

Merging Structured and Unstructured Data

Overview

Simple, intuitive access to information facilitates effective decision making and insightful thinking. Disparate data types have traditionally made integrated information access difficult if not impossible. Adding to the difficulties are well-known problems associated with accessing information spread across diverse, and sometimes proprietary data sources in different systems and applications. Acquisitions, mergers, and joint ventures compound these difficulties by introducing greater application and data diversity.

One of the biggest obstacles to integrated information access has been an inability to effectively merge structured information with unstructured information. While no single solution can solve every problem, there are ways to provide a coherent, integrated presentation of information contained in both databases and text-based content repositories.

Text analysis tools and techniques can be used to organize textual content and identify specific items contained in the text. An unstructured information organization technique using what is referred to as a taxonomy can be implemented in a way that “maps” unstructured information to a database schema. This mapping is a primary enabler for presenting information contained in both source types in a manner that is seamless to the user. The mapping can be done with very high precision. Taxonomy, as it relates to this application, can be defined as a categorization of objects and the relationships among them. Taxonomies are usually hierarchical for this kind of application, typically going from general topics at high levels to more specific further down the hierarchy.

Another text analysis technique called entity extraction can identify key items like companies, people, products, locations, and dates contained in textual content. This is very useful in identifying information in text that can be used to filter and refine the mapping to information in a database.

These techniques are further discussed below.

Text Analysis

Taxonomies

There are number of different algorithms that can be used to categorize unstructured data into a taxonomic structure. Each algorithmic approach has strengths and weaknesses, of course. Commercial software vendors have used these algorithms to create products that associate textual content with appropriate categories in taxonomic structures. This is often accomplished using tools that allow the creation of “rules” that automate the process of associating text with the appropriate categories, called “nodes,” in taxonomic structures. This process is referred to as “Publishing” content to the taxonomy.

When textual content is published to a taxonomy that has been designed to mimic the schema of a database, the content published to specific categories in the taxonomy can be mapped to areas and information in the database. As a result, information from both the structured and unstructured sources can be presented together.

Entity Extraction

When trying to gain new insight from a sea of information it is essential to identify and isolate instances of things that are known to be of interest. Things like company names, people, places, products, job titles, dates, locations, and many others can be identified. In text analysis circles, this process is called “entity extraction.” Identified items, referred to as “entities”, can also be used to refine the match between specific items in textual content and items in a database. For example, let’s say that a database query returns a record containing the name of a company, the name of a person in that company, that person’s role, the date that the person was appointed to that role, and their previous position. Entity extraction can identify items in textual content that are related to that person, role, and company from sources such as news feeds, on-line journals, articles, Web sites, Web logs, email, and other sources. Pertinent information can be pinpointed, highlighted, and presented along with related information from structured sources.

Combining entity extraction with taxonomy can provide a high degree of specificity in identifying important information contained in large bodies of textual content. The taxonomy organizes the content into categories that map to the database, and the entities identified by the extraction process provide enhanced specificity.

For example, let’s say that we have a taxonomy that uses market segments as major categories, as represented in Illustration 1, below. We’ll use the automotive segment and select a specific category further down that branch of the taxonomy. This path in the *Automotive* segment drills down through *Chassis* into *Wheels and Tires*, and on to a more fine-grained *Suppliers* node deeper in the taxonomy. This node has many documents published to it as determined by the taxonomy’s content publishing rules. These documents contain information about numerous companies that supply wheels and tires to the automotive industry. Let’s say that Goodyear tires are of particular interest. Entity extraction is used to analyze the textual content associated with the *Wheels and Tires* node. It identifies a **company** entity in three different documents named **Goodyear** who makes a tire **product** that has a specific **product name**, and a **competitor** named **Michelin**. The extracted information about Goodyear products and competitors are used to map these articles to specific structured data. By using the extracted company, product, and competitor information to identify the three documents that reference them while eliminating any other documents associated with this node of the taxonomy, overall specificity is greatly increased.

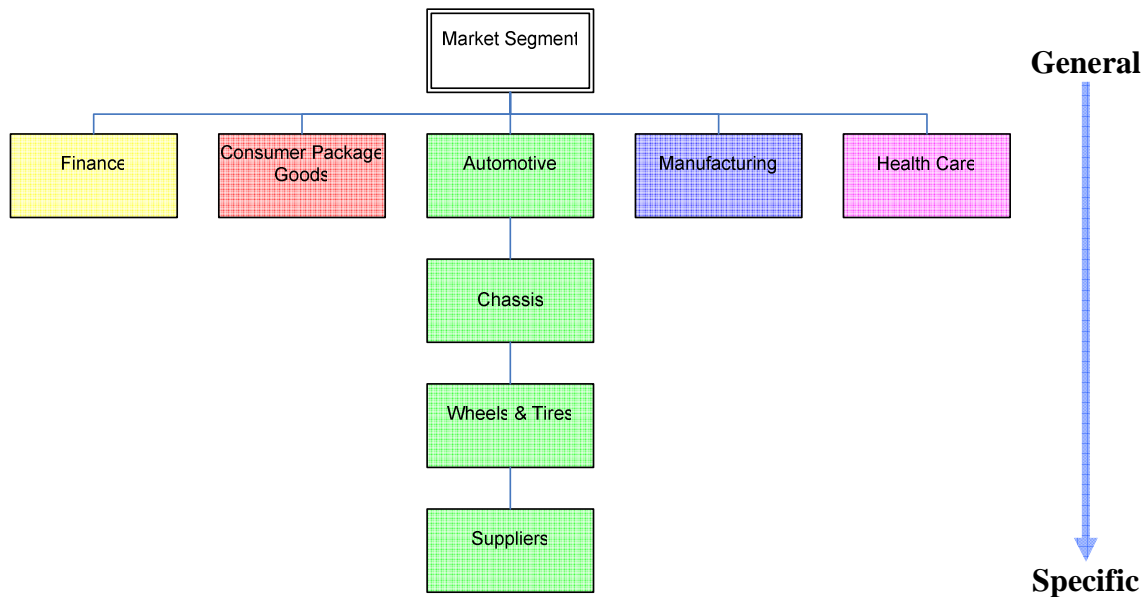


Figure 1
A path in a taxonomic structure going from general to specific topics

This illustrates the use of entity extraction to enhance specificity when mapping structured and unstructured information to each other. This example assumes that the structured data source contains entries for Goodyear as a company, a competitor field that identifies Michelin as a competitor, and a product field where product names are kept. The intersection of those three items in the structured information with the information from the taxonomy and the extracted entities creates the mapping.

Automated Summarization

A summary of textual content provides knowledge workers with a way to quickly determine if something is of interest to them. Brief summaries of textual content can be automatically generated to help identify which content might deserve further investigation. Good summarization technologies consider many different aspects of textual content. Such things as the terms in the title of a document, increased weighting of terms in the first and last paragraphs, term and phrase frequency, and other aspects are used to identify sentences that will be used in the summary. Most summarizers allow users to specify the number of sentences or some percentage of the original text to be displayed in the summary.

Conclusion

When relevant, concise information is readily available, people quickly identify and understand things that are important to them. This understanding fosters clearer thinking, better decision making. It increases the number of insightful moments that users have. Users should not have to know if the information comes from structured or unstructured sources, and they shouldn't be restricted by differences in the source of the information. Advanced information access techniques provide powerful tools for organizing and presenting high quality information.

Accessing and combining information from multiple systems and data types is a problem for every large organization. Clever application of text analysis technologies is extremely useful in unlocking the value of information across a large enterprise. Powerful text tools and technologies that enable advanced information access are available today.