Proceedings Paper:

Abolkasim, E, Lau, L orcid.org/0000-0003-1062-9059, Mitrovic, A et al. (1 more author) (2018) Ontology-based Domain Diversity Profiling of User Comments. In: Lecture Notes in Computer Science 10948. 19th International Conference, AIED 2018 – International Conference on Artificial Intelligence in Education, 27-30 Jun 2018, London. Springer Nature , pp. 3-8. ISBN 9783319938455

https://doi.org/10.1007/978-3-319-93846-2_1

Ontology-based Domain Diversity Profiling of User Comments

Entisar Abolkasim[1], Lydia Lau[1], Antonija Mitrovic[2] and Vania Dimitrova[1]

[1] School of Computing, University of Leeds, UK
[2] Intelligent Computer Tutoring Group, University of Canterbury, New Zealand

Abstract. Diversity has been the subject of study in various disciplines from biology to social science and computing. Respecting and utilising the diversity of the population is increasingly important to broadening knowledge. This paper describes a pipeline for diversity profiling of a pool of text in order to understand its coverage of an underpinning domain. The application is illustrated by using a domain ontology on presentation skills in a case study with 38 postgraduates who made comments while learning pitch presentations with the Active Video Watching system (AVW-Space). The outcome shows different patterns of coverage on the domain by the comments in each of the eight videos.

Keywords: diversity analytics, ontology, semantic techniques, video-based learning, active video watching, soft skills learning

1 Introduction

Despite the growing importance of modelling diversity, there is limited computational work on user-generated content in learning context. Despotakis et al. [4] developed a semantic framework for modelling viewpoints embedded within user comments in social platforms to understand learner engagement with situational simulations for soft skills learning [5]. Hecking et al. [10] adapted network-text analysis of learner-generated comments to capture divergence, convergence and (dis-)continuity in textual comments to characterise types of learner behaviour when engaging with videos. In both approaches, domain differences across learner groups were studied; visualisations to illustrate engagement with learning content. While visualisations were used to reveal interesting patterns, their adoption for automated profiling is not feasible.

This paper presents a novel computational approach to automatically detect the domain coverage in user comments by deriving diversity profiles for the learning objects. Our work adapts an established diversity framework developed in social science [14]. The domain knowledge is represented by an ontology [7,8]. By using semantic techniques and metrics for diversity measurement, the coverage of concepts from the interaction between the learner and the learning material is explored. Such an approach can provide useful insights for learning environment designers and, most importantly, it generates diversity profiles that can be used to broaden learner's domain knowledge.

2 Diversity Properties: Variety, Balance and Disparity

Stirling's diversity framework [14] with three basic properties (variety, balance and disparity) is used. A more thorough discussion of the model can be found in [1]. **Domain diversity** of the comments is the focus here. An ontology is used to represent the domain. Each comment is linked to a set of ontology entities (e.g. through semantic tagging). Given an ontology representing the domain **Ω**, a pool of comments linked to a set of entities **E** from **Ω**, and a class in the ontology taxonomy **T** providing the entry category for which diversity will be calculated as follows.

Variety is the number of sub-categories of T which have at least one entity from E (i.e. mentioned in the user comments). The higher the number, the higher the diversity.

$$Variety(\Omega, E, T) = |K| \tag{1}$$

where K is set of sub-categories for T, i.e. ontology classes that are sub-classes of T; each of the classes in K has at least one entity in E.

Balance calculates how much of each sub-class of T is covered by the user comments, using the set of entities E. The formula is based on Shannon Entropy Index [13]; we use only the number of distinct ontology entities covered by the comments. The higher the number, the better the domain coverage; higher diversity.

$$Balance(\Omega, E, T) = \frac{1}{n}\sum_{i=1}^{n} p_i ln(p_i) \tag{2}$$

where n is the number of sub-categories of T and p_i is the proportion of distinct entities in E that belong to the taxonomy headed by the sub-category against the total number of entities in that taxonomy.

Disparity is the manner and degree to which the elements may be distinguished. We consider within disparity, which is calculated by measuring each sub-category's dispersion, i.e. how scattered/dispersed the entities from E that belong to each sub-category of T are. The formula uses Hall-Ball internal cluster validation index [2], which gives the mean dispersion across all the sub-categories that are covered by the comments. The higher the number, the higher the disparity.

$$Disparity(\Omega, E, T) = \frac{1}{n}\sum_{i=1}^{n} dis(c_i) \tag{3}$$

where n is the number of sub-categories c_i of T. Dispersion $dis(c_i)$ is the shortest path between each of the entities in E that belong to sub-category c_i and the medoid of c_i.

3 The Ontology-based Diversity Profiling Pipeline

The steps in the pipeline are: (i) Input preparation for the **Se**mantic-**D**riven **D**iversity **A**nalytics **T**ool SeDDAT [1]. This includes getting the domain ontology **Ω** and annotate the comments with the ontology entities E. (ii) SeDDAT execution to calculate the diversity properties for specified pool of comments. (iii) Analysis to interpret the diversity characteristics. (iv) When potential interesting patterns are spotted, the analyst can specify the next level to run SeDDAT for more fine-tuned diversity profiles.

4 Case Study – Comments on Videos for Learning

Earlier work in AVW-Space indicated that constructive learners who wrote comments while watching videos were more likely to increase their knowledge [11,12]. This paper explores the relationship between diversity patterns and video usage for learning. In particular: How do the learner comments on the videos cover the domain, and are there any diversity patterns that can be related to the learning of presentation skills?

Videos used. Four tutorials (T1-T4) on presentations and four examples (E1-E4) (two TED talks and two 3-minute PhD pitch presentations) [6].

Participants. The comments were collected from 38 postgraduate students in a study conducted in March 2016. These students can be classified as constructive according to the ICAP framework [3].

4.1 The Domain Ontology – PreSOn (Presentation Skills Ontology)

A semi-automatic ontology engineering approach [10] was used to convert an initial taxonomy (extracted from surveys in study) into an ontology, then extend with the Body Language Ontology from [4]. Refinement was made after 3 presentation skills trainers (from the University of Leeds) inspected the initial ontology. This resulted in a Presentation Skills Ontology (PreSOn) with four top level categories – three related to core presentation skills (Structure, Delivery, VisualAid) and one to include terms describing quality of presentations (Presentation Attribute). The next level of core subcategories and number of entities in each are as follow: (i) Stucture: StructureApproach (5) and StructureComponent (62); (ii) Delivery: SpeakerEmotion (10), SpeakerAura (5), AudienceEmotion (1), VerbalCommunication (18), NonVerbalCommunication (55) and Preparation (10); (iii) VisualAid: VisualAidArtefact (71) and VisualAidDevice (22).

4.2 Domain Diversity for Videos

Domain variety patterns. When the entry point was ‘Thing’, all videos had variety 4 (i.e. all top categories were covered). We created profiles using each of the main categories of presentation skills (Delivery, Structure, VisualAid) as the entry point. An example of outcome: Video E4 has the highest variety for Delivery ‘5’ out of a maximum ‘6’ – this gives an indication that this video may be useful to learn about delivery.

Domain balance patterns. Top half of **Fig. 1** shows the overall balance and the proportions of the top four categories for every video. Overall, we can identify T1 as a good tutorial for VisualAid, T2 and T3 for Structure, and T4 for Delivery. For E1-E4, concepts relating to Structure seems to be most readily noticed in the comments.

Domain disparity patterns. Disparity indicates the spread of entities within a category – the higher the disparity, the broader the coverage, while low disparity indicates concentration on a small area. The bottom half of **Fig. 1** shows the dispersion of all the videos at the level Thing. Drilling down one level: Delivery scores are consistently highest (except for T1), whilst Structure tends to be lowest – i.e. a broad range of domain entities related to Delivery whereas comments on Structure are concentrated on the narrow area around StructureComponent (opening, body, closing).

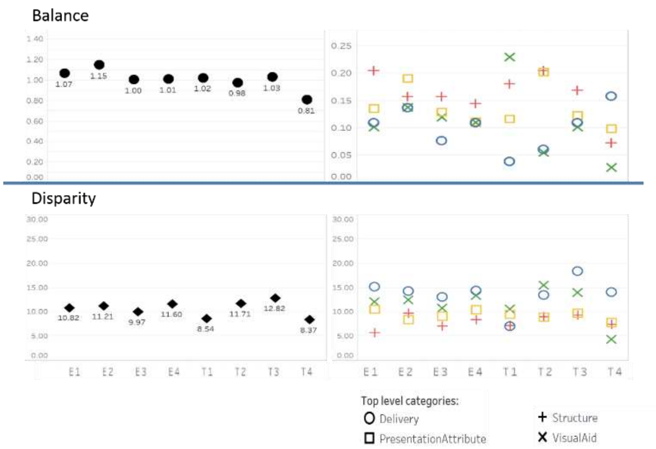

Fig. 1. Domain balance and domain disparity for Thing (left) and the proportions of the four categories (right).

Combining domain balance and domain disparity. Separately, balance and disparity yield interesting patterns for some videos only. **Table 1** summarises the possible patterns and their interpretations by combining them to provide further insight into the usefulness of the videos for informal learning.

Table 1. Possible interpretations of a combination of balance and disparity

	Low balance	**High balance**
High disparity	Lack of focus	A good diverse coverage
Low disparity	A niche or poor coverage	A good focus

Looking at the tutorial videos on the right of **Fig. 1**, several observations can be made: T1 has a good focus on VisualAid and T3 on Structure (high balance and low disparity); T2 somewhat lacks focus on Delivery (low balance and high disparity); T4 has a diverse coverage on Delivery (high balance and disparity). Observations can be made for examples, though not as prominent as in tutorials: E1 has a niche/poor coverage on VisualAid (low balance and relatively low disparity); E2 lacks focus on Delivery (low balance and high disparity; E3 gives a good focus on structure (low disparity and relatively high balance); and E4 does not show any pattern.

5 Conclusion

The contribution of this paper is a novel computational approach for detecting domain coverage automatically. The demonstrated benefit of diversity profiling for understanding learner engagement with videos indicates that the approach can be applied to other scenarios of video-based learning (e.g. MOOCs, flipped classroom, informal learning). Modelling diversity is especially valuable in soft skills learning, where contextual awareness and understanding of different perspectives is crucial.

References

1. Abolkasim, E., Lau, L., & Dimitrova, V. (2016). A Semantic-Driven Model for Ranking Digital Learning Objects Based on Diversity in the User Comments. In European Conference on Technology Enhanced Learning, Springer, Cham, 3-15.
2. Ball, G. H., & Hall, D. J. (1965). ISODATA, a Novel Method of Data Analysis and Pattern Classification. Menlo Park: Stanford Research Institute, (AD699616).
3. Chi, M. T., Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219-243.
4. Despotakis, D., (2013a). Modelling viewpoints in user generated content. PhD Dissertation. University of Leeds.
5. Despotakis, D., Dimitrova, V., Lau, L., Thakker, D., Ascolese, A., & Pannese, L. (2013b). ViewS in user generated content for enriching learning environments: A semantic sensing approach. In International Conference on Artificial Intelligence in Education, Springer, Berlin, Heidelberg, 121-130.
6. Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using learning analytics to devise interactive personalised nudges for active video watching. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization . ACM, 22-31.
7. Gruber, T. (1993) A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5, 199–220.
8. Gruber, T. (2008) Collective Knowledge Systems: Where the Social Web meets the Semantic Web. Semantic Web and Web 2.0, 6(1), 4–13.
9. Hecking, T., Dimitrova, V., Mitrovic, A., & Ulrich Hoppe, U. (2017). Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching. In ICCE 2017 Main Conference Proceedings. Asia-Pacific Society for Computers in Education. 326-335.
10. Hecking, T., Dimitrova, V., Mitrovic, A. & Hoppe, U. Using Network-Text analysis to characterise learner engagement in active video watching. In: Chen W. et al. (Eds), Proc. 25th Int. Conf. Computers in Education ICCE 2017, Christchurch, 4-9 December 2017, pp. 326-335. Asia-Pacific Society for Computers in Education.
11. Lau, L., Mitrovic, A., Weerasinghe, A., & Dimitrova, V. (2016). Usability of an active video watching system for soft skills training. In Proceedings of the 1st International Workshop on Intelligent Mentoring Systems, ITS 2016. https://imsworkshop.files.wordpress.com/2016/03/ims2016_paper_7_lau.pdf .
12. Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting constructive video-based learning: requirements elicitation from exploratory studies. In International Conference on Artificial Intelligence in Education . Springer, Cham. 224-237.
13. Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27, 379–423.
14. Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface, 4(15), 707-719.