Different views of textual “aboutness”: A recipient evaluation of the content descriptors proposed by professional indexers, authors, readers and automatic term extractors

Lynne Bowker[1], Rebecca Mackay[1], Deni Kasama[2] and Jairo Buitrago Ciro[1]

[1] University of Ottawa, [2] Universidade Estadual Paulista “Júlio de Mesquita Filho”

Abstract: As an increasing number of groups – professional indexers, authors, readers, and even automated tools such as term extractors – tackle the challenge of describing document contents, the time is right to conduct a recipient evaluation of these different key words lists to see which ones users perceive to be the most helpful for identifying relevant materials.

There are numerous longstanding challenges associated with describing the “aboutness” of a document (e.g. Hutchins 1977, 1978). Over time, the issues are becoming more complex as different players are turning their hand to this challenge. Once seen as principally the role of professional indexers (e.g. Hutchins 1977, 1978), content description is now undertaken by others as well, such as authors (e.g. Névéol et al. 2010) and users (e.g. Kehoe and Gee 2010; Woolwine et al. 2011). Moreover, there have even been attempts to automate content description to some degree, such as automatic indexing (Nazarenko and Aït El Mekki 2007).

Up to this point in time, the vast majority of research on content description has been carried out by information scientists. However, we would like to respectfully suggest that this could be an opportune time to look outside information science and to investigate what is happening with regard to content description in other fields, to see whether there are practices elsewhere that might be of interest. One potential area of overlapping interests could be the field of terminology, where automatic term extraction tools are under active development (e.g. Heylen and De Hertog 2015; Bowker and Delsey 2016).

Inspired by the work of Kipp (2011a/b), who conducted a comparative evaluation of user, author and professional indexing, we extended research in this area by including a fourth category in our own comparative analysis: keywords that were generated by an automatic term extractor called TermoStat (Drouin 2003). In this way, we sought to explore whether automated objective textual analysis could be a complement to more subjective content description techniques that draw on intuition.

Frequency can certainly be one potential indicator of a word’s importance in a text, but it is not sufficient as a sole measure of “aboutness”. TermoStat is a term extractor that works by comparing the contents of a specialized text against a much larger general reference corpus to identify those words in the specialized text that are *unusually* frequent as compared to their frequency in larger general reference corpus (i.e., a measure of the frequency of disproportionate occurrence).

The corpus used for this pilot study consisted of four chapters taken from a scholarly edited collection in the discipline of Translation Studies. The volume has a back-of-the-book index that was produced by an indexing professional. The author of each individual chapter also provided a list of key words corresponding to that chapter. A PhD student in the discipline read each chapter—without having access to either the back-of-the-book index or the author recommended key words—and supplied a list of key words for each chapter, thus representing the user perspective. Finally, TermoStat was used to generate a keyword list for each chapter automatically.

To explore whether end users found certain types of content descriptor lists to be more useful than others, we conducted a recipient evaluation where we asked 24 graduate students and professors who conduct research in the field of Translation Studies to read one of the chapters without reference to the various content descriptor lists. After reading the chapter, they were shown (in random order) the four corresponding lists as produced by:

(**A**) – author
(**R**) – reader
(**P**) – professional indexer
(**T**) – TermoStat automatic term extractor

After consulting the lists, they were asked to rank them from best to worst according to how well they perceived that each list described the contents of the chapter. Owing to the small number of participants and the variation in individual preferences, the data for the pilot study is inconclusive; however, the results do suggest some possible trends:

- The content descriptor lists produced by professional indexer were ranked last on average.
- The content descriptor lists produced by "non experts" (i.e., readers and TermoStat) were never ranked last, and so appears it appears that the non-experts in this pilot study produced more helpful lists than the "experts" (i.e. the professional indexer and the authors).
- The various lists appear to be complementary (i.e., not a lot of overlap in content), which suggests that a collaborative approach to identifying content descriptors could be beneficial.
- Perhaps most notable is the fact that lists produced by TermoStat were never ranked last, and in the amalgamated results, they placed second overall. This would suggest that automated tools have something to contribute.

Further work is needed to generate a larger set of data in order to confirm or refute the above observations.

References

Bowker, Lynne and Delsey, Tom. 2016. "Translation Studies and Information Science: Adaptation, Collaboration, Integration," in *Border Crossings: Translation Studies and Other Disciplines*. Yves Gambier and Luc van Doorslaer, eds. Amsterdam/Philadephia: John Benjamins, 73-95.

Drouin, Patrick. 2003. "Term extraction using non-technical corpora as a point of leverage," *Terminology* 9(1): 99-115.

Heylen, Kris and De Hertog, Dirk. 2015. “Automatic Term Extraction,” in *Handbook of Terminolog*y. Hendrik J. Kockaert and Frieda Steurs, eds. Amsterdam/Phildaelphia: John Benjamins, 203-221.

Hutchins, W. J. 1977. “On the problem of “aboutness” in document analysis”. *Journal of Informatics* 1(1): 17-35.

Hutchins, W. J. 1978. “The concept of “aboutness” in subject indexing”. *Aslib Proceedings* 30(5): 172-181.

Kehoe, Andrew and Matt Gee. 2011. “Social tagging: A new perspective on textual ‘aboutness’” VARIENG: Studies in Variation, Contacts and Change in English 6. http://www.helsinki.fi/varieng/series/volumes/06/kehoe_gee/

Kipp, Margaret E.I. 2011a. “Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing.” *Knowledge Organization* 38(3): 245-261.

Kipp, Margaret E.I. 2011b. “User, Author and Professional Indexing in Context: An Exploration of Tagging Practices on CiteULike.” *Canadian Journal of Library and Information Science* 35(1): 17-48.

Nazarenko, Adeline and Touria Aït El Mekki (2007) “Building back-of-the-book indexes?” In *Application-Driven Terminology Engineering*. Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellví (eds). Amsterdam/Philadelphia: John Benjamins, 179-202.

Névéol, Aurélie, Rezarta Islamaj Dogan and Zhiyong Lu. 2010. “Author Keywords in Biomedical Journal Articles.” American Medical Informatics Association (AMIA) Annual Symposium Proceedings 2010: 537–541. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041277/

TermoStat: http://termostat.ling.umontreal.ca

Woolwine, David, Margaret Ferguson, Eric Joly, David Pickup, Cristian Mihai Udma. 2011. “Folksonomies, Social Tagging and Scholarly Articles.” *Canadian Journal of Information and Library Science* 35(1): 77-92.