Combining Faceted Classification and Concept Search: A Pilot Study

Kiduk Yang (양기덕)*

Contents

ABSTRACT

This study reports the first step in the Classification-based Search and Knowledge Discovery (CSKD) project, which aims to combine information organization and retrieval approaches for building digital library applications. In this study, we explored the generation and application of a faceted vocabulary as a potential mechanism to enhance knowledge discovery. The faceted vocabulary construction process revealed some heuristics that can be refined in follow-up studies to further automate the creation of faceted classification structure, while our concept search application demonstrated the utility and potential of integrating classification-based approach with retrieval-based approach. Integration of text- and classification-based methods as outlined in this paper combines the strengths of two vastly different approaches to information discovery by constructing and utilizing a flexible information organization scheme from an existing classification structure.

Keywords: Concept Search, Faceted Classification, Knowledge Discovery

* Associate Professor, Department of Library and Information Science, Kyungpook National University(kiyang@knu.ac.kr)
논문접수일자: 2014년 10월 22일 최초심사일자: 2014년 11월 11일 게재확정일자: 2014년 11월 21일
한국문헌정보학회지, 48(4): 5-23, 2014. [http://dx.doi.org/10.4275/KSLIS.2014.48.4.005]

1. Introduction

"Digital library", like "informatics", is a popular term that is not bound to a common definition despite its prevalent use. A literal interpretation of a digital library is digital version or implementation of a physical library. Leiner (1998) defines the digital library as a collection of organized information objects and services that support its users via electronic or digital means, which highlights collection, organization, and utilization of digital information as key characteristics of a digital library. With the growth of Internet and subsequent digital information "explosion", the focus of digital library efforts shifted from creation and maintenance of digital collections to development of infrastructure and technology that facilitates organization and utilization of digital content (Gennary et al. 2003). Accordingly, classification and information retrieval are two of the main areas in digital library research.

Classification is a mechanism for both organizing and utilizing information by representing knowledge as a set of concepts and relationships. In that light, transformation of information into knowledge via classification is a comprehensive process that involves not only the construction of classification structure and categorization of information units, but also the utilization of classification data for knowledge discovery. However, there exist many challenges in applying classification-based approaches to the digital library setting, especially in terms of managing Web-based resources. Traditional classification schemes generally enumerate a fixed set of classes that are not only predefined but also static and will not be able to deal with the dynamic nature of the Web corpus. Consequently, methods of organizing Web information need to be efficient, flexible and dynamic. Moreover, post-retrieval organization of retrieved documents may be a more desirable as well as realistic approach than trying to organize the entire Web.

Based on our belief that dynamic and adaptive nature of faceted classification is well-suited for the Web, we conducted an exploratory study that investigated generation and application of faceted vocabulary as a potential approach for knowledge discovery on the Web. Construction of faceted vocabulary as well as application of faceted vocabulary and concept relationships for retrieval and knowledge discovery is a part of ongoing research to lay the groundwork for the Classification-based Search and Knowledge Discovery (CSKD) project.[1] In this paper, we describe our explorations of classification-based approach to knowledge discovery that combines information organization and retrieval approaches.

1) CSKD is an ongoing project undertaken by the author that aims to leverage an existing body of manually classified documents to enhance information retrieval and knowledge discovery on the Web.

2. Prior Research

A traditional classification scheme is a system of representation that attempts to enumerate all the knowledge of a given domain within a fixed set of static and predefined classes. Because it creates a one-to-one relationship between a class term and the individual concept the term represents, the traditional classification scheme functions as a controlled vocabulary and facilitates the transfer of knowledge across time and space without loss of information (Jacob 1994). Furthermore, the enumerative scheme serves to legitimize and reify a single ideological or sociopolitical perspective of the domain (Jacob and Albrechtsen 1997).

However, there is growing recognition among classification theorists that an enumerative classification scheme may not be the most effective access system for all users (Beghtol 2008; Broughton 2006). The inability to represent relationships other than those created by the nesting structure of the hierarchy renders traditional enumerative schemes less than effective for organizing resources in the diverse and multidisciplinary environment of the World Wide Web. Recognizing the inherent rigidity of traditional enumerative structures, Ranganathan (1944; 1959) proposed a more flexible approach to classification, where various aspects (or *facets*) and associated set of possible values (or *isolates*) can represent characteristics of a domain in a dynamically-generated hierarchy that neither prescribes a finite set of classes nor predetermines the relationships among classes.

The representational vocabulary (i.e., faceted vocabulary) of a faceted system is not only the fundamental building block of a classification scheme but also can be adopted as a post-coordinate indexing language and used to develop pre-coordinate indexing chains and subject headings. A faceted scheme can also be adapted to provide dynamic class structures capable of responding to the individual's immediate information needs by supporting user-generated ordering of facets. This offers the potential for flexible reconfiguration of the organizational structure capable of identifying new relationships between resources and thus accommodating a broader range of information needs than traditional enumerative schemes or faceted schemes using a fixed citation order (Yang and Jacob 2004; Prieto-Diaz 2003; Vickery 2008).

The typical approach to developing a faceted vocabulary involves identification and grouping of relevant values in an iterative process of inductive or bottom-up clustering of concepts where initial clusters are aggregated into progressively more comprehensive groupings to identify the baseline facets (Batty 1989). These baseline facets may then be combined to form superordinate facets. In this manner, a faceted structure of concepts and concept values provides consistency of representation and coherence of structure within individual facets, while connections between

facets remain adaptable to context and usage (Jacob and Priss 1999; Yang and Jacob 2004).

Although it is difficult to harness the inherent flexibility of faceted systems outside a computer-based medium, the application of faceted vocabularies holds potentially significant implications for the extension of user access within the Web environment. Unfortunately, these systems have not been exploited because of the perceived complexity of the faceted system itself. Thus widespread development and application of faceted systems of representation and organization has been forestalled by the intellectual effort required both to generate the faceted vocabulary and to index resources using a faceted approach. This paper reports on the development of a methodology for using an existing representational structure to seed the semi-automatic construction of a faceted vocabulary that can then be superimposed on the original structure to enhance retrieval performance.

3. Faceted Vocabulary Construction

3.1 Data

The creation of a faceted system necessarily begins with analysis of the linguistic vocabulary of the associated domain; but such analysis may not be effective if executed within a vacuum. Rather, analysis of domain content should be carried out within the existing conceptual framework of the domain, utilizing both inductive "bottom-up" and deductive "top-down" strategies (Loehrlein et al. 2005; Yang and Jacob 2004). By beginning with a top-down investigation of the conceptual framework of the domain, subsequent analysis of domain terms and term relationships will be better able both to identify the most relevant concepts that will constitute the initial set of baseline facets and to establish the meaningful relationships that obtain between those facets. Thus the first step in creation of a faceted vocabulary is necessarily "middle-out" in that it combines bottom-up acquisition of the linguistic base with top-down analysis of the domain's conceptual framework.

To assess whether this middle-out approach could be adapted to automate the process of facet generation from an existing enumerative system, it was decided to begin the process of constructing a faceted vocabulary by identifying a lexicon of concepts from a representational system currently used to index a collection of Web documents. The representational system selected for the current project was EPA Topics[2] (http://www2.epa.gov/home/az-index), an indexing scheme used by the

2) EPA website has undergone a major restructuring since the study data collection. EPA Topics page data used in the study is preserved in http://widit.knu.ac.kr/epa/topics.html.

United States Environmental Protection Agency [EPA] to provide access to a collection of high-quality resources dealing with a range of environmental issues. The harvested EPA Topics hierarchy consisted of 17 top-level categories with 184 categories at the first sub-level and 965 categories at the second sub-level, which indexed 3,708 webpages containing 14,540 outgoing links.[3] It is important to note that EPA Topics is not a true classification scheme: not all categories are mutually exclusive and any concept or category may be nested within more than one branch of the hierarchical tree structure. However, this representational system does provide a systematic ordering of nested categories with each category represented by a chain of descriptors that indicate its position within the larger conceptual structure.

3.2 Methodology

The first step in generating the faceted scheme involved the inductive, bottom-up creation of a primary lexicon base consisting of all unique, information-bearing terms in the set of EPA Topics descriptors used as category labels. To assess the conceptual framework of the domain and its influence on how domain phenomena were conceptualized and subsequently organized, all pairs of descriptors were generated in order to establish the broader context within which each unique term occurred. The analysis of unique terms within the context of their associated descriptors and descriptor pairs serves to identify unique concepts by establishing the context within which each term occurs.

Manual analysis of the automatically generated lexicon base within the conceptual framework provided by the associated terms and concept pairs allows specification of the context within which an individual concept occurs. Manual analysis also points to the lack of consistency in the existing representational system that might undermine efforts to construct the faceted vocabulary. However, the more important implication of this analysis is the need for a hybrid approach to the construction of faceted vocabularies. In an attempt to facilitate the manual process of lexicon analysis for faceted vocabulary construction, which is not only resource intensive but also prone to human error and discrepancy, we examined the thought processes involved in the manual analysis of two indexers[4] to discover a set of heuristics that can both streamline and standardize the

3) EPA Topics hierarchy can be viewed at http://widit.knu.ac.kr/epa/treeview.htm.

4) Indexers, who were graduate students specializing in classification theory, examined terms from EPA Topics (approximately 700 from category labels and 13,000 from titles and summaries of indexed webpages) to identify candidate facets and isolates.

faceted vocabulary creation process.

The examination of the manual analysis process revealed heuristics for creating an initial classification of concepts, from which a classificationist may manually create a faceted scheme. The automated component of the proposed faceted vocabulary construction process is comprised of three heuristics: *the suffix heuristic* that classifies terms based on their suffixes to group concepts according to the meaning of suffixes, *the WordNet heuristic* that utilizes term positions in the WordNet hierarchy to group related terms, and *the concept pairs heuristic* that identifies term pairs sharing a common term in the existing classification structure to group concepts that are strongly associated.

The heuristics described below organize terms that have been extracted from existing terminologies, such as enumerative classification schemes, document titles, document abstracts, and other forms of metadata relevant to the domain to be classified. Because these heuristics are mostly intended for automatic implementation, they do not make use of methods that requires an intellectual understanding of the domain to be classified. Instead, they organize the terms according to their inherent meanings and their positions in relation to other terms. It should be noted that a successful heuristic can still incorrectly classify a certain proportion of terms because the heuristics provide only the first draft of a faceted classification.

3.2.1 The Suffix Heuristic

The suffix heuristic classifies terms according to their suffixes. This approach differs from previous research into suffixes that employed stemming heuristics or machine learning in order to achieve the conflation of terms (Harman 1997; Savoy 1993; Jongejan and Dalianis 2009), or that identified a term's position within a phrase (Okada et al. 2001). Instead, the suffix heuristic groups terms according to the meaning of the suffixes themselves. These groups form the basis for potential facets.

The first step of the suffix heuristic is identification of suffixes, with which to group concept terms. An initial suffix list consisted of common word endings matching three or more terms in the EPA Topics lexicon, which are identified as suffixes in Merriam-Webster online (http://unabridged.merriam-webster.com). The initial list of suffixes was augmented with EPA word endings that were not identified as suffixes in Merriam-Webster, but that seemed likely to create meaningful classes, such as *-day* and *-man*. The suffixes in the augmented list were then conflated by meaning. For example, *-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 that end in *-ion* or *-ment* would be grouped together as potential values of the “action” facet.

Suffix meanings vary considerably in granularity, so that while some conflated meanings are as general as “action”, 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”, such as *bankruptcy*, and “offices, ranks, and functions”, such as *chaplaincy*. In most such cases, the most prevalent meaning associated with the suffix in the EPA Topics was chosen. 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 the suffix seemed to be manageable. Suffixes that are substring endings of longer suffixes (e.g., *-ar* vs. *-lar*) were not used in grouping terms. In some cases, the two suffixes may have different meanings, such as *-ess* and *-ness*. In cases where both suffixes have the same meaning, the longer suffix usually returns words at a higher level of precision. This gives the classificationist the option of increasing precision at the expense of recall by “deactivating” the shorter suffix.

We identified three flaws with the suffix heuristic. First, the most effective meaning to assign to ambiguous suffixes such as *-cy* varies between lists of terms. Therefore, it is improbable that any one designation of suffix meanings will be appropriate for every list of terms. Second, many terms do not have suffixes and therefore cannot be classified by this heuristic. Third, many terms end in strings that resemble suffixes but are not true suffixes. For example, *-ment* indicates an action, but *garment* is not an action.

3.2.2 The WordNet Heuristic

The WordNet heuristic groups terms according to their position in the WordNet hierarchy. 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 (Hane 2000; Natsev et al. 2007), or that assigned meanings to the descriptors of articles (Mock and Vemuri 1997; Suchanek et al. 2008). It is similar to the research of Burke et al. (1997), which used WordNet to associate articles that use different words but have similar meanings, except that we seek to group related terms instead of articles.

The WordNet heuristic uses WordNet 3.1, an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory (Miller et al. 1993;

Pedersen et al. 2004). WordNet assigns each meaning of each term to a class in its own enumerative hierarchical classification scheme. For example, WordNet assigns the term *networks* to the class "network, web", which is subsumed by the class "system, scheme", which in turn is subsumed by the class "group, grouping" that occupies the highest level within the WordNet hierarchy. The first step of the WordNet heuristic involved querying the WordNet with EPA Topics terms to extract the classes to which the terms belong and those classes' positions in the WordNet hierarchy. The terms with common WordNet classification hierarchy were then identified to form potential facet groups.

The groups produced by the WordNet heuristic turned out to be generally higher in both precision and recall than the groups formed by the suffix heuristic (Table 1). 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 classify *incineration* as specifically as "burning, combustion" or as generally as an "act, human action, human activity", while the suffix heuristic can classify *incineration* only as an "action".

〈Table 1〉 Sample Precision and Recall for the suffix- and WordNet heuristics

Term Class	*Term Count*	*Precision*				*Recall*			
		Suffix heuristic		WordNet heuristic		Suffix heuristic		WordNet heuristic	
Actions	150	0.96	(128/134)	0.94	(148/158)	0.85	(128/150)	0.99	(148/150)
States	41	0.79	(15/19)	0.95	(41/43)	0.37	(15/41)	1.00	(41/41)
Chemicals	25	0.89	(16/18)	1.00	(25/25)	0.64	(16/25)	1.00	(25/25)
Total	216	0.93	(159/171)	0.95	(214/226)	0.74	(159/216)	0.99	(214/216)

The WordNet heuristic also shares many of the limitations of the suffix heuristic. WordNet provides multiple meanings for many terms, from which the classificationist must select the most appropriate meaning, which is likely to vary between term lists. One disadvantage of the WordNet heuristic compared to the suffix heuristic is that it does not yet include a method for choosing between term meanings except on a term-by-term basis. In contrast, a decision regarding the meaning of a suffix affects all of the terms with that suffix. In addition, the WordNet heuristic suffers from representational problems and bias inherent in the WordNet hierarchy. For instance, classes in the WordNet hierarchy are often labeled with multiple concepts, some of which have distinct meanings (e.g. "*use, usage, utilization, utilisation, employment, exercise*"). Also, WordNet

does not classify many terms. While it provides a description and a list of synonyms for most terms, WordNet does not necessarily assign them to a class in its hierarchy. WordNet does classify some of these terms, but in a different form. For example, WordNet does not classify the EPA Topics term *innovative*, but it does classify the term's noun form, *innovation*. Stemming heuristics may be used to identify many of these alternate forms within WordNet. However, many terms that are not classified in WordNet do not have other forms, such as *solar* or *vermiculite*.

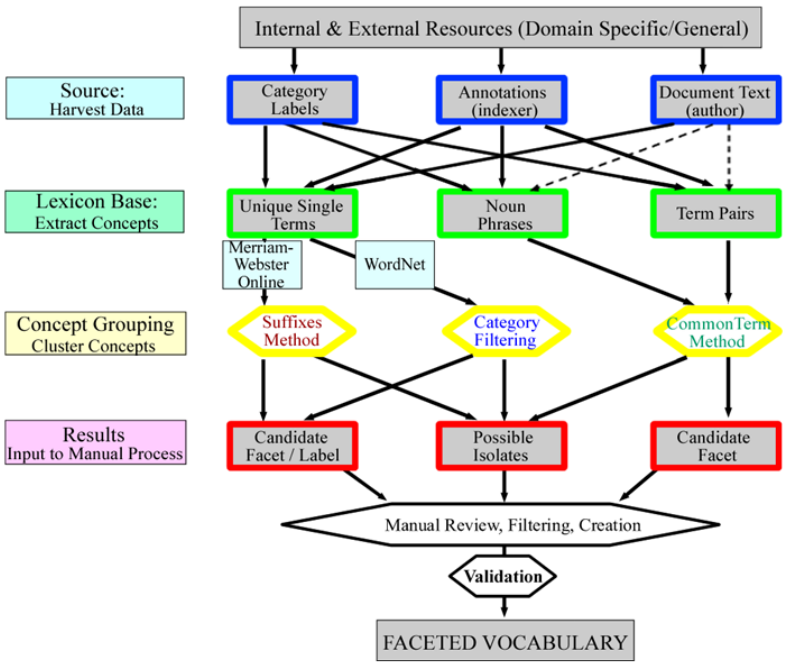
3.2.3 Concept Pairs Heuristic

The concept pairs heuristic groups together "concept pairs" that share a common term, where a concept pair consists of a pair of terms that are strongly associated in a classification structure. A list of concept pairs can be generated by identifying compound phrases in a class label, constructing term pairs from a hierarchical class path or compiling high frequency noun phrases in classified documents. In this study, we extracted term pairs from the label and class path of the EPA topics hierarchy and grouped together those term pairs that share a common term to form a potential facet group (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 the classification hierarchy, is that it mines the manually identified concept associations embedded in a classification structure, which may be missed by syntactic or linguistic approaches. In addition to leveraging human judgment about concept relationships, the concept pairs heuristic can capitalize on co-occurrence data to get at the contextual relationship between concepts. The analysis of concept pairs suggests that terms that usually only appear in the same class are likely to form a compound phrase or concept. For example, *international* and *cooperation* appear in seventy-seven and seventy-three EPA classes, respectively, and appear together in seventy-two of those classes. Therefore, it is probable that these terms are most relevant to the EPA when they are combined to represent the concept of *international cooperation*. Our analysis also indicates that, if Term A generally appears only with Term B, but Term B usually appears *without* Term A, then Term A can thought of as a qualifier to Term B. For example, "*programs*" appears in sixteen classes, of which thirteen also contain the term *cooperation*. "*Programs*" therefore appears to be most relevant to the EPA when interpreted as a type or instantiation of *international cooperation*.

3.3 Hybrid Approach

This paper describes only the first step in the iterative refinement of hybrid approach to faceted vocabulary construction, which is one of the long-term goals of the ongoing CSKD project. Figure 1 displays an overview of the proposed faceted vocabulary construction process, which depicts a hybrid approach that aims to integrate the human and machine capabilities. Not only did the manual analysis of human process seed the faceted vocabulary construction heuristics, the proposed approach also involves the manual examination of the heuristics outcome in the last phase to filter and validate the faceted vocabulary.



〈Figure 1〉 Overview of the Hybrid Approach to Faceted Vocabulary Construction

Given two concept clusters in Table 2, for example, the manual examination of the cluster 1 may lead to the creation of “chemical process” facet while cluster 2 is likely to be discarded since cluster 2 terms would group resources that have little in common in terms of useful class features. In the proposed approach, such cognitive process is combined with the automated processes of data harvest and concept clustering to leverage the best capabilities of man and machine.

〈Table 2〉 Concept clusters from faceted vocabulary heuristics

	Concept Cluster 1	*Concept Cluster 2*
Candidate Facets	Action Chemical/Molecular Changes	Action/Outcomes
isolates	aeration chlorinated combustion composting desorption incineration irradiated oxidation polychlorinated radiation tanning	measurement nonattainment performance pretreatment reimbursement requirement sediment settlement statement substance treatment

4. Faceted Vocabulary Application

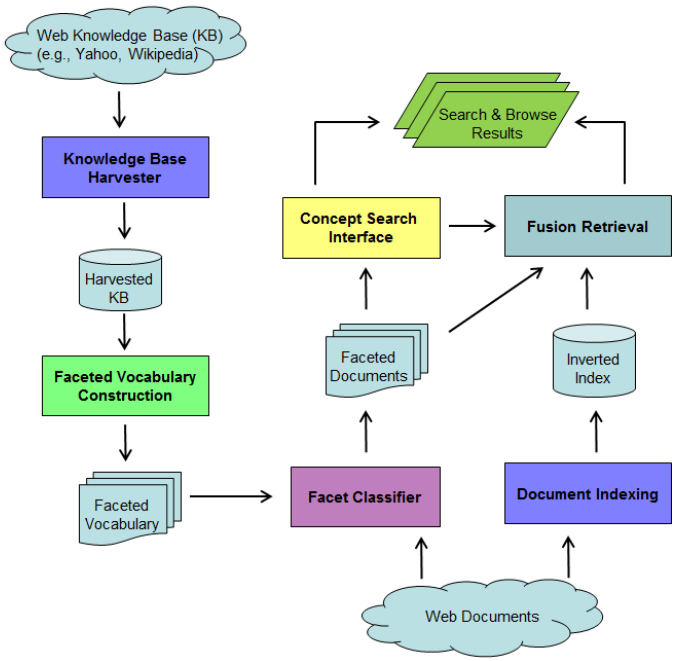
Although faceted classification addresses inherent weaknesses of conventional classification, the construction of faceted vocabulary alone is still an information organization process and does not lead well to effective information retrieval let alone knowledge discovery. Therefore, how faceted vocabulary can be applied and how it can be constructed are both sides of the same coin. Consequently, we explored the question of how to apply faceted classification in parallel with its construction.

4.1 Combining text- and classification-based Methods

Despite the success and popularity of Web search engines in recent years, full-text search does not adequately address some basic needs of information seekers. Information retrieval approach of the full-text search focuses on finding information that are likely to be "relevant" to a given query, but provides little assistance in bridging the gap between the "anomalous state of knowledge" (Belkin et al. 1982) and formulation of an effective query. Furthermore, retrieval-based approach to information discovery does not identify important concepts, let alone concept relationships, that may be useful in satisfying a more complex information need than finding specific information, thus leaving the task of knowledge discovery mostly to the user.

One of the ways classification-based approach to information discovery can address the shortcomings of retrieval-based approach is by letting the user browse a classification structure representing

important concepts and relationships, where he/she can quickly gain an overview of information landscape that may facilitate the information discovery process. The static hierarchical structure of conventional classification approach, however, can sometimes hinder information discovery process by confusing, misleading, or restricting the user with its rigidity and complexity. Faceted classification structure, on the other hand, facilitates flexible representation of knowledge using a set of concepts and relationships that can be structured dynamically to accommodate individual user's needs and perspectives. In addition, faceted vocabulary can be used in conjunction with full-text search as well as in complementing conventional classification-based approach. For example, facets can help both query refinement and focused taxonomy traversal by restricting the information domain.



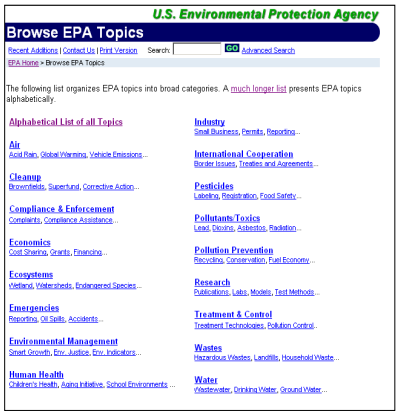
〈Figure 2〉 Concept Search System Architecture

While we investigated the methods of faceted vocabulary construction, we explored in parallel the utilization of classification data for knowledge discovery by experimenting with a "concept search" application that can effectively combine the strengths of full-text search, concept hierarchy and faceted vocabulary. The focus of concept search application is in efficient and effective identification of important concepts and concept relationships that are useful in fulfilling user's information needs. Fully implemented concept search application will consist of a knowledge base harvester, which

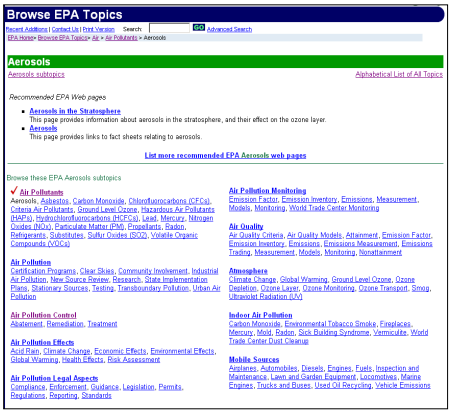
crawls and harvests an existing Web document classification structure, a faceted vocabulary construction component, which constructs a faceted vocabulary from the harvested classification structure, an automatic facet classifier, which maps faceted vocabulary to Web pages, a fusion retrieval module, which integrates text- and classification-based retrieval, and a concept search interface, which combines searching and browsing approaches to information discovery (Figure 2).

4.2 Concept Search Application

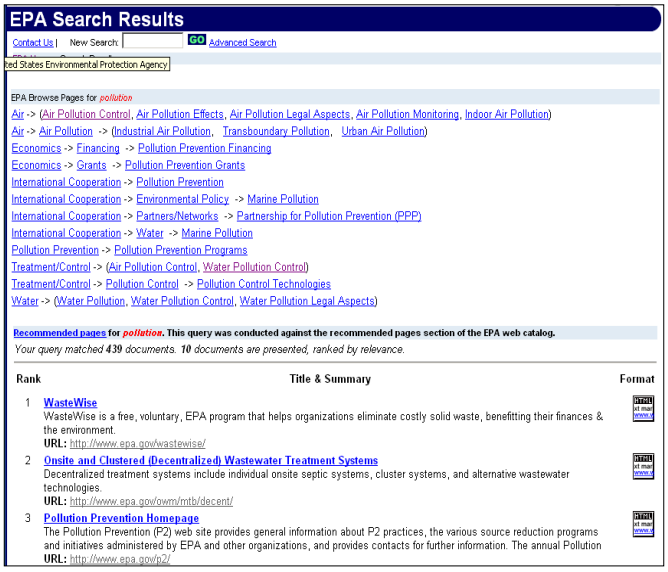
Browse EPA Topics: Home



Browse EPA Topics: Aerosols

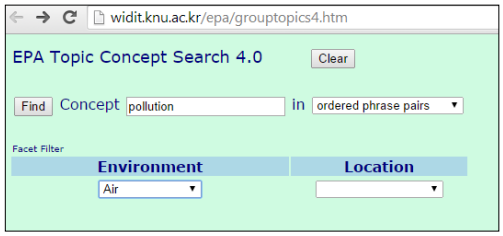


EPA Topics: Search results

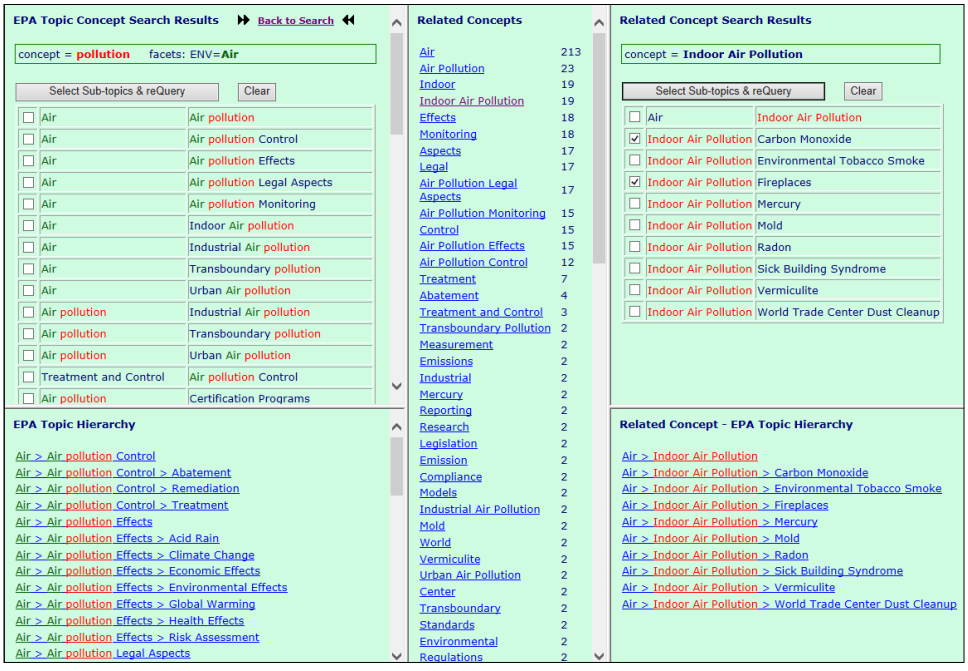


〈Figure 3〉 EPA Topics Browse & Search Interface

EPA Topics, whose hierarchical classification structure was used as data source in our study, employed both browsing and searching (Figures 3) to utilize its classification data.[5] The browse interface implemented conventional hierarchical tree navigation similar to Yahoo!, and the search interface implemented a layered search that returned multiple sets of search results returned from matching the query to different sources of evidence (i.e. Topics category labels, classified EPA pages, controlled vocabulary metadata, full-text index). The concept search application developed in our study combines browsing and searching in an integrated interface that emphasizes concept and concept relationship identification (Figures 4a, 4b, 4c).

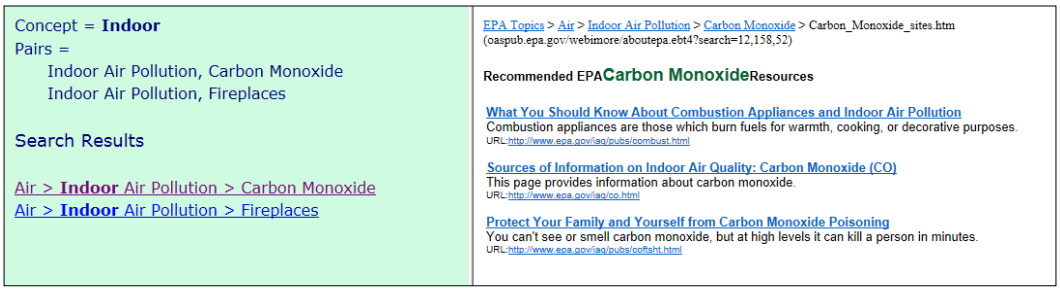


〈Figure 4a〉 Concept Search: Home



〈Figure 4b〉 Concept Search Result: Explore & Refine

5) The discussion in this section describes the EPA Topics interface before the restructuring of EPA website.



〈Figure 4c〉 Concept Search: Refined search result

The initial search screen of the concept search (Figure 4a) allows the user to search for either single word or phrase concept pairs extracted from EPA Topics, as well as providing query refinement option by facets (e.g. Environment, Location).[6] The presence of facets in the initial search screen can also serve as a guide to identifying key concept attributes (e.g. *air, soil, water* for "pollution"). The initial concept search result display consists of the portion of the EPA Topic hierarchy in which those concept pairs occur (left columns of Figure 4b), and related concept terms and their occurrence frequencies in EPA Topics (middle column of Figure 4b). Clicking the "related concepts", which are extracted from matched concept pairs and displayed in a descending order of frequency, will execute another concept search using the clicked concept to display the results in the rightmost columns. By displaying the original query results on the left and related query results dynamically generated from related concepts on the right, the concept search result interface allows the user to refine his/her query via exploring the concept hierarchy. A user can click the link of each hierarchy (bottom left for original query and bottom right for related query) or select one or more subtopics and click "Select Sub-topics & reQuery" button to display associated Web pages (Figure 4c). Different colors were used to distinguish the query terms and facet isolates.

The following example demonstrates a potential concept search session. A user, who heard about the dangers of carbon monoxide emission from fireplace but has only a vague recollection, enters "pollution" in the initial search box. After examining the concept search results, where the "Environment" facet isolates are displayed at the top of the *Related Concepts* column, or simply after viewing the isolates in the dropdown box of the "Environment" facet in the initial search screen, the user submits the "pollution" query refined with the "Environment" facet value of "air" (Figure 4a). User then sees the "Indoor Air Pollution" ranked fourth in *Related Concepts* and

6) Only two facets are integrated in the initial concept search for demonstration purposes.

clicks it to display the related concept search results in the right column (Figure 4b), where he selects “Carbon Monoxide” and “Fireplaces” to display the documents of interest (Figure 4c). As illustrated in the previous example, concept search employs faceted vocabulary to assist query formulation and refinement process as well as leveraging existing classification structure to provide context for query terms and identify concept relationships, thus facilitating both the search and knowledge discovery process.[7)]

5. Discussion

In this study, we explored the generation and application of a faceted vocabulary as a potential mechanism to enhance knowledge discovery on the Web. Our faceted vocabulary construction process revealed some heuristics that can be refined in follow-up studies to further automate the creation of faceted classification structure. Our concept search application demonstrated the utility and potential of integrating classification-based approach with retrieval-based approach. In follow-up studies, we plan to streamline the faceted vocabulary construction process by iteratively refining the facet identification heuristics, further implement the concept search application as described in the concept search system architecture, and conduct a case study to evaluate both the faceted vocabulary construction and application processes.

Integration of text- and classification-based methods as outlined in this paper combines the strengths of two vastly different approaches to information discovery by constructing and utilizing a flexible information organization scheme from an existing classification structure. Concept search application, which is designed to highlight important concepts and identify relationships among them, focuses on concept rather than document discovery and thus helps users make sense of the document collection by dynamically organizing concepts according to individual user’s information needs and requirements.

Fusion (e.g. hybrid approach) is one of the key concepts in our vision of the digital library, where integration of information organization and information retrieval approaches as well as combination of machine and human intelligence are integral components that enhance and extend the traditional library services. Our study implemented this fusion principle in several layers. The manual analysis of EPA Topics hierarchy and examination of that manual process revealed

7) CSKD concept search interfaces and associated data can be seen in http://widit.knu.ac.kr/epa/index.htm.

heuristics for identifying candidate facets and isolates, while the proposed faceted vocabulary construction approach involves manual examination of automatic heuristics outcome for filtering and validation of the faceted vocabulary. In addition, the concept search prototype combined not only text- and classification-based methods but also leveraged the computer processing power in such a way to capitalize on the human cognitive ability (e.g., pattern recognition) to facilitate information discovery.

References

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

[2] Beghtol, Clare. 2008. “From the universe of knowledge to the universe of concepts: the structural revolution in classification for information retrieval.” *Axiomathes*, 18(2): 131-144.

[3] Belkin, Nicholas. J., Oddy, R. N. and Brroks, H. M. 1982. “ASK for information retrieval: Part I.” Background and theory. *Journal of Documentation*, 38(2): 61-71.

[4] Broughton, Vanda. 2006. “The need for a faceted classification as the basis of all methods of information retrieval.” *Aslib proceedings*, 58(1/2): 49-72.

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

[6] Gennari, Jeffrey, Harrington, M., Hughes, S., Manojlovich, M. and M. Spring, M. 2003. “Preparatory Observations Ubiquitous Knowledge Environments: The Cyberinfrastructure Information Ether.” *NSF Post Digital Library Futures Workshop, Chatham, MA.*

[7] Gove, Phillip. B. (Ed.). 2002. *Webster's third new international dictionary of the English language, unabridged*. Merriam-Webster.

[8] Hane, Paula J. 2000. “Beyond keyword search - Oingo and Simpli.com introduce meaning-based searching.” *Information Today*, 17(1): 57-68.

[9] Harman, Donna. 1991. “How effective is suffixing?” *Journal of the American Society for Information Science*, 42(1): 7-15.

[10] Jacob, Elin K. 1994. “Classification and crossdisciplinary communication: breaching the boundaries imposed by classificatory structure.” *Knowledge organization and quality management: Advances in knowledge organization*, 4: 101-108.

[11] Jacob, Elin K. and Albrechtsen, H. 1997. “Constructing reality: the role of dialogue in the development of classificatory structures.” *Knowledge organization for information retrieval: Proceedings of the 6th International Study Conference on Classification Research*, 42-50.

[12] Jacob, Elin K. and Priss, U. 1999. “Application of faceted classification structures in electronic knowledge resources.” *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop*, 87-106.

[13] Jongejan, Bart and Dalianis, H. 2009. “Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike.” *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 145-153.

[14] Leiner, Barry M. 1998. “The scope of the digital library.” *Dlib Working Group on Digital Library Metrics*. <http://www.dlib.org/metrics/public/papers/dig-lib-scope.html>

[15] Loehrlein, Aaron, Jacob, E., Yang, K., Lee, S. and Yu, N. 2005. “A hybrid approach to faceted classification based on analysis of descriptor suffixes.” *Proceedings of the American Society for Information Science and Technology*, 42(1).

[16] Miller, George. A., Beckwith, R., Felbaum, C., Gross, D. and K. Miller. 1993. “Introduction to WordNet: an on-line lexical database.” *International Journal of Lexicography*, 3(4): 235-244.

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

[18] Natsev, Apostol. P., Haubold, A., Tesic, J., Xie, L. and Yan, R. 2007. “Semantic concept-based query expansion and re-ranking for multimedia retrieval.” *Proceedings of the 15th international conference on Multimedia*, 991-1000.

[19] Okada, Makoto, Ando, K., Lee, S. S., Hayashi, Y. and Aoe, J. 2001. “An efficient substring search method by using delayed keyword extraction.” *Information Processing & Management*, 37(5): 741-761.

[20] Pedersen, Ted, Patwardhan, S. and Michelizzi, J. 2004. “WordNet: Similarity: measuring the relatedness of concepts.” *Demonstration Papers at HLT-NAACL 2004*, 38-41.

[21] Prieto-Diaz, Ruben. 2003. “A faceted approach to building ontologies.” *IEEE International Conference on Information Reuse and Integration*, 458-465.

[22] Ranganathan, Shiyali. R. 1944. *Library Classification: fundamentals and procedure: with 1008 graded examples & exercises*. Madras: Madras Library Association.

[23] Ranganathan, Shiyali. R. and Palmer, B. I. 1959. *Elements of library classification*. London: Association of Assistant librarians.

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

[25] Suchanek, Fabian M., Kasneci, G. and Weikum, G. 2008. “Yago: A large ontology from Wikipedia and Wordnet.” *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3): 203-217.

[26] Vickery, Brian. 2008. “Faceted classification for the web.” *Axiomathes*, 18(2): 145-160.

[27] Yang, Kiduk and Jacob, E. 2004. “A hybrid approach to generating and utilizing faceted vocabulary for knowledge discovery on the web.” *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*.