Finding the Value in Text Analytics

By Ahren Lehnert,
FMC Technologies

The phrase "text analytics" might conjure up images of a group of computational linguists poring over complicated algorithms to reveal truth in language through some mixture of language study, advanced computer science, and alchemy. There is some accuracy to this. Because of this perceived complexity, non-academics may shy away from investigating the use and practical application of text analytics unless they are in very technical areas, e.g., business intelligence.

However, while text analytics can be complicated, the value of learning the tools and techniques involved in analyzing unstructured text can be great: in revealing the unknown, saving the organization time, and avoiding duplicated efforts.

Where to begin? As with all activities undertaken in the business, there needs to be a clearly defined problem statement with a vision of where you would like the organization to be at the end of addressing the problem. For instance, surveys are a very common instrument for getting feedback from employees or customers. Often these are easily calculated choice values which can be tallied and charted to show where the areas of interest are.

To be effective, surveys must accurately predict what matters to the target audience and frame the questions accordingly. Frequently, however, the true value in surveys comes from discovering things you could not have predicted, through unstructured, free text responses. Reading a few dozen surveys and extracting concepts manually might be worthwhile for the value which is revealed. What about a few hundred or a few thousand surveys? Are we then forced to avoid soliciting freely written responses due to the amount of painful and time-consuming analysis required? A common scenario like this may prove the value in performing text analytics.

Text analytics is particularly good at revealing the unknown and clustering together similar ideas. For instance, in my own organization, a poorly deployed search solution languished for several years as there was very little idea of what users wanted to find in search and how that content could be more effectively retrieved. Many vain attempts at assuming what users wanted to find and promoting these in the search results seemed fruitless and there was no measure of whether these efforts were successful.

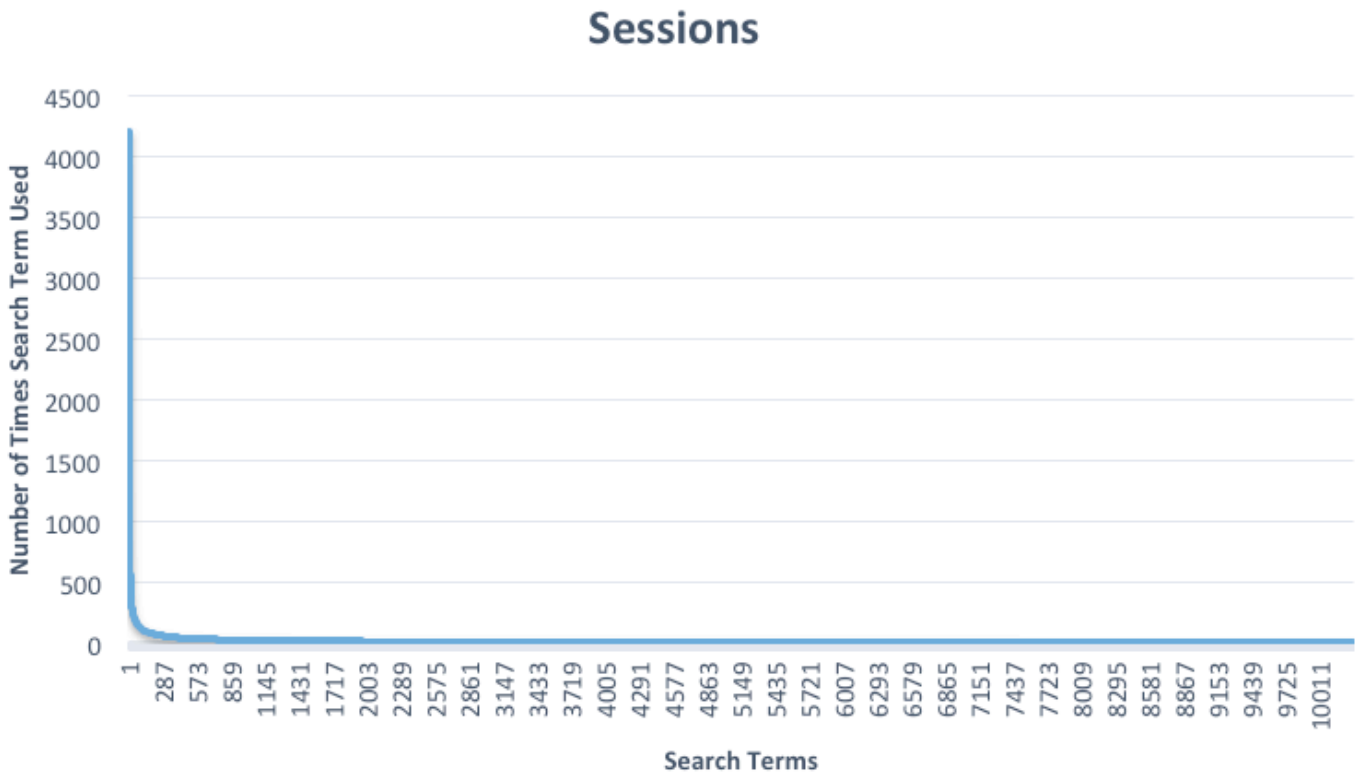
Although we had years' worth of search logs, it was pretty clear from a first-round manual analysis that the top searches weren't necessarily what users were seeking. For example, one of the top search queries was

performed by one user, revealing that his use of search was in place of bookmarking a page. While this is important, it doesn't necessarily reflect what users across the company are using search to do.

Complicating matters, we had several different versions of search running, each logging search queries separately. These logs contained character errors, multiple languages, and test queries performed in the production environment. Looking at a year's worth of logs quickly revealed that a manual cleanup and analysis could take months.

It was clear from charting the raw data that there was a very long tail of search queries which occurred fewer times than the more frequently searched concepts. While it is tempting to simply drop these from consideration, this long tail is in fact where much of the value lies. Many variants of more commonly searched phrases exist in this long tail, which, when clustered as a single group, adds up to a significant number of searches. Once clustered by similarity, the long tail of search queries paints a very different picture of users' search habits and organizational knowledge compared to simply looking at a raw count of top search terms. This was where text analytics came in.

The Long Tail of Search Terms



What were the benefits realized from performing text analytics on our search queries? One of the most common measures of search success is time saved looking for information. For our new search implementation, we were able to consolidate search terms and phrases from across the long tail to common expressions and then seed the new search feature of "type-ahead" with this list. In type-ahead, the search engine suggests search query terms based on what the user has started typing.

User history is being used to drive people to select search queries which were already used enough times to show a successful set of search results, reducing time spent in trying different search queries and getting users to useful results quickly.

We will also save time and avoid duplicated efforts with our future projects in planning and requirements gathering. The concepts we extracted and clustered showed us what content people expected to find in search. This will help us to understand what content gets indexed by our search engine and what content we

may want to include or exclude in the future. Given the potential time and expense involved in conducting user interviews, requirements gathering sessions, and scoping large search projects simply for the sake of searching everything, we can build effective search based on evidence rather than opinion or gut feel.

Clustered concepts will also be used to inform the information architecture of our new Intranet. Determining whether to use an organizational or functional navigation to allow users to find content can be informed by the topics we know are of importance to them based on the highest occurrence of thematically similar concepts in search. Search logs are essentially a raw list of things which are important to users: using text analytics on this list to group content for more effective navigation is also a time-saver in information architecture design and end user time spent looking for information.

Was the value of understanding what users were searching for worth the time it took to find and learn a tool used for text mining and analytics? We were fortunate enough to have a tool already in the organization being used by the Business Intelligence group and it only took one day of training to learn the tool well enough to begin a preliminary analysis. The time invested to get started was minimal.

While the time and cost savings is as yet to be finally determined in our own organization, the amount of time saved performing analysis in order to release new search functionality in just a few months rather than a year is both time and cost saved in resource hours. More significantly, the ability to provide analysis for the direction of the new search implementation also gained stakeholder and end user trust, which is a very hard won and valuable commodity in the enterprise.

© 2015 Ahren Lehnert. All rights reserved.