



Transformational Taxonomies: How Pandora Built a Better Mousetrap

March 15, 2010

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 meta data. Companies shaping the future of the information business are approaching taxonomies and meta data 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 sub-categories 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.

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 meta data 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. 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.

Besides the manual compilation of taxonomies by humans (either through dedicated teams or open “crowd-sourced” collaborators), user-generated data can also be used to help to create invaluable meta data 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.

When this meta data is appended to records it can improve search results dramatically thus guaranteeing fewer “dry holes,” more exact searches, and more satisfied users (and 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 is data on the searches made or links followed before and after a record is delivered and this can also be used to improve the user experience of traditional online database services.

The way these data can be used to correct misspelled company names, product names, personal names, geographic values is 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 your editorial processes.
2. Automated inferential mapping of “no results found” search string values to likely matching records. If, for example, 90% 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.

In any event, all of these examples of how taxonomies and meta data can be created and exploited require a bit of work on the part of the folks that run information services, but they are vital to beating competition and delivering a great user experience.

--

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