Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice

Wendi Pohs*, Dick McCarrick*, and Michael J. Muller**
*IBM Lotus Software Group and **IBM Research
Westford, MA
wpohs@us.ibm.com

1. INTRODUCTION

This position paper presents a hybrid view of the creation and maintenance of machine-generated taxonomies. We analyze and propose challenges in areas where information systems overlap with human social systems, and we also bring two types of experience to address the theme of the workshop. We combine a relatively conventional, formal research approach based on user studies of taxonomies in organizations with our practical experience over the past 18 months. We are subject matter experts working with machine-generated taxonomies, and we consult to customers and business partners who are creating taxonomies for their own organizations (Pohs, Thiel, Earley, & White, 2001) We hope that this hybridization (Muller, 2003) of theory and practice provides insights and reveals problems that would not be apparent if we were writing from only one of our two perspectives.

Knowledge management practitioners agree on the importance of a fine-tuned, easy-to-understand taxonomy to navigate and understand a large body of content. After all, a taxonomy is a visual representation of that most elusive yet critical concept, *what the organization actually knows*. A well-constructed, meaningful taxonomy gives users a tool for quickly finding information they need. But we have found, based on our experience with implementing taxonomies in large companies, that it's usually difficult to construct a useful taxonomy in complex organizations. In many situations, taxonomies tend to be either superficial or general, or continually out of date—often both (Pohs et al., 2001).

Software products that promise automatic clustering and categorization can facilitate taxonomy creation. But no computer program, however sophisticated, can precisely predict exactly how a particular organization might want to structure its content (Muller, Carotenuto, Fontaine, Friedman, Newberg, Simpson, et al., 1999). A taxonomy that represents a large body of constantly changing information requires humans and computer software to work together. It sometimes also requires time, patience, and creative thinking, but the result – a meaningful, easy-to-use taxonomy – is worth it.

2. THE ROLE OF THE TAXONOMIST AND ORGANIZATIONAL IMPLICATIONS

One of the more distinctive characteristics of the human intellect is its ability to join ideas—some view this as the very definition of creativity. This is what a taxonomy is, a representation of ideas and how they are joined, how they relate to each other.

Even the most powerful computer software can't join ideas in the same way that humans can. It can't read a document and understand what the document is, what its author intended to convey

ISSN: 2324-9773

Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

to readers. To computer software, a document is little more than a "bag of words." Therefore, computer-generated categories might make perfect sense in a world ruled by algorithms and abstract representations. But they might not be instantly self-evident to the users who have to navigate through them to find information. Thus a human intermediary is needed, usually in the form of a *taxonomist,* a person who understands classification and categorization of information and can evaluate the taxonomy the machine generates (Ruby, 1999; Pohs et al., 2001). Organizations also need at least one person who understands the content generated by internal business processes. In our practical experience, these two roles—taxonomist and subject matter expert (SME)—working closely with each other and in conjunction with the computer, are crucial to the process of organizing and making sense of an organization's raw information.

Analyzed at the level of the large organization, however, the team of taxonomist and SME may be in a difficult social (or even political) situation (Muller, Pohs, & Friedman, 2000; Star, 1966b). In effect, they are writing a kind of "language" that the organization will use to discover, analyze, and remember its knowledge (Muller et al., 1999). Like any language, the taxonomy will favor certain concepts over others, make connections between certain ideas but not other ideas, and will promote certain concepts and relationships at the expense of others. Creating a taxonomy becomes a matter of selecting which disciplinary or departmental viewpoint(s) will prevail in the organization, with predictable positive consequences for the stakeholders whose views are represented, and predictable negative consequences for the stakeholders whose views are not included (Albrechtsen & Jacob, 1998; Olson, 1998; see also Code's concept of *rhetorical spaces* as an institutional structuring of permissible language, 1995). The next section examines some of the fragmentary solutions that have been considered for this problem.

3. MULTIPLE TAXONOMY SCHEMES

Some people assume that for any given collection of items, there is only one way to organize them—in other words, a place for everything and everything in its place (Community Intelligence Labs, 1999). But reality is usually more complicated. In the majority of cases, there's no single right way to structure knowledge (Bowker & Star, 1999). The taxonomy selected by one organization might not be best for another. Indeed, a single organization might choose multiple taxonomies for the same data, offering its users several choices for locating information (Star, 1996a). Researchers are currently analyzing ways to map one taxonomy onto another (e.g., Eneva 2002; see http://www.cs.cmu.edu/~eneva/talks/Tax.ppt for a survey of related research), but usability studies of disparate taxonomies should also be considered. Do users want customized views of large corporate taxonomies? Or do they want to use a "See also" construction to navigate between taxonomies as needed? And do users have a fixed preference, or do they want different views or even different taxonomies depending upon what task or problem they are working on?

4. MULTIPLE CATEGORIZATION SCHEMES

In practice we also have found that it is hard to define one, all-encompassing categorization scheme for a disparate corpus of data (Albrechtsen & Jacob, 1998; Olson, 1998). A taxonomist might want to group all newsletters together, for example, regardless of what any individual newsletter is about. Rules-based classifications can work well for some categories, while

ISSN: 2324-9773

Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

statistical classifiers perform well for others. Using just one automatic classification scheme imposes constraints that can require a great deal of manual intervention.

5. STARTER TAXONOMIES

For working taxonomies, one of the first questions to consider is how much content to include in the taxonomy. There are two schools of thought on this. One approach selects all the content and lets the computer have at it (Pohs et al., 2001). A second approach starts with a small subset of the content (Graef, 2001). The "everything and the kitchen sink" approach takes longer to implement, but the result is a complete taxonomy that represents the organization's knowledge. The "subset" method allows the taxonomist to finish a draft taxonomy more quickly, control the direction the taxonomy is going in, and nip any early misclassifications in the bud.

Historically, for most large organizations it's been best to start small. Taxonomists often begin with a limited collection of documents, and create the preliminary taxonomy as a test case rather than a working taxonomy. This allows them to "get their feet wet" with taxonomy editing and produce a taxonomy that supports the work of at least a subset of users. With this iterative approach it is also easier to correct the taxonomy, if needed.

Because automated methods of taxonomy creation use training sets of relevant documents, some taxonomists prefer to start with pre-existing or packaged taxonomies that contain both categories and documents. These starter taxonomies are often generic, but they can be used as the basis for more specific work. If the training sets were created using a particular set of rules, should this same set of rules be maintained to classify data later on? Is there any room for multiple categorization schemes?

6. NAMING CATEGORIES

Category naming is where art meets science. The category name must represent the content it contains, while at the same time it must be easily understood by users (Albrechtsen & Jacob, 1998; Bowker & Star, 1999; Olson, 1998).

Statistical data mining techniques are useful ways of finding similar words in documents while maintaining impartiality. These techniques create clusters of similar documents without imposing any opinion other than an author's choice of words. The downside, however, is that individual words may not represent the concepts the author is trying to convey, so the categories may contain seemingly dissimilar documents.

In technical terms, statistical techniques treat words and phrases in documents as points in a large, multidimensional space. Each dimension corresponds to a single word or phrase and the number of times it appears. When two documents share many of the same words and phrases, they will be relatively close together in this space and will be placed into the same category.

After the clustering technique produces the initial taxonomy, a taxonomist must edit it into a final, easy-to-use form. Of particular importance at this stage is having someone with technical

ISSN: 2324-9773

Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

knowledge of the material review the categories and the content within them and judge whether they possess an accurate, real-world affinity to each other.

Commercial software products build an initial taxonomy by applying a clustering algorithm to collections of data that have been selected by users who are unfamiliar with the software's clustering techniques, but do understand the content. After it creates the clusters, the software chooses tokens from the list of tokens in the feature set as suggested label terms for the cluster. These tokens, while valid in the context of the algorithm, rarely make sense to the user, and offer little or no clues as to the *aboutness* of any of the documents in the clusters. The taxonomist makes a decision about how good (i.e., precise) the cluster is, using only his/her knowledge of the subject areas, the titles of the documents, and any summary or abstract information. A taxonomist *who is also a subject matter expert* may also have an existing classification of the subject area available (see Glaser and Chi, 1988 for an analysis of the principled ways that human experts represent knowledge), but this expertise-based classification might bear no resemblance to the algorithmically-generated clusters from the machine. This situation leaves us with a choice between algorithm-based correctness and human-expertise-based correctness – a difficult choice with no straightforward criteria for making the choice (see Floyd, 1987 for a discussion of these two approaches to software engineering).

7. AN EXAMPLE

Clustering algorithms work best on clean data; that is, data that is rich in text, about only one subject area, and that has very descriptive titles. Through applied taxonomic work with corporate business partners, we have found that there is very little clean data in the real world (Pohs et al., 2001). Data tends to be sparse, titles tend to be cryptic, and business documents often contain more than one main idea. The taxonomist has to work to select good data before the clustering algorithm is applied. This is a tedious, manual, and impractical process, especially in large corporations where data ownership is an issue.

Recently a customer of a major knowledge management software vendor created a draft taxonomy and discovered that the automatically generated category names contained long, unidentifiable numbers. The numbers remained a mystery until someone on the content team recognized that these numbers were codes for various trade publications. As it happened, many of the documents being processed were journal articles. Each had a field for identifying the journal in which it had been published. The computer discovered these numbers occurring repeatedly, so it obligingly created categories out of them. However, these categories were of limited use in helping users determine what the articles were really about. More study should be undertaken to determine ways to create meaningful labels automatically.

8. MEASURING CATEGORY VALIDITY

Once a machine-generated taxonomy is created, taxonomists examine each category to determine:

- Is this category real (in other words, is the content therein truly related in a meaningful way)?
- If this category is real, what is the best name for it?

ISSN: 2324-9773

Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

Determining the validity of categories may well be the most difficult part of the taxonomy creation process. Taxonomists will be challenged to find human-recognizable structure in some categories, which may at first appear to be totally random collections. But while taxonomists can validate categories with a little diligence and creativity (Pohs et al., 2001), we need to create and study a set of subjective measures of taxonomy validity that can be adapted to different sets of content and repeatedly used.

9. AUTOMATICALLY CATEGORIZING DOCUMENTS INTO CLUSTERS

Many commercial software products use automatic classifiers and clustering algorithms to add documents to categories, which are subsequently edited by human taxonomists. There are several problems with this approach. First, without prior knowledge about how the clusters were created the taxonomist has no idea how well the clustering algorithm matches the categorization algorithm. By moving documents from one category to another, the taxonomist might inadvertently change the cluster's feature set. Because this feature set, or set of representative tokens, is what the categorization software uses as a model, the human can inadvertently corrupt the model, causing new documents to appear in inexplicable places.

Adding new documents automatically to the set can also change the structure of the feature set, causing the composition of the category to drift, or turn toward a different set of features, thus rendering it meaningless to the human editor.

10. CONCLUSION

Creating and maintaining taxonomies in complex human organizations introduces new problems in computer-human interaction. We anticipate difficult – even painful – problems to solve, but also new understandings that can help form a new area of research into how people find and use information, and how close software can come to meaningfully categorizing information

REFERENCES

Albrechtsen, H., & Jacob, E.K. (1998). The dynamics of classification systems as boundary objects for cooperation in the electronic library. *Library Trends 47*(2): 293.

Bowker, G.C., & Star, S.L. (1999). *Sorting things out: Classification and practice*. Cambridge USA: MIT Press.

Code, L. (1995). *Rhetorical spaces: Essays on gendered locations*. New York: Routledge.

Community Intelligence Labs (1999). *Communities of practice: Issues*, retrieved from http://www.co-i-l.com/coil/ knowledge-garden/cop/issues.shtml.

Eneva, E., & Petrushin, V. (2002). Learning to change taxonomies. *Proceedings of Data Mining and Knowledge Discovery SPIE 2002*. See summary at http://www.cs.cmu.edu/~eneva/talks/Tax.ppt.

Floyd, C. (1987). Outline of a paradigm change in software engineering. In G. Bjerknes, P. Ehn, and M. Kyng (Eds.), *Computers and democracy: A Scandinavian challenge*. Brookfield, VT: Gower.

ISSN: 2324-9773

Pohs, W. (2002). Creating and Maintaining Machine-Generated Taxonomies in Human Organizations: Contributions from Research and Practice. 13th ASIS SIG/CR Classification Research Workshop, 53-58. doi:10.7152/acro.v13i1.13836

Glaser, R., & Chi, M.T.H. (1988). Overview. In M.T.H. Chi, R. Glaser, and M.J. Farr (Eds.), *The nature of expertise*. Hillsdale NJ: Erlbaum.

Graef, J. 2001. Managing taxonomies strategically. *Montague Institute Review 2001.* Retrieved from http://www.montague.com/review/taxonomy3.html.

Muller, M.J. (1997). Ethnocritical heuristics for reflecting on work with users and other interested parties. In M. Kyng and L. Matthiessen (Eds.), *Computers in context and design*. Cambridge MA: MIT Press.

Muller, M.J. (2000). *Models of the social construction of knowledge and authority in business organizations*. Plenary presentation at *Human Computer Interaction Consortium, Winter Park CO USA, February 2000.*

Muller, M.J. 2002 in press. Participatory design: The third space in HCI. In J. Jacko and A. Sears (Eds.), *Handbook of HCI.*

Muller, M.J., Carotenuto, L., Fontaine, M., Friedman, J., Newberg, H., Simpson, M., Slusher, J., & Stevenson, K. (1999). Social and computing solutions for voluntary communities of practice: Designing CommunitySpace. In *Proceedings of IEEE WET-ICE conference.*

Muller, M.J., Pohs, W., & Friedman, J. (2000). Issues in the design of software systems to support voluntary electronic communities. Position paper at CSCW 2000 workshop, *Classification schemes.* Philadelphia USA: ACM.

Olson, H.A. (1998). Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library Trends (47)*2: 233.

Pohs, W., Thiel, G., Earley, S., & White, M. (2001). *Practical Knowledge Management: The Lotus Knowledge Discovery System.* Double Oak, Texas: IBM Press.

Roberts-Witt, S.L. (1999). *Practical taxonomies: Hard-won wisdom for creating a workable knowledge classification system.* Retrieved from http://enterprise.supersites.net/kmmagn2/km199901/featureb1.htm.

Ruby, D. (1999). *Tip for taxonomists: Keep it simple, stupid.* Retrieved from http://enterprise.supersites.net/kmmagn2/km1999901/departmental.htm.

Star, S.L. (1996a). Grounded classification: Grounded theory and faceted classification. *Proceedings of Information Systems and Qualitative Research.*

Star, S.L. (1996b). "To classify is human." Keynote talk at *Hypertext'96.*

ISSN: 2324-9773