Transformational Taxonomies:

How Pandora Built a Better Mousetrap

Table of Contents

White Paper

by Matt Manning
President
Information Evolution Inc.

Last Updated: March 2010

Overview

In a world where Google is king and pinpoint searching of large databases is expected and not just hoped for, it is extremely important to associate database records with the right kinds of classification taxonomies and to append relevant metadata. Companies shaping the future of the information business are approaching taxonomies and metadata in creative and effective ways and no service better exemplifies this than Pandora.

Pandora is an information service/recommendation engine that matches musicians by the "type" of their music. Traditionally this would mean you could choose between categories like blues, jazz, classical, pop, etc., and, if you were lucky, you could drill down further into subcategories like Memphis blues, Chicago blues, barrelhouse blues, etc. Pandora took a blank slate, said "what if" and threw the traditional musical taxonomy out the window. What they did next was astounding in terms of its simplicity, its effectiveness, and the fact that it relied on a major manual effort to create a new taxonomy.

The Value of a Human Filter

Like mapping a genome, Pandora's approach was for a human being to listen to a musician and categorize their works by their music's "sound," choosing from a set of standardized metadata attributes to determine that they "sound like" another musician. Is the singer's voice "gravelly" (Kris Kristofferson, Tom Waits) or "sultry" (Celia Cruz, Cesaria Evora)? Is the tempo fast, is it loud, is it shrill? By creating a new classification system based on the melody, harmony, rhythm, instrumentation, orchestration, arrangement, lyrics, and vocal styles of particular bands—rather than on long-standing record store genre groupings—they took a bold step. The sound of the music is something automation can't determine by itself, so a human team had to painstakingly map these attributes to the bands and maintain the database.

This is what makes the system work: a human filter applying a custom taxonomy.

This is what makes the system work: a human filter applying a custom taxonomy. After that, the work done in the searching process is trivial. In other words, the human classification team allows the technology to work.

Another example of these kinds of transformational, manually compiled taxonomies is Soundex, the open source tool that maps names that sound alike to each other so a name search for "Shareef" yields matches that are spelled differently (like Sharif, Sherif, Shahareef), but sound alike. The value of this mapping tool is absolutely enormous as the case of the

erstwhile Christmas Day bomber demonstrated. This example of how very similar names in different databases are not effectively mapped to possible matches highlights the dirty little secret that most database searches, even of the most critically important types of data, are often of the simplest type—exact literal matches that can be easily thrown off by slight misspellings or even blank spaces. It is the rare (and intelligent) database manager who uses Soundex or builds a manual database of alternate names that allows them to offer accurate "did you mean" suggestions to users.

Generating Metadata

Besides the manual compilation of taxonomies by humans (either through dedicated teams or open "crowd-sourced" collaborators), user-generated data can be used to help to create invaluable metadata as well—and it has the advantage of not requiring manual compilation efforts.

For instance, aggregate data on the popularity of records (the number of times a record is delivered) and data on specific user actions like search string values and the actions taken after results are delivered to a user are very useful.

Related usage pattern data—data on the searches made or links followed before and after a record is delivered—can also be used to improve the user experience of traditional online database services.

When this metadata is appended to records it can improve search results dramatically, guaranteeing fewer "dry holes," more exact searches, and more satisfied users (and, as a result, sales). Popularity data is aggregated user data on the number of record requests, the number of records/items purchased/saved/recommended. Related usage pattern data—data on the searches made or links followed before and after a record is delivered—can also be used to improve the user experience of traditional online database services.

These data can be used to correct misspelled company names, product names, personal names, and geographic values as follows:

1. Manual mapping of "no results found" search string values to the appropriate records. This guarantees future incorrect searches will yield correct results. This effort can be a small daily part of editorial processes.

2. Automated inferential mapping of "no results found" search string values to likely matching records. If, for example, 90 percent of characters in a search string match a record's value, it can be inferred that the record "might be" the closest matching record. If there are

multiple records that match by roughly the same percentage, then the more popular of the records (based on the number of times it has been delivered) is probably the record the searcher wanted.

Conclusion

All of these examples of how taxonomies and metadata can be created and exploited require a bit of work on the part of the folks who run information services, but they are vital to beating competition and delivering a great user experience.

About the Author

Matt Manning is the president of Information Evolution, Inc., a firm that designs and implements efficient research and editorial processes for content companies.

About Information Evolution, Inc.

IEI provides human resource and technology services to companies, primarily in publishing or related industries, that manage large databases in real time. For more information, call (512) 650-1111 or visit www.informationevolution.com.