A Comparison Study of Clustering or Classification Methods for Search Results Visualization in Web Search Context

Aline Crédeville
École de Bibliothéconomie et de Science de l'Information
Université de Montréal
aline.credeville@umontreal.ca

Abstract

The amount of information on the Web is steadily growing since its beginning. The number of online information retrieval systems has increased in parallel to this amount of information. These have been designed to help the information seeking process and to perform the user final tasks from his perspective (Wilson, 1999). As a result, web search engines are massively used to allow the accomplishment of a large range of environment-dependant and goal-ended tasks (Broder, 2002; Rose & Levinson, 2004; Toms, Freund, Kopak, & Bartlett, 2003). While information retrieval techniques (indexing, organization and ranking) and interactive features have been improved since the last twenty years (Baeza-Yates & Ribeiro-Neto, 1999; Manning, Raghavan, & Schütze, 2008; Shneiderman & Plaisant, 2005), the on-line information retrieval systems still remain hard to use (Borgman, 1996; Markey, 2007a, 2007b) and don't fit the cognitive and affective processes of the information searching tasks efficiently (Ingwersen, 1996; Ingwersen & Järvelin, 2005; Kuhlthau, 2005; Kuhlthau, Heinström, & Todd, 2008). The context of the user (professionnal, scholar, or everyday life), his final tasks, his individual differences, and the Kuhlthau's stages of information-seeking process (1991) still need to be taken into account. These design problems generate a high cognitive load because of the growth of affective and cognitive uncertainty (Gwizdka, 2010) which has to be reduced to ease the learning process. Two of the reasons of this uncertainty are, on the one hand, the noise in the considered search results which overhelm the user working memory and, on the other hand, the lack of interactive features which slow down exploration, one of the critical stage of the information-seeking process (Markey 2007a; Kuhlthau 1991).

The use of information visualization could bring significant improvements to the design of information retrieval systems. Information visualization is defined as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" (Card, Mackinlay, & Schneiderman, 1999). Information visualization, cartography for example, is known to reduce redundancy in data and to facilitate the identification of meaningful patterns through large and multidimensional data (Bertin & Barbut, 1977; Larkin & Simon, 1987; Norman, 1993; Resnikoff, 1989; Tufte, 1990). It has been mainly developed in the information retrieval field as a way to display abstract information in a graphical and logical structured form (Card et al., 1999; Chen, 2004; Jin Zhang, 2008) and as a way to interact with information in an information-seeking context (Shneiderman, 1996). In 2000, after ten years of research, the information visualization field has developed largely accepted theoretical foundations. There are yet important issues (Burkhard et al., 2007; Chen, 2005; Keller & Tergan, 2005) to be solved.

- The divorce between the logical organization of the abstract information and it representation into an understandable metaphor.
- Multidimensional scaling.
- The evaluation of usability of visual information retrieval systems (Kerren, Stasko, Fekete, & North, 2007; Lin, Kerren, & Jiaje Zhang, 2009; Plaisant, 2004).

Considering these issues of traditional and visual information retrieval, we think that the gap between the ranking structure of search results and their transformation into a meaningful graphical and interactive representation could be bridged with data mining operations. More specifically, classification and clustering algorithms could extract salient structures of the retrieved set of search results in order to shape the visual representation of the results.

In the context of information-seeking with a web search engine, the goals of our research project are the following.

- Identify the organizational factors required to make the graphical representation constructed by the display algorithm a meaningful way to present the retrieved set of search results. More specifically, we seek to answer the following questions.
  - What are the constraints imposed by the classification and clustering methods on the possible graphical and interactive visual representation of search results?
  - What are the parameters to apply for each method of clustering and

classification?

- Determine which improvements, from the end-user perspective, that are made possible by the visualization of search results, for both the clustering and classification methods. For this goal, the specific questions we want to answer are the following.
    - What are the characteristics of the web search strategies enabled by each method of clustering and classification?
    - What are the graphical and interactive characteristics of the web search strategies enabled by text-listed and visual presentation of search results?
- Establish a model of the relations between the logical organization of search results, the graphical and interactive display, the end-user, and the task.

To answer these questions, a controlled experimentation is to be conducted according to the framework for Interactive Information Retrieval, designed by Borlund (2003). In our experiment, we will compare two Web Information Retrieval Systems (hereafter named WIRS); each one tested by a different sample of future librarians and domain experts. The selected end-users will have to execute a simulated search task on the Web. This comparison will take into account the variation of both organizational algorithmic method – classification and clustering – and the textual and visual presentation of the search results. The collected data will consist of the multimedia transactional logs of the web search sessions, semi-controlled user interview, and quantitative measures of subjective relevance assessment. These transactional logs will be used to determine the interactive patterns and deduce the users web search strategies, which are to be confirmed by a semi-controlled user interview. The users will be interviewed about their satisfaction, more specifically on their subjective assessment of the relevance of the graphical and interactive presentation of search results. And, relative relevance and ranked-life relevance will be the quantitative measures to compare the WIRS performance (Borlund 2003).

At Connections 2011, we would like to present in details our research design, the methodological framework used and our preliminaries results.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: Addison-Wesley Longman.

Bertin, J., & Barbut, M. (1977). *Sémiologie Graphique: Les Diagrammes, Les Réseaux, Les Cartes* (2 éd.). Paris: Éditions de l'École des Hautes Études en Sciences Sociales.

Borgman, C. L. (1996). Why Are Online Catalogs Still Hard to Use? *Journal of the American Society for Information Science*, *47*(7), 493–503.

Borlund, P. (2003). The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research*, *8*(3). [online] http://informationr.net/ir/8-3/paper152.html

Broder, A. (2002). A Taxonomy of Web Search. *SIGIR Forum*, *36*(2), 3–10.

Burkhard, R. A., Andrienko, G., Andrienko, N., Dykes, J., Koutamanis, A., Kienreich, W., Phaal, R., et al. (2007). Visualization Summit 2007: Ten Research Goals for 2010. *Information Visualization*, *6*(3), 169–188.

Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Diego: Morgan Kaufmann.

Chen, C. (2004). *Information Visualization: Beyond the Horizon*. London: Springer.

Chen, C. (2005). Top 10 Unsolved Information Visualization Problems. *IEEE Computer Graphics and Applications*, 12–16.

Gwizdka, J. (2010). Distribution of Cognitive Load in Web Search. *Journal of the American Society for Information Science and Technology*, *61*(11), 2167-2187.

Ingwersen, P. (1996). Cognitive Perspectives of Information Retrieval Interaction : Elements of Cognitive IR Theory. *Journal of Documentation*, *52*(1), 3–50.

Ingwersen, P., & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht, Netherlands: Springer.

Keller, T., & Tergan, S. (2005). Visualizing Knowledge and Information: An Introduction. Dans *Knowledge and Information Visualization* (p. 1–23).

Kerren, A., Stasko, J. T., Fekete, J., & North, C. (2007). Workshop Report: Information Visualization-human-centered Issues in Visual Representation, Interaction, and Evaluation. *Information Visualization, 6*(3), 189–196.

Kuhlthau, C. C. (1991). Inside The Search Process: Information Seeking from The User's Perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371.

Kuhlthau, C. C. (2005). *Towards Collaboration between Information Seeking and Information Retrieval*. *Information Research, 10*(2), [online] <http://informationr.net/ir/10-2/paper225.html>.

Kuhlthau, C. C., Heinström, J., & Todd, R. J. (2008). The 'Information Search Process' Revisited: Is The Model Still Useful? *Information Research*, *13*(4), [online] <http://informationr.net/ir/13-4/paper355.html>.

Larkin, J. H., & Simon, H. A. (1987). Why A Diagram Is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*(1), 65-100.

Lin, X., Kerren, A., & Zhang, J. (2009). Challenges in human-centered information visualization: Introduction to the special issue. *Information Visualization*, *8*(3), 137–138.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York: Cambridge University Press.

Markey, K. (2007a). Twenty-five Years of End-user Searching, Part 1: Research findings. *Journal of the American Society for Information Science and Technology*, *58*(8), 1071–1081.

Markey, K. (2007b). Twenty-five Years of End-user Searching, Part 2: Future Research Directions. *Journal of the American Society for Information Science and Technology*, *58*(8), 1123-1130.

Norman, D. A. (1993). *Things That Make Us Smart*. Don Mills, Ont.: Addison Wesley.

Plaisant, C. (2004). The Challenge of Information Visualization Evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces - AVI '04* (p. 109). Gallipoli, Italy.

Resnikoff, H. L. (1989). *The Illusion of Reality*. New York: Springer-Verlag New York, Inc.

Rose, D. E., & Levinson, D. (2004). Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web* (p. 13–19). New York, NY, USA: ACM.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings IEEE Symposium on Visual Languages* (p. 336–343). Boulder, Colorado: IEEE Computer Society.

Shneiderman, B., & Plaisant, C. (2005). *Designing the User Interface: Strategies for Effective Human-Computer* (4 éd.). Boston: Pearson/Addison Wesley.

Toms, E. G., Freund, L., Kopak, R., & Bartlett, J. C. (2003). The Effect of Task Domain on Search. In *Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research* (p. 303-312). Présenté au IBM Centre for Advanced Studies Conference, Toronto, Ontario, Canada: IBM Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, Conn: Graphics Press.

Wilson, T. D. (1999). Models in Information Behaviour Research. *Journal of Documentation*, *55*(3), 249–270.

Zhang, J. (2008). *Visualization for Information Retrieval*. Berlin: Springer-Verlag.