Searching for Gold

Harnessing the Power of Taxonomy and Metadata to Improve Search

EARLEY
INFORMATION SCIENCE

The Role of Taxonomy & Metadata for Search

Taxonomy and metadata are key components of search. Even when we don't have metadata, a search system actually derives metadata about content and creates a search index where that metadata is stored. Knowing that a word or phrase appears in a collection of documents is actually information about those documents.

If a search engine does that, why add more metadata?

The reason for this is some content is not as easy to index (what the search engine does) in a way that is meaningful to people. Search engines look for term occurrences, they don't tell you what a document is really about or what the value of that document is. We need humans to index the document using their judgment and experience.

Search engines are dumb. They can't infer intent. They don't know anything about you or what you are trying to accomplish. They also look for exact matches. If you call a document "brochure" and I use the term "collateral" to describe the same thing, you will not find it if I don't use your terminology. So we need to get an agreed upon set of terms to apply to documents so that everyone finds them. There are ways that a taxonomy and thesaurus can help here through mapping of synonyms. This becomes part of the approach when creating and applying the taxonomy.

Search engines are dumb. They can't infer intent. They don't know anything about you are what you are trying to accomplish.

Ambiguity and Search Relevance

There are also issues of ambiguity of terminology. For a financial services firm that specializes in investments for retirement, if a user entered "retirement" into the search box, the system would likely return a large number of terms, but they would not necessarily be about what the user was looking for. The reason is there are too many different classes of things that relate to retirement and typing in that one term does not provide any context for the query. Humans disambiguate all of the time. Any time someone asks you a question and you ask for clarification, you are disambiguating the query. Search is a conversation. You enter a search term into a search box, the system returns results, and you browse those results and search again or retrieve your desired result. The ability to refine a search using facets (like color, size, brand, price, etc. on an ecommerce site) is a process of disambiguation.

Even with machaine assists, there is no getting around the need for human intervention to ensure a quality search experience.

Taxonomy (and thesaurus which is a taxonomy on steroids) provide the ability to precisely deliver results in a variety of ways.

Can Tools Build a Taxonomy Automatically?

There are ways that tools can "extract" facets from content. This is called "entity extraction". An entity can be an address, a name of a person, product or company, or it can come from a list of terms that are more specialized. There are a couple of ways to go about extracting entities but this process does require a reference list. This is a list of names (sometimes called an authority file) or other terms (from one flavor of "flat" taxonomy or a "controlled vocabulary") are used as a reference point. In other words, you need a list of terms to compare your content with in order to have the software tell you whether they are in there or not. This means that they cannot build the taxonomy for you.

Some tool vendors claim that their software can do this, but this happens under very specialized circumstances and usually very poorly. So called "machine generated taxonomies" come up with term occurrences and are named based on frequency of occurrence and some rules within the software. In this situation, the tools are really just clustering like content and a human needs to name the cluster something more meaningful.

There are other things that can be done automatically – such as auto-categorization or machine assisted indexing. In the first case, the system places the content in a bucket using one of two approaches (or sometimes a hybrid of them) – rules based and statistical. No need to get into the technical details here. In the second case, the machine "assists" a human by making suggestions about what the content is about and what terminology should be applied to the content. The human is free to select another term or accept the selection. This can help improve the speed of classification. In any case, the system needs to have either representative content against which to compare the content to be indexed or there needs to be a "rule" to tell the system when to apply a tag (the metadata) which is actually based on the taxonomy. In fact some ontology management tools (ontology is a collection of taxonomies and thesaurus structures and all the relationships between them) actually generate a rules base when you load up the taxonomies.

A Practical Vision for Search in the Enterprise

Today's enterprise search tools can improve information access and findability out of the box better than systems costing many millions of dollars a few years ago. The algorithms are getting that much better. The problem is, information is growing that much faster so it is difficult to see the impact. We're getting lost in a sea of content and documents. Organizations are also losing control of basic information management curation processes when it comes to unstructured content. People create content willy-nilly and put that content in places where there is little control over versions. This pollutes good information with junk – by not managing content processes, information is getting further out of control. A practical vision of search includes basic content hygiene. That said, well designed search can overcome a lot of sins of poor content processes. There are many ways to leverage taxonomy and metadata - through search facets, related content, related search results, query guidance/query disambiguation, related queries, expertise location, content processing, federated queries, unified information access (to integrate structured and unstructured data), metadata triggers and other mechanisms. (Many of these approaches sound the same but differ in the nuances of their execution and when terms and relationships are presented to the user.)

The bottom line is that search can be designed as an application; it is much more than a white box.

Search is the mechanism to allow content to be integrated on the fly by a person looking for specific answers. It provides content in the context of work tasks. This can add up to exceptional ROI as in the example case study that follows.

Search can be designed as an application; it is much more than a white box.

Case Study: Search Enabled Field Service

In 2015 Applied Materials (AMAT) employed 3,000 field service technicians who serviced highly sensitive, custom equipment at customer plant locations around the world.

Applied Materials is a is an American corporation that supplies equipment, services and software to enable the manufacture of semiconductor chips for electronics, flat panel displays for computers, smartphones and televisions, and solar products. With world-wide revenue in excess of $17B US the company employs 21,000 people.

These technicians were faced with a number of job challenges that impacted their information access needs:

- Due to strict clean room constraints all information access attempts needed to be completed outside of the fabrication areas.
- Plants have equipment with a very long lifetime. Some equipment is in the field for 15 years or more
- Equipment in plants is custom – there are few standard installations
- Content is sensitive IP – both from the point of view of the manufacturer (who does not want repair information to get out to third party maintenance providers) and from that of the customer (equipment and processes are trade secrets and the company supplies companies who are competitors)

Information was stored in multiple systems and processes – these included:

- a Digital Asset Management (DAM) system that contained images, diagrams and photos of parts and equipment,
- an Enterprise Resource Planning system (ERP) that allowed for queries about the "as built" configuration at a customer site as well as the availability or backlog of parts,
- a formal Knowledge Base where answers to field problems were stored,
- a documentation database where the "official" manuals and documentation were housed,
- discussion forums, where field service reps could make comments about approaches, problems, customers, equipment, solutions (this allowed for greater context sensitive collaboration)
- a formal documentation request process
- a variety of trouble ticket and incident tracking systems

Due to the constraints of the environment and the many disparate systems that needed to be combed in order to get questions answered the service technicians were found to be spending an average of 30% of their time searching for answers and part related service information. This was verified by search log data, system logs and time sheets. Multiplied by 3,000 technicians at $100/hour the cost for just searching for information was enormous. Any kind of time savings would result in a significant economic benefit.

OPERATIONAL BENEFITS

- Integrated, intuitive, easy-to-use search application
- Intelligent routing of content ensured the right content was seen by the right users
- Cut time to get answers in half

FINANCIAL BENEFITS

- Time savings valued at $900,000 per week
- Savings allowed for additional growth as well as cost avoidance

The Solution

EIS developed a search-based knowledge portal that provided the service technician with a 'one stop' approach to finding the information needed to resolve the customer's problem.

An integrated search application tied all of source information systems together. Field service technicians entered a variety of search queries – part numbers, problems, customers, classes of equipment, types of solutions sought, field notices, etc. – and were presented with a user friendly, intuitive interface that allowed them to perform a variety of tasks including previewing of large documents while hovering over search result links (rather than clicking, waiting to download and then finding that it was not the correct document), being able to initiate requests for new documentation and submit support information at the same time, being able to enter into discussions with colleagues, being able to create their own custom notebooks from the search results, on the fly and save or print these out, and numerous other functions and capabilities. The search "technology stack" also auto-categorized content to minimize the burden on users, automatically identified and flagged sensitive and customer IP based on term occurrences, allowed for intelligent routing of content through the system so that it went to the correct person at the correct time.

The Result

After the system was deployed, the time spent on searching for information was cut in half - from 12 hours per week per technician to 6 hours per week.

Calculated conservatively, assuming three hours per week per technician saved the result was still signfiicant:

3,000 field service techs

x 3 hrs. / week

x $100/hr. loaded cost

= $900,000 per week in time savings.

Although not an immediate hard dollar savings these efficiencies allowed for additional growth and cost avoidance and allowed the same number of reps to handle higher workloads.

Further savings were anticipated as new hires are not made over the course of normal attrition. This got the attention of the CIO, the CEO, and the CFO and the news had an impact on the stock price of the company. This was an enormous win for the organization. The project had been attempted three times before over a 5 year period but the tools, methodologies and organization were finally ready to make it all happen.

The project had been attempted three times before over a 5 year period but the tools, methodologies and organization were finally ready to make it all happen.

Conclusion

The same savings are typically seen in knowledge intensive environments like support centers, transfer agencies, marketing organizations, sales processes, engineering organizations, etc. Search can no longer be considered the white box and taxonomy and metadata work together to allow search integration frameworks to function and provide extremely high ROI and value to knowledge intensive processes.

About Earley Information Science

Earley Information Science is a professional services firm dedicated to helping organizations just like yours become an AI-powered, customer-driven enterprise. We have the tools, team, and processes to design and execute a scalable, governance-driven digital roadmap, led by your customer's immediate and long-term needs. Together, we can implement a digital transformation that provides a personalized, accurate, and fulfilling customer journey, driving measurable ROI to your bottom line.

PO Box 292, Carlisle, MA 01741
P: 781-812-5551
www.earley.com