Enriching a Thesaurus as a Better Retrieval Aid

Yejun Wu
School of Library and Information Science
Louisiana State University
267 Coates Hall, Baton Rouge, LA 70803
wuyj@lsu.edu

ABSTRACT
Current thesauri do not indicate how two related (RT) terms can be related. An enriched thesaurus by differentiating the RT relationship helps people understand how two RT terms can be related, and is expected to be a better retrieval aid by facilitating information needs clarification and query formulation.

Keywords
Thesaurus, RT relationship, ERIC

INTRODUCTION
A traditional controlled vocabulary operates by choosing a preferred way of expressing a concept and then making certain that synonymous ways of expressing the concept will be connected to the preferred terminology. The preferred term is shown in relationship to its broader term(s), narrower terms(s), and related term(s), if any. These are often designated by BT, NT, and RT, respectively. A controlled vocabulary organized hierarchically for a specific subject area is a thesaurus. Here is an incomplete entry from the Thesaurus of ERIC Descriptors (Houston, 2001):

Violence

NT Family Violence

BT Antisocial Behavior

RT Aggression

Crime

The limitation of current thesauri is that the Related Term (RT) relationship is too general. It is shown that "Violence" is related to "Aggression," but not how they are related. Current thesauri are good enough as a document indexing tool, but not good enough for retrieval aiding, knowledge organization for discovery, and the understanding of the knowledge of a domain. To augment their power as a searching aid and a knowledge organization tool, I propose to enrich the RT relationship between a descriptor (e.g., Violence) and its related terms (e.g., Aggression) by mining the Web. The enriched thesaurus displays specific relationship terms (very often verbs) between two RT terms, followed by an example (usually a sentence). Here is an example of the enriched entry:

Violence

RT (contribute) Aggression (Violence plays a role in Aggression)

(increase) Aggression (exposure to media Violence increases Aggression)

(predict) Aggression (early exposure to TV Violence predicts Aggression in adulthood)

Such an enriched thesaurus can be a useful tool to help users clarify their information needs and formulate queries as well because much more information about how concepts are related is provided. Such an enriched thesaurus can also improve knowledge organization for discovery. Swanson and Smalheiser (1999) discovered numerous undiscovered implicit relationships within the biomedical literature. For example, if one article reports that substance A causes disease B and another reports that disease B causes disease C, then we can infer that substance A might cause disease C. Enriched, detailed term relationships facilitate the grouping of related terms and inference of concept relationships through the specified relationship chains.

RELATED WORK
Much work has been done on automatic thesaurus construction or term expansion (Wang, 2006; Li and Yang, 2004), and using thesauri for query expansion in information retrieval (Shiri and Revie, 2006). There are also studies on using relational thesauri for automatic query expansion in information retrieval (Green, 1996; Wang, Vandendorpe, and Evens, 1985; Wan, Evens, Wan, and Pao, 1997). A relational thesaurus is a collection of term pairs with a predefined set of RT relationships. Therefore, it is the terms rather than the term relationships that are used for query expansion. This study differs from earlier work by investigating an approach to enriching the RT relationships, evaluating the quality of such an enriched thesaurus and in the long run, its usefulness for supporting users' information needs clarification and query formulation.

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.
Copyright is held by the author/owner(s).

METHOD

The ERIC thesaurus was used as the test bed because the terms in the education domain are easier for our participants to understand, compared to those highly specialized domains (such as medicine and engineering). Since enriching a whole thesaurus is labor intensive, and the purpose of the study is to test the research idea, only a small portion of the thesaurus is enriched.

Three descriptors – violence, anxiety, and learning – were selected because each of them has many RT terms and they should be familiar to our participants in the evaluation experiment. For each descriptor, take all of its RT terms, enrich their RT relationships; then start from each of its RT terms, take all of its RT terms, enrich their RT relationships. To avoid a scalability problem, we can go only two steps down along the RT relationship chain (e.g., Violence RT Aggression RT Crime) for "violence" and "anxiety," and go only one step for "learning" because it has too many RT terms.

Violence

RT Aggression → Aggression

Bullying	RT Crime
Crime	Hostility
.....	……

A program was designed to issue a query using a pair of terms with RT relationship to Google. The top 10 PDF documents and HTML (or HTM) pages from .edu and .org Websites were crawled and processed to strip off the HTML tags to get text. A second program was written to extract a snippet of text (from a crawled document or Web page) which is composed of 1-3 sentences and includes the two query terms. Multiple snippets could be extracted from one document. The extracted snippets were manually analyzed to extract the term relationships.

To control the labeling labor, a maximum of three relationship terms were used. Since the thesaurus is a controlled vocabulary, the extracted relationship terms were controlled as well. Single-word terms (such as contribute) are preferred to multi-word terms (such as play a role in).

The quality of the enriched term relationships is evaluated preliminarily. Two graduate students were recruited to participate in a two-phase experiment – recall phase and recognition phase. In the recall phase, each student was presented with the original sections of the ERIC thesaurus related to the three descriptors used in this study, then asked to generate as many specific relationships as possible between a descriptor and each of its related terms in one and a half hour. In the recognition phase, the students was presented with the enriched sections of the thesaurus (with some noise relationships added), and then asked to determine whether the relationships are relevant in an hour.

PRELIMINARY RESULTS

The recall experiment turns out to be a difficult task (since it was a brainstorming task with scant contextual information), indicating that the original thesaurus does not help people understand the relationships of two terms. The recognition experiment turns out to be a much easier task, indicating the enriched thesaurus helps people understand how two terms can be related. This is an ongoing project, and only the preliminary evaluation experiment has been done, and more experiments are expected to be done to confirm the results.

CONCLUSION AND FUTURE WORK

Current thesauri are good enough as a document indexing tool, but not good enough for retrieval aiding and knowledge organization for discovery. Enriching the RT relationship in the ERIC thesaurus helps users understand how two RT terms can be related. The enriched thesaurus is expected to be a better retrieval aid, and is to be evaluated whether they help users with information needs clarification, query formulation, and the understanding of the knowledge of a domain.

ACKNOWLEDGEMENT

The study is supported in part by the LSU Summer Research Stipend Program.

REFERENCES

Green, R. (1996). A relational thesaurus: modeling semantic relationships using frames. *Annual Review of OCLC Research 1996*. 94-97.

Hudson, J. E. (Ed.) (2001). *Thesaurus of ERIC Descriptors (14th ed.)*. Westport, CT: the Oryx Press.

Li, K. W. and Yang, C. C. (2005). Automatic crosslingual thesaurus generated from the Hong Kong SAR Police Department Web Corpus for crime analysis. *JASIST, 56(3)*, 272-282.

Shiri, A. and Revie, C. (2006). Query expansion behavior with a thesaurus-enhanced search environment: a user-centered evaluation. *JASIS, 57(4)*, 462-478.

Swanson, D. and Smalheiser, N. (1999). Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. *Library Trends, 48(1)*, 48-59.

Wan, T., Evens, M., and Wan, Y. (1997). Experiments with automatic indexing and a relational thesaurus in a Chinese information retrieval system. *JASIST, 48(12)*, 1086-1096.

Wang, J. (2006). Automatic thesaurus development: term extraction from title metadata. *JASIST, 57(7)*, 907-920.

Wang, Y., Vandendorpe, J., and Evans, M. (1985). Relational thesauri in information retrieval. *JASIS, 36(1)*, 15-27.