Specificity and Exhaustivity of Bibliographic Classifications – A Cross-cultural comparison with Text Analytic Approach

Inkyung Choi[1] and Min Sook Park[2]

[1] University of Illinois at Urbana-Champaign, Champaign IL 61820, USA
inkyungc@illinois.edu
[2] University of Wisconsin Milwaukee, Milwaukee WI 53210, USA
minsook@uwm.edu

Abstract. The current study aims to detect sociocultural differences implied in the classification systems, employing text analytic techniques. By comparing Korean Decimal Classification (KDC) and Dewey Decimal Classification (DDC), this study probes the gaps in exhaustivity and specificity in the two classification systems developed in distant social and cultural contexts. A computer-aided quantitative approach in cross-cultural comparison of Knowledge Organization Systems (KOSs) is a relatively new attempt. Besides the finding of the study will demonstrate how to utilize classification as a large textual data set and an automated classification reading.

Keywords: Bibliographic Classification, Text Analytics, Cross-cultural Comparison.

1 Introduction

Knowledge organization systems (KOSs), such as classification systems, have the common primary goal of linking users to knowledge resources to satisfy the users' needs [19]. To achieve this, there must be sufficient commonality between the concept expressed in a KOS and the real-world object to which that concept refers [1, 22]. However, KOSs can still have variations in complexity and structure. The common characteristics of KOSs that reflect such possibilities are the following: (a) the KOS imposes a worldview on a collection and the items in it, and (b) the same entity can be characterized in different ways depending on the KOS used. In this sense, classification systems are not free from social and cultural influences [1, 23]. A classification system is a human creation and bears the imprints of its progenitors. Consequently, terminologies and classification schemes are inevitably tied to the purpose, culture, and context [9].

The embodiment of culture is more likely to be implicit and achieve its purpose when a user's semantic understanding parallels the classification structure of the system. A body of research focuses on revealing the cultural contexts embodied in classification systems. However, most of these studies focused on subjects that were of

particular interest to the researchers. This study attempts to employ computer-aided applications to classification system comparison. This relatively new quantitative process using text analytics enables us to probe the differences in terminologies and relationships in entire classification systems while also providing a bird's-eye view. As such, this study aims to explore the sociocultural differences implied in the KDC and the DDC.

2 Previous studies

In theoretical discussions of cross-cultural library classifications, Mai [14] acknowledged challenges in maintaining exhaustive and specific subject representations when the information is exported from one country to another. It occurs because subject representations are dependent on the purpose as well as the cultural and contextual circumstances in which the representation is produced. Choi's [6] study also discovered the inferred differences between cultural contexts in the Korean Decimal Classification (KDC) and the Dewey Decimal Classification (DDC) systems. Visualization of the quantity of subjects in the two classifications—social science and technology—revealed the inferred differences from their cultural context. The differences imply gaps in both exhaustivity and specificity [5]. These studies shed light on cross-cultural comparative approaches to understanding sociocultural influences in classification systems.

Diverse approaches, such as subject ontogeny [20] and discourse analysis [8], were applied to uncover cultural influences in classification systems. Recent digitalization of classification systems and advanced data analytic techniques have enabled researchers to expand their scope in exploring topical structures that embody cultural contexts underneath them. However, only a handful of studies have taken a quantitative approach. For instance, Salah et al. [17] applied a quantitative approach to examine classification systems that changed over time. They investigated and illustrated changes in the degree of complexity and composition of the Universal Decimal Classification (UDC) by counting UDC numbers. The researchers, who used relatively large amounts of data, presupposed that UDC numbers reflected the rules of classificatory structures properly. Smiraglia et al. [18] also suggested that using a quantitative approach and visualization permitted observation of changes in classification in terms of size, composition, growth, and distribution. These studies have provided useful insights for quantitative methods and related visualizations and can be used to compare different instances of classification systems. However, there is still a lack of research regarding comparison of semantic components of classification systems and the relationships among the terminologies in the systems.

We believe employing a quantitative approach using data-analytic techniques could assist researchers better probe sociocultural implications on classification systems, otherwise could be undetected. We focus in particular on probing the differences in specificity and exhaustivity. Specificity refers to the exactness with which you place a subject in the hierarchy of a classification scheme [4, 16]. Exhaustivity refers to the extent to which various parts of a compound subject are acknowledged,

indicated by the number of topics represented in a document [4, 16]. These theoretical measurement parameters in classification systems are often considered the two most important notions that measure the effectiveness of indexing and terminology sources. We also focus on the linguistic entities in a classification scheme (e.g., terminologies) that may depend on political, cultural, and moral contexts. The structure of the classification, such as hierarchy, also results from the cultural and intellectual infrastructures [15]. Findings from this in-progress study could provide insight for increasing interoperability in cross-cultural classification uses. In addition, this investigation may demonstrate how to utilize classification as a large textual data set and an automated classification reading, which would enable large-scale cross-cultural comparisons.

3 Methods

We will conduct cross-cultural analysis, which refers to a comparison of various sociological or cultural factors to assess the similarities and diversity in two or more cultures or societies [18]. The study will compare subject headings in the most recent editions of the KDC and the DDC: the sixth edition of the KDC and the 23rd edition of the DDC. The sixth edition of the KDC contains a total of 13,862 classification numbers, and the 23rd edition of the DDC contains 28,273 classification numbers and associated subject headings. However, not all classification numbers are assigned subject headings.

To compare the two classification systems thoroughly, the study will employ text-mining techniques. Text mining refers to the computer-supported process of extracting meaningful knowledge from unstructured textual documents [3, 13] as well as discovering underlying semantic structures between lexical entities (e.g., words; [7, 12]). This computer-aided approach to textual data is particularly useful when applied to different languages. This cross-language mining application enables us to explore similarities and differences in semantic topics in different languages and to use the knowledge or corpora in one language or another [2].

Once the terminologies in the KDC and the DDC are extracted from the data sets, they will be imported to the open text mining software KH coder3 [10] to compute and visualize distribution patterns of subject headings. Terminologies from the KDC and the DDC are strings associated with numeric characteristics. We will take the hierarchical structures of formal classification schemes into account for the data processes. For example, the highest hierarchical level of the decimal classification is the class level, which is represented in units of 100 (e.g. 100, 200, 300). There are 10 total classification numbers at the class level for one classification system. The second hierarchical level is division level, which is represented in units of 10 (e.g. 110, 120, 130). There are 100 total classification numbers at the division level. The same classification numbers at different levels have slightly different captions. For example, “000” at the class level is “computer science, information and general works,” but “computer science, knowledge and systems” at division level. The third is the section level, represented in units of one (e.g. 111, 112, 113). There are 1,000 total classification numbers at the section level. All classification numbers with decimal points are

listed under the section level classification numbers and form a hierarchy. The set of terminologies at each level will be processed separately based on their levels. This process will be repeated for different levels of terminologies in the two classification systems.

To gain insight about unknown variations in the two classification systems, this study will use mining algorithms for term frequency (TF) and co-occurrences. To probe specificity of a class in the KDC and the DDC, TF will be computed and compared across different levels of classes. TF indicates how frequently a term occurs in a corpus [7]. The number and variety of extracted terms will be used as indices of exhaustivity in a specific class and major topics in the current study. Comparing the two classification systems, we choose the "Technology" and "Social Science" main classes, as these examples were the cases bearing the socio-cultural influences in classificatory features (Choi, 2017). In the technology subject, the DDC found to have 6,940 subject headings whereas the KDC has 5,023 subject headings. This means the DDC found to have more details with additional 1,917 subject headings that describes technologies. TF of the extracted subject headings are compared at the section number levels in cross-tabulation tables to demonstrate differences in specificity.

Co-occurrence refers to associations of concepts or concept patterns found together in documents in a collection that reflect an underlying relationship and reveal strongly related concepts within the set of documents [3, 11]. In this study, co-occurrence techniques will be also deployed to further explore latent relationships between terminologies in the same or different levels both in the KDC and the DDC. For example, the classification numbers of "Technology" and "Social Science" at the section level and subject headings corresponding to the section levels and the lower were visualized to show distributions of subjects which accounts for exhaustivity of a certain topic addressed in the classification classes.

4 Conclusion

A bibliographic classification system is a complex language designed to organize resources in a systematic way to provide users access to information. Its complexity becomes even richer when the system evolves to be reflective of social and cultural contexts, which makes the comparison of classification systems in different languages and cultures challenging. The current study, by using a cross-language mining application, will assist in understanding the complexity by examining linguistic components of classification systems from both semantic and lexical perspectives. This suggested computer-aided approach not only takes cross-cultural and cross-language comparative studies forward in methodological advances, but it also enables utilization of large-scale classification.

References

1. Abbas, J.: Structures for organizing knowledge: exploring taxonomies, ontologies, and other schemas. Neal-Schuman, New York (2010).

2. Aggarwal, C. C., & Zhai, C.: Mining text data. Springer-Verlag, New York (2012).
3. Blake, C.: Text mining. Annual review of information science and technology 45(1), 121-155 (2011).
4. Broughton, V.: Essential classification: Facet Publishing, London (2015).
5. Choi, I.: Visualizations of cross-cultural bibliographic classification: comparative studies of the Korean Decimal Classification and the Dewey Decimal Classification. In: 6th North American Symposium on Knowledge Organization, pp. 39-55 (2017).
6. Choi, I.: Toward a Model of Intercultural Warrant: A Case of the Korean Decimal Classification's Cross-cultural Adaptation of the Dewey Decimal Classification. The University of Wisconsin-Milwaukee, Milwaukee (2018).
7. Feldman, R., & Sanger, J.: The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press, Cambridge, England (2007).
8. Fox, M. J.: Gender as an 'interplay of rules': Detecting epistemic interplay of medical and legal discourse with sex and gender classification in four editions of the dewey decimal classification. The University of Wisconsin -Milwaukee, Milwaukee (2015).
9. Gartner, R.: Metadata: shaping knowledge from antiquity to the semantic web. Library and Information History 33(2), 149-150 (2016).
10. Higuchi, K.: KH coder, http://khcoder.net/en/, last accessed 2018/09/19
11. Hotho, A., Nürnberger, A., & Paaß, G. A brief survey of text mining. In: Paper presented at the Ldv Forum (2005).
12. Khoo, C. S., & Na, J. C.: Semantic relations in information science. Annual review of information science and technology 40(1), 157-228 (2006).
13. Kumar, L., & Bhatia, P. K.: Text mining: concepts, process and applications. Journal of Global Research in Computer Science, 4(3), 36-39 (2013).
14. Mai, J.-E.: The future of general classification. In: Knowledge Organization and Classification in International Information Retrieval, pp. 21-82 (2013).
15. Olson, H.: Social influences on classification. In: Encyclopedia of library and information sciences, pp. 4204-4211 (2010).
16. Olson, H. A., Boll, J. J., & Aluri, R.: Subject analysis in online catalogs. Libraries Unlimited, California (2001).
17. Salah, A. A., Gao, C., Suchecki, K., Scharnhorst, A., & Smiraglia, R. P.: The evolution of classification systems: Ontogeny of the UDC. In: Advances in Knowledge Organization 13, pp. 51-57 (2012).
18. Smiraglia, R., Scharnhorst, A., Salah, A. A., & Gao, C.: UDC in Action. In: Classification and Visualization: Interfaces to Knowledge, Proceedings of the International UDC Seminar, pp. 259-72 (2013).
19. Svenonius, E.: The intellectual foundation of information organization. MIT press, Cambridge (2000).
20. Tennis, J.: Subject ontogeny: subject access through time and the dimensionality of classification. In: Proceedings of the Seventh International ISKO Conference, pp. 54-59 (2002).
21. U.S. National Library of Medicine.: Cross-Cultural Comparison. In MeSH (2018).
22. Weller, K.: Knowledge representation in the social semantic web: Walter de Gruyter, Berlin (2010).
23. Zollers, A.: Emerging motivations for tagging: Expression, performance, and activism. In: Workshop on Tagging and Metadata for Social Information Organization at the 16th International World Wide Web Conference (2007).