Thoughts on exploiting instability in lattices for assessing the discrimination adequacy of a taxonomy

Antony K Cooper*, Derrick G Kourie, and Serena Coetzee

Department of Computer Science, University of Pretoria, Pretoria, South Africa
acooperatcsir.co.za,dkourieatcs.up.ac.za,scoetzeeatcs.up.ac.za
http://www.cs.up.ac.za/

Abstract. Conventionally in formal concept analysis (FCA), concept stability is preferred in the lattice, because instability (i.e. low stability) represents noise that clouds the analysis of the data.
High stability means there are many objects with the same intent or many attributes with the same extent, which could be interpreted as redundant or absent objects or attributes. The differences between redundancy or absence need to be assessed quantitatively, a process that could be described as *stability exploration*. We have used FCA to analyse different taxonomies for user-generated content. For example, redundancy amongst attributes represents taxonomy classes unable to differentiate adequately the objects being classified. Absent attributes, redundant objects and absent objects can have various implications. Hence, instability in a lattice is desirable for some types of analysis.

Keywords: formal concept analysis, stability, taxonomy

1 Background on user generated content

User-generated content (UGC) in general, and *volunteered geographical information (VGI)* in particular, are becoming more important as sources for official data bases, such as those used in national *spatial data infrastructures (SDIs)*. An SDI is an evolving concept about facilitating and coordinating the exchange and sharing of spatial data and services between various stakeholders [1].

While traditional sources of official data are well understood, the same does not apply to UGC. It is interpreted in different ways, and one woman's UGC could be another man's professionally generated content. Several attempts have been made to understand UGC and the contributors of UGC, by developing taxonomies for aspects of UGC in general (eg: [2, 3]), or VGI in particular (eg: [4, 5]). VGI examples VGI include *OpenStreetMap*, a free, editable map of the world [6]; citizen-science projects such as the *Second South African Bird Atlas Project*

* Corresponding author. Current address: Built Environment Unit, CSIR, PO Box 395, Pretoria, 0001, South Africa

(SABAP2) [7]; in-car navigation systems allowing users to submit corrections; and geocoded photographs on virtual globes such as *Google Earth* [8].

We conducted an assessment amongst some geographical information professionals of their perceptions of virtual globes, VGI and SDIs [9], and we are in the process of developing a taxonomy of VGI, which we are modelling formally. We are using *formal concept analysis (FCA)* [10] to assess the characteristics of existing taxonomies of UGC, such as their discrimination adequacy. The intention is to improve the understanding of UGC, in respect of, for example, assessing UGC quality or catering for VGI in a metadata standard.

FCA uses a lattice of concepts with objects and attributes, and the linkages between them. We use the standard FCA terminology and notation: a *context* is written as: $\mathbb{K} := (G, M, I)$, where G is a set of *objects* and M a set of *attributes*. I is the binary relation between the sets of objects and attributes: $I \subseteq (G \bigtimes M)$. (A, B) represents a concept whose *extent* is $A \subseteq G$ and whose *intent* is $B \subseteq M$. For (A, B) to be a formal concept, B must contain all those attributes that the objects in A have in common, and only those attributes: denoted by $A' = B$. Conversely, A must also contain all those objects that share the attributes B, and only those objects: denoted by $B' = A$. See [10–15].

2 Stability in a lattice

The *intensional stability* of a concept indicates how much its intent depends on individual objects in the extent. It is a measure of the likelihood that removing a random set of objects from the concept's extent would change its intent. Similarly, the *extensional stability* indicates how much the extent depends on individual attributes in the intent. Formally, [15] defines the *intensional stability index*, σ_i, and *extensional stability index*, σ_e, of concept (A, B):

$$\sigma_i((A, B)) = \frac{|\ \{C \subseteq A \mid C' = B\}\ |}{2^{|A|}}$$

$$\sigma_e((A, B)) = \frac{|\ \{D \subseteq B \mid D' = A\}\ |}{2^{|B|}}$$

Each concept (A, B) has $|A|$ objects in its extent. The intentional stability index is the proportion of the $2^{|A|}$ subsets of A which have the following property: the attributes C' shared by the objects in any such subset, say C, correspond to the concept's intent, that is, $C' = B$. In a lattice built with objects in C instead of A, there will be a concept (C, B), and in this sense the intent, B, of concept (A, B) is "stable". The notion of extensional stability is similar, but with the roles of extents and intents reversed.

The more objects (or attributes) covered by a formal concept, the more likely it will be intensionally (or extensionally) stable, because of the greater likelihood of "redundant" objects (or attributes). Figure 1 shows a very stable lattice, because a lattice built from any subset of attributes (objects) would yield a concept whose extent (intent, respectively) is unchanged from that of the concept in Figure 1. That is, $\sigma_i(A, B) \approx 1$ and $\sigma_e(A, B) \approx 1$. While this

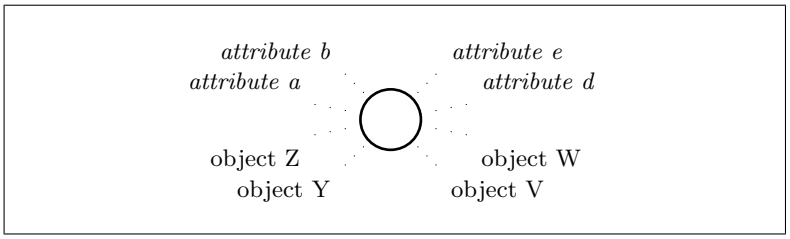

Fig. 1. A very stable, but rather boring, lattice.

is an extreme example, it illustrates why stability can mean redundancy and why it can be considered "boring" in some applications [16], because of the low information content. We appreciate that for machine learning (e.g. [16]), concepts with high stability indicate robust input data with little noise.

We have not used FCA to classify data, but to assess the *discrimination adequacy* of taxonomies for UGC [2, 3, 5, 4]. The classes in these taxonomies are the attributes for FCA, e.g. [3] provides classes for copyright issues: *User-authored content*, *User-derived content*, *User-copied content* and *Peer-to-peer as UGC*. The FCA objects are repositories of UGC, such as *in-car navigation* or an *open repository* [5].

We have used Concept Explorer (ConExp) [17][1] for FCA, because it is open-source, robust and used by our colleagues (e.g. [16, 19]), and hence has a pool of expertise readily available. ConExp provides *attribute exploration*, an interactive process to see if each *implication* (set of "linked" attributes) can also apply to objects not in the context of the implication. Questions are asked about dependencies between different attributes (i.e. the exploration), and if a dependency does not hold, the user has to provide a counterexample (effectively, add a new object) [17]. This can reveal "absent" and "redundant" attributes and objects.

3 Absent and redundant attributes and objects

FCA is applied here to determine the adequacy of taxonomies for discriminating between repositories containing UGC in general, or VGI in particular. We lack the space to provide detailed examples here, but will do so at the conference. Instead, we provide a theoretical example in Figure 2, showing absent and redundant attributes and objects. If a concept has objects, they are shown below the node and if a concept has attributes, they are shown above the node. In terms of *reduced labelling*, each concept inherits objects from its extent and attributes from its intent. The following subsections refer to Figure 2.

3.1 High intensional stability

Concept $A = (\{Obj1, Obj2, Obj3, Obj4, Obj5\}, \emptyset)$ has high intensional stability, with $\sigma_i(A) = 0.84$. Many (27) of the 32 subsets of A's extent yield a concept

[1] Note: ConExp's author requests that users cite his Russian text, [18].

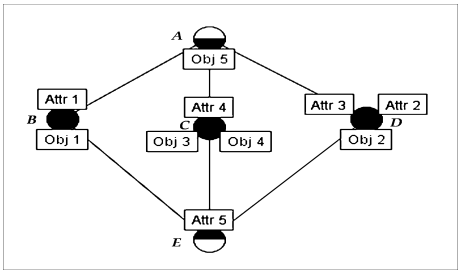

Fig. 2. Absent and redundant attributes and objects.

whose intent is also $\emptyset$. The object *Obj5* is not described or differentiated by any attributes. For our analysis, this could be addressed by adding classes to a taxonomy, so that it can differentiate better between the repositories. The taxonomy in [3] extends the taxonomy in [2] (namely, *Distribution platform* and *Type*), to cater for copyright issues. Without the classes of [3], the taxonomy of [2] does not really differentiate repositories from one another.

Concept $C = (\{Obj3, Obj4\}, \{Attr4\})$ also has relatively high intensional stability, with $\sigma_i(C) = 0.75$. The objects *Obj3* and *Obj4* are not differentiated from one another by the attributes. Effectively, $Obj3 = Obj4$ and one of them is redundant. In a comprehensive analysis of UGC repositories, one would expect this, namely repositories that are equivalent and hence direct competitors of one another. For example, referring to Figure 3, the objects shown are generic and there could be many repositories that are specific instances of each. Adding these repositories as objects would create redundancies in the taxonomy.

3.2 High extensional stability

Concept $E = (\emptyset, \{Attr1, Attr2, Attr3, Attr4, Attr5\})$ has high extensional stability, with $\sigma_i(E) = 0.84$. Again, 27 of the 32 subsets of E's intent yield a concept whose extent is also $\emptyset$. The attribute *Attr5* does not describe or differentiate any objects. This could be a weakness in the analysis, with an important type of repository omitted, or it could indicate a type of repository that does not yet exist and hence a potential "gap" in the market. While experimenting with FCA and the taxonomy of [4], we realised the value of instability in a lattice. It highlighted a potential "gap" in the market, namely repositories that do not cater adequately for privacy: a widespread problem on the Internet.

Concept D also has relatively high extensional stability, with $\sigma_e(D) = 0.75$. No objects are differentiated from one another by the attributes *Attr2* and *Attr3*. Effectively, $Attr2 = Attr3$ and one of them is redundant. This could be coincidental, could reflect a set of objects that is too narrow (eg: other types of repositories should also have been included), or could indicate that some classes should be removed from the taxonomy because they add no value or even worse, could cause confusion as users try to differentiate between classes that are, in

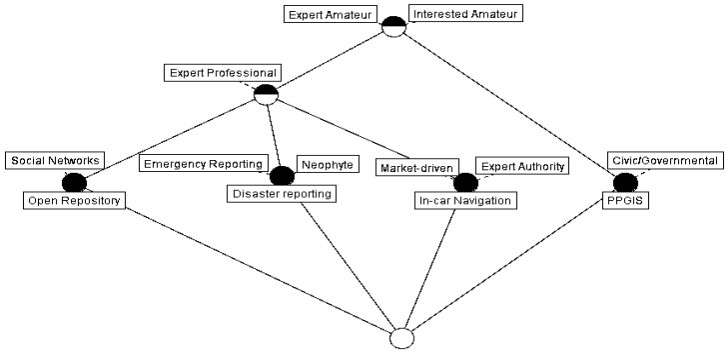

Fig. 3. A subset of the taxonomy of [5], for assessing the nature and motivation of *produsers*.

essence, equivalent. We illustrate this in Figure 3 with a subset of the VGI taxonomy from [5], for assessing the nature and motivation of *produsers* (users who are also producers). For the objects, we use the generic examples of VGI repositories given by [5]. As can be seen, there are several redundancies in the attributes, because these classes are inadequately defined, or cannot be differentiated in practice, or other types of repositories should be included in this analysis.

3.3 Stability exploration

While the approach outlined above differentiates situations of absent attributes / redundant objects from those of absent objects/redundant attributes, it does not necessarily do so within these two groupings. This is because qualitative analysis is probably required to differentiate between absent attributes and redundant objects, and between absent objects and redundant attributes. We would suggest that when assessing the discrimination adequacy of a taxonomy, the absence or redundancy of objects or attributes is undesirable. Such a taxonomy could be improved by reducing the intensional and/or extensional stability. It appears that this could be done in a manner similar to attribute exploration [17], starting with the concept with the highest stability and then moving on to the next highest. This process could be termed *stability exploration*.

4 Conclusions

We are using FCA to assess the adequacy of taxonomies in discriminating between different types of UGC repositories. In contrast to the usual FCA applications, we have shown that instability can have value for analysis. High intensional stability reveals missing classes from a taxonomy, or redundancy amongst the repositories. High extensional stability reveals missing repositories or gaps in

the market, or taxonomy classes that are redundant. We are investigating how stability exploration could be implemented to guide a user to reduce stability. Future work could involve assessing the taxonomies in detail or taxonomies in other domains, such as bloodstain pattern analysis [20].

We would like to acknowledge the fruitful discussions we have had with our colleagues and the very insightful comments of the anonymous referees.

References

1. Hjelmager, J., Moellering, H., Delgado, T., Cooper, A.K., Rajabifard, A., Rapant, P., Danko, D., Huet, M., Laurent, D., Aalders, H.J.G.L., Iwaniak, A., Abad, P., Düren, U., Martynenko, A.: An initial Formal Model for Spatial Data Infrastructures. Int J Geogr Inf Sci, 22(11), pp 1295–1309 (2008)
2. Wunsch-Vincent, S., Vickery, G.: Participative Web: User-Created Content. OECD report DSTI/ICCP/IE(2006)7/FINAL. Working Party on the Information Economy of the Committee for Information, Computer and Communications Policy (2007)
3. Gervais, D.: The Tangled Web of UGC: Making Copyright Sense of User-Generated Content. Vanderbilt JETL, 11(4), pp 841–870 (2009)
4. Budhathoki, N.R., Nedovic-Budic, Z., Bruce, B.: An interdisciplinary frame for understanding volunteered geographic information. Geomatica, 64(1), pp 11–26 (2010)
5. Coleman, D.J., Georgiadou, Y., Labonte, J.: Volunteered Geographic Information: The Nature and Motivation of Produsers. Int J of SDI Res, 4 (2009)
6. OpenStreetMap: The Free Wiki World Map. http://www.openstreetmap.org/
7. Southern African Bird Atlas Project 2. http://sabap2.adu.org.za/
8. Google Earth: Explore, Search, and Discover. http://earth.google.com/
9. Cooper, A.K., Coetzee, S., Kourie, D.G.: Perceptions of virtual globes, volunteered geographical information and spatial data infrastructures. Geomatica, 64(1), pp 333–348 (2010)
10. Wille, R.: Restructuring lattice theory: An approach based on hierarchies of concepts. In: Ordered sets, Rival, I. (ed), pp 445–470, D Reidel Publishing Co (1982)
11. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis. Preprints. 14pp (1997)
12. Carpineto, C., Romano, G.: Concept Data Analysis: Theory and Applications. John Wiley & Sons, Ltd (2004)
13. Priss, U.: Formal concept analysis in information science. Annu Rev Inform Sci, 40, pp 521–543 (2006)
14. Kuznetsov, S.O.: On stability of a formal concept. Ann Math Artif Intel, 49(1–4), pp 101–115 (2007)
15. Klimushkin, M., Obiedkov, S., Roth, C.: Approaches to the Selection of Relevant Concepts in the Case of Noisy Data. In: ICFCA 2010, Agadir, Morocco, pp 255–266, Springer (2010)
16. Kourie, D.G., Oosthuizen, G.D.: Lattices in machine learning: Complexity issues. Acta Informatica, 35, pp 269–292 (1998)
17. Yevtushenko, S., Kaiser, T., Tane, J.: Concept Explorer The User Guide. (2003)
18. Yevtushenko, S.A.: System of data analysis "Concept Explorer". In: Proceedings of the 7th National Conference on Artificial Intelligence KII-2000, Russia (2000)
19. Chan, K.S.M.: Formal Methods for Web Services: A Taxonomic Approach. In: ICSE'10, Cape Town, South Africa, 2, pp 357–360, ACM (2010)
20. Cooper, A.K. Thoughts on categorising bloodstain patterns. Technical Report 0442-0001-701-A1, CSIR (2003)