pureinsights™ bulletin

Artificial Intelligence (AI) and the future of Search in the Information Industry

By Graham Charlesworth, VP of International Operations

Over the past 25 years or so search engines have played a vital role in the information industry. Whether used via a Google-like search box at the front end or behind the scenes to help understand and enrich content, search has become a standard component of virtually every information-based product. In fact, it's so important that most large publishing houses invest heavily in building expertise in search technologies, and some have even gone as far as to build their own search engine. Back in the 1990s, the American Technology-based publisher CNET built a product called Solr. Solr was then open-sourced and went on to become one of the most widely used search engines, together with another open-source based technology which followed in its path called Elasticsearch. Publishing and search engines are synonymous.

Why is search important?

When coupled with your content and ideally integrated into your client's workflow, search can help your clients:

- Make informed and better decisions
- Save time and be more productive
- Get better results and achieve successful outcomes

From an information provider's perspective search can:

- Create new revenue generating opportunities
- Drive increased habitual usage of a service
- Provide a valuable insight into a client's informational needs and work behaviour (through studying search logs).

To quote Jason Thomson, CIO of the Market Intelligence Agency, Mintel:

'The key thing about search for us is that if it works well, it's the quickest way to get an answer to your question or find the information that you are looking for. You don't need to learn or understand the structure or nature of the data set you are querying. More than that, it's typically the first thing that anyone does when arriving at a site and then if it works well, they use it every time they come back. A great search is the lowest friction way of the navigation of a site to get straight to value'

Today search is even more important now that the majority of publishing has moved online, and products have become more sophisticated - often incorporating both structured and unstructured data with integrated analytics. In fact, many publishers have morphed into Data, Analytics and Technology companies and for good reason as their products are no longer just about the content.

Getting the basics right

Before we get into AI and the future, you need to get the foundations right and in doing so make sure that your existing search engine is performing the basic functions. This should be your 'starter for 10'. Ask your IT department or whoever is responsible for Search to take a look at the following diagram and make sure that all of the boxes in the lower left quadrant are ticked off.

Search Feature Maturity

This is basic functionality and easy to implement and is what people expect as a bare minimum. If they are not checked off and search is important to you then get this done as a priority.

The more advanced functionality in the other quadrants is something that you should probably have on your roadmap and again depending upon the relative importance of search to your business you should prioritise accordingly.

Don't be daunted though, the important thing to note here is that all of this is largely search engine independent. If you are using Solr or Elasticsearch or any of the common proprietary search engines you will be able to do most of these things, it's just a case of configuring the engine correctly.

AI in Action – 'Just make it work like Google'

People's expectations for how search should work are now higher than ever largely due to Google who have set the benchmark. People expect a search engine to not only understand exactly what they are looking for but where possible provide a direct answer to a question, rather than having to read a document and extract the information themselves.

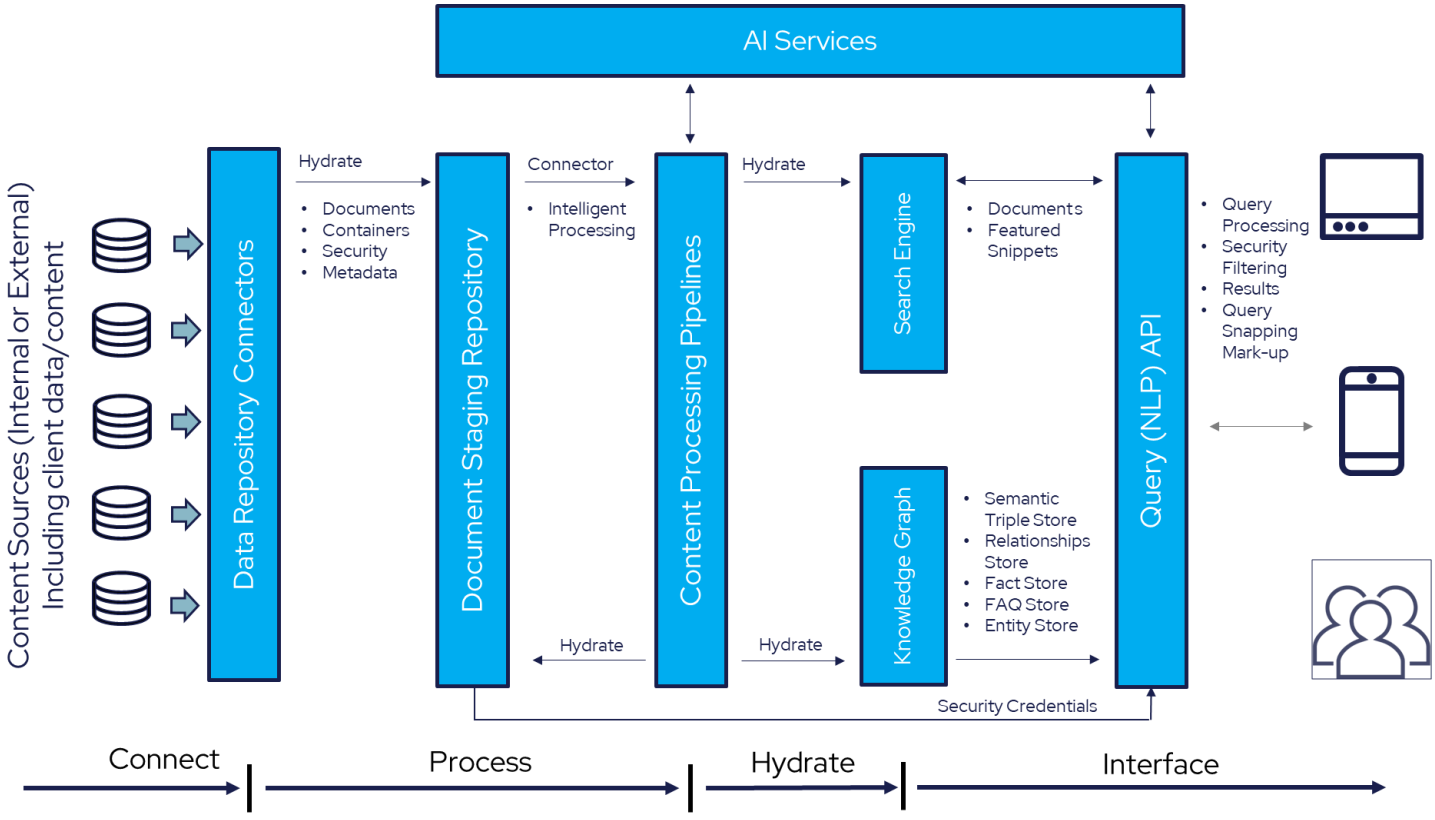
This really is the Holy Grail of search and Google have nailed it. Apparently about 15% of the searches entered into Google are in the form of a question and this trend is growing exponentially as user behaviour changes. (e.g. Question: 'How old is the Moon?'. Answer: '4.53 billion years'). It really is one of the best examples of AI in action today and the exciting thing is that the technology behind this has been open-sourced and therefore available to all.

> *"We see this as a way in which we can improve the access to EU Law and EU Publications and as such we are currently exploring ways in which we can use this technology to enhance the digital experience and provide contextualized answers to our users."*
>
> Razvan-Petru Radu, Deputy Head of Unit OP Portal, Publications Office of the European Union

Using AI to gain a competitive advantage

Now for the really exciting stuff and 'Making it Work like Google'. From here, what you should be looking to do is really raise the bar by bringing AI technology into play. This technology will complement your existing search engine and provide the Google-like functionality that was covered earlier. There is no need to replace your engine. For those who are technically minded take a look at the diagram below, for those who are not, make sure that your technical team see this. The key components and their functions are as follows:

- Search engine – for indexing and searching.
- Content Processing sub-system - for content normalisation, cleansing and enrichment (see section below for more detail)
- Natural Language Processing subsystem - for understanding the content and what the user is looking for so that the two can be matched.
- Knowledge Graph – a database for storing factual based information and the relationships between the facts.
- AI Services – external cloud-based services provided by the likes of Microsoft, Google and Amazon for performing functions such as advanced OCR, Sentiment Analysis, Language Processing, Image Recognition and Text Summarisation
- Query API – understands the question being asked and directs the question to either the search engine or the Knowledge Graph.



How does AI enhance search?

The technology that allows computers to understand humans is called Natural Language Processing (NLP). This is a branch of AI and enables machines to understand text and spoken words in much the same way as humans can. Its roots go back to the 1970s, but it has become mainstream in recent years due to the popularity of applications such as virtual assistants, chatbots and search engines and technology like Google BERT which have taken things to a whole new level. Internet Search engines use NLP algorithms to help them understand the meaning of ambiguous language by taking the context of surrounding words to establish better understanding.

Ask Google 'What is the Moon made of?' and at the top of your results page you will see something called a Featured Snippet. This is a brief excerpt from a webpage which contains the answer – '43% Oxygen, 20% Silicon, 19% Magnesium, 10% Iron, 3% Calcium etc'. Featured Snippets are an excellent example of how NLP and Machine Learning (another branch of AI) technologies can be used to extract a specific piece of text from a document that best answers a user search request.

What role do Knowledge Graphs play?

In 2012 Google added a Knowledge Graph to its search engine. This was basically a database comprising billions of facts about people, places and things and how they are interconnected and when combined with their search engine provided the question–answering capability that we experience every day. Using Machine Learning, Knowledge Graphs can be created automatically from document sets and so this process does not need to be done manually.

An example of a Knowledge Graph derived answer and how it is constructed is shown in the following diagram. It shows someone asking about the topics in a specific piece of legislation: the Knowledge Graph identifies the document and then finds the topics using the 'Topic' relationship. It would be possible to do another hop within the Knowledge Graph and from there find other pieces of legislation, or people related to the same topics of interest.

Knowledge Graph Answer

Question

"What are the topics of Regulation 2003/40/UK?"

Answer

2003/40/UK / TOPICS

- Fight Against Crime
- Legislature
- Policing
- Criminal Law
- Human Rights

Feedback

Content Processing

Poor data = poor search, it really is as simple as that! In order to combat this, Content Processing is undertaken to ensure that the data effectively 'plays nicely' with the search engine. Content processing pipelines will be set up to perform specific functions such as cleansing, normalising and most importantly enrichment of the content through the additional of metadata. So, in effect we are adding structure to unstructured data which is then used to classify, categorise and search and navigate by. Traditionally Taxonomies and Ontologies will be employed at this stage, but AI technology (NLP and ML in particular) is starting to play a role here too in terms of helping to disambiguate and as a way of automatically maintaining these resources. This can be done through external AI based Services provided by the likes of Amazon, Google and Microsoft.

Don't get left behind

Taking search seriously is really a 'must do' in most areas of publishing today and done well will help differentiate you from your competitors. People are used to Google and whether we like it or not this is the search experience that they now expect. If you don't have the people to do this, you should look to either hire in some help or even consider out-sourcing the whole management and on-going improvement of the search sub-system to a specialist company on a managed service basis. This is becoming more and more popular especially given the challenge of finding and retaining people with the necessary skills and experience, plus it can be more cost-effective. Either way, search is a living, breathing thing and as such is never finished. If it's important to your business don't let it stagnate.

And finally, and contrary to popular belief, it is no longer necessary to spend money on proprietary and often expensive search engines. Today virtually everyone in the publishing industry is using Solr or Elasticsearch, which are perfectly suited to this application and coupled with these AI-based technologies, also available as open-source, will give you an excellent foundation. Furthermore, the pace of development of these products coupled with their openness from an architectural perspective make them even more appealing. This really is an exciting time to be working with search technology and I would urge you to take full advantage.

Graham Charlesworth
Graham is a co-founder of Pureinsights and a search industry veteran with over 30 years of experience in the industry.

About Pureinsights

Pureinsights has deep expertise building search applications with conventional search engines. The company helps customers go "Beyond Search", using Knowledge Graphs, Machine Learning, and Natural Language Processing to build enterprise search applications that better understand user intent and deliver answers users want. "Just make it work like Google."

For more information visit us at www.pureinsights.com or email info@pureinsights.com