Preprint of Huvila, Isto: Mining qualitative data on human information behaviour from the Web. Griesbaum, J.; Mandl, T. & Womser-Hacker, C. *(ed.) Information und Wissen: global, sozial und frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft (ISI 2011) Hildesheim, 9. bis 11. März 2011., Verlag Werner Hülsbusch,* 2011, 322-326.

Mining qualitative data on human information behaviour from the Web

Isto Huvila

Uppsala University
Dept of ALM, Thunbergsvägen 3H
SE-75126 Uppsala, Sweden
firstname.lastname@abm.uu.se

Abstract

This paper discusses an approach of collecting qualitative data on human information behaviour that is based on mining web data using search engines. The approach is technically the same that has been used for some time in webometric research to make statistical inferences on web data, but the present paper shows how the same tools and data collecting methods can be used to gather data for qualitative data analysis on human information behaviour.

Introduction

The theoretical and methodological variety of information behaviour (IB) research is considerable (e.g. Fisher et al., 2005; Case, 2002; Wilson, 2010). In spite of the wealth of theoretical frameworks and methods, the prevalent approaches tend to focus on broad surveys of large populations or focused indepth studies of small groups of people. Especially the qualitative approaches tend to be labour intensive both during data collection and analysis phases. Large quantitative require a relatively broad understanding

of the studied phenomenon before data collection. Surveys have persistent problem with response rates that pertains to both web based and other types of surveys (e.g. Bertot, 2009). The present article discusses an approach of collecting data on IB that is based on mining web data. The approach is technically the same that has been used for some time in webometric research (Thelwall et al., 2005) to make statistical inferences on web data, but the present paper shows how the same tools and data collecting methods can be used to gather data for qualitative data analysis on IB.

Mining data from the web

Webometric mining of web data is based on the fact that large amounts of data can be collected automatically using techniques like web crawling or by exploiting the application programming interfaces (API) provided by major search engines like Google, Bing and Yahoo (Thelwall et al., 2005). The methods used to collect quantifiable data for webometric research can also be used to collect qualitative data. The present study is based on two experiments made in November 2009 and 2010 using LexiURL Searcher software (Thelwall, 2009) that can be used to retrieve various types of research data from the major search engines using their respective APIs. The experiments were based on batch searching of lists of web pages that contain a selection of IB related utterances listed in Table 1. A second experiment with search engine related phrases gave similar results to the first one. Another set of data with phrases “I tried to Google but” (532 hits, 44 valid hits in 50 analysed phrases), “I tried to search on Yahoo” (19, 13/14), “I tried to search in” (803, 49/50), “I tried to search in/on the Internet” (178, 35/50), “I tried to look for” (943, 22/50), “I searched on/in Wikipedia” (168, 50/50) and “I searched on/in Youtube” (752, 50/50) were analysed with (from methodological point of view) comparable outcomes.

Utterance	Hits returned	IB related hits	Example of data
“i asked my friends but”	75	39	“I asked my friends [about language use], but they were like, "Young, we grew up in English. It’s so hard to explain." “

"is asked my [mum/mom] but"	55	31	"Um what's a prostitute? I asked my mum but she won't tell me."
"i asked my dad but"	116	93	"I asked my dad but he couldn't explain it so I could understand."

Table 1: Examples of analysed utterances related to unsuccessful social information seeking.

The web pages that contained the utterances were analysed using *content analysis* and *close reading* to map the variety of characteristics and patterns in the information seeking situations and their contexts. Finally, the utterances were classified using the *constant comparative method*. A full analysis of the data is presented in (Huvila, 2010). The column Hits returned in Table 1 indicates the number of retrieved web pages. The column IB related hits lists the number of web pages that in the analysis were found to contain information relevant from an IB point of view. The final column short examples of the type of information that can be retrieved using the discussed method.

Even if the analysed utterances are specific phrases related to question asking and web searching types of IB, it is obvious that the proposed approach may be used with any conceivable utterances present in web pages indexed by the search engines used for data collection. The searches can be made also using standard search engine user interfaces, but the LexiURL software helps to collect the results to a single file that facilitates the analysis of the material.

Discussion

According to the observations made during the experiments, the principal benefits of using Web data were 1) that all publicly available Web data is freely accessible for research purposes, 2) Web data is relatively easy to collect and 3) Web pages contain a large corpus of heterogeneous data from all over the world. There are, however, some evident limitations with the data collection method. The material is collected from the Web and is therefore likely to represent only a very biased sample of all possible information interactions. The specific phrases tend also to be common in particular types of web pages. The utterances analysed in the present study were common in discussion forums, question and answer (Q&A) services and blogs. Besides

its contextual specificity, another pertinent aspect of the harvested data is that the study population is limited to an unknown sample of information seekers. The utterances and their contexts contain only occasional and consequential evidence of the demographics of the studied population. The problem is similar to the difference in populations between Web based, telephone and postal surveys. Only web users (and with the present method, only contributors) are represented on the web and only those with a landline telephone are able to participate in random digit dialling telephone surveys (Bertot, 2009; Deutskens et al., 2004).

Because the data collection procedure tends to retrieve data that is unrelated to the intention of the researcher, the dataset needs to be cleaned up for exclusion of invalid data. In the two described experiments, the constant comparative method Glaser & Strauss (1967) seemed to result in a reasonably confident identification of valid and invalid cases. It is, of course, possible to use other validation methods including multiple indexer approaches (Foster et al., 2008) to increase confidence to the data. There are also some specific ethical considerations that pertain to the harvested data. The data is *de facto* publicly available on the Web. Because it was not originally published with a forthcoming IB research in mind, a special emphasis should be placed on a respectful use of the data and, if necessary, anonymisation of the individual cases.

In spite of its evident limitations, the proposed data collection method has several advantages. The limitations may be considered acceptable in qualitative studies aiming to map the variety of information interactions. Most of the sampling related problems (e.g. what is known of the total population, what is the context of the data) discovered during the experiments apply also to the conventional qualitative and quantitative approaches even if sometimes to a slightly lesser extent. Even if the two experiments showed that the specificity contextual evidence tends to vary case by case, the data was rich enough to make inferences on specific aspects of information interactions (e.g. the reasons of failed information seeking). Another strength of the proposed approach is that it may be used to complement other types of data collection methods as a part of a triangulation strategy. The method makes it also possible to study (theoretically) global or semi-global populations. At the same time it is possible to restrict the sample, for instance, by selecting the language of the

search phrases or by focussing on specific top-level domains of the searched web sites. Considering its limitations, the principal asset of the approach is, however, that data collection using LexiURL Searcher and similar tools is fast and easy. The low cost of acquiring data makes it possible to experiment with a large number of phrases. The acquired data can be analysed both qualitatively and statistically, and even if the contexts and richness of the data tend to be heterogeneous, the approach can provide rich contextual descriptions of IB.

Conclusions

Mining web data using search engines APIs provides a novel approach for collecting data for qualitative information behaviour research. The principal benefits of the method are that the data is freely accessible for research purposes, it is easy to collect and the amount of collectable data from all over the world is considerable. The method and especially the resulting data have, however, several limitations. The sample is unknown, individual contexts may be hard to characterise and the results are difficult to generalise. In spite of its limitations, the approach can effectively complement other data collection methods and especially, to provide data for qualitative exploratory analysis with an ambition to map a phenomenon rather than to achieve generalisable results.

References

Bertot, J. C. (2009). Web based surveys: Not your basic survey anymore. The Library Quarterly, 79(1), 119-124.

Case, D. O. (2002). Looking for information: A survey of research on information seeking, needs, and behaviors. San Diego: Academic Press.

Deutskens, E., Ruyter, K. de, Wetzels, Met al., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. Marketing Letters, 15(1), 21–36.

Fisher, K., Erdelez, S., & McKechnie, L. E. (Eds.). (2005). Theories of information behavior. Medford, NJ: Information Today.

Foster, A, Urquhart, C., & Turner, J. (2008). Validating coding for a theoretical model of information behaviour. Information Research, 13(4).

Huvila, I. (2010). “I asked my Mum, but” and other cases of unsuccessful information seeking by asking. Proceedings of the ISIC 2010. Murcia: University of Murcia, 179-191.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. ARIST, 39(1), 81–135.

Wilson, T. D. (2010). Fifty years of information behavior research. Bulletin of the ASIS&T, 36(3), 27–34.