doi:10.2498/cit.1001924

Applying Content Analysis to Web-based Content

Inhwa Kim and Jasna Kuljis

People and Interactivity Research Centre, Department of Information Systems and Computing, Brunel University, West London, United Kingdom

Using Content Analysis on Web-based content, in particular the content available on Web 2.0 sites, is investigated. The relative strengths and limitations of the method are described. To illustrate how content analysis may be used, we provide a brief overview of a case study that investigates cultural impacts on the use of design features with regard to self-disclosure on the blogs of South Korean and United Kingdom's users. In this study we took a standard approach to conducting the content analysis. Based on our experience in using content analysis, in that study we make several suggestions on the benefits of using content analysis and on how content analysis of the material from the Web can be improved.

Keywords: content analysis, web-based content, blogs, human computer interaction

1. Introduction

The expansion of the World Wide Web (Web) and, in particular, the second generation of websites, the so-called Web 2.0 technologies, has led to a vast amount of user-generated contents being created in various forms such as blogs, podcasts, wikis, twitters, etc. Such user-generated contents may provide unprecedented opportunities for some researchers if they can access and analyse this data available on the Web.

So, instead of investing a lot of time and energy in using more traditional methods for collecting data such as interviews, surveys and focus groups, a researcher may now be able to just download data from the Web without the need to engage with users. Of course, the available web-based data is not relevant to all disciplines. However, most studies concerned with attitudes, preferences, opinions, and behaviour of users, whether in social sciences, human computer interaction or in other disciplines, can benefit from the free Web content. It is this type of material that we are concerned with in this paper. This type of data comes in a variety of formats and is mostly unstructured, so for an analysis to be undertaken, its needs are particular. In this paper we look at one such appropriate approach called content analysis.

Content analysis is a widely used research method for objective, systematic and quantitative examination of communication content [4]. The method has been employed not only in the field of traditional communication [1], but also in studies of human-computer interaction such as web based applications, norms of behaviour and cultural values [17, 20]. It can be useful for discovering and gaining insights into users' preferences and behaviours as well as into complex social and communicational trends and patterns generated by users. However, applying content analysis to Web-based content faces many challenges such as sampling and coding. The complexity of the mix of various media characteristics within the Web content affects generalisability and representativeness.

The purpose of this paper is to introduce the content analysis technique and show the potential challenges posed when it is applied to Web-based content. To support the argument, use is made of a case study that analysed blogs in order to study possible cultural differences between South Korean and British users. Based on our experience in applying content analysis to blogs, we give some recommendations on how the content analysis of such data can be improved.

This paper is organised as follows. In Section 2 we introduce content analysis including its advantages and disadvantages as a technique. Section 3 considers issues regarding applying content analysis to Web-based content. Section 4 describes how content analysis was employed to analyse blogs in order to examine cultural differences of bloggers from South Korea and UK. Recommendations regarding the application of content analysis to Web-based content are given in Section 5 and conclusions in Section 6.

2. Content Analysis and its Advantages and Disadvantages

Holsti [9] provides a broad definition of content analysis as the application of scientific methods to documentary evidence. Similarly, Krippendorff defines content analysis as "a research technique for making replicable and valid inferences from data to their context" [13].

Content analysis enables the analysis of data to be structured and may be used in both qualitative and quantitative studies [16]. Typically based on an individual's perspective, qualitative content analysis is similar to textual analysis in that it is primarily interpretive in nature, and often does not utilise statistics for data analysis.

Quantitative content analysis on the other hand is a research technique used to make valid and reliable inferences from the data to their context [13].

Krippendorff [13] identifies several advantages of content analysis such as:

- It is unobtrusive.
- It is unstructured.
- It is context sensitive and able to cope with a large quantity of data.
- It examines the artefact (e.g. text, images) of communication itself and not the individual directly.

Such benefits attract researchers who want to investigate phenomena without their investigation influencing the procedure [7]. Therefore, the outcome may be less biased compared to other techniques such as questionnaire surveys, interviews, and projective tests [10]. Another benefit is that undertaking content analysis is fairly simple and economical compared to other techniques. This is in particular true if the necessary data is readily available, like in Web-based content. The content generated by users can be reached without having to engage with users. Even the large quantities of data can be considered as an advantage since it can be employed to examine trends and patterns of Web-based content [9].

Despite all of these benefits, content analysis has some limitations like any other method. The following can be considered as disadvantages:

- Content analytic studies are sometimes considered as being devoid of a theoretical basis since the focus is on what is measurable rather than on what is theoretically significant or important [5]. Therefore, the research design must take into account whether there is a relationship with frequency of occurrence.
- Although authors can provide a number of speculative answers to the questions, content analysis alone cannot give the answers. This limitation can be lessened if combined with another method, more appropriate to measuring those aspects (e.g. experiments, surveys, interview etc) [9].
- When applied to Web-based content, the changing content can be problematic. However, some researchers claim that it can be overcome by rapid data collection [15] and downloading websites [11].

3. Content Analysis Applied to Web-based Content

The Web is a complex and rich mixture of old and new technologies. Therefore, it provides many opportunities and challenges for researchers who apply content analysis to Web-based content. In particular, the complexity of new features such as mixed multiple media (text, graphics, animation, video and audio etc), interactivity, decentralised and hyperlinked structures, and its continuously evolving nature provide challenges to the development of valid descriptive categories, recording and sampling frames for the method.

Neuendorf [16] illustrates the process of content analysis through nine steps (see Figure 1).

The continuous change of the websites' content leads to potential problems with data collection. After the analysis of 19 studies that applied content analysis on the World Wide Web, McMillan [15] found that most studies conducted data collection in one to two months. The quickest data collection reported was two days and the longest was five months.

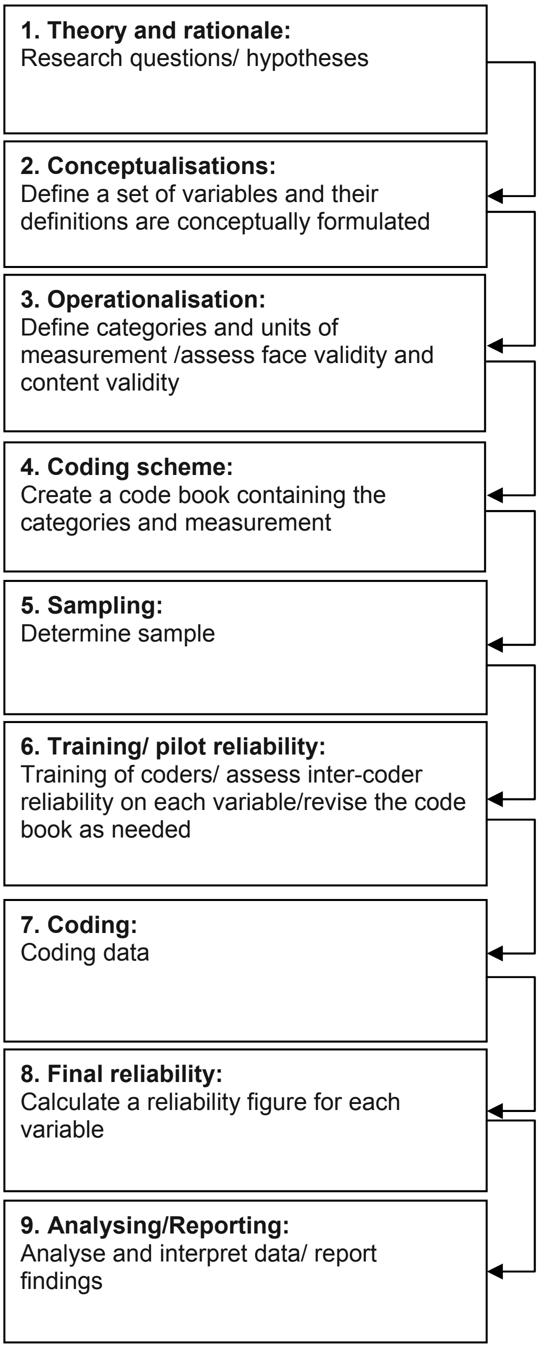
1. Theory and rationale:
Research questions/ hypotheses
2. Conceptualisations:
Define a set of variables and their definitions are conceptually formulated
3. Operationalisation:
Define categories and units of measurement /assess face validity and content validity
4. Coding scheme:
Create a code book containing the categories and measurement
5. Sampling:
Determine sample
6. Training/ pilot reliability:
Training of coders/ assess inter-coder reliability on each variable/revise the code book as needed
7. Coding:
Coding data
8. Final reliability:
Calculate a reliability figure for each variable
9. Analysing/Reporting:
Analyse and interpret data/ report findings

Figure 1. A flowchart for content analysis research adapted from Neuendorf [16].

It is important to identify the mechanism or coding scheme and categories because reliability is improved through validity of data [12]. Therefore, a careful operation of training coders and checking the reliability is of importance to overcome potential subjectivity.

Despite the challenge of applying content analysis to Web-based content, several studies have already been conducted. For example, Singh and Baack studied how cultural values are reflected in American and Mexican Websites by using content analysis [20]. The study of Callahan [6] examined cultural differences and similarities in the design of university websites using Hofstede's model of cultural dimensions [8], and found that there are correlations between graphical elements and Hofstede's index values, but these are statistically weaker than initially hypothesised.

In parallel with the Web developments, new tools that analyse Web content by automated computer programs were developed. Bauer and Scharl [3] introduced a software tool called WebAnalyzer, which automatically gathers and analyses parameters such as a site's HTML code and information about the site features, including the number of images and external links. It is well acknowledged that analysing entire websites by human coders is extremely difficult, mainly because many websites consist of thousands of pages. Therefore, researchers will have the advantage of using computer analysis content techniques to parse whole sites instead of just the home page as a unit of analysis [16].

4. Case Study: Content Analysis of Blogs

The purpose of this section is to exemplify the process of employing content analysis on Web-based content and to describe the related issues. We used content analysis to investigate whether there is any cultural impact on the design and use of the blogs. Since blogs are entirely created and managed by users themselves, we assume that their blogs then reflect their set of values and preferences that stem from their cultural background.

Cultural dimension	SK	UK
Power distance	High	Low
Individualism	Low	High
Masculinity	Low	High
Uncertainty avoidance	High	Low

Table 1. Cultural dimensions for SK and UK [8].

In our study we investigated a wide range of variables and design elements that might be affected by culture. However, to demonstrate how we used content analysis, we present only the part of the study that considered design features related to self-disclosure, as this is sufficient to illustrate the method. We compared design features related to self-disclosure on the blogs posted by users from South Korea and United Kingdom, two very different cultures according to Hofstede [8] as shown in Table 1. We analysed the actual content of the blogs.

In conducting the content analysis, in our study we followed the process proposed by Neuendorf [16] already described in Section 3. Since the concern of this paper is content analysis applied to Web content, we do not provide details of the study, but rather focus on the process itself. The following text briefly describes each step taken in our study.

Stage 1. Formulating research questions or hypotheses

People from a high uncertainty avoidance country are expected to have low tolerance for uncertainty and risk [8]. Disclosure of personal information might be considered as an example of risk-taking situation. South Korea is ranked as a high uncertainty avoidance country and the UK a low uncertainty avoidance country. We therefore expected British users to be more likely to reveal their personal information than South Korean users. Based on this observation, our hypothesis was that South Korean bloggers are less likely to disclose information about themselves than British bloggers.

Stage 2. Identifying variables

We identified variables related to the information that bloggers provided about themselves such as name, age, contact details, etc.

Stage 3. Defining categories and units of measurement

We needed to examine how bloggers identify themselves, so we indentified the author's identification information and design elements provided by the blog that allow an author to reveal or disclose their profile. These were presence or absence of design elements such as name, profile image, gender, age, location, occupation, hobby or interest, and contact link.

Defining the unit of analysis on Web-based content poses distinctive challenges due to the combined multiple media forms. Perhaps completely new context units are required to be developed [15]. In choosing which unit should be examined, we had to consider whether to analyse all the pages for each blog or just the 'home page' or opening screen of the website. We chose the profile page as the unit of analysis because it covers all the design elements we had to examine.

Stage 4. Creating coding scheme

A code book which contains the categories and their measurement was created.

Stage 5. Sampling

Potter [19] stressed the trouble of sampling because of the size and "chaotic design structure" of the Web (p.12). The sampling methods vary, depending on the specific research questions studied. Weare and Lin [21] described comprehensive sampling techniques to collect information such as Internet addresses, search engines, popular sites, randomly generated IP addresses and URLs. They argue that the use of multiple techniques may help to validate the samples drawn by one method or to ascertain the samples have been fully identified.

We already limited our study to blog sites, so the concern was which blog sites should then be selected to address our research hypothesis. We chose Blogger.com since it provides a worldwide service including South Korea and United Kingdom so that the choice of design elements by the users from the two countries can be directly compared within the same environment. We needed two sets of samples; data for South Korean and for British users. When the study was conducted, we could draw a sample of blogs created by users from South Korea and United Kingdom where the user's country is determined from the location registered in their blog. We randomly selected one hundred blogs owned by British users and one hundred blogs owned by South Korean users. To produce a coherent sample that would lead to optimally interpretable results, we did not include blogs that were affiliated with, or created by, a commercial organisation or other institution. We also did not include blogs that were inactive for more than one month. The profile page of the blogs was collected between 17 and 18 February 2008.

Stage 6. Training coders/pilot reliability

By using the codebook, two trained coders fluent in both Korean and English evaluated the sample blogs. Training sessions were used to reconcile the coding differences between the coders. Inter-coder reliability was established throughout the coder training process based on 20% ($n = 20$) of randomly selected blogs from each country from the sample. Inter-coder reliability of each coding category for each country was tested using Cohen's kappa (k) formula, which was overall above the acceptable indicator (higher than 0.75) [2].

Stage 7. Coding

Coding of the sample was processed independently, based on the code book.

Stage 8. Calculating final reliability

Using the same procedure as was used for testing the pilot reliability, intercoder reliability was tested again using Cohen's kappa (k) formula. The coder reliability of each coding category for each country was overall above the acceptable indicator (higher than 0.75) [2].

Stage 9. Data analysis

Data was analysed using κ^2 and ANOVA. The results indicated that British bloggers ($M = 4.94$) were overall more likely to reveal information about themselves than South Korean bloggers ($M = 4.86$), but this difference was not significant. The frequency of the occurrence of blogger occupation was significantly higher in the case of British blogs ($\kappa^2 = 10.00$, $p < .01$). South Korean blogs more often revealed age ($\kappa^2 = 6.610, p \leq .01$) and provided a contact link to the author than the British blogs ($\kappa^2 = 4.604, p < .05$). Therefore, the hypothesis was not supported. The result indicates that the culture may not have that much impact regarding the self-disclosure. However, when the study was conducted, there was a big difference in the number of blogs registered for each country. South Korean blogs were in a minority and therefore may not have been representative. We have therefore decided to select another sample for South Korean blogs, this time hosted on a popular South Korean hosting site. However, the results of that study are incomplete and will be reported on later.

In summary, we found that performing content analysis using the steps proposed by Nuendorf [16] was easy and useful. However, his framework does not provide guidance on how to conduct sampling of Web-based content nor how to avoid the effect of the dynamic nature of the web. We had to find out about these ourselves.

5. Issues Arising from the Study

The Web 2.0 technologies allow users to create their own contents mostly on the social networking sites leading to a huge amount of user-generated contents. Content analysis can be employed on such data in order to find out social and communicational trends and patterns as well as user's attitudes, preferences, and behaviours.

Our case study showed up some issues that need careful consideration and preparation in these situations. Firstly, sampling and sampling size pose some challenges. What units need to be identified for sampling will be determined by the research question or hypothesis. In our study, the sampling was fairly straightforward as we already restricted our investigation to blog sites and within two countries. But if one requires a more complex sample, the task may become much more complicated. Careful consideration has to be given to determine the appropriate sampling size as well as whether the sample is representative enough. Choosing an effective and efficient sampling size may save considerable unnecessary effort in analysing an enormous amount of data. Some studies selected sample size based on a certain time limit (e.g. ten consecutive days or six days in April) [14, 18]. However, we observed that the papers describing such studies do not always clearly state the rules that were used for sampling. To our knowledge, there are no sampling guidelines advising researchers on how to choose representative samples and the appropriate sampling size when examining Web-based content.

Secondly, although there are established tools (e.g. Cohen's kappa (κ), Holsti's method, Scott's Pi, etc.) that are used for checking inter-coder reliability, training coders must be done thoroughly, so there is no discrepancy in interpretations of data among different coders.

Thirdly, data collection also has potential problems. We downloaded the profile page of the blogs to get "frozen in time" versions using the software called LocalWebsite Archive to avoid possible change of the content. However, there may be a copyright issue although the profile information of the blogs is publicly available. Lastly, coders' biases must be carefully considered, especially when they are from different cultures. In our study, we took into account the coders' cultural biases that may have influenced coding; fluency in both 'languages' is not the equivalent of cultural fluency.

6. Conclusions

Regardless of its limitations, we found that applying content analysis to Web-based content is a relatively easy process that allows researchers to perform and prepare data at their convenience and to avoid lengthy ethics approval procedures. The method provides a rich opportunity to study users' styles, patterns or preferences that does not necessitate any researcher intervention. So we anticipate that more researchers will start investigating Web-based contents and that content analysis will become a method of choice in such studies.

A further benefit is illustrated by the study outcome, where the expected result (a cultural difference between users) was expected to be discernible from the way users used the design features on their blogs. However, not only was there no expected result, but the content analysis suggested that there was no difference at all. There are not many methods of analysis that will contradict the researcher's hypotheses so directly, if appropriate. So, in conclusion, it appears from the research reported on here, that Content Analysis of web content is likely to help a researcher change research direction when it is made abundantly clear that the current research direction is barren.

References

[1] F.S. AL-OLAYAN, K. KARANDE, A Content Analysis of Magazine Advertisements from the United States and the Arab World. *Journal of Advertising*, 29(3) (2000), pp. 69–82.

[2] M. BANERJEE ET AL., Beyond Kappa: A Review of Inter-rater Agreement Measures. *The Canadian Journal of Statistics*, 27(1) (1999), pp. 3–23.

[3] C. BAUER, A. SCHARL, Quantitative Evaluation of Web Site Content and Structure. *Internet Research: Electronic Networking Applications and Policy*, 10 (2000), pp. 31–43.

[4] B. BERELSON, *Content Analysis in Communication Research*. Free Press, New York, 1952.

[5] A. BRYMAN, E. BELL, *Business Research Methods*. Oxford University Press, New York, 2007.

[6] E. CALLAHAN, Cultural Similarities and Differences in the Design of University Web Sites. *Journal of Computer-mediated Communication*, 11 (2005), pp. 239–273.

[7] T. HARWOOD, T. GARRY, An Overview of Content Analysis. *The Marketing Review*, 3 (2003), pp. 479–498.

[8] G. HOFSTEDE, *Cultures and Organisations: Software of the Mind: Intercultural Cooperation and its Importance for Survival*. McGraw Hill, New York, 1991.

[9] O.R. HOLSTI, *Content Analysis for the Social Sciences and Humanities, Reading*. Addison-Wesley Publishing, MA, 1969.

[10] T.C. KINNEAR, J.R. TAYLOR, *Marketing Research: An Applied Approach*. 4th ed, McGraw-Hill Inc, London, 1991.

[11] W. KOEHLER, An Analysis of Web Page and Web Site Constancy and Performance. *Journal of the American Society for Information Science*, 50(2) (1999), pp. 162–180.

[12] R.H. KOLBE, M.S. BURNETT, Content Analysis Research: An Examination of Applications with Directives for Improvement Research Reliability and Objectivity. *Journal of Consumer Research*, 18(Sept) (1991), pp. 243–250.

[13] K. KRIPPENDORFF, *Content analysis: An Introduction to its Methodology*. Sage Publications, London, 1980.

[14] X. LI, Web Page Design and Graphic Use of Three U.S. Newspapers. *Journalism & Mass Communication Quarterly*, 75(Summer) (1998), pp. 353–365.

[15] S.J. MCMILLAN, The Microscope and the Moving Target: The Challenge of Applying Content Analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*, 77(1) (2000), pp. 80–98.

[16] K.A. NEUENDORF, *The Content Analysis Guidebook*, Sage Publications, London, 2002.

[17] S. OKAZAKI, J.A. RIVAS, A Content Analysis of Multinationals' Web Communication Strategies: Cross-cultural Research Framework and Pre-testing. *Internet Research: Electronic Networking Applications and Policy*, 12(5) (2002), pp. 380–390.

[18] K. PASHUPATI, J.H. LEE, Web Banner Ads in Online Newspapers: A Cross-National Comparison of India and Korea. *International Journal of Advertising*, 22 (2003), pp. 531–564.

[19] R.F. POTTER, Measuring the "Bells & Whistles" of a New Medium: Using Content Analysis to Describe Structural Features of Cyberspace. In *Proc. of 49th Annual Conference of the International Communication Association*, (1999), San Francisco, CA.

[20] N. SINGH, D.W. BAACK, Web Site Adaptation: A Cross-cultural Comparison of U.S. and Mexican Web sites. *Journal of Computer-mediated Communication*, 9(4) (2004).

[21] C. WEARE, W.Y. LIN, Content Analysis of the World Wide Web: Opportunities and Challenges. *Social Science Computer Review*, 18(272) (2002).

Received: June, 2010
Accepted: November, 2010

Contact addresses:

Inhwa Kim
People and Interactivity Research Centre
Department of Information Systems and Computing
Brunel University, West London
Uxbridge, Middlesex UB8 3PH
United Kingdom
e-mail: `Inhwa.Kim@brunel.ac.uk`

Jasna Kuljis
People and Interactivity Research Centre
Department of Information Systems and Computing
Brunel University, West London
Uxbridge, Middlesex UB8 3PH
United Kingdom
e-mail: `Jasna.Kuljis@brunel.ac.uk`

INHWA KIM is a PhD candidate in the Department of Information Systems and Computing at Brunel University, UK. She works as a full-time employee at Samsung SDS Europe in UK. Her research interests are in cross-cultural website design, web 2.0 technology and persuasive technology. She is researching into the manifestations of cultural differences between the United Kingdom and South Korea on the design of the web pages and blogs. Miss Kim's email address is `inhwa.kim@brunel.ac.uk`.

JASNA KULJIS is a Professor and Head of the Department of Information Systems and Computing at Brunel University, UK. She gained her Dipl. Ing. degree in theoretical mathematics from the University of Zagreb, Croatia. She gained her M.S. in information science from the University of Pittsburgh, USA, and a Ph.D. in information systems from the London School of Economics, University of London, UK. Her current research is in human computer interfaces. She is most interested in the design of graphical user interfaces and in the development of new paradigms that would further enhance the usability of interactive computer systems. She is Director of the People and Interactivity Research Centre at Brunel University. Dr. Kuljis' email address is `jasna.kuljis@brunel.ac.uk`.