Finding Meaning in Healthcare Data

The ability of the healthcare system to fully leverage the plethora of unstructured data it produces and consumes every day is very limited. Critical information is hidden within physician's notes, published research, longitudinal records and other reporting. Even though this information is accessible electronically, it is difficult if not impossible to analyze and categorize it in a systematic and scalable manner. In many cases, this information may hold answers to the industry's biggest challenges, such as designing more effective treatments, accelerating critical path, or safeguarding public health. By using new discovery technologies to query, categorize and analyze unstructured data, we can help users find the undiscovered value within this rich collection of knowledge.

Search technologies are the traditional approach to uncovering information buried within volumes of unstructured data. The challenge is that searches are typically restricted to an exact, verbatim description of the topic. This may yield some idea of what you are looking for, but the results are constrained by the search term's rigid definition (denotation) instead of a broader, more nuanced understanding of the concept (connotation). Additional complications arise from the fact that these keywords are subject to the vagrancies of synonymy (multiple words having similar meaning) and polysemy (a word having multiple meanings). While search technologies can help find specific documents, they are poor solutions for exploring and analyzing broader concepts.

Another significant limitation is that search technologies rarely look at the aggregate results to identify key themes and emerging trends. Instead, they provide a list of disparate resources with no understanding for how the information relates – are these documents contradictory? Repetitive? Is there an overriding trend? The only way to find out is by reading all of the material.

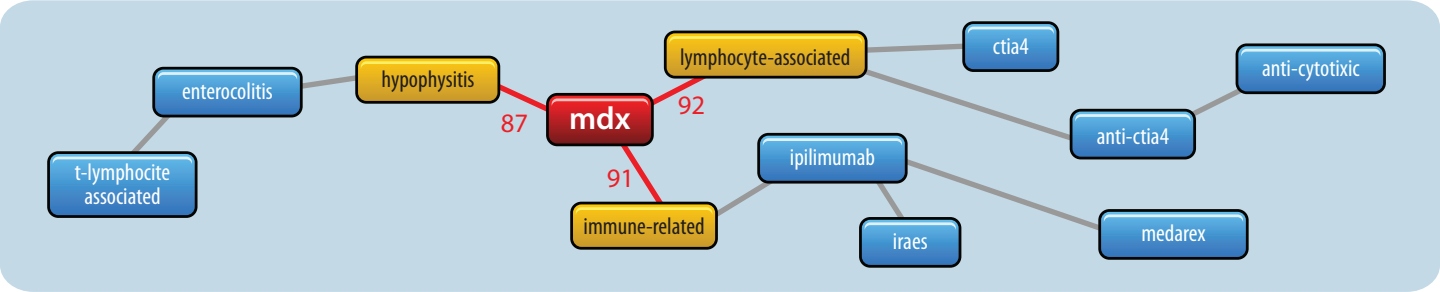
Even specialized ontologies, taxonomies and controlled vocabularies often fall short. The control and structure that they provide, which is their core strength, is also a drawback as it prevents users from discovering new insight as information is maintained within predefined silos. For example, the results of a clinical trial may not be readily available outside of its intended focus area even if these findings are relevant to other fields. Other challenges include the high cost and significant effort required to maintain these structures as well as the bias inherent in defining a standard framework.

GOOGLE IT? DISCOVER IT!

Within a number of empirical fields, including national security, legal investigations, customer service, and now healthcare, users are adopting new approaches to overcome these constraints. Exploratory in nature, discovery technologies and text analytics distinguish themselves from search through their added focus on categorizing the data and identifying relationships with similar content. The technology is often self-learning, meaning that it analyzes the content to identify specific patterns, including many that wouldn't typically be detected.

Discover technologies deliver a number of unique features, including:

- **Conceptual Search** – Instead of just searching for keywords, users can explore broader topics, even using a full document as the basis of a search. Furthermore, conceptual searches eliminate the need for word-to-word matches as they accommodate multiple synonyms, enabling all conceptually relevant material to be located. It also analyzes a document's overall focus to better assess its relevance to the original query. This avoids the constraints of Boolean search constructs.
- **Relationship Analysis and Mapping** – By understanding the distinct connections between various entities (terms) and concepts across documents, users can tap into the collective knowledge of the entire repository. For example, through examining the clinical trial results of multiple drugs new correlations between diseases and symptoms may emerge. These dynamically created maps use hyperlinks to trace, explore and recast conceptual relationships.
- **Automated Discovery with Real-time Alerting** – Once a topic is defined, the system can proactively identify new materials fulfilling similar criteria. This is particularly important in keeping abreast with new developments in areas like public health monitoring and clinical research.



Relationship maps linking specific entities and terms can be created and explored on-the-fly

A specific discovery technique is Latent Semantic Indexing (LSI). By expressing mathematically the frequency, correlation and semantic ("contextual meaning") definition of terms within documents, LSI identifies latent ("not obvious") relationships between terms, documents and the search concept. The benefit is increased flexibility as the query is not dependent on a single term and it is also more precise as it considers the contextual meaning of both the term and document.

Healthcare is global. This means that one of the key advantages of LSI is the capacity to investigate content in multiple languages, without translation tools. It can also integrate biometrics, images, spreadsheets and other forms of structured, semi-structured and unstructured data. Furthermore, LSI doesn't need a taxonomy, ontology or other mapping to operate, but rather, can create one dynamically based on the actual content.

APPLYING DISCOVERY TO HEALTHCARE

Recognizing these unique strengths, researchers have begun to apply LSI-based discovery technologies to a number of critical healthcare challenges, including:

- **Epidemiology** – An effective tool for public health monitoring, including pharmacovigilence, adverse event monitoring and biosurveillience. It can be used to integrate, access and analyze a range of traditional medical sources, news reports and even topics trending within social media to detect and pinpoint new and emerging threats.
- **Clinical Research** – To examine existing research for latent results, LSI can be applied to two discovery processes: uncovering "nearby" relationships that are necessary to initiate the literature-based discovery process; and discovering more distant relationships that may genuinely generate new discovery hypotheses.
- **Medical Ontologies** – Used to create automated ontologies for specific subjects, such as genes, and to map defined taxonomies to one another. LSI is particularly valuable due to its ability to identify relevant associations that fall outside the structured nature of these tools.
- **Grants Management** – Grant applications can be evaluated in terms of past and current literature, creating a thread of knowledge generation. Likewise, how the research is subsequently utilized can be tracked as well to determine return-on-investment (ROI) and other benefits.
- **Patient Safety** – Physician notes and other elements of electronic medical records (EMR) can be reviewed for consistency and compliance with defined standards and practices.
- **Fraud, Waste and Abuse** – By identifying textual inconsistencies or applying known FWA concepts or patterns as search criteria, LSI becomes an important tool for combating fraudulent and erroneous claims. It can also be used to identify and track ambiguous entities of interest.

What these scenarios share is an ability to utilize unstructured data proactively to advance the core mission of the organization. In terms of more immediate benefits, LSI-based discovery technologies boost productivity and confidence for research and program management professionals as it provides more conceptually relevant data to support their findings and business goals.