Search Insights 2020

The Search Network

February 2020

Contents

Editorial services provided by Val Skelton (val.skelton@blythespark.co.uk)
Design & Production by Simon Flegg - Hub Graphics Ltd (www.hubgraphics.co.uk)

Introduction

The Search Network is a community of expertise. It was set up in October 2017 by a group of eight search implementation specialists working in Europe and North America. We have known each other for at least a decade and share a common passion for search that delivers business value through providing employees with access to information and knowledge that enables them to make decisions that benefit the organisation and their personal career objectives. The Search Network is an informal community, not a hub-and-spoke network. You can talk to any one of the members and they can bring in others as appropriate.

Members of the Network have web site search, enterprise search and search application development expertise with on-premise, hybrid and cloud implementations. We can work as individuals or micro-companies. We have no commercial relationships with any search vendor or implementation partner. We often assist in identifying vendors for evaluation and consideration. In the course of our work we have gained a substantial amount of experience in application selection and application procurement and implementation which can be matched by very few IT managers.

Some of us have experience with commercial vendors (including SharePoint) and others work mainly with open source search applications. We recognise that the best option is the one that most closely meets the requirements of the organisation. Often these requirements involve members of the Search Network bringing in colleagues with specific skills or to extend our geographic scope. But successful search implementations are not just about choosing the best technology. Search is not a product or a project. It requires an on-going commitment to support changing user and business requirements and to take advantage of enhancements in technology.

Search Insights 2020 is our third annual report. Our objective in writing this report is to summarise some of the insights we have gained from these projects and make this knowledge open to the search community world-wide. That is why there is no charge for this report, and it carries no sponsorship. We have tried not to duplicate the content of the 2018 and 2019 reports so that together they represent a compendium of information and advice that you can trust.

In total the contributors to Search Insights 2020 have well over 50 years of experience in helping organisations to find business-critical information, working with enterprise search, e-commerce and web site search, and with specialised search applications. Not only do we work with different types of search applications, but we also write in our own style and from our own individual experience. As was the case with Search Insights 2019, we have invited guest writers to share their views on specific topics.

Our most significant contribution to our clients is a very good understanding of what an effective search application can deliver in terms of business benefits and employee engagement. Very few organisations have had an opportunity to see and use the range of search applications that we have worked on.

We look forward to helping you achieve search excellence.

You can download previous editions of our Search Insights report here:
Search Insights 2018
Search Insights 2019

The Search Network

David Hobbs, David Hobbs Consulting **(USA)**
David helps organisations make higher impact digital changes, especially through the early development of a strategy to best frame these initiatives before they begin. He is the author of Website Migration Handbook and Website Product Management. His clients include the Center for Internet Security, the Library of Congress, the Mideast Broadcasting Company and the World Bank. Follow David on Twitter @jdavidhobbs.

Charlie Hull, OpenSource Connections **(USA & UK)**
Charlie co-founded search consultancy Flax and recently joined OpenSource Connections where he acts as a Managing Consultant and leads operations in the UK. He writes and blogs about search topics, runs the London Lucene/Solr Meetup and regularly speaks at, and keynotes, other search events across the world. He co-authored Searching the Enterprise with Professor Udo Kruschwitz. Follow Charlie on Twitter @FlaxSearch.

Miles Kehoe, New Idea Engineering **(USA)**
Miles is founder and president of New Idea Engineering (NIE) which helps organisations evaluate, select, implement and manage enterprise search technologies. NIE works and partners with most major commercial and open source enterprise search and related technologies. He blogs at Enterprise Search Blog and tweets as @miles_kehoe, @Ask Dr Search and @SearchDev.

Helen Lippell, (UK)
Helen is a taxonomy consultant. She works on taxonomy development projects, including taxonomy audits, ontology modelling, tagging initiatives, semantic publishing, metadata training and more. Her clients include the BBC, gov.uk, Financial Times, Time Out, RIBA and the Metropolitan Police. She writes and speaks regularly and is the programme chair of Taxonomy Boot Camp London. Follow Helen on Twitter @octodude.

Agnes Molnar, Search Explained **(Hungary)**
Agnes is the managing consultant and CEO of Search Explained. She specialises in information architecture and enterprise search. She shares her expertise on the Search Explained blog and has written and co-authored several books on SharePoint and Enterprise Search. She speaks at conferences and other professional events around the world. Follow Agnes on Twitter @molnaragnes.

Eric Pugh, OpenSource Connections **(USA)**
Eric is co-founder and CEO of OpenSource Connections where he helps federal, state and commercial organisations develop strategies for embracing open source software. He co-authored Enterprise Solr Search, now in its third edition. He is interested in how Search is being invigorated by Machine Learning and exploring approaches for sharing data the way the open source movement shares code. You can follow him on Twitter at @dep4b.

Doug Turnbull, OpenSource Connections **(USA)**
Doug is CTO of OpenSource Connections and the author of Relevant Search. His goal is to empower the world's best search teams. He has assisted with search at organisations in a variety of domains. His clients include Wikipedia, Snagajob, Careerbuilder, and many search organisations. Follow Doug on Twitter @softwaredoug.

Martin White, Intranet Focus Ltd **(UK)**
Martin is an information scientist and the author of Making Search Work and Enterprise Search. He has been involved with optimising search applications since the mid-1970s and has worked on search projects in both Europe and North America. Since 2002 he has been a Visiting Professor at the Information School, University of Sheffield and is currently working on developing new approaches to search evaluation. Follow Martin on Twitter @IntranetFocus.

Guest contributors

Dr Paul H Cleverley, www.paulhcleverley.com **(UK)**
Paul is a Geoscientist and Computer Scientist who has worked on search deployments in organisations for the past three decades. He is founder of tech start-up Infoscience Technologies Ltd (Oxford), is on the Board of the non-profit GeoscienceWorld (Washington DC) and is Visiting Professor of Information Science & Technology at Robert Gordon University in Aberdeen. He has published numerous academic peer reviewed papers on enterprise search and text analytics in business. Links and further research to be found at www.paulhcleverley.com.

Elizabeth Haubert, OpenSource Connections **(USA)**
Elizabeth is a relevancy engineer and data architect. She has worked with a spectrum of data transformation needs from high-rate, high-precision, high-performance time-series sensor data to terabyte-scale text and image retrieval systems for the US Patent and Trademark Office. She has worked on person identification and classification systems both in public and private-facing systems. Her recent work with open-source search measurement and analysis has led to a number of recent conference talks and articles.

Max Irwin, OpenSource Connections **(USA)**
Max is a Managing Consultant at OpenSource Connections, which aims to empower organisations and search teams through consulting, strategy, and training. Follow Max on Twitter @binarymax or connect with him on LinkedIn.

Kurt Kragh Sørensen, Intrateam **(Denmark)**
Kurt is the CEO and intranet/digital workplace consultant at IntraTeam. He provides consultancy services, workshops and lectures on intranets, knowledge sharing, SharePoint and Office 365.

Valentin Richter, Raytion **(Germany)**
Valentin is CEO of Raytion, an internationally operating IT business consultancy with a strategic focus on collaboration, search and cloud. Prior to founding Raytion he worked for Trinkaus & Burkhardt, a private bank which is now owned by HSBC. Valentin has studied mathematics and information sciences. Raytion's clients include Global 500 companies spread around the globe from San Francisco to Tokyo and from London to Sydney. Follow Raytion on Twitter @raytion_com.

The Cambrian explosion (of search)

Paul Cleverley

Introduction

As a geoscientist and computer scientist involved in search and discovery deployments within enterprises for almost 30 years, a geological metaphor for 'search' may be apt for 2020.

The Cambrian explosion occurred more than 500 million years ago a rapid burst that diversified life. Most of the major groups of animals appeared during this radiation event and most things alive today (particularly those with hard parts) can trace their origins to this event. There is significant evidence that environmental changes caused this, such as changes in seawater chemistry. This led to a time of great body plan innovation, such as the development of hard exoskeletons - ecological change responding to these environmental factors.

Exploring parallels to enterprise search, environmental changes over the past few years such as technological advances (computer power and machine learning), exponentially increasing information creation and capture and the open source movement have arguably led to an explosion in 'search species'. This chapter aims to cover these 'search species' (Figure 1) in a light-hearted, but hopefully, informative way, providing an abstract framework to illustrate enterprise search.

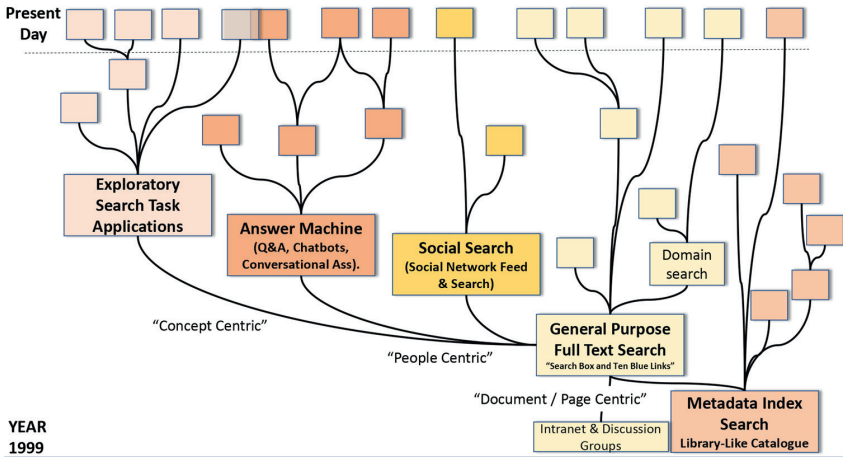


Figure 1: 'Search species'

Metadata only catalogues

Manually tagged documents of various forms have been relatively easy to index and provide search interfaces in the enterprise. Corporate library departments often curated these indexes. They were generally high precision and low on recall, as only a small amount of information was searchable. Sub-species of catalogues evolved incorporating faceted search and other scaffolding functions. The exponential increase in digital information creation (using tools such as word processors) particularly in the late 1990s meant it was no longer feasible for centralised departments to manage all of an enterprise's information. This environmental change saw the deployment of Electronic Document Management Systems (EDMS) and associated full body text search engines. Out-competed, the metadata catalogue declined in popularity and several forms became extinct. There are still niche areas where this type of search remains useful, such as hardcopy/physical asset management.

General purpose corporate Google

The corporate Intranet with its web pages was among the first drivers for a corporate Google search. Over time this often merged with the search of EDM systems, moving towards the text and image based 'enterprise search' concept we know today. This was in the form of a search box and 'ten blue links', mirroring the success seen in the Internet consumer world.

This did not meet all needs, with some functions and departments in enterprises often creating their own search deployments that were rich in functionality such as a map (spatial) interface. Without deep context and other elements that make Internet search successful, these general-purpose enterprise search deployments often had poor user satisfaction, much of which remains today. For example, the experiment to have users tag their own information (everyone becomes a librarian) has for the most part (based on the literature) failed spectacularly in the business world. Nevertheless, this 'search species' flourishes today and remains the mainstay approach for any search deployment in the enterprise. Search ranking (not user interface) being the dominant criteria by which staff judge its success and usefulness.

Social search

Enterprise 'white pages' of people and their expertise, along with discussion forums between people with common interests in organisations, were among the central planks of Knowledge Management (KM) strategies in the late 1990s and early 2000s. The popularity of personal and business social networking technologies in the mid to late 2000s as well as Instant Messaging, led many enterprises to deploy similar technologies.

There appears to be a move towards more cloud-based service adoption by enterprises. The dominance of enterprise social platforms by a very small number of technology vendors (oligopoly) with improved machine learning (ML) capabilities, may allow a reboot of this 'search species' evolving towards automated derivatives of the 'Corporate LinkedIn', where machine generated search 'push' is paramount, complementing the traditional user generated search 'pull'. Some research indicates people are spending as much time (if not more) searching within social networks on the Internet than they are using Google Internet search, where deep context significantly aids 'interestingness' of results.

Answer machine

This 'search species' marks an important evolutionary branch from the classic enterprise search deployments. The focus is answering questions not finding documents. This requires the use of natural language processing (NLP) and machine learning (ML) to convert unstructured text into structured data and information. Concepts and entities dominate, rather than 'the document'.

This structured information is typically combined with existing structured data and information (from databases) with knowledge representations such as taxonomies and ontologies. A knowledge graph is one such way to combine these data and their diverse relationships. These graphs have some 'sub-species' in terms of how the parts are organised (RDF versus Property Graphs) with lively debates on both in the literature. These Graph structures are then used to support apps such as chatbots and conversational assistants. Whilst these answer machines can be voice activated (like Siri/Alexa), they can also be text based to suit the environment. On smartphones or mobile devices, answer machines are more significant due to the real estate afforded to the user, where scrolling long lists of search results or viewing complex visualisations is problematic.

One of the criticisms of these types of search capabilities in the enterprise is the amount of engineering required per area to build them and the lack of generalisability. The Machine Comprehension OpenSource Libraries from the AllenNLP Institute for Artificial Intelligence (AI) offer some intriguing possibilities that may begin to address generalisability in some areas for niche capabilities.

Search task applications

For many high value functions in enterprises, there is a need to mine information for insights and have new information needs stimulated by applications. These often involve rich domain dashboard-like visualisations emphasising the meaningful (rather than returning a simple factual answer).

Rather than search being a 'passive' facility – meeting an existing need the user already has, search task applications are more intrusive; they are precognitive. These applications act as 'assistants', heavily curating what the user sees, notifying us, offering data driven informed opinions for important business activities based on past heuristics and information. They help create new information needs, initiate new studies through highlighting patterns and trends that the user may not have originally thought of. Why wouldn't any professional want opinions from a machine that has read every document in the company?

The new terminology of 'cognitive search' and 'insight engines' from IT market analysts may have been attempts to define this new 'search species' to differentiate it from its ancestors. The mistake some may have made is to think of these 'search species' as replacing its antecedents. As the picture shows in Figure 1, my suggestion is that they co-exist in different niches.

An interesting adaptation of this type of application is the heavy use of NLP and ML which means that they effectively operate on 'structured' data and information. Reports, presentations, papers, web pages are just the provenance of where that structured data originated from. In one sense, they are no longer text based as we know it. The distinction between these types of search driven applications and applications that perform exploratory data analysis, visualisation and analytics on existing structured data is increasingly blurred. It makes us ask questions such as 'what is enterprise search?' In many regards it is certainly the same animal as described by Hawking in 1999, but it is also very different.

Epistemology

Twenty years ago, search was a nice to have. Google did not really exist as we know it today, there was no Siri/Alexa, everyone read newspapers and we all had to carry a map around in the car. The environment has changed. Search has become an epistemology – how we come to know things – algorithms increasingly are the lens by which we encounter much information in the enterprise.

In the enterprise the difference between a search that works 'well' and one that works 'poorly' really can hit the bottom line. Search tools that show staff 'something they didn't already know' can spark ideas that lead to significant new business opportunities. Search is no longer just about 'saving time', search engines curate what we see and can influence what we know.

Digital transformation

Whilst enterprise search follows (lags behind) the consumer world, this is not necessarily the full picture. Business driven digital transformation initiatives, wishing to exploit data driven insights, are increasingly a major environmental driver for search deployments. Unstructured information has arguably been a neglected resource and a majority of organisations now realise this. Many new search deployments and initiatives in enterprises, particularly around answer machines and search task applications, are business driven. This is a major change from the IT and KM led general purpose enterprise search deployments of the past.

Summary

A rich and varied ecosystem exists for enterprise search. If enterprise search is the body, a significant amount of 'body plan' innovation has occurred in response to changing environments. Measuring the success of enterprise search is perhaps evolving past simple user satisfaction metrics of search results lists. These remain important, but enterprise search capability is so much more.

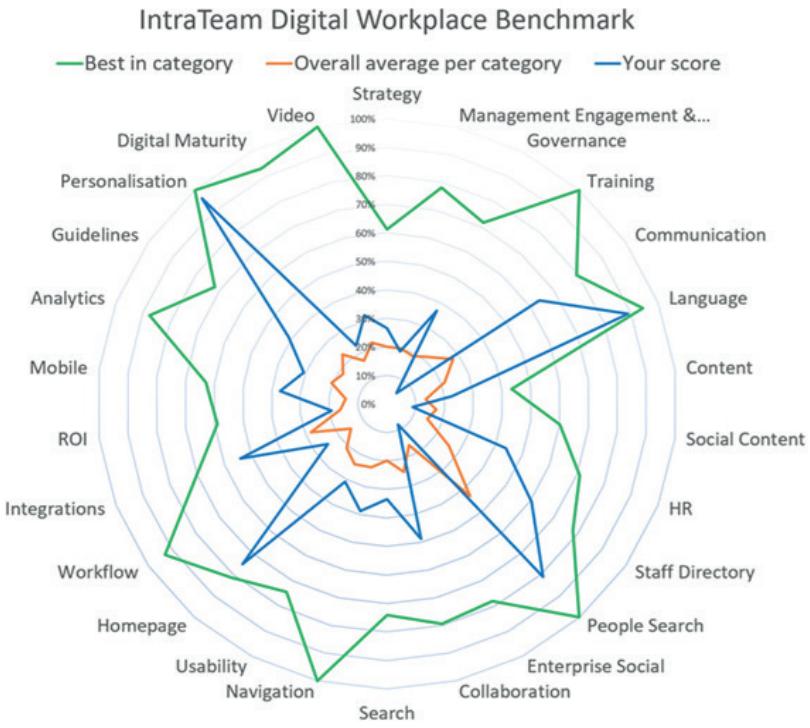
Benchmarking enterprise search – a perspective from Denmark

Kurt Kragh Sørensen

IntraTeam was established in 2000 with a vision of creating and supporting a community of intranet managers in Denmark. Every Spring members of this community come together at the three-day IntraTeam Event in Copenhagen (Denmark). 23 communities in Sweden and Denmark meet quarterly to exchange experience and ideas. There is also an IntraTeam Event held in Stockholm (Sweden) every November.

From the very beginning IntraTeam has carried out surveys among community members to help them understand the opportunities and challenges of intranet management, including search applications. The benchmarking initiative started in 2005 and over recent years has been extended to become a much wider 'digital workplace' benchmark.

There are 26 categories participants can be benchmarked against.



At present over 270 organisations participate in the survey including:

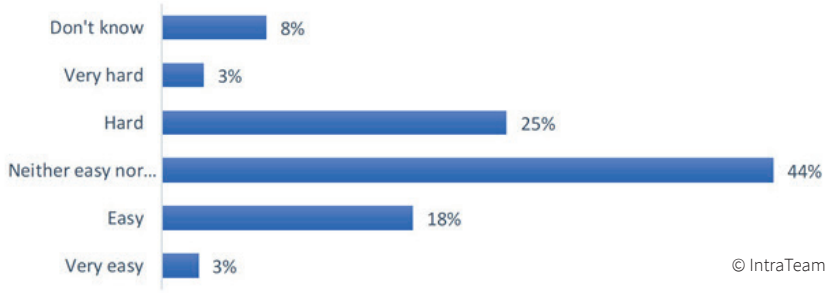
- 80 Danish companies
- 51 municipalities (local administrations)
- 23 Government departments
- A handful of not-for-profit and educational institutions
- 90 companies from other countries

Because this is a community exercise we have confidence in the quality of the information that is given by each organisation. On request we can provide benchmarks for specific industries and sectors.

In this summary the focus is on the outcomes of the search questions included in the survey.

The importance of findability

Responses to the statement "It is easy for users to find the information they need to do their job"



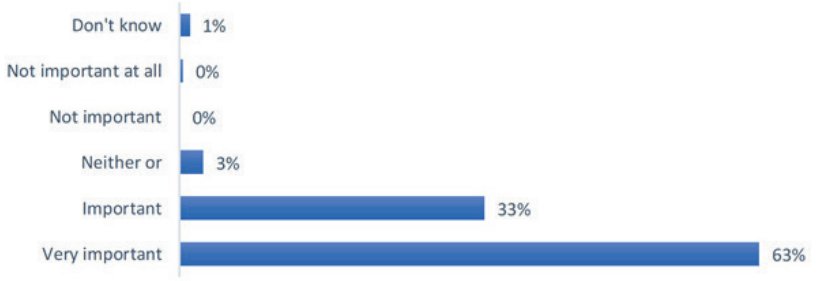
© IntraTeam

This question is not specifically about search as there are many ways in which employees can, in principle, find information. Search is likely used when all other channels have been exhausted. The core outcome of this survey question is that almost a third of respondents (28%) find it hard or very hard to find the information they need to do their job.

This outcome is similar to surveys that have been carried out by Findwise, AIIM and Net-JMC over the last few years and indicates that there is a fundamental problem within many organisations.

The good news is that the participants recognise the importance of being able to find information. The challenge is how to go about providing this improved findability.

Responses to the statement "It is important to improve the findability of the information that the employees need to do their job"

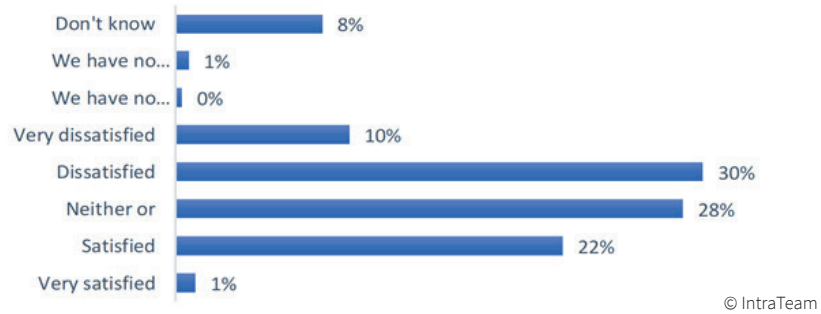


© IntraTeam

Search fails to deliver

If, as seems likely, search applications are a final resort in finding the information, they need to be trusted and easy to use.

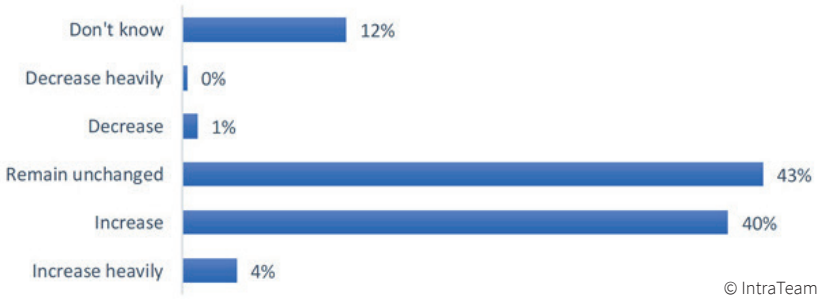
Responses to the statement "Users are satisfied with the internal search functionality"



© IntraTeam

This graph shows that in only 22% of organisations are users satisfied with their search application, with almost half (40%) being either dissatisfied or very dissatisfied. Again this is similar to the outcomes of the Findwise surveys.

Responses to the statement "Investment in search technology will:"



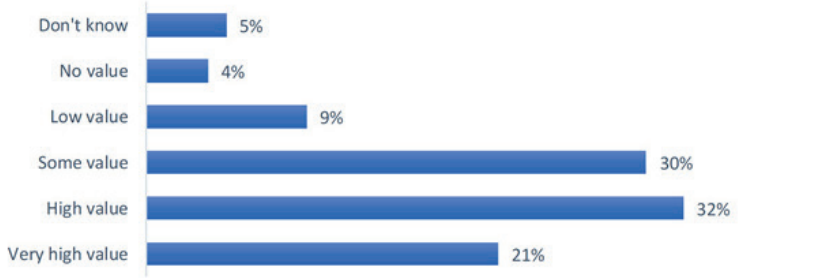
© IntraTeam

Although 44% of organisations are planning to increase their investment in search technology (which is good news for search software vendors), a similar percentage report that there will be no change in the level of investment. However, investment in technology is not a complete answer. The survey shows that where search is delivering, there is usually someone with specific responsibility for search support. None of the organisations in the survey currently have two or more people with search performance objectives.

Search and the digital workplace

When it comes to committing to search, the current trend towards a digital workplace strategy does show promise in stimulating investment in search.

Responses to the question "What value has the digital workplace created re finding relevant people inside the organisation?"



© IntraTeam

This chart shows that there is a very strong interest in finding people within the organisation that have specific skills and/or responsibilities, and this could be an important element in the business case for investing in search. However, as discussed elsewhere in this report, finding employees with specific skills is not the complete solution. Inside many organisations, there are substantial barriers to ensuring that there is a culture of knowledge sharing.

One of the interesting outcomes of the survey is that search performance is closely related to the commitment of senior managers to the digital workplace.

In conclusion

At present, our survey is mainly rolled out to Scandinavian organisations. Although the benchmarking project is gradually expanding in geographic coverage, the numbers are too small to make direct comparisons. However, the fact that these outcomes are very similar to surveys with less detail but a wider geographic coverage suggests that the differences will probably be small.

The survey shows differences between organisations in different sectors and sizes, so we provide much more detailed benchmarking of each participant up against their industry and size. This has great value in making a business case for search investment. We do not charge a fee for participation in the benchmarking survey and would welcome the participation of any organisation that would like to assess the scale and performance of their digital workplace.

You can find out more about the survey here: https://intrateam.com/benchmark/

The advent of natural language information retrieval

Max Irwin

Introduction

Search is changing faster than you'd like to think. Keeping up with big tech firms embracing this change is becoming an increasing challenge. Developments in 2018 and 2019 made significant breakthroughs in measured natural language understanding tasks and, while this has revealed promising paths for improving search using updated software and freely available base models, it won't supplant the need for search teams and relevance engineers. In fact, it will make the roles even more necessary (with increasingly diverse skills required) and more difficult to fill. The good news is that this technology is available to anyone. The bad news is that even best-practicing product teams, traditionally relying on full-text search engines, are not prepared for this change and will be caught off guard.

Natural language queries

As firms such as Google, Microsoft and Amazon push this state-of-the-art technology in more areas, this will heighten the expectations of users in a general sense. For the past two decades, the simplicity of search capability has taught users that a couple of keywords are usually a good way to get decent search results. But with these new expectations, behaviours will change, and users will become frustrated with those not meeting the new standards. The one or two keyword search habit is slowly being deprecated in favour of natural language queries. Google recently revealed that 10% of their queries are now natural language[1]. While this seems small, recent advances have shown that these queries can be responded to much more successfully. Based on this success, the percentage will grow, changing how the average user approaches a search bar.

Transformer technology

So, what is this big change? It is a fundamental way in which text is analysed and compiled into a useable model. Traditionally that has been the inverted index. But transformers[2], a newer AI language architecture, are now taking the lead in accuracy with significant results for various language tasks, including natural language search. The larger technology companies, led by Google, Microsoft, Facebook and Amazon, already have production deployments of these advances, and they are quickly being improved upon. The NeurIPS 2019 conference[3] showcased a staggering number of papers that make the architectures and models even more usable. And, in surprising ways, they have already claimed their thrones at the top of natural language understanding benchmarks[4] by significant margins.

The professional relevance community itself is now starting to catch on, focusing on variant of transformer technology known as BERT[5]. While developing natural language search training courses this summer, I stumbled upon these technologies and have been learning and practising them since. But only recently, based on the training we've given, blog posts that we have published[6,7], and some notes in the search relevancy Slack community, have people really understood the deep impact that this is bringing to the field. Transformer architectures are not the only signal of this change. Neural Search in general is a relatively new field, with the first practitioners' book published in the past year, and several researchers giving conference talks at Haystack and Activate.

The language gap

The most glaring issue is that this points at a key technology gap inherent in the popular full-text search engines. This issue has already been holding search teams back for years. The gap is, unsurprisingly, language. Search engines are notoriously bad at language comprehension of any kind. We don't learn the meaning of words in isolation. We learn words in their context and as part of larger structures in written and spoken communication. Context is always necessary for us to communicate - even with search. Many search engines lack word context, because they parse words one at a time in isolation. Being introduced to an isolated symbol is useless when trying to actually understand something. This is the biggest fundamental drawback of inverted indices, and search engines have spent decades working around this limitation. Mature search teams do their best with what they have, but search quickly fails when real, rich, verbose language is used in a search bar. We have also seen many failed promises before. To name two: semantic search and knowledge graphs both have had their day in the sun, and both have ended in rain and cancelled parties. Expensive parties. But many would now agree, this is the pivotal change that needed a new history, borrowed from the machine learning community.

We're also seeing a shift inside of full-text search engines to address this key language drawback and to be ready for the adoption of techniques like BERT, but it has been lacklustre thus far. Some search engines have adopted new features to prepare for this technology, but performance is underwhelming. Very few teams are using it openly however, so engineers don't have as large a community to fall back on for help like they do with other mainstream search engines. In the proprietary search engines, there is likely significant effort to adopt transformer technology to their language platforms. Also, interestingly, existing recommender systems[8] are being adopted to serve transformer output and augment search platforms to fill the gap. For what is a relevant result, if not a successfully recommended document in response to a query?

New skills and structures for search teams

Indeed, transformer technology is very different and difficult to grasp for existing search teams. Neural networks have not been commonplace in the relevance tuning world. The strategy, practices, knowledge, infrastructure, maintenance, and issue mitigation are all different. Expensive GPUs[9] are required to train models and reach basic performance KPIs in production. Perhaps ironically, using them in the cloud provides even more revenue to Google, Microsoft and Amazon, who are now seen as competitors to many who need to own their search quality. Learning resources for practitioners are also scarce, because the field is changing so fast, and those who venture to use BERT with search are the pioneers. 2020 is likely to see a massive uptick in adoption however, and practical knowledge from those pioneers' experience will be created and shared.

Keeping up with all of this requires a different direction for your search team. It requires bringing data science, machine learning and relevance engineering together in one team. Data science and machine learning teams can no longer exist as separate R&D groups that produce models in isolation. If learning-to-rank wasn't reason enough to bring them together, this makes it an absolute requirement for an organisation leading

against the competition. Comparing the latest 2019 AI Index Report[10] with the previous year[11], there has been a growth from 17% to 24% of companies using "Natural Language Text Understanding" technologies - either in-house or from a third-party service.

The rapid growth is not surprising, as it has been long needed given the constant expansion of content and data. The increasing volume of which is becoming harder (and pricier) to wrangle for most teams. However, many search and data science teams are still fractured in different parts of the engineering and product hierarchy. So, while some groups are working toward the same overall business goals, they are not working collaboratively on the same priorities. Many siloed teams also have different KPIs or OKRs. This misaligns priorities and increases the red tape, which slows or sometimes blocks collaboration. The other expensive side effect is pet project systems that compete internally. One common example is two or more different systems for topical classification, with different taxonomies, and in their own data pipelines.

When looking to outsource the problem, some organisations turn to using a service like IBM Watson or AWS Comprehend. However, doing this without fundamentally solid data science practices will result in poor accuracy and angry customers. Likewise, keeping a data science team separate will produce models in isolation, usually lacking the full picture and missing critical details.

The organisations I have spoken with that have already cross-pollinated (or even fully merged) these groups, are seeing much better overall collaboration towards advanced information retrieval with successful results. Both parties benefit when working together. Search teams become imbued with advanced analysis and measurement practices from data science, and data science becomes imbued with better software practices and customer experience awareness, thus the whole becomes more than the sum of its parts.

Next steps

These key actions are what I recommend in 2020:

- Monitor trends in your query logs – are people shifting away from short keyword and Boolean queries, towards verbose natural language queries? If the answer is yes, that warrants a higher priority of addressing the gap, to be ready for the larger shift.
- Keep up to date with the uptake of neural search capabilities in your information retrieval platform, and also that of your competition.
- Start looking into natural language search training for engineering and relevance teams in 2020. With some basic hands-on experience for a small investment, you'll be able to make a better decision on how to proceed and much better prepared for the future.
- Your search and data science teams are now a single language understanding team, and they are needed more than ever.

References

[1]https://blog.google/products/search/search-language-understanding-bert/
[2]https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)
[3]https://neurips.cc/
[4]https://gluebenchmark.com/leaderboard
[5]https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
[6]https://opensourceconnections.com/blog/2019/11/05/understanding-bert-and-search-relevance/
[7]https://opensourceconnections.com/blog/2019/12/18/bert-and-search-relevance-part2-dense-vs-sparse/
[8]https://erikbern.com/2018/06/17/new-approximate-nearest-neighbor-benchmarks.html
[9]https://aws.amazon.com/ec2/instance-types/p3/
[10]https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf
[11]http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf

Microsoft Search in Office 365

Agnes Molnar

In Search Insights 2019, I wrote about the confusion caused by 'classic' and 'modern' search experiences in Office 365, and the announcement of Microsoft Search. We had a promising vision, and many announcements, but the reality of Microsoft Search was still quite poor.

> *"While the potential of Graph-driven, intelligent and 'personalised' search is clear, there are still many open questions. After years of discussions with industry experts and enterprise customers, Microsoft finally concluded that a wave of significant improvements was needed. Microsoft Graph is already mature enough to support significant search upgrades.*
>
> *But first, an important decision had to be made. (…) In 2018 Microsoft publicly announced its commitment to improving the modern search experience in Office 365."*

Of course, lots has happened in the past year. Microsoft Search has evolved, and it's time to do a review, take a look at the roadmap – and to discuss what we can expect in the upcoming year and beyond.

Microsoft Search today

As of today (early 2020), we still have two out-of-the-box search options in Office 365, with an additional third one:

- 'Classic' Search
- Microsoft Search ('Modern' Search)
- PnP Modern Search web parts.

Let's review each of these options in turn, considering their strengths and weaknesses and then consider what we know about the future of Microsoft Search.

'Classic' Search in Office 365

Although Microsoft made it clear that 'Classic' Search is 'legacy' in Office 365, it's still very popular. Many organisations invested into its customisations, and they're not ready to move on (yet). The benefits of 'Classic' Search today:

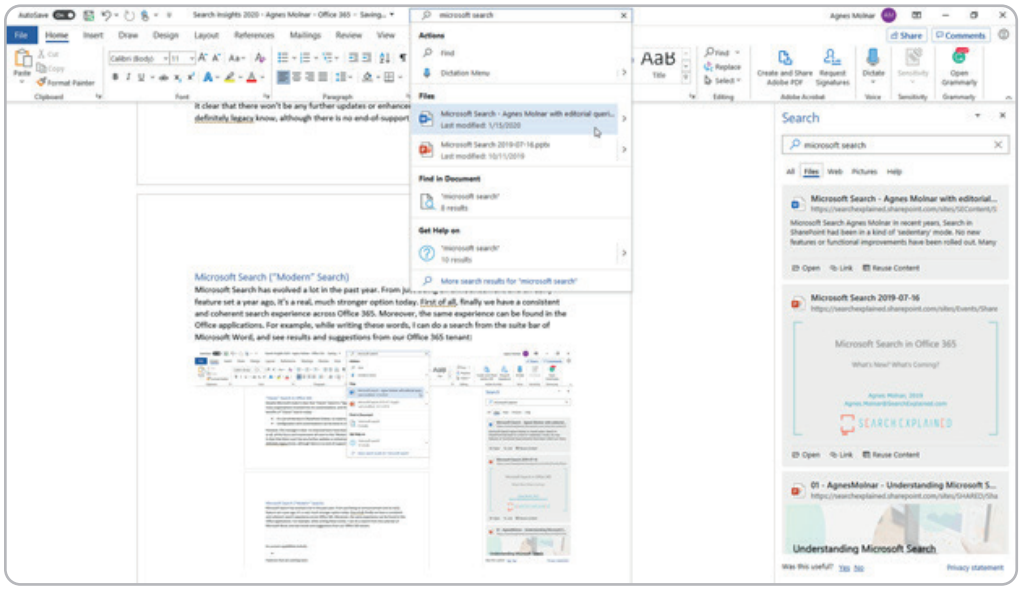
- It's out-of-the-box in SharePoint Online; no need to install or deploy anything
- Configuration and customisations can be done to make it fit your organisation's needs

However, the message is clear: no improvements have been made to 'Classic' search in the last 5+ years at all. All the focus and investment went to 'Modern' Search (see below). Also, Microsoft made it clear that there won't be any further updates or enhancements in the future. 'Classic' search is definitely legacy now, although there is no end-of-support date yet.

Microsoft Search ('Modern' Search)

Microsoft Search has evolved a lot in the past year. From just being an announcement and a first feature set a year ago, it's a real, much stronger option today. Finally, we have

a consistent and coherent search experience across Office 365. Moreover, the same experience can be found in the Office applications. For example, while writing these words, I can search the suite bar of Microsoft Word, and see results and suggestions from our Office 365 tenant:



1. Search is not a plug and play situation. It is not a case of replacing one technology with another, assuming that the default configuration will work, and that the benefits will automatically flow through. As with any change, success is driven through the combination of people, processes and technology. Only by understanding what you want to achieve, can you harness the technology to deliver the results that you are looking for.
2. Search is not a one-off activity. Making search work for your organisation requires ongoing, iterative review and changes. Even with a good implementation the way in which customers search, the content and the product set will change and as such that will change the results displayed to customers. A constant review of how search queries are performing and what changes you can make to improve them should be undertaken. And that requires dedicated support within your business, from someone who understands the business objectives, how search works and more importantly the desired customer experience.

The benefits of Microsoft Search today:

- It's out-of-the-box in SharePoint Online, no need to install or deploy anything
- It's modern
- It's updated regularly, and new features are being rolled out frequently

There are two more important things to know about this feature.

First, everything here is personalised. The suggestions and results are coming from Microsoft Graph, which respects who I am, what I've been working on recently, who I am connected to, etc.

Second, everything here is security trimmed. If you have no access to a document, there is no way you can discover it here.

Its current capabilities include:

- Bookmarks
- Q&A
- Acronyms (being rolled out)
- Locations
- Floor Plans (being rolled out)
- Search Connectors (being rolled out) and APIs.

Taking a look at the Office 365 Roadmap, we can also see that there are many Microsoft Search improvements in the 'Development' phase. A few significant updates and their Feature IDs can be found in the list below. Please note, Microsoft updates the schedule quite often. If you want an up-to-date schedule, consult the Roadmap.

- Customise search results for your organisation (32738)
- Search scoping controls with Microsoft Search (57098)
- Custom verticals and refiners in Microsoft Search (57054)
- Semantic search in Microsoft Search (57063)
- Spelling suggestions in Microsoft Search (57127)
- Query alterations using SPFx (SharePoint Framework) for custom results page (57135)

As you can see, Microsoft is working hard to add customisation features to Microsoft Search. Once these updates are rolled out (later in 2020, according to the Roadmap), especially in combination with the Search Connectors and APIs, Microsoft Search will reach its full power.

Until then, we must wait or accept current capabilities.

PnP Modern Search web parts

The two options above leave us and every organisation in a severe dilemma: invest (more) in 'Classic Search' because this is the only option that can be fully customised today; or use 'Microsoft Search' as it is today, with its limited configurations, and zero customisation.

Both options are far from ideal.

This is how the SharePoint PnP Community comes into the picture. The SharePoint Development Community (also known as the SharePoint PnP Community) is an open-source project where Microsoft and external community members are sharing their learnings around implementation practices for Office 365, SharePoint & Office. This community controls SharePoint (and Office 365) development documentation, samples, reusable controls, and other relevant open-source initiatives related to SharePoint (and Office 365) development. The PnP library uses the Microsoft Open Source Code of Conduct.

The PnP Community recognised the pain point of Microsoft Search not being customisable. The PnP Modern Search solution allows us to build custom, user-friendly search experiences in SharePoint Online, using SPFx (SharePoint Framework) in the modern user interface.

This is the option where all of these are now available:

- Query suggestions
- Custom refiners (can be "classic"-like refiners on the left side, or "modern" filters on a right-side panel)
- Custom search verticals
- Promoted results
- Result set (can be displayed as a list of results, as well as tiles – or custom!)
- Drop-down to re-order the results
- Synonyms
- Multi-lingual search
- NLP enhancements
- And much more

The PnP Modern Search solution can be considered as a bridge between 'Classic Search' and Microsoft Search. It can help organisations to customise Search to their needs, and the users to adopt Office 365 easier.

Project Cortex

In November 2019, Microsoft announced Project Cortex, the new knowledge network feature / vision in Microsoft 365. Although this is not "search" per se, it is strongly related.

According to Microsoft

> *Project Cortex uses AI to create a knowledge network that reasons over your organisation's data and automatically organises it into shared topics like projects and customers. It also delivers relevant knowledge to people across your organisation through topic cards and topic pages in the apps they use every day.*
>
> *In addition, Project Cortex enables business process efficiency by turning your content into an interactive knowledge repository—with innovations in smart content ingestion—to analyze documents and extract metadata to create sophisticated content models; machine teaching, to allow subject matter experts to teach the system how to understand semi-structured content; and knowledge retrieval, to make it easy for people to access the valuable knowledge that's so often locked away in documents, conversations, meetings, and videos. Building on the content you already have in SharePoint, Project Cortex connects content across Microsoft 365 and external systems and enables you to manage information and streamline processes with built-in security, compliance, and workflow.*

Project Cortex can also connect to content in third-party repositories and systems using the new Microsoft Search connectors (see above).

At the time of writing Project Cortex is in private beta. Once rolled out to everyone, it will be a premium feature in Office 365. I believe, with its AI and machine teaching features, Project Cortex will help us human beings to do our jobs better. Combine this with all the information architecture options we have in Office 365, add Microsoft Search, which is also promising – and you can see a really promising feature here. It'll take a while, and probably we'll see some bumps on the road, but hopefully, it'll be beneficial for all Office 365 users.

The expectations are high, Microsoft set the bar very high with the promise of Project Cortex. I am looking forward to having my hands on it and being able to write about my experiences in Search Insight 2021. One thing to keep in mind: knowledge does not organise itself. Storing, organising, curating, and managing knowledge needs - and will always need - intense human involvement, even with powerful tools like Project Cortex.

Summary

In my opinion, 2020 will finally be an exciting year in Microsoft's Search ecosystem. After a few years of being sedated, Search is an important topic again – probably more than ever. With the new features in Microsoft Search, and also the general availability of Project Cortex, I'm confident there will be a lot to share next year again. I'm already looking forward to it!

Content integration

Valentin Richter

Introduction

When I started studying mathematics and information sciences back in the late 1980s, I had some remarkable teachers. One of them told me that a deep interest in a subject often leads people to be drawn towards those who are leading thinkers in that topic. He was right. In 2004 I met Martin White in New York at the Enterprise Search Summit. I thank him for sharing a deep interest in enterprise search with me for all these years.

Putting content to work

One definition of knowledge is 'actionable information'. Knowledge is information which allows you to make better decisions. Data is no more information than hundreds of tons of cement, steel and paint are an office building. To continue that analogy, knowledge would be about the business you pursue within the building. Enterprise search allows an enterprise to turn the data it has spread across its various, and often disparate, information systems into knowledge that is readily available to support its business and drive it forward.

Now approaching its fourth decade, the Internet and its world-wide web of information is, still, a seemingly unorganised chaos. Efforts to catalogue it and put a uniform access structure on it like the Yahoo Directory and DMOZ directory are no longer available. The Yahoo Directory closed at the end of 2014. DMOZ followed that route early 2017. However, it is thanks to search that the Internet works as a source of knowledge. "Google is your friend" is commonly heard advice, if people don't know something. Search makes information, which is distributed across many sources, valuable and puts it to a useful purpose. On the Internet, search works. But searching for information in the context of an enterprise has its own peculiar set of challenges to address: the content you want to access by using search, and the user experience you need to provide in order for search to be effective for its users and your business.

Enterprise search technology and content

To start with, you need an enterprise search technology, either from an established commercial vendor or open source software. Even after the consolidation we've seen over the last couple of years, there are still several active vendors of enterprise search technology in the market, all offering to solve search for businesses. In contrast to Google on the Internet, these vendors face two specific challenges.

In an enterprise or organisation not everybody allowed to access all information or even to know that a particular piece of information exists. Enterprise search has to make sure that a user only sees the information she or he is entitled to see. GDPR regulations make this even more critical.

The second challenge enterprise search faces is: which results should be presented to a user as the answer to their query. For almost any imaginable search query, Google has matching content in its index. For Google's user experience it is enough to present some highly relevant search results. But in an enterprise a user needs and wants the right information and the exact document they are looking for. Getting that UK form from last year is not good enough, when this year's USA form is required for the process you are working on.

What do the vendors of enterprise search technology provide to approach and solve these challenges? A search solution is comprised of three major layers: the user interface, the search index, and the content integration layer. There is a fourth, but I will save that for later.

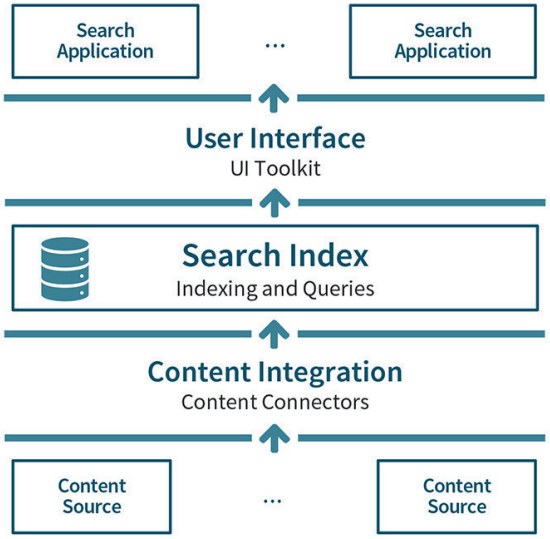


Figure 1: The three layers of enterprise search

All three layers need to be well integrated and tuned. They need to work together as one to reliably come up with timely and relevant results to individual user queries. While the vendors usually focus on providing the search technology, they often leave the actual implementation to their integration partners.

When it comes to enterprise search implementation projects, most of the time is spent on technical aspects. They concentrate on securely integrating the relevant source systems into the search index, scaling the index to the volume of documents required while providing a low latency for user queries. Admittedly, this is already difficult enough. But in order to consistently deliver search results with the quality a business requires, and which users have come to expect from using search on the Internet, the quality of the content integration and the user experience is decisive.

The user interface and the content integration are the two components which make enterprise search work. After many iterations they reflect the distilled expertise on how to put the content an enterprise has stored across its information systems to work via search. They are specific to a particular business and an organisation's culture. If done correctly, you can continue refining that distilled expertise with only marginal efforts even if you switch to the search technology of a different vendor or upgrade to a new major version of the one already in use.

In July 2018 Google announced Google Cloud Search and in November 2019 Microsoft introduced Microsoft Search, its cloud-based enterprise search offering. By running in their respective clouds, they remove the need to take care of the search index, i.e. setting it up, operating and scaling it to reach more users and index more content. This is a substantial change, because it now allows a search implementation team to focus on the aspects which are more important to the quality of the user experience and to the value enterprise search brings to a business.

Content connectors

To integrate content from systems like SharePoint, Confluence, Salesforce, Sitecore or ServiceNow (to name just a few) into their search platform, vendors offer content connectors. Some of the vendors have a broad range of connectors, others focus on a smaller standard set and leave it to their integration partners to provide what else might be required. For the search team to be able to concentrate on the quality of the search results presented to a user, the quality of the connectors is essential. Good connectors are reliable, adapt to complex enterprise requirements especially with regard to security, and have a professional service team behind them which is specialised in content integration for search scenarios.

For small scenarios with less than a hundred thousand documents in a common system, e.g. SharePoint or Confluence, and unless there are special security requirements like single-sign-on, the standard connector provided by the vendor of the search platform might work out of the box. In all other scenarios, you can expect content integration to be a dedicated project probably requiring the professional services of the connector vendor.

Professional services

Professional services are the fourth component of a successful search solution. To be more specific: enterprise search professional services. Integrating different information systems into a search application or an enterprise search platform requires special expertise and skills only gained by practice. When integrating content from a system like Documentum or SAP, the know-how about enterprise search e.g. how secure search works or what users expect is more important to make the integration work than the expertise in this particular information system. The decision is about where you want that experience to originate from. It can be either the technology vendor educating your internal search team or a dedicated content integration specialist guiding you along.

Summary

To present timely and relevant results to the users of a business search application and to offer them a consistent and effective search experience it is essential to have all the necessary content integrated into it. Any search project and proof of concept should start with making sure that the required content can be securely and reliably integrated into the search technology of choice. When selecting the content connectors for the implementation one should not only check their availability, but enquire about the team behind them, in order to validate whether it is accessible and will provide guidance, support and professional services whenever needed.

Disclaimer. Raytion is an independent IT business consultancy specialising in search, collaboration and cloud, and vendor of a family of enterprise search connectors. We started building these connectors because either a specific connector was missing, or the functionality and quality of the connector provided by the original technology vendor left something to be desired. We are on good speaking teams with all the major search technology vendors and they regularly choose or recommend us as their implementation partner. Because we provide connectors to their search platforms, we have an in-depth and sometimes inside knowledge of their solutions.

Skills for effective relevance engineering

Charlie Hull

Search engine projects have historically been viewed as the responsibility of the IT department, covering installation, configuration, content ingestion and operations. The user interface may be developed by engineers themselves or with the participation of UX specialists. What this often leads to is a search engine that is performant and reliable but not necessarily accurate and in many cases is mis-aligned with business priorities.

At OpenSource Connections (OSC) we have long believed that without an effective search team, with members drawn from all areas of the business, you can only solve part of the problem. Your search solution is there to address a business case after all – for example to sell products, provide the correct information to users, save them time – but this business case is not always clearly presented to engineers. Search engines also depend very much on good quality content. No matter how clever the technology, this is very much a case of garbage in, garbage out. New, exciting features (nowadays often incorrectly presented as AI) will not help when metadata is inconsistent and data modelling incomplete.

So when you're considering your next search project, what sort of skills should you look for in your search team? Let's consider the roles in a 'perfect' search team (in smaller organisations the same people may hold several of these roles, and it is rare in our experience to find an organisation that has all the roles covered). This list is drawn from OSC's 'Think Like A Relevance Engineer' training materials.

- Stakeholder - responsible for aligning search improvements with financial and corporate benefit
- Product Owner - responsible for ensuring search improvements meet the information needs of the customer
- Project Manager - responsible for planning and prioritising changes that are translated to features from the customer information needs
- Product Developer - responsible for design and UX implementation in the product
- Content Owner – responsible for defining the content set for the product, and coordinating development teams to arrange content access and transport
- Metadata Owner – responsible for defining and managing any metadata assets that are used to improve search, including synonyms, lemmatisation files, spelling dictionaries, word-wheels, etc.
- Architect - responsible for integration strategy and planning of technical changes across the system for cross cutting concerns and big-picture technology fit
- Search Relevance Strategist - responsible for solution strategy and planning of technical changes across the system related to search improvements
- Search Relevance Engineer – responsible for search engine tuning and delivering associated measurements and experiments
- Software Engineer - responsible for solution delivery and detail-oriented implementation of functionality and features related to search improvements
- Data Analyst/Scientist – responsible for analytical data access and transport, identifying customer trends and engagement signals, and coordinating judgement and rating data acquisition

You may already have some of these roles filled in your search team, and some can be generic across many projects – Product Owner for example. However, some of these roles need specific, specialist skills.

A Search Relevance Strategist is someone with long experience of information retrieval, search engine technology and implementation. They know what is cutting edge technology, but also what basic functionality should be built first. They know how to measure search quality effectively and how to design processes to make this happen. They probably don't write production code, but they can guide and mentor others on the team who do. They have good communication skills and can inspire others to improve search quality.

A Search Relevance Engineer has practical, up-to-date knowledge of the search engine you are using, is aware of its features (and drawbacks) and how to implement them to solve relevance challenges. They are trained in information retrieval fundamentals, know how to model source data for effective search and how to format search queries correctly. They can be a highly effective member of your development team, using their experience to build features, right first time.

A Data Analyst/Scientist working on a search project needs to know what can be measured, what is important to measure and what conclusions can be drawn from the data. They can help you create meaningful metrics and visualisations so the whole search team can see the impact of a potential change and identify potential risks and benefits. Search quality is not always easy to measure, and you may not have (or be able to gather) a full picture of how your users interact with a search application, so a pragmatic approach is best. A search data analyst/scientist will have a good grasp of the various relevance measurement tools that have appeared over the last few years, many of them open source software.

Content Owners and Metadata Owners are usually subject matter experts (SMEs). SMEs know your content intimately – in e-commerce search they know what you sell (and importantly what you don't), what your competitors sell and what is the 'right' answer to a search query. They're familiar with part numbers, content areas, what is up-to-date and what is a little stale. They're a helpful librarian who knows which shelf holds that obscure book you can half remember; a gardening expert who knows which fertiliser to use on your roses; a legal taxonomist or a medic aware of the difference between a cardiologist and a cardiothoracic surgeon. They are aware of the structure of your content – which fields are important to search and which are additional context. In your search team these SMEs can explain to engineers why a result is good, or bad for a particular query and they can be essential parts of any search engine test framework, giving expert ratings (but not always agreeing with their colleagues on these).

So now we know how to build the perfect search team, how do we make sure each member has the appropriate skills? As many have discovered when trying to recruit staff for their search project, these skills are relatively rare, and experts can command a high price. OSC's advice is to focus on empowering your search team for success, supporting them to develop their skills so eventually they can fully 'own' the search solution. There are various ways to do this:

- Expert training. OSC and other organisations provide training in search engine basics, relevance engineering and some more advanced topics such as Learning to Rank and Natural Language Processing. The quality of this training can vary, especially at the beginner level, and we recommend you consider training delivered in a practical fashion with exercises and workshops.
- Partnering and Mentoring. If you work with external partners, think about how they can help mentor your team and teach them the skills they will need while they work

on your project. It can be a bad strategy to outsource your search entirely as it can lead to over-dependence.

- Read the literature. There are many books and useful blogs on information retrieval, search engine fundamentals, relevance engineering and even advanced topics such as Deep Learning for Search. (see Search Resources: books and blogs)
- Attend events. There is a range of events where search and relevance topics are discussed, ranging from large commercial conferences such as Lucidworks' Activate[1] and Elastic's Elasticon[2], academic events such as ECIR[3] and SIGIR[4], smaller and more community-driven events such as OSC's Haystack[5], the British Computer Society's Search Solutions[6] and Berlin Buzzwords[7]. In many larger cities there is a regular Search Meetup which is usually free to attend. Encourage and fund your team to attend these events, meet others facing the same challenges and importantly, participate by presenting or even offering to host a Meetup.
- Interact online. There are mailing lists and forums for particular search engines such as Apache Solr[8] and Elasticsearch[9] and more general forums such as OSC's Relevance Slack[10] . Since the search community is widely distributed these can be a good way to keep in touch with others outside of physical events.

In conclusion

An effective search team will be drawn from areas across the business and will require members to have a wide range of skills. The specialist skills required to improve search quality are rare and we recommend that you support your own staff in gaining these, using external partners where necessary, but being aware that you should aim at eventually owning your search. Community participation is vital and helps both with skills development and also to publicise that your own organisation has committed to building effective search – which can help when recruiting and retaining staff, as well as promoting the technical excellence of your approach.

References

[1] https://activate-conf.com/
[2] https://www.elastic.co/elasticon/
[3] https://ecir2020.org/
[4] https://sigir.org/
[5] https://haystackconf.com/
[6] https://irsg.bcs.org/SearchSolutions/2019/sse2019.php
[7] https://berlinbuzzwords.de/
[8] https://lucene.apache.org/solr/community.html
[9] https://discuss.elastic.co/
[10] relevancy.slack.com/

The importance of informed query log analysis

Martin White

Introduction

Enterprise search applications are very good at producing huge amounts of statistical data about the performance of the application and the use being made by employees. In 1956 V. F. Ridgway published a research paper in Administrative Science Quarterly. In this paper Ridgway argued that performance measurements are useful tools, but indiscriminate use and undue confidence and reliance on them results because of inadequate knowledge related to the effects and consequences of their use. He also suggests that the problem with multiple measurements is that the individual is forced to judge whether an increase in effort to improve one area of performance will improve the overall performance, or reduce performance in some other area to more than offset the improvement in the first area.

Nowhere is this perhaps more evident than in the analysis of search logs!

In 1999 Professor Tom Wilson , working at what is now the Information School in Sheffield, developed a very useful schematic for the positioning of information behaviour, seeking and searching.

This positions information search behaviour, the process of using a search application, as just one element of information-seeking behaviour, and that in turn reflects organisational information behaviours. The schematic highlights that search is just one element of information seeking. To interpret search logs, it is important to appreciate that a gap in information availability can be addressed by (as examples)

- Reading through documents stored on personal or team files
- Using an enterprise application (HR, ERP, e-Learning etc.)
- Sending an email to one or more people
- Talking to a colleague or an acknowledged expert
- Posting a request on a social media channel
- Browsing through an intranet
- Checking through a department or team wiki

• Asking for assistance at the next team meeting
• Searching on the web
• Searching on a specific application
• Searching across multiple applications

The act of searching must be put into this wider context so that we not only know how employees search but why they choose search as their option and what they then do with the information they find. A corollary of this is that it is important to appreciate what topics are not being searched for. This should not be interpreted as just poor search application performance but a result of other methods of seeking proving to be more useful.

Assessing search performance

In general, search teams are under-resourced in terms of technical support, training, and the capability to look through the analytic reports that most search applications can deliver out of the box. One of the most common of these is ranking of search terms against frequency of the query being used. In my experience there is a tendency to use this frequency plot just to show the number of queries in a given period in order to demonstrate to a management team that the search application is being heavily used. Not only can this plot be misleading but also there is much more that can be gained with a little more work.

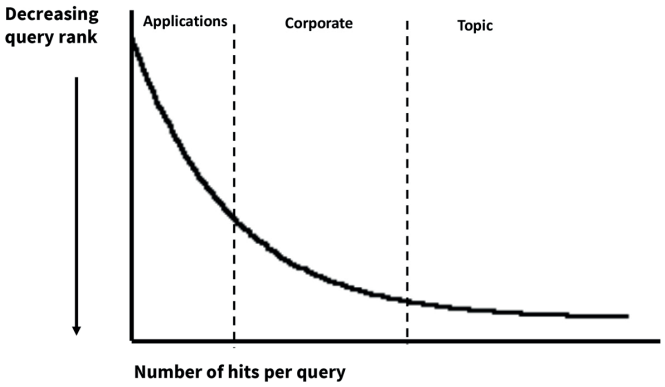
Delving into enterprise search query logs

A table of query terms ranked by frequency of use over (say) a six-month period is usually the main source of information on the queries that have been used. Organisations always seem to be proud of the number of instances of when an initial dozen or so queries were run as a means of showing the level of use of the application but rarely go beyond this level of analysis.

The first step is some initial clustering of related terms. For example, some employees may search for 'expenses' to track down the application while others search for 'Concur' which is a widely used expenses management application in larger organisations. Just to make this task linguistically more challenging Kosten, Auslagen, Spesen and Aufwendum are all potential German-language terms for 'expenses'.

Identifying these clusters is very helpful in building lists of synonyms and from these proposing alternate query terms to users.

When this analysis is completed the rank of query terms against the number of searches always ends up close to the curve shown below.



The shape of the curve is not an artefact of technology but of linguistics. It is an example of Zipf's Law, named after the linguist George Kingsley Zipf (1902-1950) who first proposed it. Zipf's Law states that given a large sample of words used, the frequency of any word is inversely proportional to its rank in the frequency table. One of the core constructs of a text search application is the term frequency.inverse document frequency (tf.idf) measure. The shape of the curve can provide a great deal of useful information if you know where to look for it and then (more importantly) how to analyse the outcomes.

In this diagram the curve is segmented into three areas.

Applications

It comes as a surprise to many that the most frequent queries are to find applications and information on how to complete a task, such as filing expense claims. Search is being used because tracking down the applications on the corporate intranet has become a nightmare, especially in larger companies where the home page is just full of news. The solution to the problem is not to try to improve search but instead work on the information architecture of the intranet.

Corporate news and policies

Where search really starts to come into its own is in helping employees track down news stories and corporate policies. These are text-rich and often search is the only way to find the most recent version. It could be argued that the intranet should be the place to find these policies but often users are looking for some specific terms ('unpaid leave', for example) that might not be reflected in the title or even the summary of a document. Corporate staff (for example in HR) may be quite surprised by the inability of perhaps thousands of employees to find global policy documents.

It comes as a surprise to many organisations that the level of use (in terms of clicks) of highly relevant (based on search algorithms) items is not as frequent as expected. The answer to this conundrum is that employees build up their own collections of core documents. The search application is often being used as a top-up to a collection of documents that may have been collected over a period of time, or perhaps a collection created by a team or project. Many of these may have been pushed to them by other applications, notably email and collaboration applications.

Attempts to improve relevance positioning (often referred to as 'precision at n') is very challenging as some documents may be in core collections for certain groups of employees but not for others.

The volume of searches is sufficiently high for AI/machine learning to be of assistance in delivering relevant results, taking into account information that the application holds on office location, language capabilities and preferences and product area.

This area looks to be where there is a significant inflection in the curve, but this is not the case. The significant number of searches is not a function of tf.idf but of application invisibility. The Zipf curve should really be starting from the most highly ranked of Corporate queries. The inflection is much less pronounced and of course the total number of queries posted by users looks much less impressive.

Topics

It is not until this point that subject-related queries dominate. There is always a very long tail. What is interesting in this area is the relative ranking of terms. Taking a six-month view should eliminate cyclic variations (for example hunting down quarterly reports) but could indicate topics that are steadily increasing in query frequency. This may require a review of 'best bets' or an increased crawl frequency on selected servers. Moreover, in recall could be very important.

Another outcome of this curve is to show just how few queries for these topics are actually run. The amount of data that is collected is almost certainly going to be below the level at which a reliance on AI/ML is going to significantly and visibly improve search relevance across all users.

Use with care

There is a great deal of information that can be gleaned from a ranked list of queries, but it is important not to rush to judgement. This is where having a wide network of users is essential to understanding not only what employees are looking for but why they are looking and were they satisfied with what they found.

Ideally there should be a panel of users and managers that reflects a broad cross-section of the subject of the queries. This panel will also be able to place the role of search within the context of other information seeking options, which may lead to changes being made to the optimisation of all these options.

Good practice in taxonomy project management

Helen Lippell

Introduction

In the four years that I've been Programme Chair of Taxonomy Boot Camp London, I've noticed a healthy increase of interest in the business use of taxonomies, ontologies and knowledge graphs. Leading the development of the Programme means I encounter many examples of good practice and excellent real-world implementations. Taxonomy is well and truly breaking out of its roots in the library and information science worlds to find uses in all sorts of digital applications.

In the private sector, taxonomies are business assets, constructed to support products and services that are intended to stand out in their marketplace and, one assumes, earn revenue for the organisation. Taxonomists want information management to be recognised as a core business process, and as something of great value. If people can't find what they're looking for on your website or app, they will go somewhere else. Very few companies have totally unique and proprietary information or data to sell.

In cases where taxonomies are used to help users inside an organisation find and use information, these are often developed because an organisation wants to reduce inefficiency. Taxonomies in the public sector may be needed for either internal or external use, but again are built in order to achieve a specific outcome that benefits the business.

Here I explore some of the reasons why taxonomy projects fail and what taxonomists and others can do to ensure they don't. The issues and mitigations discussed are equally applicable to enterprise search and ontology/knowledge graph projects. The approaches described are designed to be useful whether someone manages a taxonomy as part of a wider role, is a dedicated in-house resource, or is a consultant. I also share insights into award-winning taxonomy projects.

The stages of a taxonomy project

There are many potential bumps in the road between a taxonomy project being conceived, and the taxonomy being a long-term success. It's rarely a smooth or quick process to get a taxonomy or search project approved in the first place. Once the project is completed, the work of embedding a taxonomy properly is usually part of a wider change management initiative. Metaphorically flinging a new taxonomy 'over the garden fence' and expecting users to adopt it from day one, without helping them to understand the business benefits, rarely ends well. In the end, without an ongoing plan to keep the taxonomy for its intended purpose (or purposes), then the investment of time and money will have been squandered.

> Boot Camp's 2019 award for Taxonomy Success of the Year went to RS Components (an electronic components e-commerce brand) because they reshaped their taxonomies in order to demonstrably improve their SEO performance to attract customers.

Business approval

Starting at the beginning, there are many ways that taxonomists can bring colleagues along with them on the 'journey'. It comes down to listening, educating and advocating as much as possible within the organisation. Every conversation is an opportunity to sell the benefits of taxonomies, not just at the senior management level from where a project sponsor might emerge, but also at the peer level, where colleagues might be

wary of how a taxonomy could change their current ways of working.
(One thing that can happen during the project is that the sponsor changes, causing uncertainty and even risking the completion of the project. Business priorities may change as a result. Sponsor changes are outside a taxonomist's control. Other than continuing to cultivate understanding among a range of senior managers, there's not much we can do with this one.)

Finishing the initial phase of development

Implementing a taxonomy, whether on its own or as part of a larger technology project, is a reason for celebration. The hard work of gathering requirements, understanding user needs, understanding the domain and analysing the content or data has been done. This knowledge has been translated into a live, working taxonomy. Everything is cool now, right? Not necessarily.

The most common failure scenario is that once the taxonomy project team has completed its mission, it is disbanded without a commitment to ongoing maintenance of the taxonomy. The project checklist item 'do the taxonomy' is ticked off and not enough consideration is given to day-to-day operation and the maintenance of quality.

Moving into business-as-usual

There is a common misconception that a taxonomy is just 'finished', yet most domains can and do change over time. Terminology moves on, new entities and ideas emerge. Examples of this include how the acceptable language for certain mental health disorders has changed over time with greater social understanding. Even fairly settled vocabularies such as 'the capital cities of the world' have to be updated when countries rename or move their capitals (for those who care, there have been two such examples in the last 12 months alone, namely Kazakhstan and Burundi).

Quite apart from simple factual inaccuracy, there may be wider societal issues of bias and terminology choice to consider in many domains. Some organisations are building vocabularies that allow, for more than one preferred label, because there is no one accepted name for say, a mountain or piece of colonised land.

For these kinds of reasons, it is always advisable to have in place a role (or a part of someone's role) to oversee the taxonomy. Yet even this may cause a problem after a taxonomy is deployed. This is because the taxonomist (assuming there isn't a wider team) becomes a Single Point of Failure. When they change companies, retire, or just move job inside the organisation, their knowledge and enthusiasm may not be replaced. (Of course, many companies do keep their taxonomist role filled, but as an external taxonomist I am more likely to see those places where they haven't done this properly!)

Projects can fail even if those in charge fully understand the value of taxonomies in supporting broader business objectives. Managers can have an ambitious vision for using knowledge graphs to power data-driven products and services, yet still not care enough about the quality of the underlying vocabularies and structures. It was refreshing to hear Electronic Arts (EA) speak at Taxonomy Boot Camp in 2019. The team emphatically cares that the right taxonomies and ontologies get built. The team takes the long view so that future enterprise-wide requirements are borne in mind as well as immediate project needs. Taxonomists form a core part of a team that also works on ontologies, knowledge graphs and content models.

Models, such as the ones EA are building, depend on agreed and shared definitions of the types of entities or things the organisation cares about. For example, in a re-

cent role I was trying to model agreed definitions for digital asset types (e.g. video trailers, background images, call-to-action text). Everyone I talked to agreed there was a need to standardise and reduce the amount of types people were creating. This was because there was rampant inefficiency, duplication of effort, poor communication between production teams and poor findability. Without buy-in from the people overseeing these teams, the processes and systems won't change, and new vocabularies and models won't be adopted. Eventually this may either lead to remediation work being commissioned, or in the worst-case scenario, a total loss of business confidence in using taxonomies at all.

Longer-term sustainable maintenance

Helping others to understand that business-as-usual is as important as project work is a vital task. The taxonomy must be kept relevant and useful for its end-users, and this only has a chance of happening if care and attention are given to ongoing maintenance.

Governance is an important part of this. This can mean anything from a designed framework that encompasses committees and review schedules, to a pragmatic commitment that one person will be responsible for looking after the taxonomy and communicating with others about proposed changes.

> Ed Vald, of the Chartered Institute of Personnel and Development, won Boot Camp's Practitioner of the Year award for 2019 for his project to streamline a sprawling taxonomy down to the essential terms. A key part of this success was the implementation of metadata auditing and governance to prevent the taxonomy becoming such a mess again in future.

It can be easy to fall into the trap of focusing on the 'shiny new thing' that comes out of a project and to underestimate the value of low-key day-to-day deliverables. These include search log analyses, data analytics insights, tagging audits, reporting on taxonomy change requests, and governance processes. All of these things give precious insight into what's working and what's not and may even throw up new and surprising findings. Maybe a whole section of the website is rarely visited, or users are searching en masse for a term variant that no-one thought of. Maybe a taxonomy term is getting disproportionately tagged against content, and it's only because it's the first entry in the autocomplete list for a common word.

Working across disciplines and business siloes for long-term success

There are strategic and practical things that taxonomists can do to help the long-term success of projects. Getting and retaining business buy-in is arguably the most important. There might be one person who is project sponsor, and this is critical, but it is also important to try to build relationships with others across the organisation. After all, taxonomy (and search too!) is not something that fits neatly into one business function. It crosses disciplines such as technology, content strategy, design and user experience, product management and change management. Senior stakeholders in all of these areas should be supported to understand the value of taxonomies.

It's clear that input does not just come from the taxonomist. Other disciplines can and should be involved, e.g. content designers who understand structured content and markup, or developers who understand tagging beyond a simple view of 'stuff added to a piece of content'. I would like to see organisations treat taxonomy, metadata, search and tagging skills are as core to their overall set of digital skills.

By way of analogy, it's increasingly common to observe that professionals who aren't digital or content specialists are expected to contribute content about their particular area to digital workplaces. I would like to see organisations encourage staff to interact more with, and understand more deeply the value of, taxonomies. The best colleagues I've worked with over the years had the curiosity and imagination to deeply understand how taxonomies fitted in with their own work and thus benefitted the wider project.

It's great that things are moving in the right direction (after years of 'is taxonomy obsolete?' blog posts!) But it will be even better when people at all levels of an organisation, and across all sorts of digital disciplines, are fully on board with the excellent work that taxonomists are already doing. Every year we see this superb breadth and depth at Boot Camp; it's time the wider digital industry understood this too.

Changes in open source search

Elizabeth Haubert

The Apache Software Foundation non-profit, dedicated to supporting free, enterprise-grade, open source software, was created about 20 years ago[1]. Its sponsored products (e.g. the Apache HTTP Server, Maven, Tomcat and Spark), have defined the modern tech stack at thousands of companies, while free, open source software released with Apache Licenses has reached even more.

Apache Lucene, a library to support full-text indexing and search, was one of the first Apache projects (released in 1999). It is most commonly deployed as part of two search platforms:

- Apache Solr (2004)
- Elastic's Elasticsearch (2009).

There are not many enterprise websites that do not incorporate search in some way - ecommerce, document portals, message boards, are just some examples. For many of these sites, search is a necessary part of infrastructure. While the out-of-the-box solution is acceptable for many, achieving the relevance users expect requires some customisation, and an open source search platform can be the key to achieving this balance.

The economic models to support open source infrastructure software have changed in the last 20 years. As an Apache project, Solr is maintained by a community of individual committers, and the consortium of companies who employ them, while open-source Elasticsearch is curated almost exclusively by Elastic. There are challenges facing both support models, and there is a risk that the next generation of enterprise search might not be open source.

What does modern search need?

Search is ubiquitous across the internet. It is the primary navigation tool for users on most sites, and users expect everything to 'work like Google' - or better. This means the market for search utilities is huge.

In parallel with the explosion of search deployments, the ecosystem around search has evolved considerably. As machine learning libraries and techniques have become a commodity, the expectation is that byproducts such as entity extraction and semantic search should be available as standard features in a site search engine.

It is becoming standard practice to collect analytics from nearly every aspect of the online experience. In order to be useful, that data needs to be searchable, creating a whole new market of search experiences. There is also the expectation to monitor both user search interactions, and the performance of the search engine. Support from the search engine makes collecting those metrics easier and more performant.

Finally, the amount of data being searched by a custom engine has grown substantially over time. True, a collection of a few thousand documents for a few hundred users isn't unusual, but neither is a collection of a billion documents for thousands of users. The tooling and infrastructure to support such massively scaled applications have matured rapidly in the last ten years, and an open-source solution must support both cases. In parallel, some of the companies providing support for these large scale open-source applications have achieved a high level of financial and institutional maturity.

Open source software is not free of cost. Many open-source tools are available for download without a fee, but those tools represent thousands of hours of work by skilled developers. While the stereotype of dedicated individuals cranking out code for the public good does exist, many contributors either work in consulting around the products they support, or directly for companies founded to champion a particular product. Let's look at the relationship between three open-source search products and three affiliated companies.

Apache Solr and Lucidworks

The first major Lucene-based search engine, Apache Solr, was created in 2004, released as an Apache Incubator project in 2006, and accepted as a top-level project in 2007. It was another two years before Yonik Seeley, Grant Ingersoll and Erik Hatcher joined Lucid Imagination[2], which was later renamed Lucidworks[5]. It is not the only company offering Solr support[3], nor does it represent all Solr contributors[4], but Lucidworks employees (past and present) have been a powerful contributing force to the project.

In recent years, Lucidworks has shifted its focus. Its flagship conference, formerly Lucene Revolution, has been re-named Activate, and re-branded around search-space machine learning. Fusion, its custom toolkit built on top of Solr, is less about Solr itself and more about the ecosystem around it: support for deployment tools, query formulation, and machine learning. The marketplace rewarded this change in focus – in 2019 Lucidworks was granted $100M in funding.[6]

Elastic and Elasticsearch

The second major player in the world of Lucene search is Elasticsearch. Elasticsearch was released in 2010 as a complete re-write of an earlier Lucene-based product called Compass[7] but includes the Apache Lucene libraries as the core search technologies. Two years later, Elastic [the company] was founded to support the ELK stack (Elasticsearch, Logstash, Kibana, and Beats) - a suite of open source tools designed around the ingestion, search, and display of log-based data. This rewrite means that many of the interfaces for search and distributed maintenance can have a cleaner look and feel than the older Solr. In addition, Elastic maintains an additional toolset, XPack, to provide additional features particularly for machine learning and security. XPack was initially maintained in a private repository, but the source code was made openly available in 2018[8]. The licensing here is a little more complex. While the main Elastic stack uses the Apache 2.0 license, XPack is under Elastic's own commercial license. Note that although XPack's source code can be viewed it is not actually open source. Also, Elastic controls any new code contributed to the project, so both Apache and Elastic licensed parts are not developed 'in the open' like Solr.

The open source community has a complicated relationship with Amazon Web Services (AWS)[9]. Since Amazon launched Elasticsearch as a service in 2015[10], the company has continued to flourish[11,12].

The past year has raised new questions about the intellectual property of open source products. Elasticsearch itself includes Lucene and contributes back to the Lucene project. This year, Amazon released an open distribution of Elasticsearch[13]. Elastic (the company) has sued Amazon for trademark infringement[14]. SearchGuard, a company providing various Elasticsearch plugins, has also been sued for allegedly copying parts of the visible source code of the X-Pack codebase[15], which are under Elastic's own license.

Google and TensorFlow

As we look to the next generation of search products which natively support machine learning, analytics and distributed architecture, the change becomes even more pronounced. Let's consider a third, non-Lucene case study. Tensorflow is properly a machine learning engine but can be used for many of the search tasks we might consider for Solr or Elasticsearch. Tensorflow was developed by Google internally first, and open-sourced second. Where Lucene/Solr has had 151 committers since 2001, the Tensorflow documentation alone has accumulated 690 committers since 2016[16]. The community and usage forums are very different from the more free-wheeling Apache look and feel[17], but that may be necessary to coordinate the larger community with a wider set of use cases.

So …

What does this mean for the future of open-source search? As previously mentioned, Lucidworks is not synonymous with Apache Solr. Elasticsearch isn't the first open-source storage system that Amazon has adopted for deployment. Redis Labs is doing better than ever five years after the release of ElasticCache[19,20]. MongoDB's stock actually went up after AWS released DocumentDB (with MongoDB compatibility)[21].

The difference has to do with innovation and timelines. 2020 marks approximately 20 years of Lucene, 15 years of Solr, and 10 years of Elasticsearch. For reference, the average lifetime of a software product is 7-14 years. While there have been significant changes in every release, you don't have to look hard to find key structures that haven't changed significantly in 5-10 years. Meanwhile, the landscape of search has changed. We search more data, more often, and expect faster and better results, and that has very practical implications for the future direction of these engines. Making the changes that keep a search engine modern takes developer time, and in the Lucene space many of those changes are happening around, not in, the open-source codebase.

Whose responsibility is the maintenance and growth of open source software? The stereo-type has been motivated individuals, working on their own for the community good, and this is true in search as well. That 'community good' is very often commercial gain, and not necessarily for those writing and developing the OSS. Not all teams can field an 'extra' developer, and very few teams see it as a corporate responsibility to provide community support for the open source products they use[18]. Similarly, what does it mean to contribute to the open source community? Is it enough to balance the economies to use one tool, and not contribute back to that tool, but to another separate open-source product? These are difficult questions, and they don't have simple answers.

References

[1] www.apache.org/foundation/

[2] https://en.wikipedia.org/wiki/Apache_Solr

[3] https://cwiki.apache.org/confluence/display/solr/Support

[4] https://github.com/apache/lucene-solr/graphs/contributors

[5] https://lucidworks.com/press/lucid-imagination-changes-name-to-lucidworks/

[6] https://lucidworks.com/press/lucidworks-raises-100m-to-expand-in-ai-powered-search-as-a-service-for-organizations/

[7] https://www.elastic.co/about/history-of-elasticsearch

[8] https://www.elastic.co/what-is/open-x-pack

[9]https://www.nytimes.com/2019/12/15/technology/amazon-aws-cloud-competition.html
[10]https://aws.amazon.com/blogs/aws/new-amazon-elasticsearch-service/
[11]https://www.crunchbase.com/funding_round/elasticsearch-series-d--e681cdfb#section-overview
[12]https://www.lastweekinaws.com/blog/amazon-isnt-killing-your-business-you-just-suck-at-it/
[13]https://opendistro.github.io/for-elasticsearch/faq.html
[14]https://searchaws.techtarget.com/news/252471650/AWS-faces-Elasticsearch-lawsuit-for-trademark-infringement
[15]https://www.elastic.co/blog/dear-search-guard-users
[16]https://github.com/tensorflow/tensorflow
[17]https://www.tensorflow.org/community
[18]https://www.researchgate.net/publication/270886220
[19]https://www.redislabs.com/press/redis-labs-announces-4000-paying-customers/
[20]https://www.redislabs.com/press/redis-labs-sees-record-growth-fiscal-year-2019-advancing-instant-experience-database-market/
[21]https://www.cnbc.com/2019/03/13/mongodb-q4-2019-earnings.html

Searching for expertise and experts

Martin White

Introduction

Over the last few years there has been a steady stream of expertise search applications from search vendors (e.g. Attivio, BAInsight and Sinequa) and from specialist application vendors (e.g. Profinda and ThingMap). Both technology and market development has been stimulated by profiling software developed to track terrorists by national and international security organisations. The sales pitch from Microsoft is typical.

> *"How much time do you spend looking for someone who can help answer a difficult question… Finding expertise isn't always easy, out of date profiles, self-nominated areas of expertise, or incomplete people information makes it hard to make the connections that matter the most – finding colleagues who can help you or helping colleagues find you.*
>
> *By integrating the people answers in Microsoft Search when searching for specific topics, we can help people connect easily and efficiently. Microsoft Search is the ideal canvas to present people results for topics as it spans across mail, documents, teams, and is common and natural for people to look up topics unknown to them in search bar."*

The concept of machine-compiled profiles is certainly not new. Sopheon was just one of the companies working on this technology in the late 1990s, as was a team developing P@noptic Expert at CSIRO in 2001. In 2006 MITRE Corporation published a report on the principles of expert profiling and included descriptions of the applications from TACIT, AskMe, Autonomy IDOL K2, Endeca 25, Recommind 30, Triviumsoft's SEE-K and Entopia Expertise Location. Although this report is now fourteen years old it provides still-relevant advice on specifying and evaluating expertise finding applications.

What problems needs to be solved

The underlying business case is that employees are finding it very difficult to track down expertise, assistance and individual experts in their organisation but no evidence is presented to justify this business case. There have always been two mechanisms through which expertise and experts can be found.

The first is through the management reporting line, with employees working up through their reporting line to find either an expertise in their department or subsidiary or gaining advice on who else in the organisation might be able to help. There are two important benefits of this mechanism. The first is that at each level the employee is advised on how best to define their problem. The second is that movement up the hierarchy will be supported by a manager who in effect is vouching for the employee and ensuring that the time of an expert is not being wasted without good reason.

The second is through working in a team. The team would normally be created on the basis that all the expertise needed to find a solution is either in the team, or closely associated with it.

Both mechanisms have been used for decades. Neither are perfect but the percentage of times that neither of these mechanisms work is likely to be very low.

Profiling – from self-completion to algorithmic analysis

Developing and supporting ways to surface the expertise inside an organisation has in the past been the concern primarily of knowledge managers. The challenge has always been how best to capture expertise in a document format. This was the basis for My Site profiles in SharePoint 2007. Employees could include a profile of their expertise in a My Site and this could be searched as a defined field by the SharePoint application. IBM had been carrying out a substantial amount of research in parallel as a component of the development of Domino and subsequently IBM Connections. This work led to one of the IBM knowledge management team, Dave Snowden, publishing a blog in 2008 in which he highlighted some fundamental issues with capturing knowledge/expertise.

http://cognitive-edge.com/blog/rendering-knowledge

His observations include

- Knowledge can only be volunteered. It cannot be conscripted.
- You can't require people to share their knowledge, because you can never measure if they have.
- We only know what we need to know when we need to know it.
- The way we know things is not the way we report we know things.
- We can always know more than we can say, and we will always say more than we can write down.

The implication of these five principles is that no matter how much effort is put into persuading employees to write down what they know and what their skills are, it will only be a very partial and probably biased commentary. Searching through these self-completed profiles is not a complete solution to identifying expertise and knowledge because of the inevitable inconsistencies between profiles. Then comes the challenge of keeping these profiles current. If there is no overall policy on profile management, supported by managers, then there is no incentive for people to spend time on this process. No matter how good the technology of the application there are many situations in which the query may not be matched by the profiles of individual experts.

These include

a. People working on projects where the security clearance is limited to a need-to-know access list. The paradox here is that these could be among the most expert in the organisation as they are working on the most sensitive and innovative projects.

b. People joining the firm may not be in a position to disclose what they have been working on for (perhaps) a competitor. The paradox here is that these people have been specifically hired for their expertise. Since on average around 10% of an organisation's employees leave each year and 10% join, the question is how long it will take for the new joiners to be regarded as having equivalent expertise to colleagues who have been with the organisation for many years.

c. People working for contractors and advisors who are retained specifically for their expertise but as they are not employees, they do not show up in an expertise search.

d. People who do not want to be bothered with others asking for advice. They might well game the system by (for example) not writing blogs or not being named as the lead author on a report.

e. People with soft skills in areas such as mentoring, training and team management. These skills can be very valuable to an organisation but may well not show up on a profile where the emphasis is on technical skills.

f. People with hidden skills who may have joined the organisation from a completely different business area, perhaps as a deliberate change of career. Clearly the organisation sees their work as important but their experience to date may not be reflected in the profile.
g. People who have cultivated a range of external social media and professional networks that may not be included in an internally sourced profile but which could be important indicators of expertise.
h. People working in a language and/or an application that is not included in the indexing of the crawls and therefore do not show up in the responses to a query.

In selecting an expertise searching application the extent to which these situations are recognised and addressed must be taken into consideration.

There is also the 'definition of the problem' paradox. If an employee needs expertise to solve a problem, then they almost certainly need initial assistance to define the problem.

Solution evaluation

As with any search application, undertaking proof of concept tests and user acceptance testing is far from easy. It may be quite straightforward to test some aspects of the user interface with a small group of profiles, but as the entire purpose of the investment will be to locate expertise across the organisation the final tests cannot be carried out without a substantial load of profile information. To do this the application may need to crawl through a significant archive, which presents many technical and procedural problems.

Another consideration is how the performance of the application is going to be assessed. This is not just to ensure that it is working to specification but that it is having a measurable impact on access to expertise over and above the access provided by the two mechanisms presented above. What is almost certainly going to happen is that people will compare the expertise search with a search through documents using the enterprise search application to see if different experts are identified. If there are there need to be some good explanations if employees (and experts) are to trust the application.

Some of the evaluation criteria would be

- How will you ensure that new experts are as discoverable as those who have been working for you for a number of years?
- How will you integrate internal and external measures of expertise?
- How will you surface expertise gained in secure projects?
- How will you reassure your experts that they are indeed discoverable for what they regard as their areas of expertise?
- How will the application discover related information in multiple languages?

Above all else can the vendor present how effective its application has been in another recent customer?

Barriers to expertise sharing

If the challenges of identifying employees with relevant experience is fraught with problems, these pale into insignificance with the challenges of persuading people to share their expertise.

There have been a number of studies carried out into the barriers to expertise sharing once the seeker has located an expert. These barriers are the subject of a review paper by Morten Hertzum in which he reviews the outcomes of 72 papers on expertise finding and provides an excellent introduction to the topic.

Context

- Company size and culture does not support expertise seeking and sharing
- Management cultures (e.g. hierarchical reporting) inhibit direct access to expertise
- There is no incentive for the sharing of expertise
- There is an incentive to refrain from sharing to ensure status as the 'go-to' person

Seeker

- Time-consuming to get a response from the expert
- Expert is not willing to commit to a time when they can be available
- Expert is not willing to commit to a time by when they will reply
- Seeker does not personally know the expert and it is difficult to build a relationship
- Expert wants to know why their expertise is required before agreeing to share
- Seeker feels they are losing face by revealing uncertainty and lack of knowledge
- Seeker cannot formulate question because of a lack of expertise
- Expert only willing to give an oral answer, either because they do not have the time to write a reply or do not what to be quoted as the source of the expertise

Expert

- Expert's knowledge turns out to be incomplete or unreliable
- Expert cannot be immediately located – an issue with experts often travelling
- Expert is perceived as unapproachable or unwilling to help
- Expert is biased in their reply but the bias will not be evident to the non-expert
- Expert has concerns about sharing confidential information
- Expert may have expertise which is not visible as it was gained on a confidential project
- Expert's credibility difficult to assess
- Expert is not up-to-date with developments
- Expertise required is owned by a group rather than an individual
- Expert is external to organisation
- Expert works in a different first language than the seeker which inhibits a dialogue

Stakeholders

Expertise location and exchange should be a component of a knowledge management strategy and not regarded as a technology issue.

The primary stakeholders should be

- Corporate knowledge managers - to ensure that the application is embedded in the KM strategy
- HR - because there will be a sensitivity to who is, and how they are, defined as experts, which might lead to grading and remuneration issues
- Legal - as there could be substantial issues around data privacy guidelines and laws which vary from country to country outside of the EU
- IT - in its role of managing the application, especially in terms of which content should be crawled
- Line-of business managers – to support the business case and evaluate the performance of both the technology and its impact on business performance.

One way, not the only way

Relying on search technology to identify experts is not to be recommended. The key issue is to understand the nature of the problem, and it is likely that experts can in fact be found but that there are significant barriers to the process of expertise exchange. Until these are identified and addressed no technology is going to make any difference. The implications of an expert being overlooked on the expert, on the expertise-seekers and on the performance of the organisation all have to be taken into account. They must be worked through by the stakeholders within the context of a knowledge management strategy. The performance of the technology should be rigorously tested – given a specific problem with known experts are they all being identified by the technology?

Perhaps the major challenge is how the expertise of experienced people joining the organisation will be found using search technology. These people have been specifically recruited because of their expertise but it could take several years before they have written 'enough' documentation on their subject to be found by the search application. Before believing in the power of the technology ask to talk to organisations who are using it to see how these and other challenges mentioned in this contribution have been successfully addressed.

Search resources: books and blogs

The books listed below represent a core library which should be on the bookshelf of any manager with enterprise search responsibilities.

Designing the Search Experience
Tony Russell-Rose and Tyler Tate, 2012. (Book website) (Review)
This book takes a deeper look into information seeking models, using them to consider how best to design user interfaces.

Enterprise Search
Martin White, 2nd Edition, 2015. O'Reilly Media (Book website)
A book written for search managers without a technical background that aims to support the entire process from building a business case through to evaluating performance.

The Inquiring Organisation
Chun Wei Choo, 2015. Oxford University Press (Review)
The importance of this book is that it provides a context for search within an overall integration of the value of information and knowledge to the organisation.

Interactions with Search Systems
Ryen W. White, 2016. Cambridge University Press (Review)
Although the focus of this book is on web search, the principles also apply to e-commerce and enterprise search.

Introduction to Information Behaviour
Nigel Ford, 2015. Facet Publishing (Review)
Information seeking models are a special case of information behaviours. They form the basis of use cases for search, and the design of user interfaces.

Looking for Information
Donald O. Case and Lisa M. Given, 4th Edition, 2016. Emerald Publishing (Book website)
A survey of research on information seeking, needs, and behaviour which places search into the wider context of why people seek information and how they interact with search systems.

Multilingual Information Retrieval
Carol Peters, Martin Braschler and Paul Clough, 2012. Springer (Book website)
A good introduction to the basic principles of multilingual and cross-lingual search.

Relevant Search
Doug Turnbull and John Berryman, 2015. Manning Publications
(Book website) (Review)
The objective of all search applications is to deliver the most relevant results as early as possible in the list of results. Although based around the management of Elasticsearch and Solr this book is applicable to any search application.

Search Analytics For Your Site
Louis Rosenfeld, 2011. Rosenfeld Media (Review)
This introduction to search analytics is primarily about websites and intranets but the principles apply to enterprise search.

Searching the Enterprise
Udo Kruschwitz and Charlie Hull, 2017. Now Publishers (Review)
The authors provide an important bridge between information retrieval research and the practical implementation of search applications.

Text Data Management and Analysis
ChengXiang Zhai and Sean Massung, 2016. ACM/Morgan&Claypool (Review)
A very comprehensive handbook on the technology of information retrieval and content analytics based on a highly regarded MOOC.

Morgan Claypool and Now Publishers both offer a wide range of books on specialist aspects of information retrieval and search, though with an academic rather than a practitioner focus.

This is a list of blogs whose authors comment on aspects of search technology and implementation on a reasonably frequent basis.

Beyond Search Stephen Arnold
Complex Discovery Rob Robinson
Coveo Insights Corporate Blog
Daniel Tunkelang
Do More With Search BA Insight corporate blog
Elastic Corporate blog
Enterprise Search Miles Kehoe
Funnelback Corporate blog
Geodyssey Paul H Cleverly
Information Interaction Tony Russell-Rose
Intranet Focus Martin White
LucidWorks Corporate blog
Opensource Connections Corporate blog
Searchblox Corporate blog
Search and Content Analytics Blog Paul Nelson, Search Technologies
Search Explained Agnes Molnar
Sease Corporate blog
Sinequa Corporate blog
Synaptica Corporate blog
Tech and Me Mikael Svenson

In addition, the Special Interest Group on Information Retrieval of the British Computer Society and the Special Interest Group on Information Retrieval of the Association for Computing Machinery publish newsletters.

Enterprise search chronology

This table is an informal and certainly not definitive chronology of the development of enterprise search, with a particular focus on the mergers and acquisitions that took place between 2008 and 2012.

For a more detailed chronology on a decade by decade basis refer to this series of http://intranetfocus.com/a-history-of-enterprise-search-starting-out/

Many of the innovators in the development of search technology are profiled in the Wizards section of Beyond Search at http://arnoldit.com/wordpress/wizards-index/ There is a very detailed history of the development of information retrieval by Donna Harman at https://www.nowpublishers.com/article/Details/INR-065 but the coverage of enterprise search-related developments is limited.

Year	
1951	Master's thesis by Philip Bagley suggesting that computers could search through text
1957	H.P Luhn (IBM) sets out the fundamental characteristics of a search application
1958	Dow Chemicals sets up a pilot project to search internal documents
1958	H.P. Luhn develops a system for automatically creating abstracts from a document
1959	Maron and Kuhns introduce the concept of relevance
1964	Gerard Salton sets up the SMART project at Harvard University and begins the development of important concepts in information retrieval
1965	Rocchio and Salton consider how best to optimise the performance of retrieval systems
1965	Large-scale remote access search services established by Lockheed Dialog and SDC Orbit
1970	Launch of STAIRS (Storage and Information Retrieval System) by IBM
1974	Initial availability of in-house document retrieval systems using mini-computers, such as BASIS and INQUIRE in the USA and STATUS in the UK
1976	First assessment published of the role of AI in information retrieval
1976	Initial publication by Stephen Robertson and Karen Sparck Jones of the research that eventually led to the development of the BM25 ranking model
1980	Public release by Martin Porter of his SNOWBALL English language stemmer
1980	Thunderstone Software launched as an appliance
1983	Fulcrum Technologies (Canada) launch a client-server search application
1986	Verity established as a spin-out of Advanced Decision Systems offering a probabilistic ranking search functionality in its TOPIC software
1988	dtSearch launched by David Thede, initially as a desk-top search application

Year	
1988	Isys search software application developed by Ian Davies in Australia
1988	Latent Semantic Indexing is developed by Scott Deerwester and his colleagues
1989	Marcia Bates' paper on a berry picking model for retrieval marks the start of research into user approaches to information retrieval
1989	Peter Pirolli develops the concept of information foraging as a model for search behaviour
1992	Stephen Pollitt publishes his research into faceted navigation for search
1993	Retrievalware launched, becoming the first of many competitors to IBM STAIRS and Verity
1993	Ultraseek launched
1995	Verity has a very successful IPO
1996	Autonomy founded by Michael Lynch
1997	FAST Search and Transfer lauched
1998	Google launched
1999	Endeca launched (originally as Optigrab)
1999	Doug Cutting releases Lucene
2000	Vivisimo was founded Chris Palmer, Jerome Pesenti, and Raul Valdes-Perez
2000	Exalead launched by François Bourdoncle and Patrice Bertin (ex Alta Vista)
2000	Autonomy floated on NASDAQ
2001	CSIRO (Melbourne) launches what would become the Funnelback search software
2002	Sinequa launched
2002	Google launches its GSA search appliance
2003	BAInsight founded
2004	Solr was developed by by Yonik Seeley at CNET Networks
2005	Mindbreeze founded as a supplier of appliance search products
2006	Public launch of Amazon Web Services (AWS)
2008	Autonomy floated on the London Stock Exchange
2008	FAST Search and Transfer acquired by Microsoft
2008	Enterprise Search Summit conference launched in New York
2009	Lucid Imagination founded
2010	Exalead acquired by Dassault Systems
2010	Microsoft release FAST Search for SharePoint
2010	ElasticSearch launched
2010	Microsoft launches Azure as a cloud service
2011	Autonomy acquired by Hewlett Packard
2011	Endeca acquired by Oracle
2011	Enterprise Search Europe conference launched in London
2012	Vivisimo acquired by IBM
2012	ISYS Search acquired by Lexmark
2016	Google announces the withdrawl of its GSA search appliance

Year	
2016	Autonomy acquired from HP by MicroFocus
2017	Gartner introduces the concept of Insight Engines in its Magic Quadrant
2017	Forrester introduces the concept of Cognitive Search
2018	Google announces its cloud-based enterprise search service

List of enterprise search vendors

For Search Insights 2020 we have integrated the 'commercial' and 'open source' lists from previous reports as many of the nominally commercial search applications contain open source code elements.

No list of vendors can be comprehensive, and the Search Network would appreciate being contacted by vendors who are not on this list.

The inclusion of a search vendor on this list cannot be taken in any way as an endorsement by members of the Search Network.

Company	HQ	URL
Algolia	USA	https://www.algolia.com
Amazon	Denmark	https://aws.amazon.com/kendra/
Ankiro	USA	https://ankiro.dk/ankiro-enterprise-suite/
Aras	USA	https://www.aras.com/en/capabilities/enterprise-search
Autonomy	UK	See MicroFocus
BAInsight	USA	http://www.bainsight.com
Bloomreach	USA	https://www.bloomreach.com/en
Bonsai	USA	https://bonsai.io/
Cludo	Denmark	www.cludo.com
Copernic	Canada	http://www.copernic.com/en/products/enterprise-search-engine/
Coveo	USA	http://www.coveo.com
Datafari	France	https://www.datafari.com/en/
dTSearch	USA	http://www.dtsearch.com/
Elastic	Netherlands	https://www.elastic.co/elasticsearch
Exalead	France	https://www.3ds.com/products-services/exalead/products/
ExpertSystem	Italy	https://expertsystem.com/
Findwise	Sweden	http://www.findwise.com/en
Funnelback	Australia	https://www.funnelback.com/
Gimmal	USA	https://www.gimmal.com/records-management/enterprise-search
Google	USA	https://cloud.google.com/products/search/
Grazitti		See SearchUnify
Hyland	USA	http://www.hyland.com/en/products/enterprise-search
IBM Watson	USA	https://www.ibm.com/watson/products-services
Ilves	Finland	https://ilveshaku.fi/en/
iManage	UK	https://imanage.com/product/ravn/
Indica	Netherlands	https://indicaplatform.com
Infongen	USA	https://www.infongen.com/solutions/enterprise-search
IntraFind	Germany	https://www.intrafind.de/index_en
Knowlia	Belgium	https://www.knowliah.com/
Lucene	Community	https://lucene.apache.org/
Lucidworks	USA	http://www.lucidworks.com
Microfocus	UK	https://www.microfocus.com/en-us/products/

Company	HQ	URL
Microsoft SharePoint	USA	https://docs.microsoft.com/en-us/sharepoint/dev/general-development/search-in-sharepoint
Microsoft Azure	USA	https://azure.microsoft.com/en-us/services/search/
Mindbreeze	Austria	http://www.mindbreeze.com
Nalytics	UK	https://www.nalytics.com/
Netowl	USA	https://www.netowl.com/enterprise-search
Onna	USA/Spain	https://onna.com/enterprise-search/
Open Source Connections	USA	http://opensourceconnections.com/
OpenText	Canada	https://www.opentext.com/what-we-do/products/discovery
SAP	USA	https://blogs.sap.com/2019/08/16/enterprise-search-the-new-user-experience-for-enterprise-information-processing/
SciBite	UK	https://www.scibite.com
Searchblox	USA	https://www.searchblox.com/
Searchunify	USA	https://www.searchunify.com/
Sinequa	France	http://www.sinequa.com
Solr	Community	http://lucene.apache.org/solr/
Squirro	Switzerland	https://squirro.com/
Swiftype	USA	https://swiftype.com/
Tantivy	Community	https://github.com/tantivy-search/tantivy
Terrier	UK	http://terrier.org/
Thunderstone	USA	https://www.thunderstone.com/lp/enterprise-search/
Vespa	Community	http://vespa.ai/
Voyager	USA	http://www.voyagersearch.com
Yippy	USA	https://yippy.com/

Enterprise search integrators

This list of companies offering enterprise search implementation support has been compiled by harvesting the companies listed on the web sites of major search vendors. As far as we are aware there is no other published list.

It should be noted that many companies specialise in the support of specific vendors. No list of companies offering this capability can hope to be comprehensive, and the Search Network would appreciate being contacted by companies who are not on this list.

The inclusion of a company on this list cannot be taken in any way as an endorsement by members of the Search Network.

Company	HQ Location	Company HQ URL
Accenture	USA	https://www.accenture.com/gb-en/services/applied-intelligence/search-content-analytics
ATOS	France	https://atos.net/en/
Avalon	USA	www.avalonconsult.com
CTC	Japan	https://www.ctc-g.co.jp/en/
Dahu	UK	www.dahu.co.uk
Devoteam	France	https://www.devoteam.com/
Digital Group	USA	http://www.thedigitalgroup.com
DTI	Switzerland	https://www.dti.ch/
Ekimetrics	France	https://www.ekimetrics.com/
Findwise	Sweden	www.findwise.com
Fishbowl Solutions	USA	https://www.fishbowlsolutions.com
FranceLabs	France	www.francelabs.com
Incentro	Netherlands	www.incentro.com
Injenia	Italy	https://www.injenia.it/
Innovent Solutions	USA	www.Innoventsolutions.com
Join	Germany	https://www.join.de/en/
KBenP	Netherlands	https://kbenp.nl/en
KBQuest	Hong Kong	https://www.kbquest.com/
Netmail	Canada	https://netmail.com/
NewSync Technologies	Netherlands	http://www.newsynctechnologies.com/
Nextbrick	USA	https://www.nextbrick.com/
Noovle	Italy	https://www.noovle.com/it/#
Norconex	Canada	https://www.norconex.com/
NRX	France	http://www.nrx.fr/
OpenSourceConnections	USA	www.opensourceconnections.com
Perficient	USA	www.perficient.com
Raytion	Germany	www.raytion.com
Search Explained	Hungary	www.searchexplained.com
Sematext	USA	www.sematext.com
SHI	Germany	https://www.shi-gmbh.com/

Company	HQ Location	Company HQ URL
Sword-Group	Luxembourg	https://www.sword-group.com/en/
Tieto	Finland	www.tieto.com
Uptime	Norway	https://uptime.eu/
Wabion	Germany	https://www.wabion.com/en/
WIPRO	India	https://www.wipro.com/

Glossary

Absolute boosting
Ensuring that a specified document always appears at the same point in a results set, or always appears on the first page of results.

Access control list (ACL)
Defines permissions to access a specific repository, a set of documents, or a section of a document.

Advanced search
The provision of a search user interface which prompts the user to enter additional terms to assist in ranking results, often using Boolean operators.

AI
Artificial Intelligence, in search this often means Machine Learning.

Apache
The Apache Foundation provides support for a wide range of open source applications, including Lucene and Solr.

Appliance
A search application pre-installed on a server ready for insertion into a standard server rack.

Auto-categorisation
An automated process for creating a classification system (or taxonomy) from a collection of nominally related documents.

Auto-classification
An automated process for assigning metadata or index values to documents, usually in conjunction with an existing taxonomy.

Average response time
An average of the time taken for the search engine to respond to a query, or the average end-to-end time of a query.

Best bets
Results that are selected to appear at the top of a list of results that provide a context for other documents generated and ranked by the search application.

BM25
A ranking function developed in the 1990s but still widely used. It has its origins in the tf-idf ranking function.

Boolean operators
A widely used approach to create search queries; examples include AND, OR, and NOT—for example, information AND management.

Boolean search
A search query using Boolean operators.

Boosting
Changing search ranking parameters to ensure that certain documents or categories of documents appear in the results.

Categorisation
The placing of boundaries around objects that share similarities (e.g., taxonomy).

Clustering
A process employed to generate groupings of related words by identifying patterns in a document index.

Cognitive search
A description loosely applied by search vendors to applications using machine learning and AI techniques to determine the work context of the user and deliver personalised results. (See also Insight engine)

Collection
A group of objects methodically sorted and placed into a category.

Computational linguistics
The use of computer-based statistical analysis of language to determine patterns and rules that aid semantic understanding.

Concept extraction
The process of determining concepts from text using linguistic analysis.

Connector
A software application that enables a search application to index content in another application.

Controlled vocabulary
An organised list of words, phrases, or some other set employed to identify and retrieve documents.

COTS
Commercial off-the-shelf software.

Crawler
A program used to index documents.

Cross-language search
A query in one language is translated into other indexed languages (often using a multi-lingual thesaurus) so that all documents relevant to the concept of the query are returned no matter what language is used for the content.

Description
A brief summary, generated automatically, that is then included as a description of a document in the list of results. See also Key sentence

Document
A structured sequence of text information, but often used as a generic description of any content item in a search application.

Document processing
The deconstruction of a document into a form that can be tokenised and indexed.

Document repository
A site where source documents or other content objects are stored, generally a folder or folders. See also Information source

Early binding
A search conducted only across documents that a user has permission to access. See also Late binding

Entity extraction
The automatic detection of defined items in a document, such as dates, times, locations, names, and acronyms.

Exact match
Two or more words considered mutually inclusive in a search, often by enclosing them in quotation marks—for example, “United Nations”.

Facet
Presentation of topic categories on the search user interface to support the refinement of a search query.

Fallout
A quantity representing the percentage of irrelevant hits retrieved in a search.

Federated search
A search carried out across multiple repositories and/or applications.

Field query
A search that is limited to a specific field in a document (e.g., a title or date).

Filter
A function that sets specific criteria for search results.

Freshness
The time period between a document being crawled and the index being updated so that a user will be able to find the document.

Fuzzy search
A search allowing a degree of flexibility for generating hits (i.e., matches that are phonetically or typographically similar).

Golden set
A set of documents used to benchmark search performance that is representative of content that will be searched on a regular basis.

Guided search
A search in which the system prompts the user for information that will refine the search results.

Hit
A search result matching given criteria; sometimes used to denote the number of occurrences of a search term in a document.

Hybrid search
The use of a combination of on-premise and cloud technologies.

Index
List containing data and/or metadata indicating the identity and location of a given file or document.

Index file
A file that stores data in a format capable of retrieval by a search engine.

Ingestion rate
The rate at which documents can be indexed, usually specified in Gb/sec.

Insight engine
A term used by some search vendors to denote a search application which makes extensive use of AI and other software applications to personalise results for each individual user. (See also Cognitive search)

Inverse document frequency (IDF)
A measure of the rarity of a given term in a file or document collection.

Inverted file
A list of the words contained within a set of documents, and which document each word is present in, so acting as a pointer to a document.

Inverted index
An index whose entries identify a given word and the documents in which it appears.

Iterative calculation
A calculation utilising a recursive and self-referential algorithm.

Key sentence
A brief statement that effectively summarises a document, often employed to annotate search results.

Keyword
A word used in a query to search for documents.

Keyword search
A search that compares an input word against an index and returns matching results.

Language detection
The indexing process identifies the language (or languages) of the content and assigns it to appropriate language specific indexes.

Late binding
Access permission checking carried out immediately before the presentation of the document to the user. (See also Early binding)

Lemmatisation
A process that identifies the root form of words contained within a given document based on grammatical analysis (e.g., run from running). (See also Stemming)

Lexical analysis
An analysis that reduces text to a set of discrete words, sentences, and paragraphs.

Linguistics
The study of the structure, use, and development of language.

Linguistic indexing
The classification of a set of words into grammatical classes, such as nouns or verbs.

Machine learning
The use of algorithms that can 'learn' – in search, this can be used to automatically classify or re-rank data, or to extract patterns from data.

Meta tag
An HTML command located within the header of a website that displays additional or referential data not present on the page itself.

Metadata
Data that provides information about other data (i.e., is data about data).

Morphologic analysis
The analysis of the structure of language.

Natural language processing (NLP)
A process that identifies content by attempting to adhere to the rules of a given language.

Natural language query
A search input entered using conventional language (e.g., a sentence).

Parametric search
A search that adheres to predefined attributes present within a given data source.

Parsing
The process of analysing text to determine its semantic structure.

Pattern matching
A type of matching that recognises naturally occurring patterns (word usage, frequency of use, etc.) within a document.

Phrase extraction
The procurement of linguistic concepts, generally phrases, from a given document.

Precision
The quantification of the number of relevant documents returned in a given search.

Proximity searching
A search whose results are returned based on the proximity of given words (e.g., 'pressure' within four words of 'testing').

Query by example
A search in which a previously returned result is used to obtain similar results.

Query transformation
The process of analysing the semantic structure of a query prior to processing in order to improve search performance.

Ranking
A value assigned to a specific result returned for a query—the first item listed has a ranking of 1, the second has a ranking of 2, and so on.

Recall
A percentage representing the relationship between correct results generated by a query and the total number of correct results within an index.

Relevance
The value that a user places on a specific document or item of information. Both precision and recall are defined in terms of relevance.

Relevance engineering
The practice of improving the relevance of search results using a combination of technical knowledge of the search engine's capabilities and the overall business requirements.

Search results
The documents or data that are returned from a search.

Search terms
The terms used within a search field.

Semantic analysis
An analysis based upon grammatical or syntactical constraints that attempts to decipher information contained in a document.

Search orchestration
Used by some vendors as an alternate to Federated Search.

Sentiment analysis
The use of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in documents.

Soundex search
A search in which users receive results that are phonetically similar to their query.

Spider
An automated process that provides documents to a data extraction or parsing engine. (See also Crawler)

Stemming
A process based on a set of heuristic rules that identifies the root form of words contained within a given document (e.g., run from running). (See also Lemmatisation)

Stop words
Words that are deemed to have no value in an index. (See also Word exclusion)

Structured data
Data that can be represented according to specific descriptive parameters—for example, rows and columns in a relational database, or hierarchical nodes in an XML document or fragment.

Summarisation
An automated process for producing a short summary of a document and presenting it in the list of results.

Synonym expansion
Automatically expanding a search by adding synonyms of the query terms derived from a thesaurus.

Syntactic analysis
An analysis capable of associating a word with its respective part of speech by determining its context in a given statement.

Taxonomy
In respect to search, the broad categorisation of objects (typically a tree structure of classifications for a given set of objects) in order to make them easier to retrieve and possibly sort.

Term frequency
A quantity representing how often a term appears in a document.

TF.IDF
The term frequency.inverse document frequency formulation gives a score that is proportional to the number of times a word appears in the document offset by the frequency of the word in the collection of documents. (See also BM25)

Thesaurus
A collection of words in a cross-reference system that refers to multiple taxonomies and provides a kind of meta-classification, thereby facilitating document retrieval.

Tokenising
The process of identifying the elements of a sentence, such as phrases, words, abbreviations, and symbols, prior to the creation of an index.

Truncation
Removal of a prefix or suffix.

Unstructured information
Information that is without document or data structure (i.e., cannot be effectively decomposed into constituent elements or chunks for atomic storage and management).

Vector space
A model that enables documents to be ranked for relevance against a query by comparing an algebraic expression of a set of documents with that of the query.

Weight
A value applied to a given area of a search system (e.g., term weighting, which represents its importance with respect to other factors).

Wildcard
A notation, generally an asterisk or question mark, that when used in a query, represents all possible characters (e.g., a search for boo* would return book, boom, boot, etc.).

Word exclusion
A list containing words that will not be indexed. This is usually comprised of words that are excessively common (e.g., a, an, the, etc.).