicating this is the fact that data held in different repositories is not typically cross-referenced, e.g. there are no hyperlinks from an email repository to a content management system [Mukherjee and Mao, 2004]. This silo-based structure with all its related challenges turns out to be one of the defining features of enterprise search and has implications on the data gathering step and the search architecture to be employed as will be discussed in more detail later.

Email servers and file shares are the most important repositories to search according to Miles [2014], other important resources include documents residing in content management systems, structured database content, the corporate intranet and staff directories (*unlike* blogs and internal social streams which still seem to represent only a very small part of the overall information infrastructure).

Duplicated data represents another core feature of enterprise search not least because email records are an important part of the information architecture. Enterprise-based email corpora that have been made available demonstrate this point, e.g. the Enron corpus was cut down to one third of the original number of messages when duplicates and irrelevant folders were removed [Klimt and Yang, 2004] and a third of the emails in the Avocado email collection turned out to be duplicates [Oard et al., 2015]. Both collections will be introduced in more detail in Chapter 4.

More generally, a good proportion of data growth within an organisation is attributed to data duplication and while not all of it is considered waste (such as copies made for caching and intentional duplication), a lot of it is [Forman et al., 2009]. This includes entire directories being copied. Document versions as commonly found in an enterprise environment give rise to further issues with duplication (or near-duplication), e.g. [Cherkasova et al., 2009, Khan et al., 2016].

Duplicated information at a database level might highlight problems that reflect more on the *data quality* in general. For example, if particular services are available from different repositories, e.g. two systems offering similar services such as information about daily bed utilization within a hospital but both systems having frequently different values, or within a health maintenance organisation, inconsistent data values between internal patient records and the bills submitted by hospitals for reimbursement [Strong et al., 1997]. Inconsistencies in the information architecture do not just lead to user frustration but can also result in extensive use of email and telephone communication [Lund and Ørnager, 2016].

Let us finish with a general point. Data quality is highly important and even the best search technology cannot provide a good experience if data quality is not addressed. While this is a general problem that many types of search applications face, the fact that content quality is not adequate can in fact be seen as a major reason why enterprise search in particular so often fails [White, 2015a]. Part of the solution is to identify which repositories to include in the first place thus cutting down the content [Morville and Callander, 2010, p.38] and making sure the right data sources are selected and indexed, and this approach hints at the importance of customisation and continuous maintenance expanded on in Chapter 5.

3.2 Collection Gathering

Unlike in more homogeneous environments, collection gathering, maintenance and curation in enterprise search is more complex with a multitude of data sources of varying structure. We will discuss some key concepts around gathering the collections that should be accessed by the search application.

3.2.1 Connectors

The concept of *connectors* is omnipresent in enterprise search. Connectors are the interfaces between individual data silos, e.g. structured content stored in databases, and the search engine and they do not just enable a search engine to download the content but integrate the database schemas as indexable entity attributes therefore allowing the capture, access and exploitation of rich semantic metadata encoded in database structures [Grefenstette and Wilber, 2010]. Apart from interfacing with databases there are file system connectors (e.g. for accessing enterprise file servers) and messaging connectors (to connect to the enterprise email systems).

Connectors tend to be version-specific and need constant attention to make sure they work correctly [White, 2015b] providing a strong argument for White's point of view that enterprise search is not a project but a continuous process.

The fact that the connectors also manage the required security protocols when accessing repositories means that there is always some latency introduced into the delivery of results [White, 2015a].

3.2.2 Crawling

Keeping the available data sources up to date is another issue. A typical approach to making them searchable is in principle no different to any Web search engine – via a crawler. This requires knowledge of where to find information in the first place [Mukherjee and Mao, 2004], but due to the nature of an enterprise search architecture such a crawler faces additional complexities including manageability and maintenance (e.g. a multitude of APIs), authentication and security, synchronisation, and the problem that the crawler will collect data in an application-agnostic way [Delgado et al., 2005]. In addition to that server locations do change and documents might not always be accessible when a server fails or has been taken down for service [White, 2015b].

Servers that should be indexed need to be identified together with update cycles. Practical considerations such as bandwidth requirements of crawling and indexing play an important role [White, 2015b]. It might also be desirable to exclude certain types of content from the crawl altogether to improve the overall quality of the index, e.g. by removing old content [White, 2015a].

Alternative models to crawling (a 'pull model') such as publishing and syndication ('push models') might become more prevalent [Delgado et al., 2005] but it does not look likely at this point that these will completely replace the crawler any time soon.

However, a full crawl is not always possible preventing the creation of a central index leaving federated search as an alternative. Reasons for this scenario include, for example, commercial software with license restrictions or undocumented data formats, as well as pay-per-access databases run by third parties [Li et al., 2013].

There is a more general point though, namely that crawling and indexing are bandwidth and processor intensive both of which have cost implications [White, 2015b, p.32]. With potentially terabytes of data on a file share, scanning the file share to find new or changed content will become impractical due to load imposed on the system and the time taken. Similarly, processing millions of PDF documents which might be held in a repository of scanned letters to customers would require significant computing power and significant time. Bandwidth restrictions and limitations of the repositories could make this process take months. Incremental updates are one way of addressing this at least in the short term but over time an incremental index can grow rapidly in size while access speed declines [Hawking, 2010]. The only viable way of keeping a search index up to date could be by choosing a search technology which supports continuous document ingestion e.g. by inserting a 'T junction' in the document creation pipeline which sends each new document to the search engine as well as to the repository. Another condideration is that the contents of the search index are valuable because at the outlined scale they cannot easily be re-created. In cases like this it is essential to have redundancy and resilience in the search system.

3.2.3 Security

Specific requirements the repositories have to satisfy are that enterprise search systems are to be used by named users who have been assigned specific roles in the organisation and – in contrast to Web seach – security control via access permissions are an essential part in such an environment [Mukherjee and Mao, 2004]. This does not just apply at search time but equally at crawl time which is further complicated by a potential range of different access protocols [Delgado et al., 2005]. The crawler itself may have to assume a 'super user' role which gives it access to all content – but a role which may be difficult to grant in an enterprise context. It is also common that the very act of a wide crawl of enterprise content reveals flaws in the security model as content becomes accessible via search.

3.2.4 Deposition versus Publishing

It has already been pointed out that information on an intranet might be created for pure dissemination rather than published with the aim to attract the attention of users. Actually, documents might be created with neither of these two aims in mind, and within an enterprise environment we need to distinguish between data that is simply *deposited* from data that is actually *published*. It frequently happens within enterprises that a newly created document is deposited rather than published. An employee might write a policy or a report, send a letter to a customer, type up case notes, or write a record of an interview, then deposit this on a shared drive and leave it there. They might create it on their own private drive first and later enter it into a record management system. Documents may be created in the course of an employee's duties with no thought that others may want to read it. There is no sense of disseminating or publishing, rather the fulfilment of an obligation to keep records or simply that the document has to live somewhere.

This is in stark contrast to the GOV.UK example we picked as a motivating example in the introduction to represent site search. In that case all content can be assumed to be published rather than just deposited.

This is also very different to publishing content on the Web where companies frequently apply search engine optimisation techniques to make sure the content can be found whereas enterprise authors may make no effort whatsoever to facilitate subsequent discovery. The problem of course is that the documents such as letters to a particular customer or the record of an interview still need to be found when needed.

From a collection gathering perspective it is then the task of the search administrator to decide what data needs to be indexed so that it can be found.[2]

[2] We will get back to this issue in Section 5.2.3 that will argue that *not* indexing something can be desirable in certain circumstances.

3.2.5 Other Issues

There are many other considerations to be taken into account in the collection gathering step, some more content-specific and others more technical.

One common problem is that employees might not even be aware of what information sources are available in the enterprise whether they have access to them or not [Laqua et al., 2011, Feldman and Sherman, 2001]. Noisy legacy data will always be an issue no matter the data quality improvements through enhancing consistency and automated processing of content [Mukherjee and Mao, 2004].

Given the nature of structured backend databases there is a need to consider transient/virtual documents which are generated on the fly [Delgado et al., 2005]. This is comparable to the 'deep Web' but within an organisation the presence of such documents might be a lot more pronounced (and perhaps more critical to uncover).

An organisation may be faced with significant changes in collection and query characteristics through a merger or acquisition [Manning et al., 2008, p.134], something that happens at a high rate [Grefenstette and Wilber, 2010].

Finally, Berners-Lee motivated his proposal for what eventually became the World Wide Web by illustrating how, due to high turnover of people in an organisation like CERN, information is constantly being lost or is being recorded but cannot be found [Berners-Lee, 1989]. This still remains a challenge today, and capturing the 'corporate memory'[3] [Brooking, 1999], conducting proper role hand-overs and training a new employee ('on-boarding') are examples of important management processes directly impacting the effectiveness of the search system.

3.3 Search Architectures

Figure 2.1 introduced a general architecture of enterprise search. At the heart of it lies a typical IR processing pipeline but at closer inspection

[3] Part of the collective memory is likely available (implicitly) in email repositories and linking them up to the enterprise search system should be included in the collection gathering process.

the picture also reveals a number of components not necessarily part of a standard IR system such as heterogeneous data sources residing in repositories of varying structure, the need of connectors emerging from this setting, and the core requirement of access control. Together with the discussion so far it should have become clear that employing simply commonly applied search architectures without catering for the specifics of an enterprise context will not suffice.

3.3.1 Basics

To motivate the need for a search architecture that is different to an out-of-the-box search framework that could be applied to any collection we return to the axioms put forward by Fagin et al. [2003], investigating IBM's intranet as a case study, to spell out what is different on an *intranet* compared to the Web: (1) documents created for simple information dissemination[4]; (2) a large fraction of queries with a small set of correct answers (often unique); (3) essentially no spam; and (4) large portions of the intranet are not search-engine-friendly.

One implication of the first point is that enterprise documents are much sparser than Web documents, perhaps also due to a lack of incentives to create textual content [Cortez et al., 2015].

There are additional challenges that need to be accounted for. For example, the use of legacy software is common in this context. Enterprises often use old software versions which is even more pronounced once the software is embedded in third-party enterprise applications that come with extended release cycles [Mukherjee and Mao, 2004].

Apart from differences between enterprise search and other search architectures it is worth pointing out some commonalities such as the need to keep the index up to date and hence to be able to rapidly re-index the document collection (assuming this is feasible as discussed earlier on). IBM's Trevi intranet search engine, for example, observed already more than a decade ago about 500,000 daily changes which includes newly added documents as well as updates to existing ones [Fontoura et al., 2004].

[4] To qualify this claim, please do note the discussion earlier on about documents being deposited as opposed to published/disseminated.

3.3.2 Search Algorithms

It has long been known that link-based ranking methods which work well on the Web such as PageRank [Brin and Page, 1998] and HITS [Kleinberg, 1998] do not necessarily work well in enterprise search contexts [Mukherjee and Mao, 2004, Hawking, 2004, Fagin et al., 2003]. This is still true today, e.g. [Schymik et al., 2015]. One of the obvious reasons for this is the sparsity of links between documents within an enterprise document collection which makes this a much weaker signal than on the Web [Chaudhuri et al., 2011]. That does not exclude the use of other static rank algorithms, e.g. IBM's Trevi search engine serving the company's global intranet assigns a *hostcount* to each document, a static rank reflecting the number of different hosts pointing to the document [Fontoura et al., 2004]. It also means that a static score of a document might take other features into account such as the frequency of access to a resource, the recency of publication, the document type and information about the repository the document resides in [Hawking, 2010].

More sophisticated search algorithms like *learning-to-rank* (LTR) have yet to find their way into enterprise search. Data sparsity issues are certainly a bottleneck in applying any machine learning algorithms in this context. In Chapter 5 we will explore how existing search algorithms are best deployed and tuned in an enterprise search environment.

3.3.3 Aggregation

Search results within an organisation commonly need to be aggregated, e.g. as in expert search across multiple domains [Pal, 2015], content repositories [Venkateshprasanna et al., 2011], as well as in federated search in heterogeneous environments [Li et al., 2013].[5]

Aggregating results from different silos is non-trivial due to the difficulty of comparing relevance across sources [Chaudhuri et al., 2011, Hawking, 2010], somehow comparable to the problem of aggregating re-

[5] There is similarity with desktop search in this respect [Elsweiler et al., 2010].

sults from different verticals in a Web search environment.[6] Federation of search poses another problem, namely that of maintaining document level security [Hawking, 2010]. Note that unlike in common meta search systems the data sources employed in enterprise search typically have no documents in common and employ different ranking and scoring schemes [Mukherjee and Mao, 2004].

Optimising the presentation of results coming from many different sources presents a major challenge also from a human-computer interaction perspective. For aggregated search on the Web click logs have been proposed to learn which portals to select and how to present them [Arguello, 2017, p.405-420], but the availability of training features and the availability of common features across different data repositories presents a serious challenge in particular as the number of sources grows. In enterprise search there will be infrequently used repositories, e.g. perhaps a source code repository for a long-completed project, and best practice in faceted search appears suitable here when deciding which repositories to select and how to present them to make best use of the screen space and provide a strong 'information scent' to the user [Russell-Rose and Tate, 2013, p.174-175].

3.3.4 Security

Security issues such as document-level access control have always been key requirements in enterprise search systems and apart from having to cater for them this represents efficiency and response time challenges [Bailey et al., 2006], so much so that the cost of effectively imposing and managing access control by using the standard approach of creating a separate index per user (early binding) or using a centralised index and doing the access control check at query time (late binding) can become prohibitive for large enterprise environments [Singh et al., 2009].

The primary security concern is that of *data confidentiality* leaving *integrity* to the parent's application (such as a relational database) and *availability* to broader network defences and architectures [Grefenstette and Wilber, 2010]. As such, confidentiality is another defining feature

[6] Although there are also some notable differences, e.g. [Arguello, 2017, p.370-372].

that distinguishes enterprise search from Web search (see Table 2.6).

Like in general enterprise network management [Yu and Wang, 2013], access to documents is commonly granted to users via Access Control Lists (ACLs) [Hawking, 2010] but the drawback is that these are tied to individuals making it difficult to maintain when users, for example, take on different responsibilities [Ferraiolo et al., 1999]. Role-Based Access Control (RBAC) provides a higher level of abstraction and is therefore commonly adopted. RBAC does not directly link users to their permissions but via a role they are assigned to in the organisation [Sandhu et al., 1996]. Roles and role hierarchies reflect the organisational structure of the enterprise/organisation. Access to resources can then be defined through a job function or job title and if a user leaves or moves to a different position, then the user's assignment to his or her role(s) is updated without having to amend any ACLs [Ferraiolo et al., 1999]. The identification of roles however might be a complex issue on its own [Roeckle et al., 2000].

No matter which approach is being taken to impose access control there are non-trivial security issues, mainly information leaks, which arise in an enterprise search context. For a start, the ordering and relevance scores of documents presented to users might reveal general collection properties, i.e. the presence or frequency of keywords in documents not presented to the user [Singh et al., 2007]. Furthermore, query suggestions might leak information if, for example, a user searching for *redundancy* gets offered a query suggestion like *redundancy strategy planning* even though the user has no access to this document [White, 2015b, p.41]. The same applies to autocomplete suggestions.

Finally, the difficulty of managing metadata within an organisation has already been discussed at length earlier with the conclusion that the quality often lags behind expectations. Applying and maintaining access controls suffers from the same problem and it is not uncommon that employees are routinely denied access to material they should be able to see. On the other hand, deploying an enterprise search system often exposes content which should have been protected.

3.4 Information Needs and Applications

Information needs are closely coupled with strategies to satisfy these needs and many approaches to classify them and to suggest models for information seeking have been proposed addressing, for example, the general information search process, e.g. [Kuhlthau, 1991], exploratory search [Marchionini, 2006], collaborative search [Shah, 2012] and search and discovery patterns within enterprises [Russell-Rose et al., 2011]. Our approach to classify information needs and their corresponding applications will be driven very much by an analysis of the refereed literature. This will complement our earlier discussion around Table 2.4 for which we had adopted a handful of use cases that have been identified as commonly occurring in enterprise search.

The lack of publications on enterprise search in the refereed literature will necessarily make this analysis incomplete but should generally provide a good overall picture of what problems *typical* enterprise search systems need to be able to serve.

First of all, the information needs of a user in an enterprise context are closely linked to the requirements of the user's job. The obvious implication is that these needs tend to differ from what a casual user might submit to a Web search engine. For example, within the enterprise employees of PricewaterhouseCoopers (PwC) UK search for people, tools, employee information, office information, policies, help manuals, IT support, technical delivery materials, client advice, industry information, sales materials, social network groups, and that is not even all [Findwise, 2016]. An analysis of the internal searches within the Microsoft intranet identified searching for definitions, persons, experts and homepages as being among the most important types of information needs applying a log analysis[7] study followed by a survey among

[7] We would like to add a *general* note here. Observations on user behaviour patterns obtained through search logs or user surveys are heavily coloured by the capabilities and performance of the search service actually in use, or more accurately, user perceptions of capability and performance. People do not search for things with an enterprise search tool which they perceive or believe to be hopeless. They limit their use of such a tool to types of search which they have confidence will succeed. A poorly performing search engine limits the value of click or browse data because users may not ever reach the best answer in order to click it. We simply want to

employees [Li et al., 2005]. It could even be argued that staff will only use the search application when they need to make a decision [White, 2015a], although it should also be considered that users might conduct business and non-business activities on an enterprise network [Carter et al., 2014].

A fairly typical (but not ideal) scenario appears to be that users access multiple search systems with multiple taxonomies designed specifically for those systems [Doane, 2010]. Going beyond this silo-based approach via some *Unified Information Access* still remains one major challenge in enterprise search [White and Nikolov, 2013, Grefenstette and Wilber, 2010]. In this vision one would have a user-friendly interface that sits on top of a hybrid system accessing data of varying structure from heterogeneous data sources, strikingly similar to the vision a decade earlier [Feldman and Sherman, 2001]. However, establishing a managed search environment that ensures employees find the information they need to achieve their organisational and personal objectives does not per se imply a unified framework [White, 2015a].

Let us now look in more detail at some common search types within an organisation. We will structure the discussion according to different search application types which represent particular information need categories.

3.4.1 People Search

Quite clearly, a good number of the different information needs cases are concerned with various aspects of people search. This is simply because searching for other individuals is considered one of the most fundamental scenarios within an enterprise [Guy et al., 2012, Hertzum, 2014, Findwise, 2016]. People search is also prominent on the Web, e.g., via a people search engine, but there it is dominated by the search for persons in news-related events and known ‘celebrities’ [Weerkamp et al., 2011], while enterprise users commonly aim at finding contact details, roles, expertise *and* this is possibly conducted with incomplete information at hand, e.g. as in “Alice whose last name starts with an ‘H”’ [Guy et al., 2012]. The differences are further highlighted by

state this here as a potential caveat of any enterprise search study.

the fact that users on a Web-based people search engine tend to click through to social networking sites [Weerkamp et al., 2011] which is in contrast to the data sources accessible in an enterprise search system, e.g. a telephone directory [Li et al., 2013].

Identifying colleagues that know about specific projects appears to be a core information need among people-related searches [Hertzum and Pejtersen, 2000]. It does not even need to be project-specific information, the social network accessed through face-to-face conversation, by phone, or by email is a crucial backbone for employees with information needs and contacting colleagues can be the main fall-back strategy to obtain required information [Lund and Ørnager, 2016, Laqua et al., 2011]. Expert and expertise finding are particular problems in large organisations [Hawking, 2010, Balog et al., 2012]. One of the main reasons that search for people and expertise is centre-stage in enterprise search is because a large proportion of an organisation's intellectual capital has always been tacit knowledge [Baumard, 1999]. Another reason could be the frequent rotation of staff between, e.g. the organisation's headquarters and field stations. Examples are large UN organisations like the World Food Programme (WFP) which employs 14,000 people worldwide and finding a topic expert has been found to be an imperative information need among staff of the WFP [Lund and Ørnager, 2016]. The need of an expert or of expertise might also arise in an industrial setting in cases of *real-time* collaborative troubleshooting needs, as is the case, for example, for power plant equipment maintenance [Paul, 2016] or to help the mobile work force of a telecommunications company [Albakour et al., 2013].

Finding people responsible for certain tasks [White, 2015b], finding name, location, email, job role, phone number of an employee etc. are all common within an organisation [Guy et al., 2012]. A study of three different companies and organisations of differing sizes confirmed that lookup of staff contact details and employee information feature prominently as among the most common intranet information seeking categories [Stenmark, 2010]

Enterprise-search related information needs come with their own problems that need to be addressed. For example, a simple name lookup might be complicated by language issues [White, 2015b]. It might also be that employees do not really know the exact name of a person they have in mind and whose contact details they want to obtain. A study investigating more than a million logged queries collected over four months on *Faces*, a large-scale people search application embedded within the IBM enterprise, found that more than a third of all submitted search tokens did not have an exact match suggesting the need of some sort of fuzzy search[8] [Guy et al., 2012], this is partly attributed to the global enterprise environment with names that are difficult to spell or pronounce. A survey conducted as part of the same IBM study uncovered that among the most common scenarios of people search were to locate people that sent emails to the searcher's inbox, appear in their calendar meetings or participate in chats or phone calls.

There are various approaches to solve such problems including customisation and low-level tuning. Hashing-based people search algorithms that can be employed to learn hash functions that map similar names to similar binary code words in a language-independent space have been proposed to support fuzzy search of people names in an email context [Ramarao et al., 2016]. This goes beyond simple lookups in dictionaries, gazetteers or synonym lists.

Additional support might also be used to enhance the search experience. For example, IBM's employee directory was reported to have links to audio files to help with pronouncing names [Pernice et al., 2006].

A study investigating the information-seeking behaviours of engineers concluded that engineers search for documents to actually find people that are the right point of contact, they search for people to obtain relevant documents from them, and they interact socially [Hertzum and Pejtersen, 2000]. Two product-development organisations were studied, *Novo Nordisk*, the then world's largest producer of industrial enzymes, and *Danfoss*, a large Danish manufacturer for heating and refrigeration systems. In the *Novo Nordisk* case study it was found that engineers had a strong preference for obtaining new information from

[8] This type of problem has obviously become much less prominent in search in general with the introduction of auto-complete suggestions in recent years.

people while documents were considered important but "usually concomitant". In the *Danfoss* context an important information-seeking activity was to obtain information about people so that the expertise of other companies and potential sub suppliers can be determined. The authors also conclude that "while concrete product information can be found in documents, context information must be obtained from people."

The enterprise search setting of these common search scenarios provides a clear contrast when compared to other types of search such as Web search, desktop search, or site search.

3.4.2 Email Search

Email overload has long been recognised as a pattern referring to the use of email for functions that it was not designed for, i.e. as a plain asynchronous communication tool [Whittaker and Sidner, 1996]. This includes storing personal names, addresses but also the use of email for task management. Not much has changed since then in principle although overload is different between personal and work-related email [Grevet et al., 2014]. It then looks like a natural conclusion that email search in general represents a large proportion of enterprise searches [Guy et al., 2012]. Email search can actually be considered one of the biggest requirements in an enterprise search setting, not necessarily to identify a specific person or piece of information but often driven by legal discovery. At the same time few organisations fully satisfy this requirement [Miles, 2014].

There are of course inherent problems with emails as a business communication and management tool which include the circulation of attachments and if multiple versions are circulated, then overloaded inboxes, for example, might lead to access and editing of an outdated version [Massey et al., 2014]. This is more of a problem when conducting collaborative work but affects also general search.

Email search has found its way into the academic research community primarily through the TREC Legal Tracks 2010 and 2011 [Cormack et al., 2011, Grossman et al., 2012]. The track was aiming to model real use cases and the findings are therefore not just of academic interest. We will talk more about this in Chapter 4 which looks at evaluation.

3.4.3 Business Intelligence

Searching for business knowledge or intelligence followed by search across emails and search for customer-related content appear to describe the most prevalent information needs for *advanced* search according to the AIIM survey with some variation such as 'Freedom of Information' requests being among the top needs in government organisations and public services [Miles, 2014]

This is where enterprise search needs to be seen as part of a bigger framework [Chaudhuri et al., 2011, Ben-Yitzhak et al., 2008]. More broadly speaking, enterprise search goes beyond traditional finding problems and is also used for information integration, discovery, collaboration and knowledge management, compliance, and records management [Delgado et al., 2005].

3.4.4 Exploratory Search

Lookup-based systems are not always most suited for an information need at hand. These needs might for example not be well-defined, they might describe a broad subject area or be aimed at decision-making in which case exploratory search as an interactive activity of querying and collection browsing is more appropriate [White and Roth, 2009, Marchionini and White, 2009]. The information needs in exploratory search might involve multiple query iterations or even multiple sessions, and apart from being potentially open-ended the information needs can be persistent and multi-faceted. Put differently, search might at its best be a conversation, an iterative, interactive learning process [Morville and Callander, 2010, p.9].

Exploratory search by information professionals plays an important role in an enterprise environment, e.g. [Stenmark, 2008, Cleverley et al., 2017]. Cleverley and colleagues conducted a user study for which they identified some realistic and multifaceted information needs in the oil and gas industry around gravity and magnetics in a particular region which would typically form part of a much larger set of search tasks related to the topic. Stenmark on the other hand started with the search engine log files from a corporate intranet and then clustered searchers according to their interaction behaviour identifying different types of users including those that tend to be more interested in recall and are characterised by long interactions with the search engine, i.e. 'explorers'.

However often little attention to supporting exploratory search is paid in enterprise search [White, 2015b].

3.4.5 Discovery

E-discovery [Oard and Webber, 2013] might not be what one considers a typical search need within an enterprise but it very much is or at least should be given that legal discovery cases which require an organisation to uncover all electronically held records about specific customers, suppliers, contracts, cases etc often come out of the blue, yet, 74% of organisations report they do the process manually [Miles, 2014].

Data discovery more generally is a common problem users in enterprises face: finding information in relational databases, e.g. [Cortez et al., 2015]. We do not expand on this area here as it is, unlike e-discovery, mainly a DBMS issue.

3.4.6 Other Information Needs

There are many other information needs within an organisation and they vary depending on the type of business, the size, the location etc. There could be the need to find out about the history of purchases and support calls related to a specific customer [Chaudhuri et al., 2011] as well as item and product search [White, 2015b, Johansson and Westerling, 2009] etc. Just to illustrate the broad range of other possible problems to be addressed by enterprise search we outline some additional scenarios here:

- *Remote area mining operations*: assume a multi-national mining company established a large mine in a remote area in a developing country, they installed expensive and complicated mining machinery (conveyor systems, ore crushers, transport systems, hole boring and explosive control systems) which must be maintained and perhaps repaired. In the absence of an internet connection, maintenance and repair relies on very extensive manuals in electronic form. Effective search over this documentation has the potential to significantly improve productivity.

- *Petroleum exploration*: oil and gas exploration companies typically accumulate drilling reports in heterogeneous formats [Hawking, 2004]. These describe (usually in informal and non-standard terms) the features and hazards encountered at various depths during a drill. When a new well is being planned, knowledge of the experiences in nearby wells can save time and trouble. There is a search problem complicated by the heterogeneity of formats and language, the geographical proximity dimension and the importance of depth.

- *Help desk call centres*: in cases where the help desk agents rely on search to respond to queries, e.g. [Albakour et al., 2013], call centres are a case where it might be possible to evaluate the real world impact of better search (by assessing the quality of the answers or the time taken for the call).

- *Tender responses, bidding*: responding to tenders and bidding for work are critical activities for consulting and contracting organisations. Proposals may run to hundreds of pages [Hawking, 2010] and putting one together is usually done under intense time pressure. Effective enterprise search can increase bid quality and reduce time taken by finding staff with required expertise, capability statements, customer references, and other material to include in the bid.

- *Customer relationship management (CRM)*: although many organisations use a specific CRM system, it is often the case that information which is incomplete or not up-to-date in the CRM can be found in email or internal discussions.

3.5 Search Context

Having surveyed the basic characteristics of enterprise search we have identified some rich contextual information and information about the searcher being readily available to the search system. This is a clear advantage over Web search, for example, where the available context is more limited and the user information typically needs to be guessed via implicit signals. Let us look at how *context* and *user* information may then be usefully incorporated in the search algorithms. We start with context.

Contextualising search has become a popular research area, much of it concerned with making use of the user's local context as well as past search patterns [Melucci, 2012, Ruthven, 2011]. However, enterprise search offers a great advantage over Web search and other applications as a user is defined by his or her role within an organisation and therefore much is known about them and the tasks they are likely to perform even before submitting anything at all [Hawking, 2010]. Furthermore, the smaller size and well-defined topic and task domains compared to the Web make enterprise information spaces particularly suited to implementations of contextual search [Freund et al., 2005]. Contextual information comes in different flavours such as user context and task context.

Let us start with the *user* context. In exploratory search, subject familiarity, job role, and personality are all potential contextual factors affecting the searcher's perception of navigational aids such as query suggestions [Cleverley and Burnett, 2015b]. Contextual information derived from a user's email communication and calendar within an enterprise setting provides an information-rich environment that can be used to build user models for automated contextual search across different sources such as the user's work space, external sources and emails [Lu et al., 2011]. Moving beyond the individual user and considering the user as an employee in an enterprise management structure, knowledge about a user's membership of a particular group (such as sales department or production department) can be used to select a group-specific thesaurus when performing automatic query reformulation [Hawking, 2010]. More generally, domain-specific dictionaries (e.g. acronyms and person names) have indeed been shown to improve precision for search within an organisation [Zhu et al., 2007].

We could also take the *task* context to derive useful information to improve the search process. Within an enterprise users are employees that conduct tasks related to the business at hand, and the information sources to be accessed are dominated by the business operation [Hawking et al., 2005]. In fact, the nature of a user's task might actually be derived from the application a user is currently running [Hawking, 2010].

Contextual information can also be captured at *document* level. For example, the location of the document as defined by the data structure such as its place in a corporate taxonomy, the structure of the underlying database table, or a site map, all provide additional metadata applicable in the search process [Delgado et al., 2005].

A more heterogeneous context is *time*. Time is generally considered an important contextual factor in search systems but has been underappreciated until recently [White, 2016, p. 295]. Utilizing temporal information has attracted more attention recently, in particular the extraction of temporal information from documents but equally information related to document creation and document focus time [Kanhabua et al., 2015]. Document creation time has already been pointed out as a major issue in enterprise search, to identify the latest version of a document, for example. A separate issue related to time is the workflow in organisations which differs from queries submitted to a Web search engine. One way of how this surfaces are the query patterns that are closely linked to the organisation at hand, e.g. seasonality of certain queries submitted to a university site search engine which reflects the annual student admission cycle among other things [Dignum et al., 2010]. In line with typical enterprise search-related problems such as sparseness of link structure and anchor text and use of domain-specific jargon, there is another common feature – a strong presence of dynamic terminology [Bao et al., 2012b]. An example of a topic that changes over time within an enterprise is the query *"benefits"* submitted to the IBM intranet for which the most important page at the time of publication was *netbenefits* [Vaithyanathan, 2011]. A seasonal example is the query *"timetable"* reported to be frequently submitted within an intranet of an academic institution and which in autumn appears to be aimed at the teaching timetable, whereas in spring to be more likely to be aimed at the exam timetable [Kruschwitz et al., 2013].[9] Examples from the site of a different university support this point as searches for the library spike when term papers are due and searching for ways around the campus are popular when the semesters begin [Rosenfeld, 2011, p.47].

Other contexts that are more in line with Web search are the detection of the class of devices a searcher is using so that, for example, the display of results can be adjusted to a mobile device or telephone [Hawking, 2010]. The user's search *session* can also be interpreted as some form of contextual information [Rosenfeld, 2011, p.83]. Context is particularly important in supporting exploratory search [White and Roth, 2009]. Such searches naturally go hand in hand with longer sessions.

There is a lot of potential to apply contextual information but in reality very little of this gets utilised in enterprise search systems.

3.6 User Modelling

Information retrieval systems are becoming increasingly personal and contextual [Hofmann et al., 2016], and Figure 2.2 made a strong case for the need to distinguish different users involved in the overall enterprise search ecosystem. In fact, the very setup of an enterprise environment offers the chance to make use of a lot of information about the users as the user base is defined by the enterprise. Users do not simply 'opt in' by submitting searches as they would when accessing a Web or a news search engine. This user base can still be substantial in size, e.g. 400,000 employees in a multi-national company like IBM [Guy et al., 2013] that need to be served around the clock, although typically an enterprise search user base is smaller than that as illustrated in Table 2.5.

[9] This could of course also be interpreted as an example of finding the most up-to-date *version* of a document.

User models and personalisation have been studied extensively, mostly in a Web search context [Teevan et al., 2010, Pitkow et al., 2002] but much of it appears to be applicable in an enterprise setting too. Within an enterprise context one can distinguish persistent profiles that exploit a user's role in addition to a profile that captures past access patterns and session-based profiles [Mukherjee and Mao, 2004]. History-based (including session-based) profiles that capture a user's query and click behaviour have been shown to offer substantial benefits in predicting the relevance of documents [Bennett et al., 2012], and these profiles are now commonly applied in Web search. Role-based profiles on the other hand are very specific to the organisational structure of an enterprise, although the organisational structure on its own does not necessarily best reflect the users' needs as expressed in their search behaviour [Carter et al., 2014].

More broadly speaking, i.e. looking at the personalisation literature beyond enterprise search, a user model aims at capturing a user's or a user group's interests [Teevan and Dumais, 2011]. There are a number of common methods for structuring such models, e.g. [Gauch et al., 2007]. Models can be built from queries that users submit to search the collection by building query flow graphs, for example [Deng et al., 2009, Boldi et al., 2009], from anchor text [Kraft and Zien, 2004], from mining term association rules [Fonseca et al., 2003], or by extracting term relations from documents, for example [Kruschwitz, 2005, Sanderson and Croft, 1999]. They can aim at modelling individual user's interests [Teevan et al., 2010] or cohorts of users [Yan et al., 2014]. Models can be explicit, where users input topics of interest, or implicit, where those interests are inferred from their actions [Teevan and Dumais, 2011]. However, explicit models have several drawbacks such as the time it takes to build them and their static nature. The types of implicit data used to construct profiles can vary [Teevan and Dumais, 2011], e.g., the analysis of log records, has been shown to be good at approximating explicit feedback, and query log analysis has developed into a very active research area [Jansen et al., 2009, Silvestri, 2010].

Cohort modelling has been effective for Web search [Yan et al., 2014] as well as for exploration/navigation in a site search context [Alhindi et al., 2015], in the latter case the issues arising in building models that represent user groups' search interests are similar to enterprise search, including sparsity and recall as an important metric in a context with little or no redundancy.

Personalisation applied to people search within an enterprise setting (via identification through cookies) has been used to boost persons in the result set found in the searcher's network, management chain, location, organisational unit, country etc. [Guy et al., 2012]

The content and structure of the email communication network within an enterprise offers a specific way of modelling users. If we treat recipient recommendation for emailing in an enterprise as a specific enterprise search problem, it is interesting that the utilisation of the content and structure of the (email) communication network in addition to the actual email content outperforms approaches that only use the email content [Graus et al., 2014]. Similarly, in a content recommendation scenario in which a user is typing an email it was found that topic models representing the user's interactions with others (persons, groups and email threads) and contextual information yield better performance than a recommendation based on only one of these or the email message alone without a user model [Lu et al., 2011].

A lot of the signals used to model users and to personalise the search experience are generic enough to be collected in various search environments, also, the explicit organisational structure offers potential to model individual users or cohorts of users. Despite all this, in practice very little user profiling tends to be applied in enterprise search environments, e.g. as reported by 78% of respondents of the '2015 Enterprise Search and Findability Survey' [Findwise, 2015]. Part of the reason is data sparsity as there is the problem that even within the largest enterprise the user population tends to be relatively small and, for example, many exploratory queries will not have been posed to the system at all before [Cleverley and Burnett, 2015b]. Nevertheless, given the growing adoption of user models in other search systems [White, 2016], there is huge potential for enterprise search as well in this area.

The organisation of users around roles that they are assigned to in an enterprise context has already been discussed and different roles within an enterprise may well give rise to different profiles when modelling users [Hawking, 2010].

3.7 Tools, Frameworks and Resources

The enterprise search landscape has been transformed in recent years, in particular due to a thriving open source community that provides powerful frameworks such as *Apache Lucene/Solr* and *Elasticsearch* which are highly scalable and can be deployed in a distributed environment. At the same time these frameworks are feature-rich and customisable and are actually *not* necessarily working well when installed out of the box [Turnbull and Berryman, 2016, p.6]. The last point sounds like a paradox but we will expand on this in more detail in Chapter 5 where we argue that enterprise search simply will not work when just applied in an out-of-the-box fashion.

However, after a period of consolidation in the enterprise search field, *Microsoft SharePoint* now appears to dominate the market. In fact, Sharepoint has developed into a search-based application in which much of the functionality is driven by search [White, 2015a].

Given the fast pace in practical developments we refer the reader to online resources such as KMWorld[10] and the blogs by Miles Kehoe[11], Steve Arnold[12] and Martin White[13]. Extensive usability advice on enterprise search is provided by Jakob Nielsen[14].

Before we discuss how to make enterprise search work we will however explore another important aspect of search which again lends itself to identifying striking differences between enterprise search and other application areas, namely evaluation.

[10] `http://www.kmworld.com`

[11] `http://www.enterprisesearchblog.com`

[12] `http://arnoldit.com/wordpress/`

[13] `http://www.intranetfocus.com/blog`

[14] `https://www.nngroup.com/people/jakob-nielsen/`

4

Evaluation

The discussion so far has uncovered a number of distinguishing factors that characterise enterprise search and which have a direct impact on how systems are to be evaluated. One conclusion that we can draw at this point and which we will expand on in this chapter is that standard IR evaluation metrics are not necessarily directly applicable to enterprise search – at least not without making them fit the enterprise context and considering a range of additional metrics. In practice, enterprise search evaluation is performed as part of a continuous search testing cycle which will be discussed in more detail in the next chapter that explores how to make enterprise search work.[1]

This chapter will start with the concept of *relevance* in enterprise search and discuss how this drives the choice of metrics used to evaluate enterprise search systems. We will then look into different existing evaluation paradigms and how they can be applied in the enterprise search context. Each such paradigm can naturally only address certain aspects of a system. We will discuss evaluation campaigns for those

[1] This is more of an idealised view. A fairly common scenario is in fact that enterprise search is seldom tested, rather developments are driven by complaints from users.

aspects of enterprise search that have attracted academic interest. We also look at existing test collections (again only covering aspects of the full enterprise search application scenario). Lessons learned are drawn mainly from the more academically driven evaluation campaigns.

For readers who are less interested in the scientific perspective on evaluation and who are, for example, faced with the decision as to what enterprise search solution to acquire, we recommend the practical guide to specifying and selecting a search application in [White, 2015b, p.171-190].

4.1 Relevance and Metrics

Standard Web search relevance measures do not necessarily apply in enterprise search. For example, employees need to find the best/correct and not the most popular information [van der Lans, 2013],[Morville and Rosenfeld, 2006, p.431]. In addition to that, the notion of a 'good answer' is different from Web search and the target of a search within a work space environment is actually often to find the 'right answer' [Fagin et al., 2003]. Finally, from the perspective of the users, relevance is often relative [Rosenfeld, 2011], influenced less by the user's personal interest but more by his or her domain knowledge [Wu et al., 2014].

Alongside a different definition of what it means to be *relevant*, we also observe that this applies equally to *metrics* that are commonly used as benchmarks in Web search.

Our main focus is on technical measures, more specifically on effectiveness rather than efficiency, which aligns well with the approaches dominating the research community. If we want to adopt precision and recall as two commonly applied (families of) metrics in IR research, then recall appears to be the more important measure in an enterprise context, a feature enterprise search shares with many 'niche' application areas but which is in contrast to precision-driven Web search as illustrated in Table 2.7.[2] Having said this, precision and recall are highly context-dependent in an enterprise search setting [White, 2015b]. At a

[2] Obviously, within an enterprise there will at the same time always be certain types of queries that are clear-cut *high precision* queries, e.g. trying to find a document that describes the process for hiring a new employee.

high level these metrics are still applicable in certain enterprise search application contexts. For example, for compliance purposes it is essential to achieve high recall so that no documents remain undiscovered which could affect the outcome of a court case [White and Nikolov, 2013].

Manning et al. suggest 'user productivity' as a more relevant metric in an enterprise context, i.e. the time spent on looking for information they need [Manning et al., 2008, p.156].

White, very much framing enterprise search as a problem that needs to be addressed not just theoretically but will have to work in realistic applications, suggests five components to enterprise search evaluation: technical performance, query performance, usability and accessibility, search satisfaction, business impact [White, 2015b]. To do this properly, it is important to go beyond search log analysis and user satisfaction surveys and assess the impact on the business performance [White, 2015a]. Some of the technical metrics proposed by White that illustrate the pragmatic aspect of enterprise search and that distinguish it from general Web search include:

- Percent of searches that return zero results
- Percent of sessions that use search
- Average time spent after searching
- Average time spent before searching.

A non-technical metric that might be used to supplement other benchmarks is uptake, i.e. proportion of employees who use the intranet, the Nielsen Norman Group recommend it to be at least 75% [Pernice et al., 2006].

More general considerations for information-seeking in an enterprise setting include the observation that “easy access is of paramount importance” as “the cost associated with using an information source is the most important determinant of its use” [Hertzum and Pejtersen, 2000]

4.2 Evaluation Paradigms and Campaigns

There are many reasons why evaluating an enterprise search system might be desirable including scientific enquiry, product testing and for internal purposes of the company [Hawking, 2010]. Our main interest is in evaluations that are based around controlled experiments, i.e. that provide the rigour of an academic approach.

Evaluation has always been an integral part of information retrieval research, much of it in an academic context. While the general idea is to assess how 'well' a search engine might perform there are many different angles and aspects one could consider which then leads to a wide range of different approaches to conduct the evaluation. This has resulted in a number of commonly used evaluation paradigms starting with the Cranfield paradigm of the 1960s [Cleverdon, 1997]. Here we will distinguish between *technical evaluations* (or system-focussed evaluations) which look at measures such as effectiveness and efficiency of a system, and *user studies* that either involve real users or make an assessment based on real or simulated user behaviour.

An alternative would have been to distinguish *online* and *offline evaluation* approaches [Hofmann et al., 2016], with the Cranfield paradigm being a typical offline approach and where online evaluation can be defined as the evaluation of a fully functioning system applied in a natural usage environment and measurements made based on implicit user signals such as clicks and dwell time. This type of evaluation is commonly applied in industry, for example by Web search engines, e.g., through A/B testing [Kohavi et al., 2007] and interleaving [Joachims, 2002, Radlinski and Craswell, 2010] as long as the pool of potential users and their search activity is on a large enough scale. In an enterprise this is much less common[3] which makes this type of evaluation more difficult to employ effectively. As a result, evaluation within an enterprise is typically not conducted in the same way and evaluations often look at specific aspects of the search infrastructure [Wu et al., 2014]. The implication is that any such evaluation only approximates the natural usage environment. For technical evaluation metrics this

[3] Except for larger-scale companies or organisations.

would normally mean that queries are sampled with the aim of obtaining a representative set of information needs. For user studies the implication is that any analysis of their behaviour as derived from a study will have a slightly artificial touch to it no matter how hard one tries to make the setting as natural as possible. Even if the analysis is conducted using realistic log files of a search system one can only work on assumptions made by analysing the implicit user signals obtained from the logs.

4.2.1 Technical Evaluation Campaigns

Technical evaluations of enterprise search systems are not commonly reported in the academic literature. However, some specific aspects of enterprise search have attracted interest from the wider research community when they were the subject of investigation as part of the annual Text Retrieval Conference (TREC) series[4]. The TREC *Enterprise Search track* had a distinct focus on search over emails and search for experts within an organisation. The TREC Legal Track was primarily concerned with an e-discovery scenario. We will briefly discuss both campaigns.

TREC Enterprise Search Track

The TREC Enterprise Search track ran from 2005 until 2008. It was introduced to conduct experiments with enterprise search data, namely intranet pages, email archives and document repositories, that reflect realistic search settings within organisations [Craswell et al., 2005]. *Email search* and *expert search* were the tasks in 2005 and 2006. While expert search continued in 2007 and 2008, email search was replaced by *document search* with the introduction of a new document collection [Bailey et al., 2008a]. Originally, the document collection comprised a crawl of the public pages of the World Wide Web Consortium (W3C)[5] which in 2008 was replaced by a crawl of the public-facing web site of the Australian Commonwealth Scientific and Industrial Research Or-

[4] `http://trec.nist.gov`

[5] `*.w3.org`

ganisation (CSIRO)[6] (both discussed below). Both collections suffered from the fact that they only contained publically available pages which makes the collections only a rough approximation of a realistic enterprise setting. However, the track has managed to generate a lot of interest in the research community, primarily the expert finding task, with rapid progress in algorithms, modelling and evaluation having been made throughout the four years [Balog et al., 2009].

TREC Legal Track

The TREC Legal Track ran from 2006 until 2011 with a focus on *e-discovery* of business records and other materials [Baron et al., 2007]. Despite some variations in the task definitions throughout the years the main focus remained to identify as many as possible (ideally all) relevant documents from a collection that are considered responsive to a legal request [Grossman et al., 2012].

Tasks included both interactive settings as well as batch processes. The main change throughout the six years was perhaps the introduction of a new test collection based on *emails* which in 2010 replaced a collection of largely *scanned documents* used in the early years. We look at these test collections in more detail further down.

The TREC Total Recall track[7] (introduced in 2015 after the Legal track had finished) adopted a similar idea in trying to identify *nearly all* (or a "reasonable" number of) relevant documents for a task at hand, e.g. [Grossman et al., 2016].

4.2.2 User Studies

In line with system-focussed evaluations, there is a noticeable lack of studies of user-based experiments in the refereed literature which investigate enterprise search systems, in particular such studies that are based on controlled experiments with well-defined experimental settings, e.g., as outlined in Kelly [2009]. Of the evaluations that have been conducted, most rely on a fairly small sample of users and tasks.

[6] `*.csiro.au`

[7] `http://trec-total-recall.org/`

For example, Freund and Toms conducted a task-based study of enterprise search behaviour of software engineers [Freund and Toms, 2006]. Simulated search tasks were used to study the search behaviour of employees in a professional government setting (the Danish tax authorities) [Svarre and Lykke, 2014], and Hansen and Järvelin studied real work tasks conducted by ten professional patent engineers [Hansen and Järvelin, 2000].

Exploratory search was investigated by a sample of employees within a large oil and gas operator [Cleverley et al., 2017]. Barriers of enterprise search systems were explored through guided observations of engineers working in research and development of an organisation in the vehicle industry [Stocker et al., 2015]. Participants conducted search tasks and their search experience was recorded.

Log-based studies based on the search logs of varying enterprise settings have been reported by Stenmark and colleagues. These include studies of the search in a large manufacturing company (SwedCorp) [Stenmark, 2005b,a, Stenmark and Jadaan, 2006], and an unnamed large manufacturing company [Stenmark, 2007].

Other user studies include surveys conducted among employees, e.g. practicing petroleum engineers from different organisations who were asked to assess the quality of query suggestion terms in exploratory search [Cleverley and Burnett, 2015b], employees of different types of organisations to obtain insights into people's work-related information seeking behaviour [Stenmark, 2010], as well as studies into the general use and attitudes towards an internal company's intranet, e.g., [Stenmark, 2006].

All these studies provide an interesting insight into enterprise search but all of them come with the caveat that they might not easily be generalisable beyond the specific enterprise setting, the chosen set of users and the type of domain-specific tasks identified as being relevant for the given setup.

There is certainly room for much more research as any additional study will provide another piece of what appears like a giant jigsaw.

4.3 Test Collections

Progress in IR has benefited enormously from the availability of test collections that allow the comparison of different search algorithms using a commonly agreed evaluation framework. Test collections have become such a core feature of evaluation in information retrieval that evaluation using test collections is now a research area in its own right [Scholer et al., 2016].

A typical test collection consists of a set of documents, a set of topics/queries and a set of relevance judgements (or 'qrels', query relevance sets) that specify the relevance of each document given a query [Sanderson, 2010]. Test collections have become widely available for a large number of different search scenarios as exemplified by the broad spectrum of search tracks in major evaluation campaigns such as TREC, CLEF[8], NTCIR[9] and FIRE[10]. These tracks cover application areas ranging from Web search to spam identification, from genomics to chemical IR and also include enterprise search.

However, while test collections have been made available for the TREC Enterprise Search track, it has to be pointed out that these collections are far from representative examples of enterprise repositories in general. Hawking points out that standard test collections and evaluation metrics for enterprise search are not easily available [Hawking, 2004]. This should come as no surprise given that there is one major obstacle which White and Nikolov highlight in their analysis of the enterprise search market in the European Union when they argue that it is difficult to push forward the state of the art in enterprise search as "it is impossible to construct, or use with permission, enterprise-type collections of information. Companies are not willing to provide access to what is regarded as confidential information and, even if a collection could be constructed, the range of queries that would be suitable to use as test queries would be constrained by the information content." [White and Nikolov, 2013]. It is more complicated than that as organisations do not even want to let their competitors know what their

[8] http://www.clef-initiative.eu/

[9] http://research.nii.ac.jp/ntcir/

[10] http://fire.irsi.res.in/fire/

employees might be searching for let alone what sets of documents they might be searching [Hawking, 2010]

The problem is very similar to a desktop search context, i.e. the difficulties of working with personal collections, the difficulties in building up personal collections, the lack of established or standardized baselines and evaluation metrics, and partly as a result of this the lack of commonly available test collections [Elsweiler et al., 2010].

While enterprise search test collections are not freely available for research purposes, they are being constructed and applied *within* enterprises to continuously tune the performance of the search framework in place. Just like with any test collection the aim needs to be to build a representative and reusable set of queries and judgements faithfully modelling a realistic enterprise search setting, and using log data to bootstrap this process is one approach in which a uniform random sample of queries might be drawn and corresponding answers to the information need assumed to be triggering the query are identified [Hawking, 2010]. An alternative is to build domain-specific test files capturing actual user needs in the enterprise domain at hand without reverse-engineering via log files, e.g. [Hawking et al., 2009].

Let us go one step further by putting ourselves into the shoes of a search administrator within a company. Considering that enterprise repositories are changing over time and that fixed test collections are therefore of limited use, a new paradigm of test-driven relevance tuning has been proposed to allow content owners/creators and search administrators/developers to collaborate on improving relevance scores for particular test queries [Turnbull and Berryman, 2016]. We simply use this example to hint at the rather complex picture of enterprise search evaluation and will discuss such practical issues in the next chapter when we explore how to make enterprise search work.

Collections resembling aspects of enterprise search have nevertheless found their way into the public domain. They tend to focus on either expert search or email search – both common enterprise search needs.

4.3.1 Expert Search

Expert search describes a core set of information needs in an enterprise setting as we had already identified. For this specific type of problem test collections have been made available to the research community. The only ones that have been made freely available are:

- The World Wide Web Consortium Enterprise Search Test Collection (W3C) is a crawl of the public W3C sites, described as not comprehensive but still representing a significant proportion of the public W3C documents [Craswell et al., 2005]. In 2005, *expert* judgements are derived directly from the information contained in the documents (working group membership was used as ground truth). In 2006, the annotation was done by the TREC Enterprise Search track participants [Soboroff et al., 2007].

- The CSIRO Enterprise Search Test Collection (CERC) presents a gold standard for *document search* and *expert search* [Bailey et al., 2007] but with fairly limited structure. In fact, the collection is based on a crawl of the CSIRO Web site together with information need statements and relevance judgements for some real tasks that arose from communicating the organisation's science to the public and potential partners [Hawking, 2010]. As such one of the novel points of this collection was that it realistically modelled an actual task. In addition to providing gold standard judgements by CSIRO science communicators who originally proposed the topics, comparisons were made with other annotations, namely silver standard (science communicators from outside CSIRO) and bronze standard (TREC participants with neither task nor topic expertise) [Bailey et al., 2008b].

- The Tilburg University Expert Collection (UvT) [Bogers and Balog, 2007] represents a Web site with four main features, namely (1) it is clean, heterogeneous, structured, and focussed, but comprises a limited number of documents; (2) it contains information on the organizational hierarchy; (3) it is bilingual (English and Dutch); and (4) the areas of *expertise* assigned to individuals (i.e., the experts) are provided by the employees themselves [Balog et al., 2007]

However, as a reminder we should reiterate that anyone wanting to use these collections to improve the performance of a search system employed on a different enterprise will face the problem that the queries are obviously constrained by the information content [White and Nikolov, 2013]. More importantly, no two enterprises are the same, which complicates things even further.

4.3.2 Email Search

Emails collected via the inboxes of employees *within an enterprise* represent a substantial aspect of the wider area of enterprise search as discussed in Chapter 3. Both the everyday search across emails as well as search triggered by legal discovery demands were highlighted as important enterprise search applications.

Collections that emerge from email repositories are hence a valuable source to identify patterns within an enterprise context. Two substantial collections of this type have so far been made available (one freely available and the other one through subscription).[11] Note however, that neither of the two collections is a *test* collection as for that to be the case they would require a representative set of topics and judgements, not just a corpus of emails. They nevertheless represent valuable resources that have been used, for example, in the TREC Legal Track (Enron) or to predict enterprise email reply behaviour (Avocado) [Yang et al., 2017].

Enron Corpus

The Enron corpus is a good example of a more specialised dataset which only became available as a result of external circumstances, in this case at the conclusion of the investigation into the collapse of Enron. The collection was made available by the Federal Energy Regulatory Commission (FERC).

[11] We should also point out that the W3C corpus has been annotated for email search and email discussion search.

The Enron corpus comprises a total of 619,446 messages belonging to 158 users in the raw corpus which were then processed to result at 200,399 messages (by 158 users) [Klimt and Yang, 2004]. This processing step involved the removal of certain folders that were computer-generated and a de-duplication step.

The TREC 2010-2011 Legal Track used a processed version of the Enron corpus that identified 455,449 *canonical* messages, e.g. duplicates were removed; in addition to emails there are 230,143 attachments giving a total of 685,592 documents [Cormack et al., 2011, Grossman et al., 2012].

Avocado

Similar to the Enron corpus, the *Avocado Research Email Collection* is a corpus of emails and attachments of communication within a now-defunct IT company ("AvocadoIT" is the pseudonym used to refer to the company). The collection is distributed via LDC [Oard et al., 2015]

The Avocado corpus is bigger than Enron, the total number of emails being 938,035 of which there are 323,574 duplicates, hence 614,461 non-duplicate emails. In addition to emails there are 110,023 attachments and 298,022 extracted files. The collection also contains contact details, appointments, stickynotes etc. giving a total of 869,777 non-duplicated items.

4.3.3 Other Enterprise Collections

The TREC 2006 Legal track (and subsequent tracks in the following years) used a document collection that itself emerged from real legal cases. The collection, IIT CDIP 1.0, comprises documents related to a set of smoking and health-related lawsuits which were released under the tobacco 'Master Settlement Agreement' (MSA) [Baron et al., 2007]. The Illinois Institute of Technology (IIT) produced a snapshot of a subcollection with a total of 6,910,192 documents containing the scanned images, OCR-processed text and metadata. The wide range of document genres, including email, reports, memos, budgets, minutes, letters among others, resembles a typical enterprise search setting (as discussed in Section 2.5.1) but is also typical for a legal/discovery setting [Baron et al., 2007].

4.4 Lessons Learned

There are a number of important findings from the discussion in this chapter that provide an explanation as to why enterprise search has attracted little attention in the academic community. These include:

1. Due to the very nature (and value) of a document collection within an organisation or company it is hardly possible to get hold of realistic test collections. Exceptions are those that have been constructed for very focussed information needs and collections that have been made available due to some legal rulings.

2. Even if it was possible to distribute a complete enterprise information infrastructure, it would be difficult to generalise any findings obtained from working with this distribution as each enterprise collection will differ from any other.

3. Evaluation metrics that are relevant in enterprise search do not easily align with those proposed in academic IR research.

The findings presented here may sound a bit ‘pessimistic’ but what this tells us is that evaluation needs to be tailored towards a specific use case, needs to be conducted on a continuous basis and, concluding from this, if these requirements are not addressed properly, enterprise search will eventually fail.

What would help are methodologies for evaluation that are developed in collaboration between academia and industry which can easily be applied in an enterprise setting by search administrators who will have access to suitable test data. The methodologies could be developed using public collections.

The next chapter will investigate what needs to be done to make enterprise search work.

5

Making Enterprise Search Work

"Recognize that enterprise search is an approach and not a technology." [White, 2015b]

Having laid the foundations of what defines enterprise search in Chapter 3 and having contextualised it with the evaluation literature in Chapter 4, we will now focus on what needs to be done to make it actually work in practice. Enterprise search is more than just a search engine, i.e. a purely technological issue. First of all, to work properly, it needs to be seen as a continuous process rather than a one-off project [Findwise, 2015]. It relies on a multi-disciplinary search team [White, 2015b]. Finally, making enterprise search work is in many ways different to making other search applications work, for example the existence of an explicit strategy for enterprise search within an organisation has a significant (positive) effect on user satisfaction in regards to the search engine [Stenmark et al., 2015, Findwise, 2016].[1] More generally speaking, achieving search excellence should be seen as a journey and not just a project [White, 2015b]. This is all the more important given the time an average knowledge worker is estimated to spend searching for

[1] yet only a small fraction of businesses appear to have such a strategy in place [Miles, 2014]

information which is typically in the region of 20% to 30% [Feldman and Sherman, 2001, Doane, 2010].

Having said all this, there appears to be a startling contrast between the requirements and expectations on what an enterprise search system should be able to do and what is actually being done to make enterprise search work. The AIIM survey captures this conundrum nicely by summarising that 71% of polled organisations consider enterprise search to be vital or essential to productivity and effectiveness but 58% show little or no search maturity [Miles, 2014]. Search maturity here comprises features such as an agreed corporate taxonomy or vocabulary of terms, a metadata standard across different repositories, a dedicated budget, dedicated and trained search staff, an owner of search, and a search strategy. Even in the largest organisations (about a third of the more than 400 respondents representing companies with more than 5,000 employees) more than half had not a single one of these features present. Some core issues of how enterprise search can actually be made to work are discussed in more detail in the following sections.

The discussion so far makes a strong case for continuous support and customisation of the search environment. What can in part be left to the automatic indexing tools in other search contexts is simply not sufficient to make enterprise search work. Tools to support the maintenance of an enterprise search system are essential and so is appropriate support for users accessing the system.

This chapter will discuss both administrator and end user support with a stronger focus on the latter as we will embed the support of the administrator in the wider context of relevance tuning as a core requirement. Our interests are more centred around the technical issues than management issues which are not the focus on this monograph.

5.1 Putting the User in Control

There are many ways in which the user can be put in control and much of it depends on the specific audience, the type of search and other factors. We will again focus on issues specific to enterprise search here and refer the interested reader to Morville and Callander [2010] and

Russell-Rose and Tate [2013] for a much broader and more detailed discussion of how to design user-centred search applications.

5.1.1 Basics

Putting the user in control starts by making sure the user perceives the search system as something valuable that will contain information which can satisfy a specific user need as otherwise users might be reluctant to use the system at all [Stenmark, 2006, Hertzum and Pejtersen, 2000].

Data silos (often going hand in hand with different responsibilities across departments) have already featured prominently in our discussion but an even more fundamental step is to digitise data in the first place so that it can be searched at all. While this might seem obvious, this first step needs to be seen as a prerequisite for all follow-on steps, and the McKinsey report on big data explicitly identifies the public sector in which they discovered "cases where departmental personnel were spending 20 percent of their time searching for information from other government departments using non-digital means (e.g., paper directories and calling people), and then obtaining that information by traveling to other locations and picking up data on physical media" [Manyika et al., 2011, p. 97]

There are other basic steps that should be considered to be included in any enterprise search context such as query suggestion and auto-completion, e.g. [Hawking and Griffiths, 2013]; this can be essential in people search [Guy et al., 2012]; also query-rewriting which in an enterprise search context may require domain knowledge to decide whether a query rewrite rule can be considered 'reasonable' or 'sensemaking' [Bao et al., 2012a].

Another basic but important consideration is based on the observation that many enterprise search failures are due to the desired document being outside the default search scope. Jakob Nielsen considers scoped search dangerous in general and recommends that the default search scope should always include the entire site with suggestions being offered for narrowing down the scope if appropriate. [2]

[2] https://www.nngroup.com/articles/search-visible-and-simple/

Due to the limited user base of an enterprise search engine as opposed to a Web search engine, there is the challenge that the collected log files are of a much smaller scale than Web logs and they therefore suffer more from data sparsity which will affect the effectiveness in generating any query suggestions in the first place. One solution is to focus on frequent queries [Kruschwitz et al., 2013], an alternative is not to use log files at all but to exploit occurrence of terms and phrases in the document collection which has been shown to be effective in suggesting auto-completions in an enterprise context [Bhatia et al., 2011]. After all, organisations have structured sources of data like staff directories, product catalogues or domain-specific taxonomies which can be used to drive effective auto-completion.

As discussed, a lot of information in an enterprise setting is hidden in email repositories. Much of the research on email search relates to personal email, but there is a wealth of knowledge in the email sent to distribution lists or to functional email addresses. When new employees start work, or employees change roles, the corporate memory associated with their role is entirely absent from their personal mail box. This lack can be overcome if mail sent to functional addresses such as `sales@x`, `dvc-r@uni-x`, `support@x`, etc. is archived and made searchable. Such *corporate email* repositories also help addressing the problem of knowledge being lost due to staff turnover (corporate memory). Future research will have to find out whether conclusions drawn from search over personal email will also apply to corporate email, such as the preference of a date-based ordering of search results over a relevance-based one [Dumais et al., 2003].

5.1.2 Supporting Exploration

Navigational searches might make up the bulk of frequently submitted queries, nevertheless, exploratory search has been identified as a key activity among information workers e.g. [Cleverley and Burnett, 2015b, Stenmark, 2008].

Adding support for exploration in addition to standard search is therefore essential to cover typical information needs in an enterprise environment. Russell-Rose and Tate propose the analogy of ‘search as a journey' in which the search process is an ongoing exploration [Russell-Rose and Tate, 2013], and, although very generic, this surely fits a typical organisational context in which repositories need to be discovered, content assessed and ideally by the end of the journey a much clearer picture of the information space has emerged in the searcher's head. They identify four specific dimensions that need to be taken into account to properly implement this paradigm, namely the *type of user* including their level of knowledge and expertise, their *goal*, their *context*, and their *search mode*.

White and Roth provide a list of features that are appropriate and necessary, i.e. must be present, for exploratory search systems [White and Roth, 2009]. Some of these features map directly into the information needs identified for enterprise search discussed in Chapter 3 and include the need for *facets* and *metadata-based* result filtering, the need to leverage search *context*, support querying and rapid query refinement, facilitate *collaboration* and support task management, among other features. In fact, contextual differences should be taken into consideration when generating suggestions proposed via facets [Cleverley and Burnett, 2015b].

Looking at more specific types of exploratory search, take the example of email search. Users can be pro-actively supported in exploring related internal and external content as part of their email communications, e.g. the current email can be used as context to retrieve and display related information. This has been applied in a seven-week trial in a large IT enterprise with the most commonly used feature of the tool being the uptake of automatic recommendations of related corporate information with participants stating that searching for information within the corporate repositories being made significantly more efficient [Laqua et al., 2011].

One might also go a step further, i.e. beyond search by providing support in exploring and analysing relational enterprise data [Zouzias et al., 2014].

The combination of automatic and manual knowledge organisation methods, e.g. manually constructed thesauri in combination with co-occurrence-based clustering of words, is a promising paradigm to not just help exploring a collection but also facilitate *serendipitous* information discovery [Cleverley and Burnett, 2015a].

Ideally, search and browsing should both be supported [Morville and Rosenfeld, 2006, p.35-37]. In this respect, many ideas to provide users with more control have been proposed for *Web* search but they seem equally (if not more) applicable in enterprise search, e.g., Olston and Chi [2003] combine the strengths of searching and browsing in a single interface. This guides users towards search results by highlighting relevant hyperlinks on the Web pages that they are browsing. Another approach to combine the two interaction modes is proposed by Freyne et al. [2007] in an attempt to harness and harvest community wisdom by incorporating social search and social browsing. White et al. [2007] enhance Web search by suggesting links to Web sites frequently visited by other users with similar information needs — in addition to the regular search results. This exploits the searching and browsing behaviour of previous users.

As outlined in the discussion of related work, Web sites and intranets, types of collections which we consider to fall within the scope of this review, can be difficult to navigate [Karim et al., 2009]. Apart from utilizing a domain-specific taxonomy for navigation support, e.g. [Lund and Ørnager, 2016], another common approach to add assistance to a Web site or an intranet is to use an overlay window or hover text, essentially adding a "layer" on top of an existing site. This can be used for presenting search results, e.g. [White et al., 2002, Dumais et al., 2001], or for navigation by introducing automatically acquired summaries [Alhindi et al., 2015] or links and suggestions to commonly visited pages, taking advantage of the collective search and navigation effort of other users [Karim et al., 2009, Saad and Kruschwitz, 2011].

5.1.3 Applying Analytics

The information gained from search logs is an important aspect of getting insights into what employees are searching for, and combining the knowledge of the business of the organisation with the information obtained from the search logs on a regular basis provides some key analytics [White, 2015a]. Applying analytics to tune the overall relevance ranking of a search engine will be the focus of the next section, but here we look at how users can be better supported in the actual search process by applying analytics.

Obviously, different employees will conduct different searches depending on their role in the enterprise. However, rather than starting with the organisational structure one might exploit the search behaviour as recorded in the log files to automatically cluster user groups, e.g. [Stenmark, 2008]. Such process uncovers that the user population is not a homogeneous group of information seekers but that they have very different approaches to searching (e.g. fact finding *vs.* more holistic information seeking) and this clustering stage can be used to help developers provide more targeted solutions instead of the currently predominant one-size-fits-all approaches. Interestingly, applying topic models based on the users' search behaviour in a mid-sized enterprise it was found that the communities inferred by the topics showed significant differences from the pre-defined organisational structure suggesting the application of such analysis to get a more truthful representation of users' shared interests [Carter et al., 2014].

Applying analytics to identify common types of search can also assist in moving away from (or enhancing the process of) simply returning a list if matching *documents* to presenting actual snippets of *information*, e.g. by supporting 'search by type' using information extraction methods to identify the type of query, e.g. a person, manual or product [Li et al., 2005].

Analytics should also be applied to help the search administrator by graphically or textually explaining why a particular desired result was not returned in response to a query. Examples are *'It is not in the index'*, *'It is security or robots protected'*, *'It does not contain any of the query words'*, *'More than ten other documents rank more highly than yours - here is a graphical display of the component ranking scores'* etc.

Finally, looking beyond search, analytics tools that support discovery need to be continuously maintained not just due to the ever-growing size of the collections but also due to new regulations and new trends in litigation discovery [Cherkasova et al., 2009].

5.1.4 Taking the User Seriously

Offering the user to control the settings helps making the system more transparent. It may well be that a user never ever changes any of the default settings as commonly observed, e.g. [Markey, 2007]. However, it can also help avoid confusion. If a user or group profile is being applied in the search system for example and a user conducts a broad range of searches, one might offer a small set of group profiles such as ‘sales’, ‘HR’, ‘R&D’, or alternatively ‘plain vanilla’ [Hawking, 2010].

An important general aspect of search systems – enterprise search or other applications – is to make sure the user is indeed in control by being able to switch options on and off [White, 2016, p. 301–303].

Offering a single point of access for employees to all information sources, whether internal and external-facing, should be desirable for any enterprise search application although the access permission security issues to be addressed might not be trivial, e.g. [Best et al., 2007].

Intranet users may actually be willing to cooperate with the search engine to improve the search quality not just for them but also for colleagues [Dmitriev et al., 2006], this can be exploited to acquire explicit page annotations that are treated like anchor text (which tends to be more sparse in such environments). The willingness of users within an enterprise to provide explicit feedback has also been demonstrated in a log study of an enterprise social media platform employed in a large organisation [Guy et al., 2016]. The study just looked at the ‘liking behaviour’, i.e. an employee pressing the ‘like’ button of a post, and observed that *information need* was a dominant reason for such activity.

5.1.5 Aggregation and Facets

Aggregation and ranking of results coming from different sources (a key feature of enterprise search) has already been pointed out as a challenge – also due to the fact that sources might vary in coverage and authority. A possible approach is to avoid merging results altogether [Hawking, 2010]. A study of a government metasearch context concluded that users were much more receptive to interfaces which supported the users' choices when navigating the results than having a merged interface which was rated poorly despite it being the most familiar one [Thomas et al., 2010] – not that surprising though given the problem with merging from different federated sources. Exposing the different sources and offering navigational aids to explore the results puts the user in control. The recommendations on intranet usability by the Nielsen Norman Group back this up by, e.g., proposing to clearly differentiate intranet site search and employee directory search [Pernice et al., 2007, p.66] and by recommending to treat special repositories that most employees would not normally want to search separately from the usual search functions [Pernice et al., 2007, p.25].

Standard faceted search can be extended to not just return counts of documents across facets but to allow richer aggregation which supports better decision making [Ben-Yitzhak et al., 2008]. This is an example where enterprise search meets business intelligence and the user is put in control as he or she explores the data collection.

The preferred lookup mode might not actually be the use of a search engine but the use of menus as uncovered by a survey among three different organisations, a large manufacturing company, a medium-sized manufacturer and a municipality [Stenmark, 2010]. Regardless of organisation or role, menus were preferred over search engines, the use of bookmarks and notification services. The same observation was made in a larger-scale study of intranet usage in 27 organisations ranging-ing from about 100 employees to about 160,000 [Pernice et al., 2007, p.11-13].

5.2 Relevance Tuning and Support

This section will look at the more technical issues that need to be addressed to make enterprise search work. Unlike Web search engines, enterprise search is still largely managed in an ad hoc fashion [Li et al., 2014]. Despite this, we can distinguish a number of different ways in which relevance tuning and support can be applied and we will discuss them in turns.

5.2.1 Relevance Tuning

Tuning a ranking algorithm for an enterprise search system to the information it actually indexes can make a great difference [Hawking, 2010]. In enterprise search, tuning is therefore an essential requirement, the search tool should be monitored, evaluated and tuned [Miles, 2014]. However, in line with other discrepancies between expectation and reality as we discussed earlier, Miles also reports that only 18% of organisations monitor ongoing results, 30% running the search system out of the box and 38% not even having tuned their search tools at all, all of which makes the divergence between expectation and perception less surprising. Rosenfeld observes that site search analytics still does not receive much attention no matter whether it is a small setting or a more advanced one with entire business units devoted to Web analytics and user research [Rosenfeld, 2011, p.25].

Search and access logs are a rich source of information to identify important/typical user needs [Hawking, 2010] based on which the search system can be tuned but these might not be available at sufficient scale to be exploited for ranking, for example [Chaudhuri et al., 2011].

5.2.2 Domain-customisation

The use of metadata and domain-specific taxonomies to organise and classify content can help users find the right information, in particular in a controlled space like an enterprise, e.g. [Schymik et al., 2015]. In fact, the majority of organisations surveyed by Findwise in 2016 make use of such organisational structures [Findwise, 2016]. A practical problem that arises however is that the information explosion has reached the point where many information architects no longer have a full grasp of the themes and topics covered in the collection [Mukherjee and Mao, 2004] which makes it even more crucial that different teams being in charge of such knowledge structures coordinate their efforts [Findwise, 2016]. Another problem is that every domain is changing and this should apply to the domain's vocabulary and its descriptive metadata as well but typically it lags behind [Rosenfeld, 2011, p.150].

While taxonomies offer huge potential benefits they also need to be seen from an economic perspective, no matter whether they are flat term lists, thesauri, classification schemes or organised in other ways. Apart from being an asset they can also be a liability if, for example, they are incomplete, out of date, confusing to users or simply not used at all [Bedford, 2014].

Vocabulary mismatch is a particular example addressed by domain-customisation. The mismatch between vocabulary of users and that of authors is particularly striking on intranets where corporate policy dictates that certain terms are avoided in official documents or on Web pages, e.g. [Dmitriev et al., 2006]. In line with that and as noted earlier, domain dictionaries like acronyms and person names have significant value in improving precision, e.g. [Zhu et al., 2007]; also synonym lists can be a powerful tool for expanding terms into known alternatives i.e. company-wide vocabularies, acronyms etc., e.g. [Lund and Ørnager, 2016].

Issues around multilinguality need to be seen as part of the customisation step. International organisations might well choose English as the corporate language as a default but in reality there will be a mix of other languages being used, e.g. United Nations (UN) agencies have to support six different languages as a matter of policy [White and Nikolov, 2013]. Equally, even in multinational organisations there will be local communication that will not necessarily be in the corporate language.

Multi-national enterprises require additional customisation, e.g. the catching of phonetic misspellings when looking up names [Guy et al., 2012]

5.2.3 Quality Control Mechanisms

Enterprise search relies heavily on manual intervention to assure that certain common/important information needs can be guaranteed to be served no matter how the ranking algorithm might affect the order of resulting documents.

A simple approach is to target frequent queries, after all, enterprise search queries tend to follow a power law distribution and the 200 most frequently submitted queries may well make up 20% of the overall query traffic [Norling and Lamb, 2017, p.336–337].

Among the most popular approaches in this respect are ‘best bets’ and ‘query boosting’ as well as the tuning of top-N queries. One does however need to be careful with sampling of top queries to have confidence in the expected performance [Rowlands et al., 2007]. Best bets are very powerful while being simple but care needs to be taken in deciding which queries should be considered. Rather than simple frequency, a combination of *popularity* and *persistency* is sensible [Rosenfeld, 2011, p.127]. Over-use of boosting is also common: it is likely that different groups of users will have different views of the relevance of certain results, and if one group is able to impose their own views on boosting, then this may have a significant negative effect on relevance for others.

It might be the case that the best answers to a query are not found because of a terminology mismatch or because the document only exists as a scanned copy. In such cases it might be better to improve the way information is published rather than tuning the system [Hawking, 2010]. If these are very common information needs, then ‘best bets’ might well be the best solution.

There are a few important considerations to take into account when applying best bets. One issue is the integration of the manually chosen matches with algorithmic results. Best bets should be removed from algorithmic results to avoid wasting valuable space due to redundancy [Morville and Callander, 2010, p.91]. Furthermore, incorporating best bets is necessarily labour intensive and without proper maintenance the search system can become an embarrassment.

The selection of data sources to be searched and indexed is important, e.g. [Morville and Rosenfeld, 2006, p.151]. Obviously there is a trade-off, as a large quantity of searchable information offers potentially greater benefit to the business but it also complicates the possibility of finding the right information [Norling and Lamb, 2017, p.328]. In devising an enterprise search strategy PwC (PricewaterhouseCoopers) have focussed on making business critical content searchable rather than all content resulting in a dramatic improvement in usage [Findwise, 2016]. This is a practical example of something more fundamental, namely applying the removal of ROT – redundant, outdated, trivial data [Morville and Callander, 2010, p.38].

The collection also needs to be checked for missing content, e.g. using search logs as in [Jhamtani et al., 2017], as not finding a document is a common cause for complaints from end users in an enterprise and reasons could include an actually missing document, missing connectors/filters, access restrictions, and an out-of-date index [Hawking, 2010]. A task-based study of the search behaviour of software engineers within a large hi-tech company found that they failed to identify any useful documents in about a quarter of all searches [Freund and Toms, 2006]. Applying *site search analytics*, i.e. the regular analysis of query logs and how the search engine responded to user queries, should be able to flag up such cases and should in any case be a standard tool for the site manager but all too often receives little or no attention [Rosenfeld, 2011].

5.2.4 The Human in the Loop

A practical difficulty is that enterprise search is typically managed by administrators who are domain experts but not search experts which means that translating the domain knowledge into tuning an underlying retrieval model is non-trivial if not impossible [Bao et al., 2012a,b]. To put it differently, domain experts in an enterprise tend to be very knowledgeable and experienced in their specific domain with a deep insight into the problem area and the contents of the documents but they are likely to have less formal training on how to formulate search strategies [Wu et al., 2014]. Bao *et al.* address this by offering search architectures for the administrator that feature two principles, namely 'comprehensibility' of the ranking mechanism which makes the ranking transparent and 'customizability' of the search engine by means of adjustable rules for reranking and query rewriting [Bao et al., 2012b]. Wu and colleagues demonstrate that having domain experts involved in improving the search results by providing explicit relevance assessments, in particular for frequent queries, leads to more accurate results and as a result offers substantial cost benefits [Wu et al., 2014]. The idea of 'test-driven relevance tuning' addresses the exact same point, namely the specific problem of relevance in enterprise search by connecting business users (who know what results are relevant) with technical people (who know how to adjust the search algorithms accordingly) [Turnbull and Berryman, 2016], a *general* enterprise-search-specific problem.

Offering search administrators and domain experts the ability to customise the process of interpreting user queries then turns into a requirement [Fagin et al., 2011]. Fagin *et al.* report that offering a query rewriting administration tool proved to be a powerful and effective mechanism with substantial uptake.

Much of the highlighted support for search administration is done by direct intervention or *manual customisation* but there is also scope for *machine learning*-based approaches, e.g. the supervised classification of a query rewrite rule being considered as ‘natural’ [Bao et al., 2012b].

Figure 5.1 sketches the typical cycle of relevance tuning in enterprise search.

A more detailed discussion of the more hands-on issues can be found elsewhere [White, 2007, 2015b].

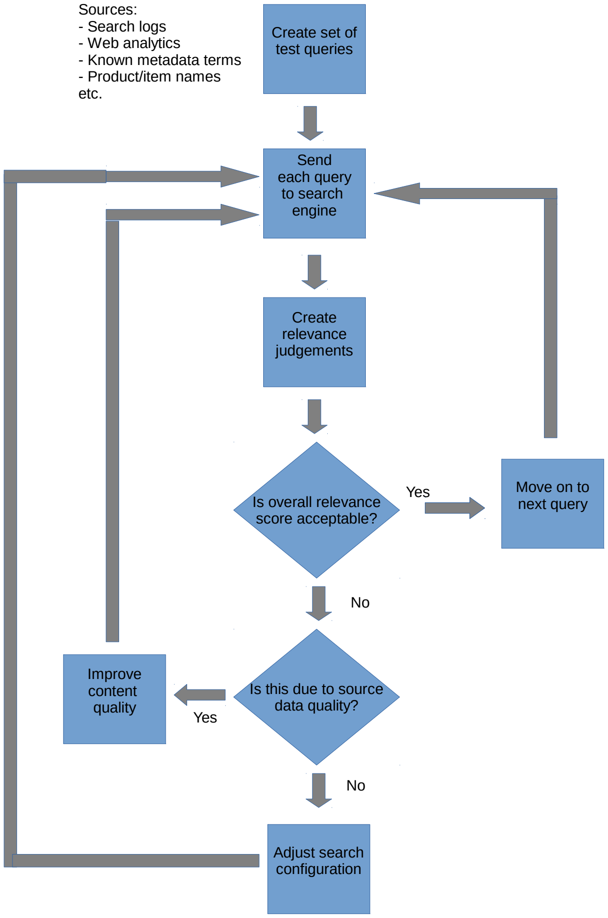
Sources:
- Search logs
- Web analytics
- Known metadata terms
- Product/item names
etc.
Create set of test queries
Send each query to search engine
Create relevance judgements
Is overall relevance score acceptable?
Yes
Move on to next query
No
Is this due to source data quality?
Yes
Improve content quality
No
Adjust search configuration

Figure 5.1: A search relevance testing flow chart for enterprise search

6

The Future

The future certainly looks interesting with a number of developments already shaking up the enterprise search market. To pick just two, think of some general trends such as the move towards cloud solutions and open-source platforms. Here we will look into some emerging trends and potential future developments. We also identify some research challenges in enterprise search which are mainly derived from progress made in Web search. There is no claim for completeness as this is our very own take on where things might be moving and where potentially interesting avenues open up.

6.1 General Trends

The adoption of social media in enterprises is rapidly gaining pace. This does and will have an impact on the information seeking processes within organisations and an example area in which this can already be observed is the way in which new employees tap into knowledge about the company they are joining, as well as its culture and values [Treem and Leonardi, 2012]. In analogy to interpreting enterprise search as a platform rather than a technology, one could treat the different social media tools employed in an organisation as part of an integrated enterprise social media platform [Leonardi et al., 2013].

Security issues are already a major concern with 31% of respondents in the AIIM survey reporting that security and permission concerns would be a 'show-stopper' and a further 41% treating it as a major concern [Miles, 2014]. This is only going to become more of an issue with the rise of cloud-based solutions.

A more general and hardly surprising observation is that the amount of data generated, stored and consumed in enterprises and beyond has been shown to grow exponentially [Manyika et al., 2011] and this rise of *Big data* offers potential but also serious challenges if the full potential of this data is to be captured. The McKinsey report estimated that in 2009 each company across all sectors of the US economy was sitting on 200 terabytes of stored data on average – though much of this will comprise sensor and behavioural rather than text data and as such will not be directly applicable to enterprise search.

We will also see more of a convergence of natural language processing and information retrieval techniques when it comes to any form of text analytics. The identification of named entities, relation extraction, sentiment analysis are all becoming standard processing steps. The recent text book by Zhai and Massung is the best example to demonstrate this development [Zhai and Massung, 2016].

6.2 Technical Developments

Deep learning approaches to search are rapidly finding their way into search algorithms and this will certainly have an impact on the underlying technology employed in enterprise search engines (with some delay). One only has to look at the proceedings of recent major information retrieval conferences such as SIGIR, WSDM and ECIR to see how rapidly neural-network-based approaches have become the paradigm of choice (just compare the frequency of terms like *neural networks*, *embeddings*, *deep learning* in this year's proceedings with the same conference ten years ago). This development started even earlier in the computational linguistics community (e.g. see the ACL and EMNLP proceedings).

Cloud applications are everywhere now. The enterprise market is already well-represented in the ‘cloud’ – even the concept of ‘search-as-a-service’ has been established [Singh et al., 2009], but there are some interesting problems that come with this move beyond simple scalability and availability such as legal aspects, access control, etc. Many common challenges such as security and fine-grained access control get amplified in the context of moving a service to the cloud [Chaudhuri et al., 2011]. Cloud collaboration, for example through *Dropbox* and *GoogleDrive*, is now a well-established paradigm in an industrial setting with participants often using more than one shared repository [Massey et al., 2014]. It will also be very interesting to see the effect of the demise of the Google Search Appliance (GSA), a now-deleted product, still installed on many customer sites. This is likely to further push the move towards cloud solutions, but do note the points made above – cloud solutions are not appropriate for everybody, there are concerns about security (especially across borders), privacy etc. which are very important.

The more wide-spread use of *open-source* applications seems like a natural progression given the rising popularity of platforms such as Apache Lucene/Solr and Elasticsearch and the general appetite within organisation for open source solutions [Miles, 2014].

The acquisition and curation of *taxonomies and metadata* will likely remain a core task in enterprise search for some time. Automatic approaches tend to be fast and scalable and the level of noise might be an acceptable trade-off when compared to the manual effort a knowledge engineer might have to invest in creating such resources. A range of methods have been developed to turn document collections as well as query logs into structured knowledge that can be utilised for search support but also to support users exploring a collection, e.g., see [Clark et al., 2012] for an overview. Similarly, metadata can also be automatically mined in order to improve access to relational databases. Cortez and colleagues aim to enrich database schemas with descriptive keywords. To do this they first mine enterprise spreadsheets to find candidate terms that are then automatically assigned to corporate databases [Cortez et al., 2015].

Compliance issues are making enterprise search more essential (as discussed in the related work section, compliance does already form a major part of typical enterprise search needs in bigger companies). There is evidence that compliance failure and major litigation issues frequently trigger the re-evaluation of search tools within an organisation [Miles, 2014].

A major issue to be addressed is to move away from a more ad hoc *evaluation* approach to enable automatic evaluation and generally cut down manual involvement [Li et al., 2014].

Other technical developments that are clearly happening in search but might have an impact more in the longer term on enterprise search include the move to mobile search, speech-driven search, even image-driven search and new developments in graph search [Makhani, 2015][1]. The last point in particular will have an impact not just on the architectural design of search applications but also on evaluation methods.

6.3 Moving towards Cooperative Search

Computer-Supported Cooperative Work is an area with a lot of potential, e.g. [Morris et al., 2008]. It has to be acknowledged though that despite recent progress there is still work to be done to develop systems that are capable of supporting effective intentional collaborative searching [White, 2016, p. 251].

Collaboration via tools like *Dropbox*, *SharePoint*, *Slack* and *GoogleDrive* is now well-established but effective search mechanisms still need to be employed, although it also needs to be recognised that in a qualitative study that Massey and colleagues conducted looking into how co-workers overcome major co-organizational barriers found that less than a third of their subjects needed to conduct search to find the information they were after [Massey et al., 2014]

For enterprise search there is even less progress and White summarizes that in an enterprise setting “search remains a solitary exercise” [White, 2015b]. Having said this, there are examples of collaborative

[1] given the structured nature of much of the searchable content in an enterprise information infrastructure

search in provenance-related contexts, e.g. in the construction industry, where collaborative search efforts may last for days or weeks, e.g. [Khan et al., 2016].

The development of more extensive virtual teamwork can however be seen as one factor that will make collaborative search work more common. An important practical aspect is of course that tools for collaborate search need to be designed for that purpose [Hansen and Järvelin, 2000] and search vendors will then have to build support into the products. One first step towards supporting more collaborative work might be to identify latent groups of users that appear to share common interests as expressed by their search behaviour but which are opaque to ground truth enterprise structure [Priest and Carter, 2014, Carter et al., 2014].

Beyond enterprise search there is work on computer-supported cooperative work that could be employed, e.g. group-based information appears to be a promising route for a community of users with common concerns. Such communities are formed of individuals — e.g. employees of a company or members of a university — that, over time, collectively acquire knowledge about a resource such as a local Web site. The idea is to tap in to this knowledge, and facilitate the sharing of search and navigation experiences among community members [Smyth, 2007]. This bears some resemblance with the idea of "trait-based groups" as people who "may be highly likely to repeat or augment tasks already accomplished by other group members, have interests in the same queries and results as other group members." [Teevan et al., 2009]. The idea is that learning from one user should benefit future users with similar information needs, an idea also shared with other approaches of assisting users in navigating a collection, e.g. [Alhindi et al., 2015, Kantor et al., 2000, Wexelblat and Maes, 1999].

6.4 Some Research Challenges

We would like to point out a number of research challenges and directions that in our view offer the potential to make significant progress on the technological side of enterprise search. Much of this has to do with trying to adopt recent progress made for Web search and adjust it accordingly so that enterprise search applications could benefit. These challenges complement the directions and open questions we mentioned in passing as we reviewed the field.

6.4.1 Transferring from Web Search

A huge amount of work has been done on learning ranking functions for Web search. This is also true for spelling correction systems, query rewriting, knowledge graphs, and answer panes. Is it possible to transfer learnings from the Web search domain to the enterprise? This is an interesting question in general but one of specific importance to companies like Google, Microsoft, Yandex and Baidu. How can learnings from user log data collected on a Web search engine be used to overcome the sparsity of user behaviour data in enterprise clouds, email services, etc.?

6.4.2 Closing the Vocabulary Gap

We noted that there is often a vocabulary gap in enterprise search – students search for *courses* but the university intranet only knows *modules*; users search for *gun license* but the relevant document talks about *permit to acquire a long arm*; the search is for *fiscal outlook* but the user actually wants documents that discuss the *budget situation*. Now, this vocabulary gap is equally present in Web search, but the existence of large-scale query logs and user behaviour data means that auto-complete and query suggestions can address this problem quite effectively. Devising techniques for closing such gaps in the absence of large-scale logs is an interesting challenge for the research community.

Similarly, how do you provide accurate and useful spelling suggestions in the absence of large-scale log data? Dictionary-based approaches are very limited, in particular in a multi-lingual environment, word-by-word comparisons are crude, and two specific features of enterprise search complicate this process. First of all, one needs to avoid making suggestions that have no answer, and secondly, the access rights of users need to be considered when applying suggestions so that no information is given away beyond what that user is allowed to see.

6.4.3 Research without Test Collections

We noted that there are essentially no shareable test collections available for enterprise search, a significant shortcoming when conducting research compared to other areas of information retrieval. Again, research developments in Web search point at some possible directions here, namely a simulation approach to automatically construct test collections or approaches that work without the need of test collections altogether.

With a shortage of usable test collections, simulation studies represent an appealing route to conduct research. Simulating the choice of user queries and their judgement can be used to build a test collection which could be used just like a manually created one. This is now a well-established paradigm for Web search as it has been demonstrated that simulated topics can be generated that are comparable to real topics – at least for known-item search [Azzopardi et al., 2007]. Simulation studies have also been extended to interactive information retrieval settings, e.g. [Maxwell and Azzopardi, 2016]. Closer to enterprise search – in terms of variety of data structures, sparseness of hyperlinks, lack of test collections – simulated test collections have been constructed for desktop search in a similar fashion, resulting in a *pseudo-desktop* [Kim and Croft, 2009], and the challenge is to apply this stream of work to the more complex setting of enterprise search.

Conducting evaluation experiments with *real* users presents another challenge but here again we might benefit from progress made in Web search. We have already pointed out that the use of online evaluation has become a de facto standard for evaluating Web search engines but could they not also be employed in enterprise search? Hofmann and colleagues provide explicit tips for doing online evaluation with just tens of users or hundreds of queries [Hofmann et al., 2016], which could make evaluation methods like interleaving [Joachims, 2002, Radlinski and Craswell, 2010] or a side-by-side result panel comparison [Thomas and Hawking, 2006] a practical option that does not distract the user in his or her day-to-day work. Enterprise search-specific features like access rights and user roles will have to be addressed though in order to draw any generalisable conclusions from these experiments.

Apart from applying test collections and online evaluation, the third commonly applied evaluation type – user studies – also offers a lot of potential for enterprise search. As discussed, most user evaluations that have been reported for enterprise search have been designed for a specific use case and do not generalise. Adopting a solid experimental setup for task-based evaluations, e.g. [Kelly, 2009, Järvelin et al., 2015], would allow comparisons across experiments and hence push forward our understanding of enterprise search.

Let us conclude with a more general point. Being able to meaningfully interpret *implicit* feedback provided by searchers has had a major impact on the state of the art in Web search, be it by simply interpreting result set navigation [Joachims et al., 2005] or taking a more long-term behavioural angle, e.g. [Kelly, 2004]. Finding out how this could be adopted to enterprise search is a worthwhile challenge to tackle.

6.5 Final Words

The future of enterprise search will also need to be driven by a rising awareness of what enterprise search is, what the user needs are, what the fundamental problems are but also what technical solutions exist, in short, an awareness of challenges and potential benefits. We hope that this monograph helps moving one step in that direction so that we will move away from this commonly observed scenario:

> “When users complain about the quality of search there is anecdotal evidence that the decision is to ‘upgrade’ the search application, on the basis that clearly the current search implementation is not adequate. Since the underlying issue is one of a lack of support post-implementation the results from the replacement search application are usually no better” [White and Nikolov, 2013].

7

Conclusion

We conclude this survey with a short list of take-home messages as follows:

1. Enterprise search is an area that is hugely important in industry yet has attracted relatively little academic interest.
2. There are substantial differences between enterprise search and other types of search such as Web search which include heterogeneous data sources, silo-based repositories, and users defined by roles in the enterprise.
3. Enterprise search will not work out-of-the-box, and the human in the loop is essential, e.g. for customisation, continuous relevance assessment and tuning.
4. People search and email search are among the dominant search types in an enterprise. Information needs are driven by business needs, e.g. people search is mainly aimed at finding experts or expertise and not for entertainment.
5. Evaluation in enterprise search is essential and in some way very different to other search areas.

6. Much progress in other search areas has not found its way yet into standard enterprise search applications.

7. Given the controlled environment of an enterprise there is much scope to utilise user interactions with the search system to improve search and exploration for individuals or groups of users. Little of this is currently being employed.

8. There are plenty of research challenges worth exploring academically to push forward the state of the art in enterprise search.

9. The reality is that even now enterprise search systems are falling short of expectations.

Acknowledgements

There are a number of people we would like to thank for their support throughout this venture. First and foremost we express our gratitude to Martin White and David Hawking for providing extensive feedback on earlier versions of the manuscript. We surely did not manage to address all their recommendations and suggestions but we tried very hard. So if there is anything in the monograph that does not look right to you, it is entirely our own fault. We also would like to thank the anonymous reviewers for their insightful comments, for their concerns and their critical questions. Again, we tried to address these – and it did take us a lot of time to do so. Thanks go also to Tony Russell-Rose with whom we never stopped discussing this monograph. We are particularly thankful to Mark Sanderson for making this happen in the first place and for pushing us to get it out of the way. We also thank Maarten de Rijke and Doug Oard for feedback on the original outline and early versions of the manuscript and to James Finlay without whom we would not have been able to publish this monograph at all. There are many more people who have contributed in one way or another, and rest assured that we will buy you all a pint of the finest beverage.

References

M-D. Albakour, G. Ducatel, and U. Kruschwitz. The Role of Search for Field Force Knowledge Management. In *Transforming Field and Service Operations: Methodologies for Successful Technology-Driven Business Transformation*, Theory and Applications of Natural Language Processing, pages 117–132. Springer, 2013.

D. Alberts, C. B. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. Introduction to Patent Searching. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, pages 3–43. Springer, 2011.

A. Alhindi, U. Kruschwitz, C. Fox, and M-D. Albakour. Profile-based summarisation for web site navigation. *ACM Transactions on Information Systems (TOIS)*, 33(1):4:1–4:39, March 2015.

J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, Challenges, and Opportunities for Information Retrieval: Report from SWIRL 2012 the Second Strategic Workshop on Information Retrieval in Lorne. *SIGIR Forum*, 46(1):2–32, 2012.

E. Amitay, D. Carmel, N. Har'El, S. Ofek-Koifman, A. Soffer, S. Yogev, and N. Golbandi. Social Search and Discovery Using a Unified Approach. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 199–208. ACM, 2009.

J. Arguello. Aggregated Search. *Foundations and Trends in Information Retrieval*, 10:365–502, 2017.

L. Azzopardi, M. de Rijke, and K. Balog. Building Simulated Queries for Known-item Topics: An Analysis Using Six European Languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 455–462. ACM, 2007.

R. Baeza-Yates and B. Ribeiro-Neto, editors. *Modern Information Retrieval.* Addison-Wesley, 2nd edition, 2010.

P. Bailey, D. Hawking, and B. Matson. Secure search in enterprise webs: Tradeoffs in efficient implementation for document level security. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 493–502. ACM, 2006.

P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO Enterprise Search Test Collection. *SIGIR Forum*, 41(2):42–45, 2007.

P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2007 Enterprise Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, NIST Special Publication: SP 500-274. NIST, 2008a.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance Assessment: Are Judges Exchangeable and Does It Matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 667–674. ACM, 2008b.

K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad Expertise Retrieval in Sparse Data Environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 551–558. ACM, 2007.

K. Balog, I. Soboroff, P. Thomas, P. Bailey, N. Craswell, and A. P. de Vries. Overview of the TREC 2008 Enterprise Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, NIST Special Publication: SP 500-277. NIST, 2009.

K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise retrieval. *Foundations and Trends in Information Retrieval*, 6(2-3):127–256, 2012.

Z. Bao, B. Kimelfeld, and Y. Li. Automatic suggestion of query-rewrite rules for enterprise search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 591–600. ACM, 2012a.

Z. Bao, B. Kimelfeld, Y. Li, S. Raghavan, and H. Yang. Gumshoe quality toolkit: Administering programmable search. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2716–2718. ACM, 2012b.

J. R. Baron, D. D. Lewis, and D. W. Oard. TREC 2006 Legal Track Overview. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*, NIST Special Publication: SP 500-272. NIST, 2007.

P. Baumard. *Tacit Knowledge in Organizations*. SAGE, 1999.

D. A. D. Bedford. Understanding and managing taxonomies as economic goods and services. *Bulletin of the Association for Information Science and Technology*, 40(4):15–22, 2014.

S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 321–328, Sheffield, 2004. ACM.

S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman. Temporal Analysis of a Very Large Topically Categorized Web Query Log. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(2):166–178, January 2007.

O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 33–44. ACM, 2008.

P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the Impact of Short- and Long-term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 185–194. ACM, 2012.

B. Berendt and M. Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9(1): 56–75, March 2000.

O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved Search Engines and Navigation Preference in Personal Information Management. *ACM Transactions on Information Systems (TOIS)*, 26(4):20:1–20:24, October 2008.

T. Berners-Lee. Information Management: A Proposal. Technical report, CERN, 1989.

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 5:34–43, May 2001.

B. Best, J. Jürjens, and B. Nuseibeh. Model-Based Security Engineering of Distributed Information Systems Using UMLsec. In *Proceedings of the 29th International Conference on Software Engineering*, ICSE '07, pages 581–590. IEEE Computer Society, 2007.

S. Bhatia, D. Majumdar, and P. Mitra. Query Suggestions in the Absence of Query Logs. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 795–804. ACM, 2011.

M. Bishop. *Computer Security: Art and Science.* Addison-Wesley, 2003.

D. C. Blair and M. E. Maron. An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System. *Communications of the ACM*, 28(3): 289–299, March 1985.

T. Bogers and K. Balog. UvT Expert Collection documentation. ILK Research Group Technical Report Series no. 07-06, 2007.

P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 56–63. ACM, 2009.

C.P. Bourne and T.B. Hahn. *A History of Online Information Services, 1963-1976.* MIT Press, 2003.

S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International World Wide Web Conference (WWW7)*, pages 107–117, Brisbane, 1998.

A. Broder. A Taxonomy of Web Search. *SIGIR Forum*, 36(2):3–10, 2002.

A. Brooking. *Corporate Memory: Strategies for Knowledge Management.* International Thomson Business Press, 1999.

V. Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, 1945.

Cabinet Office. Government ICT Strategy, March 2011.

L. Carata, S. Akoush, N. Balakrishnan, T. Bytheway, R. Sohan, M. Seltzer, and A. Hopper. A Primer on Provenance. *Queue*, 12(3):10:10–10:23, March 2014.

K. M. Carter, R. S. Caceres, and B. Priest. Latent Community Discovery Through Enterprise User Search Query Modeling. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 871–874. ACM, 2014.

C. Castillo and B. D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, 4(5):377–486, 2011.

R. Cattell. Scalable SQL and NoSQL Data Stores. *SIGMOD Record*, 39(4): 12–27, 2010.

M. Chau, X. Fang, and O. R. L. Sheng. Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology (JASIST)*, 56(13):1363–1376, November 2005.

S. Chaudhuri, U. Dayal, and V. Narasayya. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98, 2011.

L. Cherkasova, K. Eshghi, C. B. Morrey, J. Tucek, and A. Veitch. Applying Syntactic Similarity Algorithms for Enterprise Information Management. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1087–1096. ACM, 2009.

M. Clark, Y. Kim, U. Kruschwitz, D. Song, M-D. Albakour, S. Dignum, U.C. Beresi, M. Fasli, and A. De Roeck. Automatically Structuring Domain Knowledge from Text: an Overview of Current Research. *Information Processing and Management*, 48(3):552–568, 2012.

C. Cleverdon. The Cranfield Tests on Index Language Devices. In K. Sparck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–59. Morgan Kaufmann Publishers Inc., 1997.

P. H. Cleverley and S. Burnett. The best of both worlds: highlighting the synergies of combining manual and automatic knowledge organization methods to improve information search and discovery. *Knowledge Organization*, 42 (6):428–444, 2015a.

P. H. Cleverley and S. Burnett. Retrieving haystacks: a data driven information needs model for faceted search. *Journal of Information Science*, 41(1): 97–113, 2015b.

P. H. Cleverley, S. Burnett, and L. Muir. Exploratory information searching in the enterprise: A study of user satisfaction and task performance. *Journal of the Association for Information Science and Technology*, 68(1):77–96, 2017.

G. V. Cormack and M. R. Grossman. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 75–84. ACM, 2016.

G. V. Cormack, M. R. Grossman, B. Hedin, and D. W. Oard. Overview of the TREC 2010 Legal Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2010)*, NIST Special Publication: SP 500-294. NIST, 2011.

E. Cortez, P. A. Bernstein, Y. He, and L. Novik. Annotating Database Schemas to Help Enterprise Search. *PVLDB*, 8(12):1936–1939, 2015.

N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 enterprise track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, 2005.

B. Croft, D. Metzler, and T. Strohman, editors. *Search Engines: Information Retrieval in Practice.* Pearson Education, international edition, 2010.

J. Delgado, R. Laplanche, and V. Krishnamurthy. The new face of enterprise search: Bridging structured and unstructured information. *Information Management Journal*, 39(6):40–46, 2005.

H. Deng, I. King, and M.R. Lyu. Entropy-biased models for query representation on the click graph. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2009.

S. Dignum, U. Kruschwitz, M. Fasli, Y. Kim, D. Song, U. Cervino, and A. De Roeck. Incorporating Seasonality into Search Suggestions Derived from Intranet Query Logs. In *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence (WI'10)*, pages 425–430, Toronto, 2010.

P. Dmitriev, P. Serdyukov, and S. Chernov. Enterprise and desktop search. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1345–1346. ACM, 2010.

P. A. Dmitriev, N. Eiron, M. Fontoura, and E. Shekita. Using annotations in enterprise search. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 811–817. ACM, 2006.

M. Doane. Cost-benefit analysis: Integrating an enterprise taxonomy into a SharePoint environment. *Journal of Digital Asset Management*, 6(5):262–278, 2010.

C. Doctorow. Metacrap: Putting the torch to seven straw-men of the meta-utopia. `https://www.well.com/~doctorow/metacrap.htm`, 2001.

S. Dumais, E. Cutrell, and H. Chen. Optimizing search by showing results in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 277–284. ACM, 2001.

S. Dumais, E. Cutrell, JJ Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 72–79. ACM, 2003.

D. Elsweiler, G. Jones, L. Kelly, and J. Teevan. Workshop on Desktop Search. *SIGIR Forum*, 44(2):28–34, 2010.

R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the Workplace Web. In *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, pages 366–375, Budapest, 2003. ACM.

R. Fagin, B. Kimelfeld, Y. Li, S. Raghavan, and S. Vaithyanathan. Understanding queries in a search database system. In *Proceedings of the Twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 273–284. ACM, 2010.

R. Fagin, B. Kimelfeld, Y. Li, S. Raghavan, and S. Vaithyanathan. Rewrite rules for search database systems. In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '11, pages 271–282. ACM, 2011.

S. Feldman and C. Sherman. The High Cost of Not Finding Information: An IDC White Paper. Technical Report 29127, IDC, 2001.

D. F. Ferraiolo, J. F. Barkley, and D. R. Kuhn. A Role-based Access Control Model and Reference Implementation Within a Corporate Intranet. *ACM Transactions on Information and System Security*, 2(1):34–64, February 1999.

D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C.A. Welty. Building Watson: an overview of the Deep QA project. *AI Magazine*, 31 (3):59–79, 2010.

Findwise. Enterprise Search and Findability Survey 2014. `http://www2.findwise.com/findabilitysurvey2014`, 2014.

Findwise. Enterprise Search and Findability Survey 2015. `http://www2.findwise.com/findabilitysurvey2015`, 2015.

Findwise. Enterprise Search and Findability Survey 2016. `https://findwise.com/Enterprise-Search-Findability-Report-2016`, 2016.

B.M. Fonseca, P.B. Golgher, E.S. De Moura, B. Pôssas, and N. Ziviani. Discovering search engine related queries using association rules. *Journal of Web Engineering*, 2(4):215–227, 2003.

M. Fontoura, E. Shekita, J. Y. Zien, S. Rajagopalan, and A. Neumann. High performance index build algorithms for intranet search engines. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 1122–1133. VLDB Endowment, 2004.

G. Forman, K. Eshghi, and J. Suermondt. Efficient Detection of Large-scale Redundancy in Enterprise File Systems. *SIGOPS Operating Systems Review*, 43(1):84–91, January 2009.

E. A. Fox and O. Sornil. Digital libraries. In *Encyclopedia of Computer Science*, pages 576–581. John Wiley and Sons Ltd., Chichester, UK, 2003.

L. Freund and E. G. Toms. Enterprise Search Behaviour of Software Engineers. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 645–646. ACM, 2006.

L. Freund, E. G. Toms, and C. L.A. Clarke. Modeling Task-genre Relationships for IR in the Workplace. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 441–448. ACM, 2005.

J. Freyne, R. Farzan, P. Brusilovsky, B. Smyth, and M. Coyle. Collecting community wisdom: Integrating social search & social navigation. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, pages 52–61. ACM, 2007.

M. E. Frisse. Searching for information in a hypertext medical handbook. *Communications of the ACM*, 31(7):880–886, 1988.

S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The Adaptive Web*, pages 54–89, 2007.

J. C. Gomez and M.-F. Moens. A Survey of Automated Hierarchical Classification of Patents. In *Professional Search in the Modern World - COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, volume 8830 of *Lecture Notes in Computer Science*, pages 215–249. Springer, 2014.

M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Transactions on Information Systems (TOIS)*, 22(2):270–312, 2004.

D. Graus, D. van Dijk, M. Tsagkias, W. Weerkamp, and M. de Rijke. Recipient recommendation in enterprises using communication graphs and email content. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1079–1082. ACM, 2014.

G. Grefenstette and L. Wilber. *Search-Based Applications.* Morgan & Claypool Publishers, 2010.

C. Grevet, D. Choi, D. Kumar, and E. Gilbert. Overload is Overloaded: Email in the Age of Gmail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 793–802. ACM, 2014.

M. R. Grossman, G. V. Cormack, B. Hedin, and D. W. Oard. Overview of the TREC 2011 Legal Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2011)*, NIST Special Publication: SP 500-296. NIST, 2012.

M. R. Grossman, G. V. Cormack, and A. Roegiest. TREC 2016 Total Recall Track Overview. In *Proceddings of the 25rd Text Retrieval Conference (TREC 2016).* NIST, 2016.

C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.

I. Guy, S. Ur, I. Ronen, S. Weber, and T. Oral. Best Faces Forward: A Large-scale Study of People Search in the Enterprise. In *Proceddings of CHI 2012*, pages 1775–1784. ACM, 2012.

I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen. Mining Expertise and Interests from Social Media. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pages 515–526. ACM, 2013.

I. Guy, I. Ronen, N. Zwerdling, I. Zuyev-Grabovitch, and M. Jacovi. What is Your Organization 'Like'?: A Study of Liking Activity in the Enterprise. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3025–3037. ACM, 2016.

S. Han, S. Sörås, and O. Schjødt-Osmo. Governance of an Enterprise Social Intranet Implementation: The Statkraft Case. In *23rd European Conference on Information Systems, ECIS 2015, Münster, Germany, May 26-29, 2015*, 2015.

P. Hansen and K. Järvelin. The Information Seeking and Retrieval process at the Swedish Patent- and Registration Office: Moving from Lab-based to real life work-task environment. In *Proceedings of the ACM SIGIR 2000 Workshop on Patent Retrieval*, 2000.

D. Hawking. Challenges in enterprise search. In *Proceedings of the 15th Australasian Database Conference - Volume 27*, ADC '04, pages 15–24. Australian Computer Society, Inc., 2004.

D. Hawking. Enterprise Search. In R. Baeza-Yates and B. Ribeiro-Neto, editors, *Modern Information Retrieval*, pages 645–686. Addison-Wesley, 2nd edition, 2010.

D. Hawking and K. Griffiths. An Enterprise Search Paradigm Based on Extended Query Auto-completion: Do We Still Need Search and Navigation? In *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS '13, pages 18–25. ACM, 2013.

D. Hawking and J. Zobel. Does Topic Metadata Help with Web Search? *JASIST*, 58(5):613–628, March 2007.

D. Hawking, C. Paris, R. Wilkinson, and M. Wu. Context in enterprise search and delivery. In *Proceedings of ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, pages 14–16. Royal School of Library and Information Science, Copenhagen, 2005.

D. Hawking, P. Thomas, T. Gedeon, T. Jones, and T. Rowlands. New methods for creating testfiles: Tuning enterprise search with C-TEST. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Boston, 2009.

M. Hertzum. Expertise Seeking: A Review. *Information, Processing and Management*, 50(5):775–795, September 2014.

M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: Searching for documents as well as for people. *Information Processing and Management*, 36(5):761–778, 2000.

K. Hofmann, L. Li, and F. Radlinski. Online Evaluation for Information Retrieval. *Foundations and Trends in Information Retrieval*, 10:1–117, 2016.

B. J. Jansen, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.

B. J. Jansen, A. Spink, C. Blakely, and S. Koshman. Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(6):862–871, April 2007.

B. J. Jansen, A. Spink, and I. Taksa, editors. *Handbook of Research on Web Log Analysis.* IGI, 2009.

K. Järvelin, P. Vakkari, P. Arvola, F. Baskaya, A. Järvelin, J. Kekäläinen, H. Keskustalo, S. Kumpulainen, M. Saastamoinen, R. Savolainen, and E. Sormunen. Task-Based Information Interaction Evaluation: The Viewpoint of Program Theory. *ACM Transactions on Information Systems (TOIS)*, 33(1):3:1–3:30, March 2015.

H. Jhamtani, R. Saha Roy, N. Chhaya, and E. Nyberg. Leveraging Site Search Logs to Identify Missing Content on Enterprise Webpages. In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 506–512. Springer, 2017.

T. Joachims. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142. ACM, 2002.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161. ACM, 2005.

M. Johansson and L. Westerling. Designing for Enterprise Search in a Global Organization. In *Third Workshop on Human-Computer Interaction and Information Retrieval (HCIR)*, pages 74–77, 2009.

N. Kanhabua, R. Blanco, and K. Nørvåg. Temporal information retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2015.

P. B. Kantor, E. Boros, B. Melamed, V. Meñkov, B. Shapira, and D. J. Neu. Capturing human intelligence in the net. *Commununications of the ACM*, 43(8):112–115, August 2000.

J. Karim, I. Antonellis, V. Ganapathi, and H. Garcia-Molina. A dynamic navigation guide for webpages. In *CHI 2009*, pages 1–4. Stanford InfoLab, September 2009.

D. Kelly. *Understanding implicit feedback and document preference: A naturalistic user study*. PhD thesis, Rutgers University, 2004.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.

S. Khan, U. Kanturska, T. Waters, J. Eaton, R. Bañares Alcántara, and M. Chen. Ontology-assisted Provenance Visualization for Supporting Enterprise Search of Engineering and Business Files. *Advanced Engineering Informatics*, 30(2):244–257, April 2016.

J. Kim and W. B. Croft. Retrieval Experiments Using Pseudo-desktop Collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1297–1306. ACM, 2009.

J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9^{th} ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677. ACM, 1998.

B. Klimt and Y. Yang. The Enron Corpus: A New Dataset for Email Classification Research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *ECML*, volume 3201 of *Lecture Notes in Computer Science*, pages 217–226. Springer, 2004.

R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 959–967. ACM, 2007.

R. Kraft and J. Zien. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, pages 666–674. ACM, 2004.

D. R. Krathwohl. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218, November 2002.

U. Kruschwitz. An Adaptable Search System for Collections of Partially Structured Documents. *IEEE Intelligent Systems*, 18(4):44–52, July/August 2003.

U. Kruschwitz. *Intelligent Document Retrieval: Exploiting Markup Structure*, volume 17 of *The Information Retrieval Series*. Springer, 2005.

U. Kruschwitz, D. Lungley, M-D. Albakour, and D. Song. Deriving Query Suggestions for Site Search. *Journal of the American Society for Information Science and Technology (JASIST)*, 64(10):1975–1994, October 2013.

C. C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science (JASIS)*, 42(5):361–371, 1991.

S. Laqua, M. A. Sasse, S. Greenspan, and C. Gates. Do You KnowDis?: A User Study of a Knowledge Discovery Tool for Organizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2887–2896. ACM, 2011.

N. Leavitt. Will NoSQL Databases Live Up to Their Promise? *Computer*, 43 (2):12–14, February 2010.

P. M. Leonardi, M. Huysman, and C. Steinfield. Enterprise social media: Definition, history, and prospects for the study of social technologies in organizations. *Journal of Computer-Mediated Communication*, 19(1):1–19, 10 2013.

H. Li, Y. Cao, J. Xu, Y. Hu, S. Li, and D. Meyerzon. A New Approach to Intranet Search Based on Information Extraction. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 460–468. ACM, 2005.

P. Li, P. Thomas, and D. Hawking. Merging Algorithms for Enterprise Search. In *Proceedings of the 18th Australasian Document Computing Symposium*, ADCS '13, pages 42–49. ACM, 2013.

Y. Li, Z. Liu, and H. Zhu. Enterprise Search in the Big Data Era: Recent Developments and Open Challenges. *Proceedings of the VLDB Endowment*, 7(13):1717–1718, August 2014.

X. Liu, H. Fang, C.-L. Yao, and M. Wang. Finding Relevant Information of Certain Types from Enterprise Data. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 47–56. ACM, 2011.

X. Liu, F. Chen, H. Fang, and M. Wang. Exploiting entity relationship for query expansion in enterprise search. *Information Retrieval*, 17(3):265–294, 2014.

J. Lu, S. Pan, J. C. Lai, and Z. Wen. Information at your fingertips: Contextual ir in enterprise email. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 205–214. ACM, 2011.

H. Lund and S. Ørnager. Company Taxonomy development: The case of an international emergency response organization. *Aslib Journal of Information Management*, 68(2):193–211, 1 2016.

M. Lupu and A. Hanbury. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7(1):1–97, 2013.

A. Makhani. Structure, Personalization, Scale: A Deep Dive into LinkedIn Search. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1081–1081. ACM, 2015.

T. Mandl, C. Womser-Hacker, and N. Gätzke. Has Retrieval Technology in Vertical Site Search Systems Improved over the Years? A Holistic Evaluation for Real Web Systems. *Journal of Information Science Theory and Practice*, 3(4):19–34, 2015.

C. Manning, R. Prabhakar, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 2008.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, May 2011.

G. Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, April 2006.

G. Marchionini and R.W. White. Information-seeking support systems. *IEEE Computer*, 42(3):30–32, 2009.

G. Mark, I. Guy, S. Kremer-Davidson, and M. Jacovi. Most Liked, Fewest Friends: Patterns of Enterprise Social Media Use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '14, pages 393–404. ACM, 2014.

K. Markey. Twenty-five years of end-user searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology (JASIST)*, 58(8):1071–1081, June 2007.

C. Massey, T. Lennig, and S. Whittaker. Cloudy Forecast: An Exploration of the Factors Underlying Shared Repository Use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2461–2470. ACM, 2014.

C. Mathieu. Practical Application of the Dublin Core Standard for Enterprise Metadata Management. *Bulletin of the Association for Information Science and Technology*, 43(2):29–34, 2017.

D. Maxwell and L. Azzopardi. Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 731–740. ACM, 2016.

M. Melucci. Contextual search: A computational framework. *Foundations and Trends in Information Retrieval*, 6(4-5):257–405, 2012.

I. Mergel. *The Social Intranet : Insights on Managing and Sharing Knowledge Internally.* IBM Center for the Business of Government, 2016.

D. Miles. AIIM Industry Watch: Search and Discovery - Exploiting Knowledge, Minimizing Risk. AIIM, 2014.

M. R. Morris, J. Teevan, and S. Bush. Enhancing collaborative web search with personalization: Groupization, smart splitting, and group hit-highlighting. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 481–484. ACM, 2008.

P. Morville and J. Callander. *Search Patterns.* O'Reilly, 2010.

P. Morville and L. Rosenfeld. *Information Architecture for the World Wide Web*. O'Reilly Media, Inc., 2006.

L. Muchemi and G. Grefenstette. Rapid Induction of Multiple Taxonomies for Enhanced Faceted Text Browsing. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 7(4):1–13, 2016.

R. Mukherjee and J. Mao. Enterprise search: Tough stuff. *Queue*, 2(2):36–46, April 2004.

K. Norling and B. Lamb, editors. *Intranets – Handbook for Intranet Managers*. Intranätverk, 2017.

D. W. Oard and W. Webber. Information retrieval for e-discovery. *Foundations and Trends in Information Retrieval*, 7(2-3):99–237, 2013.

D. W. Oard, B. Hedin, S. Tomplinson, and J. B. Baron. Overview of the TREC 2008 Legal Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, NIST Special Publication: SP 500-277. NIST, 2009.

D. W. Oard, W. Webber, D. Kirsch, and S. Golitsynskiy. Avocado Research Email Collection LDC2015T03. DVD, 2015. Philadelphia: Linguistic Data Consortium.

C. Olston and E.H. Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Transactions on Computer-Human Interaction*, 10:177–197, September 2003.

A. Ortiz-Cordova, Y. Yang, and B. J. Jansen. External to internal search: Associating searching on search engines with searching on sites. *Information Processing and Management*, 51(5):718–736, September 2015.

A. Pal. Discovering Experts Across Multiple Domains. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 923–926. ACM, 2015.

S. A. Paul. Find an Expert: Designing Expert Selection Interfaces for Formal Help-Giving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3038–3048. ACM, 2016.

K. Pernice, M. Schwartz, and J. Nielsen. Intranet Design Annual: The 10 Best Intranets of 2006. `https://www.nngroup.com/reports/10-best-intranets-2006/`, 2006.

K. Pernice, A. Schade, and J. Nielsen. Intranet Usability Guidelines, vol. 6: Searching the Intranet and the Employee Directory. `https://www.nngroup.com/reports/intranet-searching/`, 2007.

J. Pitkow, H. Schütze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel. Personalized search. *Communications of the ACM*, 45(9): 50–55, September 2002.

J. Pokorný. Nosql databases: a step to database scalability in web environment. *International Journal of Web Information Systems*, 9(1):69–82, 2013.

A. Popescu, O. Etzioni, and H. Kautz. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI'03)*, pages 149–157. ACM, 2003.

J. Prager. Open-domain question-answering. *Foundations and Trends in Information Retrieval*, 1(2):99–231, 2007.

B. Priest and K. M. Carter. Characterizing Latent User Interests on Enterprise Networks. In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014, Pensacola Beach, Florida, May 21-23, 2014.* AAAI Press, 2014.

F. Radlinski and N. Craswell. Comparing the Sensitivity of Information Retrieval Metrics. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 667–674. ACM, 2010.

P. Ramarao, S. Iyengar, P. Chitnis, R. Udupa, and B. Ashok. InLook: Revisiting Email Search Experience. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 1117–1120. ACM, 2016.

H. Roeckle, G. Schimpf, and R. Weidinger. Process-oriented Approach for Role-finding to Implement Role-based Security Administration in a Large Industrial Organization. In *Proceedings of the Fifth ACM Workshop on Role-based Access Control*, RBAC '00, pages 103–110. ACM, 2000.

L. Rosenfeld. *Search Analytics for Your Site: Conversations with Your Customers.* Rosenfeld Media, Brooklyn, New York, 1st edition, 2011.

T. Rowlands, D. Hawking, and R. Sankaranarayana. Workload sampling for enterprise search evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 887–888. ACM, 2007.

T. Russell-Rose and J. Chamberlain. Real-World Expertise Retrieval: The Information Seeking Behaviour of Recruitment Professionals. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings*, pages 669–674. Springer, 2016.

T. Russell-Rose and T. Tate. *Designing the Search Experience: The Information Architecture of Discovery*. Elsevier (Morgan Kaufmann), 2013.

T. Russell-Rose, J. Lamantia, and M. Burrell. A Taxonomy of Enterprise Search and Discovery. In *Proceedings of the European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR)*, volume 763, pages 15–18. CEUR Workshop Proceedings, 2011.

I. Ruthven. Information retrieval in context. In R. Baeza-Yates and M. Melucci, editors, *Advanced Topics in Information Retrieval*, chapter 8, pages 195–216. Springer, 2011.

S. Z. Saad and U. Kruschwitz. Applying web usage mining for adaptive intranet navigation. In *Proceedings of the 2^{nd} Information Retrieval Facility Conference*, volume 6653 of *Lecture Notes in Computer Science*, pages 118–133. Springer, 2011.

M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213. ACM, 1999.

R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-Based Access Control Models. *Computer*, 29(2):38–47, February 1996.

R. L. T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. *Foundations and Trends in Information Retrieval*, 6(1):1–125, 2012.

F. Scholer, D. Kelly, and B. Carterette. Information Retrieval Evaluation Using Test Collections. *Information Retrieval*, 19(3):225–229, June 2016.

G. Schymik, K. Corral, D. Schuff, and R. D. St. Louis. The Benefits and Costs of Using Metadata to Improve Enterprise Document Search. *Decision Sciences*, 46(6):1049–1075, 2015.

C. Shah. *Collaborative Information Seeking - The Art and Science of Making the Whole Greater than the Sum of All*, volume 34 of *The Information Retrieval Series*. Springer, 2012.

J. Shin, S. Wu, F. Wang, C. De Sa, C. Zhang, and C. Ré. Incremental Knowledge Base Construction Using DeepDive. *Proceedings of the VLDB Endowment*, 8(11):1310–1321, July 2015.

M. Shokouhi and L. Si. Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102, 2011.

C. Silverstein, M. Henzinger, and H. Marais. Analysis of a Very Large AltaVista Query Log. Digital SRC Technical Note 1998-014, 1998.

F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2010.

A. Singh, M. Srivatsa, and L. Liu. Efficient and Secure Search of Enterprise File Systems. In *2007 IEEE International Conference on Web Services (ICWS 2007), July 9-13, 2007, Salt Lake City, Utah, USA*, pages 18–25, 2007.

A. Singh, M. Srivatsa, and L. Liu. Search-as-a-service: Outsourced Search over Outsourced Storage. *ACM Transactions on the Web*, 3(4):13:1–13:33, September 2009.

B. Smyth. A community-based approach to personalizing web search. *Computer*, 40(8):42–50, August 2007.

I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2006)*, NIST Special Publication: SP 500-272. NIST, 2007.

D. Stenmark. One week with a corporate search engine: A time-based analysis of intranet information seeking. In *Proceedings of the Eleventh Americas Conference on Information Systems*, Omaha, Nebraska, 2005a.

D. Stenmark. Searching the intranet: Corporate users and their queries. *Proceedings of the American Society for Information Science and Technology*, 42(1), 2005b.

D. Stenmark. Corporate intranet failures: interpreting a case study through th elens of formative context. *International Journal of Business Environment*, 1(1):112–125, 2006.

D. Stenmark. Analysing terms, pairs, triplets and full queries used in intranet searching. In *Proceedings of the 3rd International Conference on Web Information Systems and Technologies (WEBIST 2007)*, Barcelona, 2007.

D. Stenmark. Identifying clusters of user behavior in intranet search engine log files. *Journal of the American Society for Information Science and Technology*, 59(14):2232–2243, 2008.

D. Stenmark. Information Seeking in Organisations: A Comparative Survey of Intranet Usage. In *Proceedings of the 16th Americas Conference on Information Systems (AMCIS)*, Lima, Peru, 2010.

D. Stenmark and T. Jadaan. Intranet Users' Information-Seeking Behaviour: A Longitudinal Study of Search Engine Logs. In *Proceedings of ASIS&T*, Austin, TX, 2006.

D. Stenmark, F. Gardelöv, and V. Larsson. Why should Organisations Govern Enterprise Search? In *Proceedings of the Twenty-first Americas Conference on Information Systems*, Puerto Rico, 2015.

A. Stocker, M. Zoier, Selver S. Softic, S. Paschke, H. Bischofter, and R. Kern. Is Enterprise Search Useful at All?: Lessons Learned from Studying User Behavior. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '14, pages 22:1–22:8. ACM, 2014.

A. Stocker, A. Richter, C.hristian Kaiser, and S. Softic. Exploring barriers of enterprise search implementation: a qualitative user study. *Aslib Journal of Information Management*, 67(5):470–491, 2015.

D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, May 1997.

T. Svarre and M. Lykke. Simulated Work Tasks: The Case of Professional Users. In *Proceedings of the 5th Information Interaction in Context Symposium*, IIiX '14, pages 215–218. ACM, 2014.

M. Taghavi, A. Patel, N. Schmidt, C. Wills, and Y. Tew. An Analysis of Web Proxy Logs with Query Distribution Pattern Approach for Search Engines. *Computer Standards & Interfaces*, 34(1):162–170, January 2012.

J. Tait. An Introduction to Professional Search. In *Professional Search in the Modern World - COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, volume 8830 of *Lecture Notes in Computer Science*, pages 1–5. Springer, 2014.

J. Teevan and S. Dumais. Web retrieval, ranking and personalization. In I. Ruthven and D. Kelly, editors, *Interactive Information Seeking, Behaviour and Retrieval*, pages 189–203. Facet Publishing, 2011.

J. Teevan, M. R. Morris, and S. Bush. Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 15–24. ACM, 2009.

J. Teevan, S. T. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction*, 17(1):4:1–4:31, April 2010.

P. Thomas and D. Hawking. Evaluation by Comparing Result Sets in Context. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 94–101. ACM, 2006.

P. Thomas and D. Hawking. Experiences evaluating personal metasearch. In *Proceedings of the Second International Symposium on Information Interaction in Context*, IIiX '08, pages 136–138. ACM, 2008.

P. Thomas, K. Noack, and C. Paris. Evaluating interfaces for government metasearch. In *Proceedings of the Third Symposium on Information Interaction in Context*, IIiX '10, pages 65–74. ACM, 2010.

J. W. Treem and P. M. Leonardi. Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association. *Communication Yearbook*, 36:143–189, 2012.

D. J. Tunkelang. *Faceted search.* Morgan & Claypool Publishers, 2009.

D. Turnbull and J. Berryman. *Relevant Search.* Manning Publications, 2016.

M. Upshall. Text mining: Using search to provide solutions. *Business Information Review*, 31(2):91–99, 2014.

S. Vaithyanathan. Building search systems for the enterprise. `http://www.slideshare.net/YunyaoLi/sigir-keynote`, 2011.

A. van der Lans. Enterprise Search and Retrieval (ESR): The Binding Factor. In P. Baan, editor, *Enterprise Information Management: When Information Becomes Inspiration*, pages 175–209. Springer, 2013.

H. M. Venkateshprasanna, R. D. Gandhi, K. Mahesh, and J. K. Suresh. Enterprise search through automatic synthesis of tag clouds. In *Proceedings of the 4th Bangalore Annual Compute Conference, Compute 2011, Bangalore, India, March 25-26, 2011*, 2011.

E. M. Voorhees. The TREC Medical Records Track. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, pages 239:239–239:246. ACM, 2013.

W. Weerkamp, R. Berendsen, B. Kovachev, E. Meij, K. Balog, and M. de Rijke. People searching for people: Analysis of a people search engine log. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 45–54. ACM, 2011.

S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core Metadata for Resource Discovery. RFC 2413, 1998.

A. Wexelblat and P. Maes. Footprints: History-rich Tools for Information Foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 270–277. ACM, 1999.

G. J. White. *The Medieval English Landscape, 1000-1540.* Bloomsbury Academic, 2012.

M. White. *Making Search Work: Implementing Web, Intranet and Enterprise Search.* Facet Publishing, 2007.

M. White. Critical success factors for enterprise search. *Business Information Review*, 32(2):110–118, 2015a.

M. White. *Enterprise Search.* O'Reilly, 2nd edition, 2015b.

M. White and S. G. Nikolov. Enterprise Search in the European Union: A Techno-economic Analysis. Scientific and Policy Report by the Joint Research Centre of the European Commission, 2013.

R. W. White. *Interacting with Search Systems.* Cambridge University Press, 2016.

R. W. White and R. A. Roth. *Exploratory Search: Beyond the Query-Response Paradigm.* Morgan & Claypool Publishers, 2009.

R. W. White, I. Ruthven, and J. M. Jose. Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 57–64. ACM, 2002.

R.W. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166. ACM, 2007.

S. Whittaker and C. Sidner. Email Overload: Exploring Personal Information Management of Email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 276–283. ACM, 1996.

M. Wu, A. Turpin, J. A. Thom, F. Scholer, and R. Wilkinson. Cost and benefit estimation of experts' mediation in an enterprise search. *JASIST*, 65(1):146–163, 2014.

W. Wu, D. Kelly, A. Edwards, and J. Arguello. Grannies, Tanning Beds, Tattoos and NASCAR: Evaluation of Search Tasks with Varying Levels of Cognitive Complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 254–257. ACM, 2012.

J. Yan, W. Chu, and R. W. White. Cohort modeling for enhanced personalized search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 505–514. ACM, 2014.

L. Yang, S. T. Dumais, P. N. Bennett, and A. H. Awadallah. Characterizing and Predicting Enterprise Email Reply Behavior. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2017, Tokyo, Japan.* ACM, 2017.

K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 401–408. ACM, 2003.

B. Yu and R. Wang. Research of Access Control List in Enterprise Network Management. In W. Du, editor, *Informatics and Management Science VI*, pages 121–129. Springer, 2013.

C. Zhai and S. Massung. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* Association for Computing Machinery and Morgan & Claypool, 2016.

W. Zheng, H. Fang, C. Yao, and M. Wang. Search Result Diversification for Enterprise Data. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1901–1904. ACM, 2011.

H. Zhu, A. Löser, S. Raghavan, and S. Vaithyanathan. Navigating the intranet with high precision. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 491–500. ACM, 2007.

A. Zouzias, M. Vlachos, and V. Hristidis. Templated Search over Relational Databases. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 21–30. ACM, 2014.