In notitia i confide - Enterprise search and information quality

Martin White [1]

[1] *Visiting Professor, Information School, University of Sheffield*

Abstract

Enterprise search implementations first started to be undertaken in the early 1970s. Although there may well be around 100,000 research papers published on information retrieval there is only one that provides a detailed enterprise search case study. Related research indicates that content quality, along with technology limitations and a lack of training, contribute to a significant lack of satisfaction with the performance of enterprise search applications. This paper is based on experience gained by the author from enterprise search projects undertaken by the author between 2009 and 2019, highlighting the impact of information quality on enterprise search performance and satisfaction.

Keywords

Enterprise Search, Information Quality, Information Management

1. The enterprise search research gap

Enterprise search applications have a long history [1]. A convenient date to mark the beginning of commercial enterprise search is 1970 and the launch by IBM of STAIRS (Storage and Information Retrieval System). This was an evolution of the AQUARIUS software that IBM developed to cope with the internal enterprise documentation for the defence of an anti-trust suit in the USA that started in 1969. Enterprise search is now ubiquitous as a component of Microsoft 365 and there are over 80 other vendors of enterprise search applications, many of them based on the Apache Lucene/SOLR open-source search applications.

The academic discipline of search is ‘information retrieval’, a term which emerged in the mid-1950s. There is an IR Anthology site [2] which currently lists over 62,000 research and conference papers on information retrieval, but this is most certainly an understatement as the number of journals included is very limited, for example the Journal of Information Science. Just as a guess for the purposes of this paper there could be as many as 100,000 research papers.

However, there are no more than 10 that consider in detail the information seeking behaviours of enterprise employees, and many of these focus on a specific community, such as software engineers and patent agents. [3]

There is just one paper that provides a detailed case study of the corporate-wide use of enterprise search using a mixed-methods methodology. [4]

There have been a number of surveys published over the last decade that indicate perhaps only 1 in 5 organisations offer employees a very satisfactory search experience.[5]

ISKO UK Conference, July 24–25, 2023, Glasgow, Scotland

EMAIL: martinswhite@outlook.com (A. 1)
ORCID:

This paper considers the reasons for this lack of satisfaction, and the extent to which information quality is a factor. It is based on the author's personal experience with around 40 enterprise search projects between 2009 and 2019.

2. What do we mean by 'search'?

Given the title of this paper it is important to understand the diversity of 'search' applications and processes. The recent arrival of ChatGPT has led to vendors and observers of search applications to announce either that search is now dead or that AI will significantly improve the performance of search applications. The use of 'search' in this context is similar to using 'car' to describe everything from a Fiat 500 up to a Maclaren supercar. In some situations (the design of car parks) this may be a valid use but not in terms of fitness for purpose (taking the family on holiday).

It is important to understand that there is a difference between 'fitness to specification' and 'fitness to purpose' and any gap between them will almost inevitably result in a workaround by the employee to enable employees to meet both their personal objectives and those of their organisation. There are eight categories of 'search'.

1. Web search – Publicly available information with effort being paid to quality, metadata and links.
2. Website search - Publicly available information with effort being paid to quality, metadata and links.
3. Intranet search - Highly curated web-server based enterprise-specific information searchable with an internal search application. [6]
4. Academic search - Research services for academic users with highly curated content on special-purpose commercial and open-source applications. [7]
5. E-commerce search - Highly curated content accessed through a specialist website or a vendor-specific search application. [8]
6. Professional search – Specialist collections of curated content for lawyers, clinicians, patent agents and other professional groups who use search intensively. [3]
7. Systematic search - Highly curated content used by experienced researchers where a very high degree of recall and reproducibility are important. [9]
8. Enterprise search - Structured and unstructured content, often in multiple languages and very little of which is curated. [10]

3. The characteristics of enterprise content

Enterprise content is invariably written for a defined audience that the author is either familiar with personally or has a good knowledge of the potential readership of the document. [11] It will almost certainly contain (for example) internal project names, trade and internal names for products, alphanumeric product tags and short-hand expressions for offices ('the team in Boston report…').

In many cases there will be no author identified, just that the document has been prepared by HR. There will be multiple versions of similar documents which may vary in title and scope. Date tagging is very important. Knowing the date of the most recent modification can be very misleading if a document written in 2019 has now had a spelling correction made.

An important issue for multinational organisations is that content can be in multiple languages. Employees searching for information may be doing so in their second or even third language, which may mean that they do not have an appreciation of synonyms that could be used to improve search quality and that the content items retrieved could be in a language of which they have only a limited ability to read.[12]

4. Content quality

Just because a document has been deemed relevant by an algorithm does not mean to say that it is useful. Content quality is a major issue in search implementation because it is often very obvious that the most ‘relevant’ references listed on the SERPs (Search Engine Results Pages) are widely different in terms of quality and value to an employee.

Research on the use of Enterprise Content Management applications (ECM) [13] (which invariably have good search functionality) strongly indicates that that the problems lie not with the functionality of the ECM with regards to either adding content or finding it but with the quality of the content that is retrieved. The research suggests that there are two aspects to enterprise content quality assessment.

To quote Laumerm Maier and Weitzel:

“The first is representational information quality. Our analysis of the interviews indicates that the format of information is an important influencing factor for user satisfaction and a unique dimension in our additional analysis. This dimension reflects the way information is presented to the user and subsumes related characteristics of information including conciseness, presentation, and understandability. Conciseness reflects the rigor and the sententiousness of information, presentation refers to the format and the way information is designed to make it understandable to users, and understandability is the extent to which information is clear, unambiguous, and easily comprehensible. All these characteristics have in common that they focus on the way information is presented to the user and reflect the requirement that information needs to be represented in an appropriate format that accentuates its meaning. They are independent of the use of information in a specific context.

The second dimension we identified in our interview analysis was contextual information quality, an important influencing factor for user satisfaction. This dimension reflects the extent to which information fits the needs of the task the information is used in. In our analysis, we identified completeness, relevance, timeliness, and usability as information characteristics which we subsumed into the contextual information quality dimension”.

Some examples of poor information quality the author has encountered include:

- missing versions of documents.
- not being able to be sure that the version found is the current version.
- no specified date of initial authorship.
- authorship attributed to a department and not to an individual, making verification very difficult.
- no context about why a document has been prepared and any restrictions on its scope.
- references in a document to related documents but without the information needed to be able to locate and obtain them.
- no information about when a PowerPoint presentation was given and whether it has been modified following presentation to correct errors
- the scope of Excel spreadsheets and a lack of a ‘last updated’ comment
- titles on documents that bear no relationship with the content, a particular issue with PowerPoint presentations.

The fundamental issue is that few organisations have developed a set of information quality policies and even fewer have implemented an effective governance structure to achieve conformance.

As a result, employees have to place their trust (and reputation) in information that they cannot validate.

5. Enterprise searching – why and how?

The image of the lone employee faced with a challenging problem and having to rely on a search application to find an expert (as proposed by many search application vendors) is questionable. Enterprises were full of supportive teams even before the advent of wide-scale remote working as a reaction to the Covid pandemic. Moreover, employees are in receipt of data and information from many database applications, email and social media. When confronted with the need to locate information in order to make a decision, employees will invariably have a collection of information to hand but need to verify, update and expand this collection of information. The result is that the search query is often along the lines of 'More like this' and the employee already has a good vocabulary of query terms.

That has implications for relevance, precision and performance metrics. The outcomes of a search could include a significant number of relevant documents on the first two SERPs but these may well duplicate information the employee has already acquired. Relevance and value are not synonyms. Efforts by an enterprise search team to improve the click-through on the first two SERPS may do nothing more than increase the effort involved in doing so with no visible benefit to the employee.

For a significant number of enterprise queries a search application will return a substantial number of results defined broadly by the business scopes of the enterprise. Research on professional searching suggests that different communities of professionals make use of different aspects of the user interface to filter the results and this pattern of use is similar across employees in the enterprise as each becomes a 'professional searcher' in their various roles, responsibilities and teams.

This is a good point to consider the balance in enterprise search between precision and recall. Many (too many!) search vendors claim that their applications deliver high precision results on the initial SERP. In the enterprise environment there are many situations where a reasonable degree of recall is of value in validating the initial query and perhaps an initial collection of documents. The employee may then wish to either narrow the scope to improve precision or expand the scope to improve recall. This links into the issue of stopping strategies, which is of significant importance in enterprise search but lies outside the scope of this paper. [14]

6. The role of snippets

In enterprise search situation relevance is obviously important but equally so is the quality of each result. This cannot be directly assessed but it is likely that the user will be considering a range of clues from the result snippet. These might include:

- the quality (especially clarity) of the title
- the name of the author
- their position in the organisation
- the department for which they were working when they authored the document
- the origination date of the document
- the file format where it might have an impact on accessibility
- the language
- the size of the document and therefore the challenge of locating the position in the document of the information satisfying the query terms.

In effect the employee is seeing the extent to which there is an audit trail that enables them to check on the quality of the information they have found. This is a contributory factor to the very high level of search queries for people in the organisation. Some might be to find an 'expert' but many more will be to check out the authority of an unknown employee as a means of assessing the veracity, value and quality of the information.

Although there has been research undertaken into the value of various snippet formats [15] none of this research has been on enterprise search use.

7. Search dissatisfaction

Despite over almost sixty years of enterprise search deployment a number of surveys conducted over the last two decades indicate that perhaps in only 1 out of 5 organisations are employees very satisfied with the performance of enterprise search. [5]

Cleverley and Burnett [16] indicate that issues around technical performance, content quality and training are the root causes of search dissatisfaction.

A fundamental problem is that enterprise search is implemented on the basis that it can be used (in principle) by any employee without the need for training and support. All too often the assumption of both senior IT and business managers is that search is intuitive. Research indicates that search training makes a substantial difference to employee search performance. [17]

The reality is that very few employees have the expertise and experience to construct effective queries and to assess the value of the results that are delivered. Each employee has their own domain knowledge and expectations and has multiple information seeking options of which search is just one of many.

8. Impact of AI

Despite the claims that Large Language Model (LLM)-based applications mark the end of 'search', no result has yet been carried out on information discovery in the enterprise. Much attention has been paid to the use of generative artificial intelligence (AIGC) in the form of summaries of documents and machine translations, and what I refer to as 'faux-search' when a short summary is given in response to a prompt-based query. [18]

With (at this stage) very little focus on the use of private LLMs in the enterprise, it is difficult to do more than highlight the extent to which an employee is going to be able to validate the content of any AIGC outputs. This challenge will be even greater when the multiple languages prevalent in an enterprise are taken into account in the design of training sets and the modification of these to reflect changes in the scope of the enterprise.

It is also important to accept that just changing the search technology is not going to make any significant improvement to employee satisfaction with enterprise search.

9. Five steps to achieving enterprise search satisfaction

There are four steps that need to be taken to ensure that employees can use enterprise search to find information of the highest quality in order to make decisions of the highest quality.

1. Adopt an information management strategy and related policies within a pragmatic governance structure.
2. Integrate the information management strategy with the AI strategy.

3. Select search technology software on the basis of both functional and non-functional requirements.
4. Identify content categories where the risk to the organisation from consistently poor information quality puts the organisation at risk and take remedial action that is then carried forward as examples of good practice.
5. Develop training and mentoring schemes for all employees (but especially for newcomers) that is specific to the technology and the use cases of the content.

10. References

[1]M. White, A History of Enterprise Search 1938-2022, University of Sheffield, 2022. URL: https://sheffield.pressbooks.pub/eshistory1/

[2] IR Anthology, URL: https://ir.webis.de/anthology/

[3] T. Russell-Rose, J. Chamberlain, L. Azzopardi, Information retrieval in the workplace: A comparison of professional search practices, Information Processing & Management 54(6) (2018) 1042-1057.

[4] M. Lykke, A. Bygholm, L.B. Søndergaard, K. Byström, The role of historical and contextual knowledge in enterprise search, Journal of Documentation, (2022) 78(5), 1053-1074.

[5] M. White, Achieving Enterprise Search Satisfaction, 2023. URL : https://searchresearch.online/wp-content/uploads/2023/01/Achieving-enterprise-search-satisfaction.pdf

[6] A. Kayley, Intranet Search Essentials, Nielsen Norman Group, 2022. URL: https://www.nngroup.com/articles/intranet-search/

[7] O. Hoeber, D. Patel, D Storie, A study of academic search scenarios and information seeking behaviour, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, pp. 231-235.

[8] M. Tsagkias, T.H King, S. Kallumadi, V. Murdock, M.de Rijke, Challenges and research opportunities in ecommerce search and recommendations. ACM SIGIR Forum 54(1), (2021) 1-23.

[9] P. Levay, J. Craven (Eds) Systematic Searching. Facet Publishing 2019, ISBN 978-1-78330-374-8

[10] P.H. Cleverley, S. Burnett, Enterprise search: A state of the art, Business Information Review, 36(2) (2019) 60–69. URL: https://doi.org/10.1177/0266382119851880

[11] C. Mathieu, Defining knowledge workers' creation, description, and storage practices as impact on enterprise content management strategy, Journal of the Association for Information Science and Technology, 73(3) (2022) 472-484.

[12] M. Harvey, D. Brazier, E-government information search by English-as-a Second Language speakers: The effects of language proficiency and document reading level, Information Processing & Management, 59(4) (2022) 102985.

[13] S. Laumer, C. Maier, T. Weitzel, Information quality, user satisfaction, and the manifestation of workarounds: A qualitative and quantitative study of enterprise content management system users. European Journal of Information Systems 26 (2017) 333–360.
URL: https://www.tandfonline.com/doi/full/10.1057/s41303-016-0029-7

[14] D.M. Maxwell, Modelling search and stopping in interactive information retrieval, Doctoral dissertation, University of Glasgow, 2019. URL: https://theses.gla.ac.uk/41132/

[15] M. Bink, S. Zimmerman, D. Elsweiler, Featured Snippets and their Influence on Users' Credibility Judgements, in: CHIIR '22, Proceedings of the 2022 conference on Human Information Interaction and Retrieval, March 14–18, Regensburg, Germany, 2022.
URL: https://doi.org/10.1145/3498366.3505766

[16] P.H. Cleverley, S. Burnett, 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) (2017) 77-96.

[17] Y.-L Lee, E.A. Chu, Chu, S. K.-W., Lee, M. M.-L. Chiu, R. C. H. Chan, Scaffolding in information search: Effects on less experienced searchers, Journal of Librarianship and Information Science, 48(2) (2015) 177-190.

[19] N. F. Liu, T. Zhang, P. Liang, Evaluating verifiability in generative search engines, 2023, arXiv preprint arXiv:2304.09848. URL: https://doi.org/10.48550/arXiv.2304.09848