Yang, K., Jacob, E., Loerhlein, A., Lee, S., Yu, N, (2004). The Best of Both Worlds: A Hybrid Approach to the Construction of Faceted Vocabulaires. Proceedings of the Asia Information Retrieval Symposium. Beijing, China. 18-20 October 2004.

The Best of Both Worlds: A Hybrid Approach to the Construction of Faceted Vocabularies

Kiduk Yang, Elin K. Jacob, Aaron Loehrlein, Seungmin Lee, Ning Yu
School of Library and Information Science
Indiana University
1320 East 10th Street, Bloomington, Indiana, USA
{kiyang, ejacob, aloehrle, seungmin, nyu}@indiana.edu

ABSTRACT
In this study, we explore a semi-automatic construction of a faceted vocabulary, which can be used as a mechanism for organizing Web-based resources. Based on the analysis of the manual process of faceted vocabulary construction, we modeled a hybrid approach to facet generation that integrates the strengths of manual and automatic methods.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, system issues, user issues*.

General Terms
Algorithms, Theory.

Keywords
Faceted Vocabulary, Classification, Information Organization, Digital Library, Semantic Web.

1. INTRODUCTION
Enumerative classification schemes have long provided an effective tool for organizing a collection of resources by assigning each resource to a single class in a set of predefined and mutually exclusive classes. Although librarians have traditionally relied on classification schemes such as the Library of Congress Classification [LCC] and the Dewey Decimal Classification [DDC] to provide access to physical resources, the use of machine-based full-text searching undermined the perceived utility of classification for information discovery and retrieval. However, growing frustration with the huge retrieval sets and numerous false drops that accompanied do-it-yourself searching on the Web has generated renewed interest in classification, categorization and the power of controlled vocabularies.

There are many challenges to a classification-based approach to organizing the Web. For example, it is impossible to "organize" the whole Web due to its massive size and the diversity of Web resources. Even if such a feat were feasible, clustering approaches are not incremental and text categorization approaches are based on a static classification scheme, rendering them unable to deal with the dynamic nature of the Web corpus. A highly variable and dynamic environment such as the Web requires an organizational approach, which can not only accommodate the dynamic nature of human knowledge but also respond to the information needs of a diverse and interdisciplinary population.

Because traditional classification schemes attempt to enumerate all knowledge in a given domain within a fixed set of predetermined classes, they are ill-suited for organizing resources in the diverse and multidisciplinary environment of the Web. Recognizing the inherent rigidity of traditional enumerative structures, Ranganathan [8] proposed a more flexible approach to organization that represented knowledge not as a set of static classes but as a set of concepts and relationships. This approach identifies the various aspects (characteristics or *facets*) of a given domain so as to derive a set of independent concept hierarchies that represent the range of characteristics relevant to that domain. Each such concept hierarchy is populated by the set of possible values (or *isolates*) that are used to describe that aspect for a given resource. Classes are created by combining isolates from this controlled vocabulary according to an established citation order, assuring collocation of related resources within a dynamically-generated hierarchy [4]. Thus, construction of a faceted organizational scheme neither prescribes a finite set of classes nor predetermines the relationships among classes. Rather, it establishes control over the formal semantics underlying the scheme and, in so doing, provides a conceptual basis for both the formation of classes and the establishment of relationships among the classes that comprise the resulting classification structure.

The dynamic and adaptive nature of a faceted vocabulary is more effective in organizing Web documents than traditional classification schemes that establish a fixed set of predefined and static classes. However, manual construction of a faceted vocabulary is a resource-intensive process requiring considerable intellectual effort and its implementation on the Web is impractical. The goal of this research is to discover a semi-automated method of faceted vocabulary construction that will make such an approach more viable for organizing the Web. This paper describes work in progress that investigates automated methods for streamlining and standardizing the process of constructing a faceted vocabulary.

2. CONSTRUCTION OF A FACETED VOCABULARY
The fundamental organizing principles underlying the development of a faceted system are the grouping of that which is related and the separation of that which is unrelated. Unlike the fixed structure of classes produced by enumeration, faceting provides for the organization of concepts in modular hierarchies by splitting (separating) unrelated or dissimilar concepts and lumping (grouping) related or similar concepts. Relevant concepts are identified by partitioning domain terminology into mutually-exclusive baseline facets [7] that are subsequently combined to form higher-order facets. Typically, development of the faceted vocabulary is an iterative process of analyzing a domain vocabulary and identifying clusters of relevant values [1]: initial clusters of values are aggregated into progressively more comprehensive groupings that identify general concepts and provide the initial set of baseline facets. These baseline facets are then combined to form modular hierarchies of superordinate facets. To create a classification scheme, values from this modular vocabulary are joined according to a standardized combinatorial order, generating a hierarchical structure of classes. In this way, a faceted structure of concepts and concept values ensures consistency of representation and

coherence of structure within individual facets while assuring that the facets and the relationships between facets remain adaptable to context and usage [7].

2.1 A Hybrid Approach to Faceted Vocabulary Construction

The process of constructing a faceted classification scheme is generally described as "analytico-synthetic". Because construction of such a scheme begins with the collection and subsequent grouping of linguistic terms specific to a given domain, the process is generally described as “bottom-up”, distinguishing it from the “top-down” process of division employed in the construction of enumerative classification schemes. The development of a faceted vocabulary necessarily begins with analysis of the linguistic terminology of the associated domain; but this analysis may not be effective if executed within a vacuum. For this reason, analysis of domain content should combine inductive (or “bottom-up”) acquisition of the linguistic base and deductive (or “top-down”) analysis of terms and term relationships based on the domain’s conceptual framework. By employing a "middle-out" strategy that integrates bottom-up and top-down approaches by analyzing the terminology of a domain within its existing conceptual framework [7], the resulting vocabulary only identifies the most relevant concepts for the initial set of baseline facets but also maintains the relationships between concepts and concept hierarchies that are most meaningful within the domain context.

Bottom-up creation of a faceted vocabulary is prone to human error and inconsistency. And, because facet creation is intellectually labor-intensive, automation of the development process has not seemed feasible. However, we theorized that using a hybrid, middle-out approach could support automation of facet generation by integrating the processing capabilities of the machine with the analytical and evaluative capabilities of the human. This hybrid approach to facet generation would begin with identification of the heuristics or basic sorting strategies used by humans in the grouping process. Analysis of these heuristics would then indicate which strategies could be handled automatically by the machine to generate a set of candidate facets and values.

2.2 Analyzing the Faceted Vocabulary Construction Process

To assess the viability of an integrated, hybrid approach, we decided to begin the process of constructing the faceted vocabulary by identifying a lexicon of concepts from an existing representational system currently used to index a collection of Web documents. The representational system selected for this project was EPA Topics (http://www.eap.gov/eaphome/topics.html), an indexing scheme used by the United States Environmental Protection Agency [EPA] to provide access to a collection of high-quality resources dealing with a range of environmental issues.

The first step in generating the faceted scheme involved creating a primary lexicon base consisting of all unique, information-bearing terms in the set of descriptors used in the EPA Topics category labels. To assess the conceptual framework of the domain and its influence on how domain phenomena were conceptualized, all pairs of descriptors were generated automatically to establish the broader context within which each unique term occurred. Manual analysis of each individual term by examining its function in associated descriptor pairs identified unique concepts by establishing the conceptual context(s) within which each term occurred. Analysis of the automatically generated lexicon base within the conceptual framework provided by term pairs allowed specification of the context within which an individual concept occurs and highlighted any consistencies in the existing indexing system that could undermine efforts to construct a faceted vocabulary. However, the important aspect of this phase was investigation of the sorting heuristics. Specification of the analytic strategies used by humans in analysis of a domain's lexicon would point to heuristics that could be automated to augment the manual process. Accordingly, we examined the analytic strategies used by two indexers to discover a set of heuristics that can both streamline and standardize the process of creating a faceted vocabulary.

2.3 Automating the Process of Faceted Vocabulary Construction

Examination of the analytic strategies employed by two indexers revealed complementary heuristics that could be handled automatically to create an initial set of baseline facets. These heuristics organize terms extracted from an existing structure of terms and term relationships such as an enumerative classification schemes, a thesauri or other forms of metadata relevant to the domain to be organized. Because these heuristics were identified for automatic implementation, they do not include methods that require an intellectual understanding of the domain to be classified. Instead, these methods rely on organizing terms according to their inherent meanings and their positions in relation to other terms. It should be noted that, because an otherwise productive heuristic may group a certain proportion of terms incorrectly, these heuristics are used to generate a preliminary set of candidate facets with their associated facet values.

2.3.1 The Suffix Heuristic

The suffix heuristic classifies a term according to its suffix. This approach differs from previous work with suffixes that employed stemming heuristics to achieve the conflation of terms [3, 10] or that identified a term’s position within a phrase [6]. In the hybrid approach to facet construction, the suffix heuristic is used to automatically group terms according to the meaning of the term's suffix to create a set of preliminary candidate facets.

The first step in the suffix heuristic is identification of those suffixes which will be used to group concept terms. An initial list of suffixes was generated consisting of common word endings identified as suffixes by *Webster's Third New International Dictionary* [11] that matched the endings of three or more terms in the EPA Topics lexicon. This list was augmented with EPA word endings that were not identified as suffixes in Merriam-Webster, but seemed likely to create meaningful classes (e.g., *-day* and *-man*). The suffixes in the augmented list were then conflated by meaning. For example, the suffixes *-ion*, which indicates an “act or process; result of an act or process”, and *–ment*, which indicates an “action, process, art, or act of a (specified) kind”, were grouped under the general class of “action”, so that terms ending in *–ion* or *–ment* would be grouped together as potential values of an “action” facet.

Suffix meanings vary considerably in granularity: some conflated meanings are as general as “action”, while others are highly specific, such as “doctrines, theories, and sciences”, which applies to *-logy* and *-science*. In addition, many suffixes have multiple meanings. For example, the suffix *-cy* indicates both “states, qualities, and conditions” (e.g., "bankruptcy") and “offices, ranks, and functions” (e.g., "chaplaincy"). In such cases, the most prevalent meaning associated with the suffix in the EPA Topics was selected. A few suffixes were grouped under more than one meaning if it appeared that terms with that

suffix would contribute equally well to both classes of meanings and if the number of terms with that suffix seemed manageable. Suffixes that are substring endings of longer suffixes (e.g., *-ar* is the substring of *-lar*) were not used to group terms. In some cases, two very similar suffixes may have different meanings, such as *-ess* and *-ness*. In cases where both suffixes have the same meaning, the longer suffix generally returns words at a higher level of precision. This provides the option of increasing precision at the expense of recall by "deactivating" the shorter suffix.

2.3.2 The WordNet Heuristic

The WordNet heuristic groups terms according to their position in the WordNet category hierarchy available at <http://www.cogsci.princeton.edu/~wn/>. The groups formed by this heuristic form the basis for potential facets in a manner similar to the suffix heuristic. This approach differs from previous research that used WordNet to assign specific meanings to the terms of a query [3] or that assigned meanings to the descriptors of articles [5]. Our work is similar to that of Burke [2], who used WordNet to group articles using different words with similar meanings; but our approach is to group related terms based on WordNet categorization.

The first step of the WordNet heuristic involved submitting EPA Topics terms to the WordNet database to extract the category structure of each individual term within the hierarchy. Terms that share a common WordNet category hierarchy were subsequently grouped to form a potential candidate facet. The groups produced by the WordNet heuristic were generally higher in both precision and recall than the groups formed by the suffix heuristic. Another advantage of the WordNet heuristic is that it allows the granularity of class meanings to be modified more easily than the suffix heuristic. For example, WordNet can identify *incineration* as specifically as "burning, combustion" or as generally as an "act, human action, human activity", while the suffix heuristic identifies *-ation*, and thus "incineration", only as an "action".

2.3.3 The Concept Pairs Heuristic

The approach used in the concept pairs heuristic is to group pairs of terms that share a common term. A concept pair consists of two terms that are "paired" based on their association in the primary resource from which the lexicon base is drawn. In this study, we extracted term pairs from category labels and from the hierarchy of EPA Topics. Term pairs that shared a common term were then grouped on the basis of the terms function (e.g., noun or modifier) to form potential facets (e.g. "air" and "water" from *air pollution*, *water pollution*).

The strength of the concept pair heuristic, especially when it generates the concept pairs from an existing category hierarchy, is that it mines manually identified concept associations embedded in a organizational structure that may be missed by syntactic or linguistic approaches. In addition to leveraging human judgment about concept relationships, the concept pairs heuristic capitalizes on co-occurrence data that identifies contextual relationships between concepts. The analysis of concept pairs suggests that terms that generally appeared in association (e.g., in the same EPA Topics category label) are likely to form a compound phrase or concept within the domain.

2.4 Creating the faceted scheme

After concept terms have been grouped through automation of one of the three heuristics discussed above, the validity of each candidate facet must be assessed manually. Each potential facet is checked against the base lexicon for conceptual and contextual (domain-based) consistency both within the individual facet and across the set of candidate facets. This will identify duplication of concepts across facets as well as inclusion of irrelevant concepts that may have occurred from splitting of meaningful phrases. During the process of checking for internal and external consistency, individual terms may be shifted from one facet to another. In some cases, an entire facet may be eliminated when all of its terms are moved to other facets. In extreme situations, the initial set of candidate facets may be rejected and the process of automated facet generation may be repeated by re-applying the heuristics in a different manner. The result of validity checking should be a set of potential facets whose values (isolate terms) demonstrate maximum intension and minimum extension.

After candidate facets have been validated, they are labeled. The facet label must capture the most specific superordinate concept represented by the terms nested within the facet. For example, the isolates *shirt*, *trouser*, *sock*, and *skirt* would be labeled *clothes*, since the concept of "clothes", consisting of characteristics such as "is a thing" and "worn by people", applies to each of the values and has no characteristics that are not shared by all. At this point, the decision may be made to organize a facet's isolate terms into subfacets. A subfacet is a grouping of isolate terms by characteristics that are shared by a subset of terms in a facet. Subfacets can help users comprehend a complex list of values in a single facet and provide the indexer more flexibility in representing individual resources.

Once a facet and its subfacets and/or isolate terms have been established, the internal ordering must be established for values within a facet or subfacet and for subfacets within a facet. This is known as *order in array*. The order in array used to arrange the subfacets themselves follows the principle of increasing complexity. While users will be able to locate a known item in a large array that is ordered alphabetically, isolate terms that are arranged alphabetically are likely to have less in common with their immediate their immediate neighbors than with other values that may appear elsewhere in the listing. Although there is generally no single best principle by which to order values in an array, the arrangement of isolate terms should follow a recognizable principle and allow users to predict the location of different types of values.

After the order in array has been established for each facet, the citation order must be established. This is the order in which facets will be combined to produce a class or category label for each resource. More importantly, it generates a hierarchical structure by collocating related classes. Because one of the primary advantages of using a faceted scheme with digital collections is that it allows a user to reorganize the collection by simply redefining the order in which facets are combined, the citation order established during construction of the faceted scheme serves as a default organizational structure. Nonetheless, the default structure should be useful to the widest possible range of domain users so that the user need not specify a citation order to search the collection.

3. A GENERALIZED APPROACH TO FACETED SCHEME CREATION

We have described a semi-automatic process that integrates machine processing and human intelligence to facilitate the construction of a faceted scheme. Although our hybrid process is based on research utilizing an existing hierarchical category structure (i.e., EPA Topics), it provides a generalizable approach to construction of a faceted vocabulary that can be applied to a Web corpus without an existing indexing structure.

3.1 Data Source Selection

The first step in a semi-automated process of constructing a faceted vocabulary involves identifying the data source from which to extract the key concepts and concept relationships. Concepts and relationships can be mined from the classificatory structure and/or category labels of existing organizational systems, domain-specific thesauri, document surrogates annotated by an indexer or the document texts themselves. Existing category data can be internal to the corpus to be organized (e.g., EPA Topics), external to the corpus but about the same domain (e.g., the EPA category of Yahoo!) or external to both the corpus and the domain (e.g. WordNet).

An existing organizational scheme, especially when it is about the corpus, is likely to be the richest data source since it contains the distilled efforts of system creator(s) and indexer(s) to organize the corpus. Despite its richness, however, this data is typically constructed manually and is liable to be influenced by the bias and subjective view of the indexer. Combining multiple sources of information, which has shown to be effective in the retrieval setting [12, 13], is preferred because concepts and concept relationships can be harvested across the multiple views of the system creators, indexers, and authors.

3.2 Lexicon Base Generation

Once a data source has been selected, the next step is to generate a lexicon base of concept terms and term pairs from the selected data source. The lexicon base, which will provide input data for the concept grouping methods, is comprised of three lexicon subsets. The first lexicon subset consists of the unique single terms from the data source, whether category labels, annotations, or document text. When the data source is noisy, as is the case with document text, only statistically significant terms should be selected. The second lexicon subset consists of noun phrases. A noun phrase is defined as a noun-noun, noun-noun-noun, or adjective-noun term pair whose component terms appear adjacently in a phrase window identified by punctuation. The third lexicon subset consists of noun-noun or adjective-noun term pairs that are identified based on co-occurrence in the data source. Terms that co-occur frequently (but not next to each other) in category paths, annotations, or document texts are good candidates for this third lexicon subset of concept term pairs.

3.3 Concept Group Identification

Having generated a lexicon base that contains potential concepts and the concept relationships occurring in the organizational scheme or corpus, the automated concept grouping methods described in section 2.3 are applied to generate the concept groupings that will constitute the preliminary faceted vocabulary. The basic strategy here is to identify groups of related concepts that could be potential facet values. Application of concept grouping methods employing various data sources and lexicon subsets will generate different concept groupings, which can then be compared and evaluated to create a more comprehensive set of candidate facets.

The WordNet heuristic is used to group individual terms that share a common hierarchical structure to identify candidate facets. Application of the suffix heuristic not only groups single terms based on a shared suffix, but also provides a potential facet label (e.g., *-ing = action*). The concept pairs heuristic is used to group word pairs that share a common term and can be applied to noun phrases or concept pairs.

3.4 Faceted Scheme Construction

These first three steps are designed to generate automatically a set of concept groupings or candidate facets that are subsequently evaluated manually by the system builder or indexer to validate the facet structure by comparison with one or more external resources (e.g., domain-specific thesauri or other representational structures); to assign (or validate) a potential facet label; to establish the order in array of facet values and subfacets; and to determine the default citation order for resource indexing and category organization. If appropriate, an associated metadata scheme can be created, based on the facet structure, that will define metadata elements for the corpus.

4. CONCLUSION

In this paper, we have described a study that explored the feasibility of constructing a faceted vocabulary using an existing hierarchical classification structure. We have also generalized the findings from that study to outline a hybrid, semi-automatic approach to faceted scheme creation that combines the strengths of the human with the strengths of the machine: the intelligence, context awareness and evaluative judgment that the human brings to the construction of high-quality faceted schemes with the speed of processing, unlimited memory and consistency in repetition of the machine.

5. REFERENCES

[1] Batty, D. Thesaurus construction and maintenance: a survival kit. Database 12, 1 (1989), 13-20.

[2] Burke, R.D., Hammond, K.J., Kulyukin, V., Lytinen, S.L., Tomuro, N., and Schoenberg, S. Question answering from frequently asked question files: Experiences with the AQFINDER system. *AI Magazine 18*, 2 (1997), 57-66.

[3] Harman, D. How effective is suffixing? *Journal of the American Society for Information Science, 42*, 1 (1991), 7-15.

[4] Jacob, E. K., and Priss, U. Non-traditional indexing structures for the management of electronic resources. In *Advances in classification research, vol.10*. Information Today for the American Society for Information Science, Medford, NJ, 2001, 73-90.

[5] Mock, K.J. and Vemuri, V.R. Information filtering via hill climbing, WordNet, and index patterns. *Information Processing & Management, 33*, 5 (1997), 633-644.

[6] Okada, M., Ando, K., Lee, S.S., Hayashi, Y., and Aoe, J. An efficient substring search method by using delayed keyword extraction. *Information Processing & Management, 37*, (2001), 741-761.

[7] Priss, U., and Jacob, E.K. A graphical interface for faceted thesaurus design. In *Proceedings of the 9th ASIS SIG/CR Classification Research Workshop* (Pittsburgh, PA, October 25, 1998). American Society for Information Science, Silver Spring, MD, 1998, 107-118.

[8] Ranganathan, S. R. (1945). *Elements of library classification: based on lectures delivered at the University of Bombay in December 1944*. N.K. Publishing House, Poona, 1945.

[10] Savoy, J. Stemming of French word based on grammatical categories. *Journal of the American Society for Information Science, 44*, 1 (1993), 1-9.

[11] *Webster's Third New International Dictionary, Unabridged*. Merriam-Webster, 2002. http://unabridged.merriam-webster.com (18 Jun. 2004).

[12] Yang, K. Combining text- and link-based retrieval methods for Web IR. In *Proceedings of the 10th Text Retrieval Conference (TREC2001)* (Gaithersburg, MD, 2001). U.S. Dept. of Commerce, Technology Administration, Washington, D.C., 2002, 609-618

[13] Yang, K. Information retrieval on the web. In *Annual Review of Information Science and Technology*, 39 (in press).