Seungmin Lee
Proposal for presentation

Thesaurus integration for generating faceted vocabulary

Abstract

With the rapid growth of the content of digital information, classification (information organization) and information retrieval have been the two of the main areas in digital library research. The traditional classification schemes which enumerate fixed classes (categories) cannot cope with organization of digital information. Constructing faceted vocabulary with dynamic nature might be one of the solutions to retrieve Web contents effectively and efficiently. To construct dynamic faceted vocabulary, it is necessary to have lexicon with clear relationship among terms and hierarchical structures. Applying thesaurus to the construction of faceted vocabulary can provide the necessary resource. However, applying one specific thesaurus to the faceted vocabulary construction process might cause several potential problems because each thesaurus has its own characteristics inherited from its construction process and there are serious differences among thesauri, such as the use of synonym and antonym, different hierarchical levels, and the scope of thesaurus. Therefore, integration of thesauri with keeping their own structure and conceptual relationship can be one of the appropriate approaches for the construction of faceted vocabulary, because we can expand the range of the thesaurus through thesaurus integration without loosing structures and relationships.

This current research selected physics field as a domain and two prominent thesauri in the domain: PACS and PIRA. PACS is a thesaurus which covers the whole domain, but it lacks of the basic concepts of the domain because it is constructed to deal with high level knowledge in physics field. Whereas, PIRA is for the instructional purpose, so it only contains basic concepts of physics field. By integrating these two thesauri, it is possible to construct a hierarchical structure which covers the whole concepts related to physics from basic to high level.

To integrate both thesauri, this research took bottom-up approach. All terms and/or phrases used as labels were extracted, and the extracted terms from two thesauri were mapped according to the meaning of them. It shows a huge difference that 9 top-level categories, 61 subcategories, and 484 terms are included in PIRA, whereas PACS contains 10 top-level categories, 66 subcategories, and total 6,281 of terms. However, PIRA contains concepts and terms which are not included in PACS thesaurus. The result of the mapping between two thesauri shows that 22 terms in category labels (for both main and sub categories) are exactly matched, and 136 terms are conceptually matched over the whole structure. Based on the analysis and mapping stage, integrated structure was constructed. Among the terms matched, we selected terms that can be used as facets and potential facets which can have subordinate. The result of the mapping shows 12 numbers of facets, 136 of potential facets, and 3,406 specific terms. It can provide the faceted vocabulary with clear hierarchical relationship. Based on this structure, faceted vocabulary can serve as a full-fledged lexicon base for retrieval and organization of concepts in physics field.

By constructing the integrated structure from thesaurus integration, it supports lexicon base with hierarchical structure and clear relationship with dynamic nature. It can be used to utilize faceted vocabulary for retrieval and knowledge discovery which is more efficient approach to organizing the Web.

References

1. Yang, Kiduk & Jacob, E.K (2004). Organizing the Web: Semi-Automatic construction of a faceted scheme. The International Association for Development of the Information Society (IADIS) WWW/Internet 2004 Conference. Madrid, Spain. 6-9 October 2004.
2. Doerr, Martin (2001). Semantic problems of thesaurus mapping. Journal of Digital Information 1 (8).
   Available at: http://jodi.ecs.soton.ac.uk/Articles/v01/io8/Doerr/?printable=1
3. Okada, M., Ando, K., Lee, S.S., Hayashi, Y., & Aoe, J. (2001). An efficient substring search method by using delayed keyword extraction. Information Processing & Management, 37, 741-761.
4. Hane, P. (2000). Beyond keyword search – Oingo and Simpli.com introduce meaning-based searching. Information Today, 17(1), 57.
5. Jacob, E.K. & Priss, U. (1999). Application of faceted classification structures in electronic knowledge resources. Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, 87-106.
6. Priss, U., & Jacob, E.K. (In press). Conceptual representation of faceted thesauri. Advances in classification research, vol.9. Medford, NJ: Information Today for the American Society for Information Science.
7. Leiner, B. 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
8. Priss, U., & Jacob, E.K. (1998). A graphical interface for faceted thesaurus design. Proceedings of the 9th ASIS SIG/CR Classification Research Workshop, 107-118.
9. Burke, R.D., Hammond, K.J., Kulyukin, V., Lytinen, S.L., Tomuro, N., & Schoenberg, S. (1997). Question answering from frequently asked question files: Experiences with the FAQFINDER system. AI Magazine 18(2), 57-66.
10. Mock, K.J. & Vemuri, V.R. (1997). Information filtering via hill climbing, WordNet, and index patterns. Information Processing & Management, 33(5), 633-644.
11. Jacob, E. 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.
12. Miller, G. A., Beckwith, R., Felbaum, C., Gross, D., Miller, K. (1993). Introduction to WordNet: an on-line lexical database. International Journal of Lexicography, 3(4): 235-244.
13. Batty, D. (1989). Thesaurus construction and maintenance: a survival kit. Database 12 (1), 13-20.