Using Ancillary Text to Index Web-Based Multimedia Objects

Lyne Da Sylva *(lyne.da.sylva@umontreal.ca)*
EBSI, Université de Montréal
James Turner *(james.turner@umontreal.ca)*
EBSI, Université de Montréal

PériCulture is the name of a research project at the Université de Montréal which is part of a larger project based at the Université de Sherbrooke. The parent project aimed to form a research network for managing Canadian digital cultural content. The project was financed by Canadian Heritage and was conducted during the fiscal year 2003-2004. *PériCulture* takes its name from *péritexte* and culture, *péritexte* being one of a number of terms used (in French, our working language) to mean ancillary text associated with images and sound. It is a sister project to *DigiCulture*, another part of the same larger research project which studied user behaviours in interactions with Canadian digital cultural content. The general research objective of *PériCulture* was to study indexing methods for Web-based nontextual cultural content, specifically still images, video, and sound. Specific objectives included:

1. identifying properties of ancillary text useful for indexing;
2. comparing various combinations of these properties in terms of performance in retrieval;
3. contributing to the development of bilingual and multilingual searching environments;
4. developing retrieval strategies using ancillary text and synonyms of useful terms found therein.

In computer science, research into indexing images and sound focuses on the low-level approach, performing statistical manipulations on primitives in order to identify semantic content. This approach is also referred to as the 'content-based approach' (e.g. Gupta and Jain, Lew). In information science, research into indexing images and sound focuses on associating textual information with the nontextual elements, and this often involves manipulating ancillary text. This approach is referred to as the 'high-level' or 'concept-based approach' (e.g. Rasmussen, O'Connor, O'Connor, and Abbas). A number of factors militate in favour of automating the high-level approach as much as possible. These include the very large volume of Web-based materials available, the disparity among cataloguing and indexing methods from one collection to another, and the high cost and relative inconsistency of human indexing.

Our work in this project focuses on text associated with Web-based still images, and builds on previous work in this area of information science (e.g. Goodrum and Spink, Jörgensen, Jörgensen et al., Turner and Hudon). We identified a number of Web sites that met our criteria, i.e., that contained multimedia objects, that had text associated with these objects that was broader than file names and captions, that were bilingual (English and French), and that housed Canadian digital cultural content. We identified keywords that were useful in indexing and studied their proximity to the object described. We looked at indexing information contained in the `Meta` and `Alt` tags, and whether other tags contained useful indexing terms. We studied whether standards such as the Dublin Core were used. We identified Web-based resources for gathering synonyms for the keywords.

Our study found that a large number of useful indexing terms are available in the ancillary text of many Web sites with cultural content. We evaluated various types of ancillary text as to their usefulness in retrieval. Our results suggest that these terms can be manipulated in a number of ways in automated retrieval systems to improve search results. Cross-language comparison of the results reinforces our previous research results, which suggest that indexing in other languages can be generated automatically from a single language using Web-based tools.

Rich information that can be used for retrieval is available in many places on Web sites with cultural content, from the file name to explicit information in captions to descriptive information in surrounding text to the contents of various HTML tags. Algorithms need to be developed to exploit this information in order to improve retrieval.

Finally, we feel that our work is useful because of the synergy created by the approaches we use. We are both interested in image indexing, but come from different fields. Lyne Da Sylva's expertise is in linguistics and James Turner's in information science. By working together, we are able to pool our knowledge and develop richer methods than would otherwise be available to either of us for approaching the question of automating indexing for images and other multimedia objects.

Bibliography

Goodrum, A., and A. Spink. "Image searching on the Excite web search engine." *Information Processing and Management* 27.2 (2001): 295-312.

Gupta, A., and Ramesh C. Jain. " Visual information retrieval." *Communications of the ACM* 40.5 (71-79): 71-79.

Jörgensen, Corinne. *Image attributes: an investigation.* PhD thesis, Syracuse University, 1995.

Jörgensen, Corinne. "Image attributes in describing tasks: an investigation." *Information Processing and Management* 34.2/3 (1998): 161-174.

Jörgensen, Corinne, Alejandro Jaimes, Ana B. Benitez, and Shih-Fu Chang. "A conceptual framework and empirical research for classifying visual descriptors." *Journal of the American Society for Information Science and Technology (JASIST)* 52.11 (2001): 938-947.

Lew, Michael S. *Principles of visual information retrieval.* New York: Springer, 2001.

O'Connor, Brian C., Mary K. O'Connor, and June M. Abbas. "User reactions as access mechanism: an exploration based upon captions for images." *Journal of the American Society for Information Science* 50.8 (1999): 681-697.

Rasmussen, Edie M. "Indexing images." *Annual Review of Information Science and Technology* 32 (2004): 169-196.

Turner, James M., and Michèle Hudon. "Multilingual metadata for moving image databases: preliminary results." *L'avancement du savoir : élargir les horizons des sciences de l'information, Travaux du 30e congrès annuel de l'Association canadienne des scicnces de l'information.* Ed. Lynne C. Howarth, Christopher Cronin and Anna T. Slawek. Toronto, 2002. 34-45.