Semantic Approach to Model Diversity in a Social Cloud

Entisar Nassr Abdulati Abolkasim

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Computing

January 2019

Dedication

To my family

Declarations

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Some of the work in this thesis has been published prior to thesis submission.

The following paper contributes to some of the work in chapter 3 in section Formal Model for Diversity in a Social Cloud. It also provided preliminary findings for chapter 4 in the section Domain Diversity Profiling.

Abolkasim E., Lau L., Dimitrova V.: A Semantic-Driven Model for Ranking Digital Learning Objects Based on Diversity in the User Comments. In: proceedings of the 11th European Conference on Technology Enhanced Learning (EC-TEL 2016).

The following paper contributes to some of the work in chapter 5 in sections related to Presentation Skills Domain Ontology and Domain Profile for Videos.

Abolkasim E., Lau L., Mitrovic A., Dimitrova V.: Ontology-Based Domain Diversity Profiling of User Comments. In: proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018).

The following paper contributes to some of the work in chapter 3 in the section related to the Proposed Approach Overview, and chapter 5 in section Domain Profiling for Users.

Abolkasim E., Lau L., Dimitrova V., Mitrovic A.: Diversity Profiling of Learners to Understand Their Domain Coverage While Watching Videos. In: proceedings of 13th European Conference on Technology Enhanced Learning (EC-TEL 2018).

The candidate confirms that the above jointly-authored publications are primarily the work of the first author. The role of the other authors was mostly editorial and supervisory.

Acknowledgments

First and foremost, I praise my God, Allah, who supported me throughout my PhD journey.

My deepest appreciation and gratitude go to my supervisors Dr Lydia Lau and Dr Vania Dimitrova for all their support and advice in my academic and sometimes personal life. Thank you for sharing your knowledge and personal experiences. Your guidance helped me to pursue my studies and research.

I have no words to describe my gratitude for the support I received from my mother. Mum, your encouragement, love and compassion made me who I am today. I love you forever and ever.

My deepest love and thanks to my brother Ahmed. Thank you for your support my dear.

I would like to express my gratitude and love to my husband Ahmad and beautiful daughters Hala, Toqa and Ruba for their love and patience during my PhD journey. I will make it up to you. I promise.

Warm thanks to my fellow PhD students in the School of Computing, and particularly to my dearest friend Mashael, Ahlam, Amira, Lolo, Areej, Deema, Diana, Marwan, Alicja, Ibrahim and Aziz. My PhD experience would not be the same without you.

To my friends, Naima, Monica and Suzan, many thanks for your support and help during my stay her in the UK.

I would like to thank School of Computing staff, especially Mrs. Judi, Mrs. Gaynor and Dr Richard. Thank you for your smiles and will to help.

Abstract

Understanding diversity is important in our inclusive society to hedge against ignorance and accommodate plural perspectives. Diversity nowadays can be observed in online social spaces. People from different backgrounds (e.g. gender, age, culture, expertise) are interacting every day around online digital objects (e.g. videos, images and web articles) leaving their social content in different format, commonly as textual comments and profiles. The *social clouds* around digital objects (i.e. user comments, user profiles and other metadata of digital objects) offer rich source of information about the users and their perspectives on different domains. Although, researchers from disparate disciplines have been working on understanding and measuring diversity from different perspectives, little has been done to automatically measure diversity in social clouds. This is the main objective of this research. This research proposes a semantic driven computational model to systematically represent and automatically measure diversity in a social cloud. Definitions from a prominent diversity framework and Semantic Web techniques underpin the proposed model. Diversity is measured based on four diversity indices - variety, balance, coverage and (within and across) disparity with regards to two perspectives – (a) domain, which is captured in user comments and represented by domain ontologies, and (b) user, which is captured in profiles of users who made the comments and represented by a proposed User Diversity Ontology. The proposed model is operationalised resulting in a Semantic Driven Diversity Analytics Tool (SeDDAT), which is responsible for diversity profiling based on the diversity indices. The proposed approach of applying the model is illustrated on social clouds from two social spaces - open (YouTube) and closed (Active Video Watching (AVW-Space)). The open social cloud shows the applicability of the model to generate diversity profiles of a large pool of videos (600) with thousands of users and comments. Closed social clouds of two user groups around same set of videos illustrate transferability and further utility of the model. A list of possible diversity patterns within social clouds is provided, which in turn deepen the understanding of diversity and open doors for further utilities of the diversity profiles. The proposed model is applicable in similar scenarios, such as in the social clouds around MOOCs and news articles.

Table of Content

List of Tables

List of Figures

Conventions

In this thesis, categories (sub-categories) and classes (subclasses) are used interchangeably to refer to the same thing, which is a concept in an ontology usually under a given entry point within the ontology.

Diversity in social clouds and social clouds diversity are used interchangeably to refer to the same thing.

When disparity is mentioned with no discrimination of type, it means both of its types for this research - within disparity and across disparity.

Chapter 1 : Introduction

1.1 Motivation

“In the last decades diversity and its management has become a feature of modern and postmodern organizations” [1]. Our globalised and networked world increased our interactions with people who are different on various attributes, such as age, gender, culture and expertise. As “diversity is all about *difference* and *inclusion*”[1], there has been an increasing awareness of the importance of understanding diversity and adapting to diverse individuals and groups in disparate disciplines. This became essential for “hedging against ignorance” and “accommodating plural perspectives” [2]. Recently, in the Interactions Magazine, an article titled “Diversity Computing” emphasised the importance of embracing diversity and avoiding “normative ordering”. Authors, who are from diverse specialities - psychology, philosophy and engineering, explained that diversity computing builds on denying that one group represent the norm against which others are measured. For that, they proposed a vision where they urge researchers from multidisciplinary fields to work hand to hand to better understand diversity and adopt to diverse individuals in our inclusive society.

Diverse individuals bring diverse knowledge, perspectives and expertise. Psychological theory suggests that stereotypes and biases tend to reduce the cognitive resources required to interact and understand new people and contexts[3]. On the other hand, a diverse crowd is argued to be a wise crowd [4] and research shows that diverse crowds outperform experts in solving problems as they exhibit wider range of knowledge, skills and abilities than homogeneous groups [5]–[7]. Such wisdom or what is referred to as collective intelligence can be valuable for different purposes, such as learning.

Nowadays, social media mediated the interactions of diverse individuals and groups. Diverse crowds are available and engaging on online social spaces, where users from different (e.g. cultural, educational, and professional) backgrounds interact with digital objects (e.g. videos, images, web news articles) on these social spaces. They leave their social interactions around these digital objects, commonly in textual comments and associated user profiles.

These interactions around online digital objects form, as coined in this research, a *social cloud* (i.e. user comments, user profiles and other metadata associated with those digital objects) that can facilitate the understanding and measurements of diversity. For instance, the users' comments can serve as a proxy of their diverse knowledge, perspectives, expertise and real life experiences on a subject domain as argued by the recently finished European Project ImReal[8]. In other words, the social cloud offers rich source of information about the users, the digital objects and users' perspectives and knowledge on different subject domains, which in turn can facilitate understanding and measuring diversity of the users and their domain exposure/coverage.

The understanding and measurements of diversity in social clouds can be useful for variety of purposes and applications. Consider the following two example scenarios in the learning and news domains respectively:

Scenario 1: The enormous user generated content on online social spaces offer opportunities for learning, but also come with challenges. For example, for a learner it is challenging to find the "right" content [9].

Videos are considered one of the main resources for learning - formal or informal [10]. For example, YouTube is one of the widely used platforms by tutors and learners for learning. It was ranked the second most popular social resource that has been used for informal learning by students[11], [12]. User comments on these videos provide access to diverse perspectives on the topic discussed in videos, where other users can learn vicariously by reading these comments. One of the challenges that faces the learners and tutors is the enormous number of videos available in social spaces (e.g. 300 hours of video are uploaded to YouTube every minute[1]). Finding the right learning videos can be time consuming, especially if the learner is seeking knowledge in ill-defined domains, such as culture or body language. Assuming the aim is providing a diverse coverage of a subject domain in order to diversify the learners' perspectives, the question here is: which videos will trigger the learner to notice diverse aspects about the domain? Also, it can be beneficial to know how diverse is the crowd around the videos i.e. how diverse are the users who wrote the comments around the videos? One useful approach for a learner or a tutor is to have a pool of videos (e.g. YouTube

[1] https://www.omnicoreagency.com/youtube-statistics/

videos) characterised based on the level of diversity in the social clouds around these videos. This helps to identify which videos trigger diverse domain coverage and diverse users to comment on them.

Scenario 2: Nowadays, majority of people are reading news online[13], where many news websites or social media spaces allow the readers (e.g. via social media accounts like Facebook profiles) to interact with the news article and leave their comments about the articles' content, so who sees what?

Consider an online news article about a crucial political event like election of the United States' president. It will be interesting to know whether the readers are talking about diverse aspects related to this event or just focussing on a particular one aspect. Also, it will be interesting to know who the readers are and how diverse (e.g. with regards to age groups, educational level and cultural background). An indication of the level of diversity in the comments can assist an inspection of the topics covered and the nature of this coverage. Same applies for the readers. The level of their diversity can show, for example, whether certain age groups are more interested in this event or it is an across-generation event. It will be interesting to link the readers' diversity to their perspectives' diversity for further explorations (e.g. whether a certain aspect is discussed in the comments by certain readers due to their cultural background) and patterns detection (e.g. who is the dominant cultural group that is driving the discussion in the comments).

Measuring diversity in a social cloud is a multifaceted process and little has been done to automatically and systematically tackle this issue. There is a need for a computational model that offers a systematic representation and automatic measurement of a social cloud diversity. This is the main objective of this research.

Research shows that Semantic Web techniques offer a great potential to explore social clouds for diversity representation and measurement.

"… social media streams pose a number of new challenges, due to their large-scale, short, noisy, context dependent, and dynamic nature". … "Semantic technologies have the potential to help people cope better with social media-induced information overload" Bontcheva & Rout (2014) [14]

Furthermore, diversity has been investigated in various disciplines, such as management, cultural diversity studies and social science, resulting in rich theoretical

resources that can serve as the backbone for diversity measurements and understanding.

This research aims to integrate findings from the multidisciplinary diversity research as well as Semantic Web techniques for the proposal of a semantic driven approach for modelling diversity in social clouds. The research questions and methodology to solve them are presented next.

1.2 Research Questions and Methodology

This research aims at exploring the user textual comments, associated user profiles and other related metadata around online digital objects to measure diversity (*diversity profiling*) in a social cloud. This is conducted by answering the following two research questions:

RQ1: Identifying and computing diversity metrics: *How to compute metrics for measuring diversity in a social cloud around digital objects?*

RQ2: Detecting diversity patterns: *What diversity patterns can be detected in a social cloud?*

To answer the research questions, an iterative approach was conducted. Initially, via utilising finding from the diversity-related research and Semantic Web techniques, a formal diversity model was established. Then, the formal model was operationalised to be applied on different social clouds. A case study was designed to test the applicability of the model and extend it with related components and identified diversity patterns. After that, another case study was conducted to test the applicability of the model with another social cloud with different domain and different user attributes i.e. test its transferability. The model was extended once more with findings from the second case study. The phases and detailed steps associated with this iterative approach are as follows:

I. **Define computational diversity model:** This involves number of steps utilising the multidisciplinary diversity research on understanding and measuring diversity as well as Semantic Web techniques. These steps are as follows: (**a**) identify a generic diversity framework to underpin the proposed computational model; (**b**) identify suitable diversity perspectives that are captured in social clouds; (**c**) identify suitable theories and/or models that assist the exploration of

the identified diversity perspectives; (**d**) identify suitable diversity indices for measuring diversity with regards to the identified diversity perspectives; and (**e**) identify suitable methods for connecting/linking the diversity perspectives for measurements and patterns detection.

II. **Evaluate Applicability of the proposed model:** This is informed by Semantic Web techniques and it requires to (**a**) operationalise the computational model to measure social clouds diversity. This is to implement a Semantic Driven Diversity Analytics Tool (SeDDAT) that facilitate measuring diversity for the diversity perspectives i.e. generate diversity profiles for the perspectives; (**b**) collect/use a social cloud for diversity measurements (case study one); (**c**) identify suitable Semantic Web techniques - tools and ontologies to underpin the process of understanding and measuring diversity. Propose and implement ontologies if no suitable ones exist; (**d**) identify suitable profiling levels to generate diversity profiles of the collected social cloud (e.g. profile social cloud per digital object or per a collection/pool of digital objects); (**e**) identify suitable entry points in the used ontology for the diversity profiling (i.e. utilise all or a branch of selected ontologies); (**f**) identify suitable methods for analysing and detecting any possible diversity patterns and update/extend the proposed model accordingly.

III. **Evaluate transferability of the proposed model:** This requires to (**a**) collect/use other social clouds (case study two); (**b**) identify suitable ontologies for the new social clouds. Propose and implement ontologies if no suitable ones exist; (**c**) identify suitable profiling levels and ontology entry points to generate diversity profiles of the collected social cloud; (**d**) apply and detect diversity patterns within the new social clouds and update/extend the model if required.

1.3 Contributions

Answering the above-mentioned research questions contributes to *three main research areas*:

I. **Computational models of diversity**. This research adds to the ongoing research in different disciplines on understanding and measuring diversity. This research proposes a computational model that facilitates a systematic representation and an automatic measurement of diversity in social clouds,

where for a given diversity perspective and an underpinning ontology, diversity profiles can be generated. Lists of possible levels for diversity profiling and diversity patterns within social clouds are provided, which enable further insights into the diversity of social clouds. Moreover, this research provided a refinement for one of the diversity indices by providing an index for measuring coverage against ontologies. The model is applicable with variety of similar social clouds and extendible to cater for all different contents within social clouds.

II. **Diversity profiling and patterns for learning**. The diversity profiles and patterns can be useful for personalisation and adaptation in the domain of learning. Both case studies are selected and designed within this domain, where they illustrate possible utilities of the profiles. This adds to and enforces the vision of technology enhanced learning (TEL). The diversity profiles and detected patterns can be utilised for personalised recommendations to support learners. The learners' diversity profiles facilitate better understanding of their diversity based on different attributes (e.g. their cultural background), and based on their domain knowledge and exposure, which in turn can assist the identification of potential and limitations. Also, research shows that social influence like social comparison of individuals or groups can be useful for motivation and encouragements to enhance learners' performance[15], [16]. Identification of diverse learners (e.g. learners with diverse domain coverage) and allowing other learners to come in contact with this diversity (e.g. read comments of the domain diversified learner's) can encourage them (e.g. nudge them to notice diverse aspects about the domain).

III. **Application of semantic technologies.** This research supports the potential provided by Semantic Web techniques. A step-by-step semantic approach and formal model that pave the way for modelling diversity in social clouds are proposed. Underpinned by Semantic Web techniques, the model is operationalised resulting in a Semantic Driven Diversity Analytics Tool (SeDDAT), which is responsible for generating diversity profiles that can be used to indicate diversity levels in social clouds. This tool illustrates the potential of semantic technologies for modelling diversity. Also, two ontologies have been implemented during this research. A User Diversity Ontology is proposed to facilitate the representations of user diversity attributes, such as

demographics (i.e. surface-level attributes), educational level (i.e. knowledge attributes) and other hidden characteristics (i.e. deep-level attributes), such as values, beliefs and attitudes. This ontology was extended twice for each case study used with this research. The ontology is proposed to complement the model, yet the model can still work if different ontology for the user is to be used. The ontology is available to be used for relevant research. Also, a domain ontology, the Presentation Skills Ontology (PeSOn), was extended and implemented for this research and made available for related research. Both ontologies and SeDDAT (with sample input files and instructions to use it) are available from here[2].

1.4 Thesis Structure

The remaining of this thesis is structured as follows:

Chapter two. Positions the conducted research in four main areas including (a) *social clouds* - definition, types, content and example applications; (b) the multidisciplinary *diversity research*, where selected theoretical studies are adopted as the backbone of the proposed computational diversity model; (c) related work on *social cloud diversity* highlighting potential and limitations (d) Semantic Web techniques for measuring and analysing diversity in social clouds

Chapter three. Proposes the semantic approach for modelling diversity in social clouds, where a computational model to represent and measure diversity is introduced. Possible diversity profiling and diversity patterns are discussed. It also introduces the operationalised model, the Semantic Driven Diversity Analytics Tool (SeDDAT), highlighting its input, preparation and processes.

Chapter four. Illustrates the applicability of the proposed model. It discusses the application of the proposed semantic approach on an open social cloud collected from an open online social space (YouTube) presenting results for two diversity perspectives as well as detected patterns.

Chapter five. Illustrates transferability of the proposed model. It discusses the application of the semantic approach on other social clouds collected from a closed

[2] University of Leeds Repository at: https://doi.org/10.5518/560

online social space - Active Video Watching (AVW-Space) to explore different domain, richer user profiles and different ontologies.

Chapter six. Introduces this research conclusions highlighting the contributions, limitations and possible future directions.

Chapter 2 : Literature Review

There are four main areas of research that this thesis builds upon: (**a**) *social clouds* to serve as the source for social content for diversity measurements and explorations; (**b**) *diversity research* highlighting related work from diversity research literature to theoretically underpin the proposed model for diversity understanding and measurement in a social cloud; (**c**) *social cloud diversity* to highlight related work that attempted understanding and/or measuring diversity; and (**d**) *semantic web techniques* with a focus on ontologies to assist the measurements and analysis of the social cloud diversity. This chapter reviews these areas as follows.

2.1 Social Clouds

For this research, a **social cloud** is a term that is coined to refer to a collection of digital objects (e.g. image, video, web article, etc.) uploaded onto any social media platform for user consumption, the user comments written on those digital objects in online social spaces, the profiles of users who wrote the comments, and any metadata associated with the digital objects. For example, a video on YouTube is uploaded to serve a certain purpose (e.g. educational) and it triggers users, who have accounts with YouTube with associated profiles, to write comments about the content of this video. The video has some metadata, such as a title and a unique URI. The video with associated metadata, the comments, the user profiles form a social cloud.

Social media, in particular user generated content (UGC), is the main source of social cloud collection. This section discusses the rationale behind using social media to represent and measure diversity, content and possible types of social clouds (open and closed social clouds), and applications with social clouds.

2.1.1 Social Media and User Generated Content

Social media penetrated people's lives creating a networked and globalised society. The ease of access and availability of these online social spaces allowed users from various age generations, professions, cultures etc. [9] to generate content and socially interact with different digital objects, such as videos, posts, images and web articles for different purposes such as learning[17], [18]. UGC denotes any form of content, such as blogs, discussion forums, posts, comments, tweets and other forms of media that was created by users in an online social space[19]. As a result, an enormous, rich

and diverse pool of content containing the opinions, experiences and expertise of those users is created every single day. For example, 30 million users are active daily on YouTube and 50 billion videos have been created since the start of YouTube (statistics last updated June 2018)[3].

This pool of socially generated content is considered to be a fertile source of innovations and genius[20]. Researchers consider this content as “collective intelligence” or “wisdom of crowds”. These are phrases used to describe the value created by the user (UGC) on the social web [21]. The French philosopher and media scholar Pierre Lévy, who conceived the term collective intelligence back in 1994, defines it as “form of universally distributed intelligence, constantly enhanced, coordinated in real time, and resulting in the effective mobilization of skills”[22]. James Surowiecki [4] and Scott E. Page [7] agree that wisdom of crowds as where under the right conditions groups can outperform the best or expert individuals.

Researchers and organisations have been harvesting this wisdom for variety of applications. For example, via crowdsourcing, where a random and large number of individuals are asked to perform an activity with a shared goal[23]. Another way is via learner sourcing, which is inspired by the concept of crowdsourcing. This tem emerged as more domain-specific and complex problems require domain knowledge and expertise beyond what a random crowd could possess or handle. Whether it was passive sourcing (where a system collects learners’ interactions passively with no interfering/interruption) or active sourcing (where a system interrupts and asks the learners for specific information), the fundamental difference is that users (learners) are motivated and engaged in their learning, which leads to more tailored content due to better quality control and scoping of tasks [24].

2.1.2 Types and Content of Social Clouds

Social Clouds can be characterised based on the types of social media spaces used. Social spaces can be described based on the type of connection between users, how the information is shared, and how users interact with the media [25]. Based on that social spaces can be grouped into two broad types - open and closed. The former refers to social platforms that are publicly available where any user can interact with the digital objects available, while the latter refers to more specialised spaces where

[3] https://www.omnicoreagency.com/youtube-statistics/

only certain users can have access. In other words, the former involves random crowds (e.g. different age, expertise, and education levels), whereas the latter can be restricted to special crowds, such as students. The fundamental difference here is the scale, quality and scope of the content captured in the social cloud. Closed social spaces offer less noisy and tailored social content, where usually specific users are targeted for social cloud collection (e.g. learner sourcing [24]), but they may miss the openness and scale offered by the open spaces.

Users tend to leave their interaction trails as content in different formats that is usually tailored to the social space they are using, but mostly in the format of textual content and user profiles [25]. Some social spaces allow users to leave short text in the form of comments (e.g. on YouTube videos), opinions (e.g. around news articles), reviews (e.g. on restaurants or movies). Other spaces provide services for microblogging, such as Twitter, where users can create hash tags, tweet (i.e. write short blogs), and retweet previously written tweets. These spaces limit the length of posts to 140 characters. This limitation impacts the writing style of microblogs[26]–[28]. There are spaces (e.g. forums) that provide question and answer (Q&A) discussions like Stack Overflow, where learners write a question and experts answer this question. Most social spaces require the user to create some sort of a profile[19]. This profile captures some attributes about the user e.g. age, gender, interests and location[25].

The UGC in online social spaces around digital objects i.e. *textual content* and *user profiles* form a social cloud around these digital objects, where *metadata* about the digital objects (e.g. title, type and URI) is captured as well.

2.1.3 Applications using Social Clouds

There is a growing body of research that has been utilising social clouds for different purposes. This section highlights example applications based on different content of social clouds discussed above.

Microblogs, specifically tweets from Twitter have been used for several applications, such as opinion mining and sentiment analysis [29], [30]. For crisis monitoring and detection like utilising tweets to create flood maps during floods [31] and searching for real world incidents [32] and building maps of the most at-risk areas[33].They have been also used to measure social diversity of urban locations via connecting geo-social networks i.e. venues and visitors' social media content, in particular Twitter[34].

For user profiling and modelling [35][36]. For creating labelled datasets like pet ownership and diabetes[37]. To detect customer satisfaction using personality traits and emotion detected from their tweets[38].

Textual comments of hotel reviews have been used to improve hotel recommendations[39], [40]. Also, comments on videos (e.g. on YouTube) have been used for variety of applications. To improve performance of video retrieval [41]. For the identification and assignment of relevant categories to the videos [42]. To propose a re-ranking method for producing a new ordered list of videos that are originally produced by the traditional YouTube recommender [43]. To identify relevant comments for ranking[44]. For filtering and predicting useful comments [45]. In domain of learning, comments have been used for deriving group profiles that can support the design decisions when implementing simulated environments for learning[46].To facilitate and support informal learning [47]. To predict the learners personality and its impact on the learning process [48].To identify learning challenges during video-based learning[49]. To investigate students' learning [50] characterise user's engagement[51] while watching learning videos for soft skills.

Textual content and user profiles in Q&A forums, have been gaining more interest; they have been used to model user collaboration using the comments on the posted questions[52]. It also has been used to explore aspects related to learning. For instance, textual content on Stack Overflow have been used to identify learning challenges that face computer science learners [53]. Also, for helping educators and stakeholders to identify and recommend programming languages to learners by using textual content in questions and answers and users profiles of expert users/programmers[54].

The user profiles captured in different social spaces have been used to deepen the understanding of users and their behaviour on online social spaces. For example, detecting users' empathy from their Facebook profiles (e.g. from status updates) [55]. For user modelling and personalised Massive Open Online Courses (MOOCs) recommendations to overcome cold start situation[56]. To increase MOOCs completion rates via social comparison with successful users[15].

2.1.4 Discussion

For this research, a *social cloud* is a term used to refer to *a collection of digital objects* in online social spaces, the *user comments* written on those digital objects, the *profiles of users* who wrote the comments, and any *metadata* associated with the digital objects.

The *social clouds with textual comments* (e.g. around videos, images, reviews or web news articles) are the *main focus of this research*. Although social clouds created on microblogging spaces like Twitter have been widely used for different purposes, the tweeting experience (i.e. writing short blogs, retweeting and hash tagging) impacts the writing style of microblogs, which in turn (a) results in the requirement of rigours pre-processing of this text to enable diversity measurements, and (b) the text might not be rich enough for diversity exploration. Similarly, the social clouds from Q&A forums might require a different type of analysis (e.g. discourse analysis) to understand the sequential nature of the questions and answers in the comments.

In terms of modelling diversity *both types of social clouds, open and closed*, should be explored to gain more insights into social clouds diversity. Each type of social clouds brings opportunities and limitations. Open social clouds offer openness and scale, but this wisdom is created by random crowds where no qualification or expertise are required. Such content has limitations including being noisy and might not be domain specific (e.g. contains swearing, sarcasm, advertising). Closed social clouds encounter less noise and their content is tailored to the social space and other users. However, it misses the opportunities offered by the open social spaces.

2.2 Diversity Research

The term diversity has been used in various disciplines interchangeably with many synonyms such as heterogeneity, rational demography, disparity, distance, and variation [57], [58]. Konrad et al. argue that a dictionary definition of diversity is insufficient and researchers in different disciplines have been attempting to conceptualise it. They explain that “At its core, the concept of diversity is all about *difference* and *inclusion*”[1]. Researchers from different disciplines have been attempting to understand and measure diversity, where some argue that diversity is not one thing[58]–[60].

An overview of this body of research is discussed next. First, the section starts with an overview of research that is linked to efforts for understanding diversity with a focus on related work from computing. It highlights two *main perspectives of diversity*. The second section discusses research concerned with quantifying diversity.

2.2.1 Understanding Diversity

Philosophers argue that “a concept of diversity is an understanding of what makes a group diverse that may be applicable in a variety of contexts” [60]. In diversity lietrature, there are two perspectives of diversity that have been the main focus for diversity understanding - individuals’ (or groups’) diversity and domain diversity. Individuals’ diversity is concerned with comparing individuals based on their attributes, while domain diversity is concerned with differences with regards to a subject domain (e.g. music or body language) captured in e.g. content, perspectives, or opinions.

This section reviews discipline-based related work on understanding individuals’ diversity, which has been the main focus in different fields. Then, it reviews work that explored domain diversity.

2.2.1.1 Individuals’ Diversity

Individuals’ diversity refers to the variations of traits visible (e.g. age and gender) or not (beliefs and values) of groups of two or more people [61]. In other words, diversity can operate between people on known attributes, such as age, gender, and ethnicity, or based on non-visible and deeper attributes like mood, health and personal experiences [62].

Some disciplines focussed on one type of attributes (e.g. visible), while others investigated diversity with regards to variety of attributes. Nationality-based cultural (or national culture) variations have been one of the widely used visible attributes for understanding individuals’ diversity, especially in social science, which in turn informed research conducted in computing.

Understanding diversity based on visible attributes. Some visible attributes have been used to compare and understand individuals’ differences. Below is work that focussed on national culture, gender and combination of these attributes with other visible attributes.

National Cultutre. Many researchers linked diversity of individuals and groups to their cultural variations. Great body of rsearch focussed on national culture i.e. cluster and compare individuals and groups based on their nationality. In psycology, using data from 33 nations, Gelfand et al. ranked countries based on two dimensions - tight versus loose or (tightness-looseness). They define tight cultures as individuals with "strong norms and a low tolerance of deviant behaviour", whereas loose cultures are those with "weak social norms and a high tolerance of deviant behavior"[63].

Social scientists established cultural models to facilitate comparisons between different cultures. Examples of prominent models, especially in computing are: (**a**) Hofstede's model which is based on six dimensions. Example dimensions are Individualism versus Collectivism, which are related to the relationship between and integration of individuals and the groups, best example is extended families[64], [65]. (**b**) Another widely used model is *GLOBE* (Global Leadership and Organizational Behaviour Effectiveness). GLOBE is an extension of Hofstede's Model[66]. GLOBE consists of ten cultural or societal clusters[66], [67]. These clusters gather similar cultures by nationality. Example clusters are Anglo (e.g. Canada and the United Kingdom) and Middle East (e.g. Turkey and Egypt). Unlike Hofstede's cultural model, GLOBE proposes cultural dimensions for individual and group levels[68], [69]. This model was extended by adding more countries to the relevant GLOBE clusters[70]. (**c**) The Lewis Model was developed to assist trainees to behave in a more productive manner and communicate successfully in "multi-cultural situations". The model consists of three categories of "cross-cultural communications behaviour" - Linear-active, Multi-active and Reactive. Each describes individuals with regards to certain characteristics including politeness, use of body language and the way they talk. For example, Linear-active categorise individuals who are polite but direct, use limited body language and talks half the time (e.g. Germany and Switzerland). Lewis explained that each country could be a mix of the three categories but mainly present in one or two categories[71], [72].

In computing, as research moved from stereotyping to adapting to diversity, various attributes have been used to understand individuals' diversity. Culture was one of the main attributes for understanding diversity. Many studies have been informed by the aforementioned studies as follows.

The awareness of the importance of user diversity based on their culture and its effect on their preferences and performance resulted in culturally-aware systems [73]. A culturally-aware system “refers to any system where culture-related information has had some impact on its design, runtime or internal processes, structures, and/or objectives” [68]. Examples of such systems are: TLCTS (Tactical Language and Culture Training Systems), which is a virtual environment that trains learners to gain cultural knowledge related to face-to-face communication skills in a foreign language and culture[74]. TRAVELLER (TRAining for Virtually Every Location for Learning Empathic Relationships), which is following a story-driven approach that aims at educating young adults (18-25) cultural sensitivity based on the Hofstede’s Model [75]. MOCCA, which is a culturally-aware interface that adapts its appearance based on the user cultural background[76]. Hofstede’s six dimensions were correlated with International Large-scale Assessments data of 81 countries to identify principles for designing “culturally-appropriate” Educational Assessment Technology[77]. Same model was used to design and build a culture-aware music recommender system[78]. Culture was the main attribute for the design decisions for building a culture-based persuasive technology that promotes physical activity among university students[79].

Culture was also used for understanding differences of user behaviour. Collectivism and Individualism dimensions from Hofstede’s Model were used to explore users’ behaviour with regards to social questioning and answering between users from 4 countries - the United States and United Kingdom as individualist and China and India as collectivist[80]. Same dimensions were used for investigating predictors of competitive behaviour across participants from two countries, Canada and Nigeria[16].

Gender. This attribute has been widely used across different disciplines for understanding differences among individuals. In computing, gender was the main attribute for comparing individuals from different countries in terms of their perception of online products or services in the mobile domain[81]. It was used to compare differences with regards to cooking behaviour (e.g. commenting and uploading recipes) captured in online cooking website[82]. Gender differences were investigated in the domain of learning, for example, it was used to detect difference between learners in terms of their facial expression during learning[83]. Gender was used to investigate how learners behave (e.g. perceive, interact, and engage) with artificial pedagogical agents and games for learning. School students were compared on the

way they perceived animated pedagogical agents that showed emotional and motivational support while learning math[84]. Similarly, students' engagement with female agents was examined between students in terms of their gender. Gender was also a key feature to compare the performance of students with an educational game[85].

Combination of visible attributes. Some studies used more than one visible attribute to compare individuals. Age, gender and culture based on Hofstede's model were used to explore country-based music diversity[86]. The influence of age and gender has been explored to identify difference in accepting persuasive strategies[87]. Also, cultural models in conjunction with other attributes were used to understand individuals' differences in terms of how they perceive social influence (e.g. rewards, social comparison, and social learning). Collectivism and Individualism dimensions were used to explore whether culture influence how users from different age and gender perceive social influence in persuasive technology. This involved participants from North America, Africa and Asia [88]. An investigation was carried out on whether social comparison between learners promotes effective "self-regulatory" behaviour and achievement in MOOCs. This investigation was linked to culture including individualism dimension from Hofstede's Model and tightness from Gelfand et al [15]. A comparison of users based on their programming skills (expert vs. novice) and education (first degree vs. graduate degree) was conducted in terms of their responses to persuasiveness [89].

Understanding diversity based on non-visible attributes. Some studies focussed on hidden attributes related to cognition and motivation. This is a rich area in different disciplines. In computing, the influence of these attributes on individuals for learning has been investigated. Users have been compared based on their cognitive ability to evaluate their acceptance to persuasive strategies[90]. The differences between two user groups in terms of engagement while learning was conducted based on hidden attributes like task value, self-efficacy and metacognitive self-regulation[91]. The Individual differences between students in terms of performance-orientation and visual attention were compared to investigate their impact on the design of pedagogical agents[92].

Understand Diversity Based on visible and non-visible attributes. Some fields investigated individuals' diversity based on a combination of visible and non-visible

attributes. Below is work from management and organisational literature, cultural diversity field, philosophy and computing.

In management and organisational literature, studies have explored individuals' diversity within teams in order to understand diversity and its influence on the organisational commitment and outcome. Initially, the studies were based on the individuals' demographics (also referred to as demographical, categorical or relational diversity). The research developed over time to go beyond these attributes involving more hidden characteristics of individuals like personality and attitudes[93]. For example, a *diversity taxonomy* was proposed by Harrisons et al. in [94] to classify diversity as *surface-level* and *deep-level diversity*. Surface-level diversity refers to "differences among group members in overt, biological characteristics that are typically reflected in physical features", while diversity at the deep-level "includes differences among members' attitudes, beliefs, and values" [94]. Later, researchers added informational or *knowledge attributes* (e.g. experience and education) to this taxonomy. This is because, researchers argue that there is another distinction between the hidden attributes. They argue that some hidden attributes require long time of observing or interacting with individuals to be revealed (e.g. personality), while other attributes can be identified on a short period of time (e.g. level of education)[95].

Similarly, Harrison & Klein viewed individuals' diversity within an organisation as a combintaion of attributes types but in terms of three different concepts of diversity. They defined diversity from the organistional view as "the distribution of differencess among memebers of a unit with respect to a common attribute X"[58]. They proposed the within-unit diversity typology as separtion, variety, and disparity. With-in unit seperation is defined as "differences in position or opinion among unit members", variety defined as "differences in kind or category, primarily of information, knowledge, or experince among unit members", and disparity is defined as "differences in concentartion of valued social assests or resources such as pay and status among unit members".

In the field of cultural diversity, the diversity taxonomy discussed above, deep and surface levels, was adapted to propose two types of cultural diversity: subjective (e.g. attitudes, values, identities) and objective (e.g. language systems, gender, political systems). Authors refer to subjective cultural diversity as a type of deep diversity, and objective cultural diversity as a type of surface diversity [96].

In philosophy, researchers distinguished two broad types of diversity concepts to compare individuals based on the two types of attributes, within group and comparative. They derived three diversity concepts that fall under those two types - egalitarian, representative, and normic diversity. The three concepts are similar in meaning to the three concepts prroposed by Harrison & Klein above, variety, seperation and disparity respectively. Authors explinaed that egalitarian diversity is a within-group concept, whereas representative, and normic diversity are comparitive concepts [60].

In computing, some studies combined visible and non-visible attributes to compare individuals. Students learning math were compared based on their emotional state, perceptions of item difficulty, and gender[97].

2.2.1.2 Domain Diversity

Some researchers focussed on diversity (e.g. in UGC) with regards to a subject domain. This is whether it was with a link to individuals (when applicable) or not i.e. whether using individuals' attributes (discussed above) to compare the domain differences. Examples are, diversity of a news article's content with regards to a crucial event, the opinions of individuals on a political event (e.g. in tweets), or perspectives of learners on a topic displayed on a learning video (e.g. reflected in comments on video). This is discussed as follows.

Diversity was used to indicate interdisciplinary based on publications i.e. diversity of subjects cited in papers' references [59]. Diversity was investigated in journalism based on archives of news articles to assist journalist to diversify the content of their news articles based on the people (e.g. refugee, politician, and policeman) mentioned in the article[98]. Also, news articles were analysed to identify the ones that can give diverse opinions on a given topic[99]. Opinions from social media on crucial events were analysed to help journalists and archivists understand the sentiment of those opinions [30]. Similarly, opinions and sentiments about companies, products, and policies expressed on social spaces were analysed to inform intelligent business applications[100]. In cultural diversity, different domains have been explored, such as diversity with regards to radio [101], policy making and regulations[102], books [103], cinema [104], television channels [105]. Differences in terms of music genres and subgenres to identify listening patterns have been investigated in many studies [78],

[86], [106], [107]. With regards to “genderification of cooking and eating”, UGC including comments related to cooking recipes from males and females were analysed to detect difference[82]. In the domain of learning, learners’ perspectives reflected in their comments while watching videos for learning soft skills were analysed to detect differences - when learning soft skills for job interviews with a focus on body language and emotions[108] and aspects related to pitching presentations[109].

2.2.2 Measuring Diversity

A common method found in diversity research literature for measuring diversity is based on three properties of diversity - variety, balance, and disparity. These properties were derived over time by researchers from different disciplines to quantify the concept of diversity. Different indices with different approaches were used to measure diversity based on variety, balance and disparity. A general diversity framework was proposed to measure diversity based on these properties. This framework is widely applied in various domains. This is discussed next.

2.2.2.1 Diversity Properties and their Measurements

Diversity is a concept that is prominently used in a variety of disparate disciplines, such as in economy [2], ecology [110], cultural diversity [111] and energy[112]. There have been several studies concerned with quantifying the concept of diversity. This has been conducted with regards to three diversity properties emerging overtime from different disciplines. These properties are **variety**, **balance**, **and disparity**. Some research used each property as a synonym for diversity, others combined or aggregated at least two to quantify diversity, where they argue that diversity with regards to one of them is insufficient[103], [105], [113]–[115]. The related general definitions and indices for properties’ measurements are discussed next.

Variety is concerned with the number of types or categories of given elements or individuals of a given population, such as number of species in ecology [116]. It is a term that is widely discussed in ecology as a synonym of diversity[2]. It is considered a basic but important diversity property [78]. It is a positive integer that is measured based on counting the "nonempty” and “well-defined” categories[113], [117]. One can choose to exclude empty categories from diversity measurements when these categories are given but not covered [117].

Balance is referred to as relative abundances of elements [118] and the nature of apportionment of elements across their categories [78]. For example, it was used to measure the relative abundances of species [114]. It is a set of positive fractions that sum to unity[119]. Balance has been referred to as evenness in ecology[120], [121] and concentration in economics[122], yet all terms have similar meaning with regards to diversity[77]. Balance has been measured via Shannon Entropy Index (also referred to as Shannon and Weaver) [123], Shannon Evenness Index [120], Gini Index [124] and Simpson Index[125]. Shannon Entropy is one of the widely used and most robust indices for balance[2], [59].

Disparity indicates the distance between elements [2], [112], [113]. This property emerged as researchers argue that variety or balance do not incorporate differences between elements[2], [59], [103], [118], [126]. For example, it is argued that variety and balance do not include ecological differences between species or “inter-species differences”[118]. Disparity seems to be the trickiest property to interpret and measure. Researchers used different and tailored indices to their field to measure this property. For example, average dissimilarity index was used to measure disparity of interdisciplinary domain based on publications. This was based on the Cosine Similarity Index to measure similarity/dissimilarity of referenced subjects [127]. In their study to evaluate whether public television channels are more diverse than the private ones, authors used their own methodology by selecting attributes to distinguish between different programme categories. They used the Euclidean Distance to measure disparity of programme categories based on these attributes[105].

Variety, balance and disparity have been mostly measured separately (as discussed above) to quantify the concept of diversity, but some researches aggregated them for one overall value for diversity. The index proposed by Stirling [113] and the Rao Diversity Index from Biology [128] aggregate variety, balance and disparity to give an overall value for diversity. More is discussed in the next two sections on the way the diversity properties have been used for measuring diversity.

2.2.2.2 Approaches for Diversity Measurements

In diversity research literature, diversity measurement has been based on two approaches using the diversity properties (variety, balance, and disparity) discussed above. These are as follows (inspired by the review in [2]):

One-concept diversity, where only one diversity property is used to quantify the concept of diversity i.e. diversity is measured based on one value that is used as an indicator of the level of diversity. An example of this diversity is research conducted in ecology where species diversity is measured based on variety only (as discussed above) i.e. the value of variety indicated whether species diversity was high or low.

Multi-concept diversity i.e. diversity is quantified based on a combination of properties (e.g. dual-concept diversity or triple-concept diversity), where two or three of the diversity properties are used to measure diversity. This can be either: (**a**) measuring the properties separately but use them together as multi indicators of diversity level i.e. diversity is indicated with more than one value like the work in interdisciplinary domain [59]; or (**b**) aggregating the diversity properties, where two or three properties are aggregated as one overall value for diversity. For instance, work in the journalism domain [98], which used the index proposed by Stirling [113] to indicate whether the content of a journal article is diverse or not. Some researchers classify Shannon Entropy Index and Shannon Evenness Index (mentioned in section above) as indices for dual-concept diversity i.e. measure diversity based on an aggregation of variety and balance [2].

2.2.2.3 Framework for Diversity Measurements

Based on the diversity properties mentioned above, Stirling [113] proposed a general "interdisciplinary" framework for analysing diversity in science, technology and society, where he proposed that diversity is three "basic" properties: variety, balance and disparity. Each is a necessary but insufficient property of diversity. He defines the diversity properties as follows:

Variety "is the number of categories into which system elements are apportioned ". This property answers the question: "how many types of thing do we have? … All else being equal, the greater the variety, the greater the diversity".

Balance "is a function of the pattern of apportionment of elements across categories". This property answers the question: "how much of each type of thing do we have? … All else being equal, the more even is the balance, the greater the diversity".

Disparity "is the manner and degree in which the elements may be distinguished". Disparity answers the question: "how different from each other are the types of thing

that we have? … All else being equal, the more disparate are the represented elements, the greater the diversity".

Based on Stirling's Diversity Framework, measuring diversity for a given system requires identifying the **system elements** (e.g. a user) and the main **categories** of these elements (e.g. culture groups).

He also proposed a diversity index, referred to sometimes in literature as Stirling, Rao-Stirling or Quadric Stirling Index, which aggregates the three properties to come up with one diversity value for a given system. He also proposed a more generalised index adding weight to balance and disparity in the index, which can be adjusted subjectively.

Unlike his indices, his definitions of the properties and how they can be measured are adopted widely for diversity measurements in different disciplines. Researchers mostly used the second type of diversity measurements (section 2.2.2.2), where they measured each property separately. In cultural diversity studies, Stirling's definitions of the three properties have been used to measure: radio diversity in terms of music broadcasted in radio channels [101]; diversity of policy making and regulations[102]; books diversity in terms of diversity of consumed published book [103]; to evaluate cinema diversity[104] and to compare public television channels with privet ones[105]. Also, variety, balance and disparity were used to identify experts on social media, specifically from Flicker using the annotations users generate to tag photos[129]. In interdisciplinary domain [59], researchers measured each property in isolation and then aggregated by Stirling Index to indicate the level of diversity (interdisciplinary) in publications. This allowed them to have multi-diversity indicators where by using these properties they constructed overlay maps to visualise the disciplines' citation diversity[59]. Park et al. [106] used Stirling Index to measure music diversity based on music genres.

It seems that the aggregated diversity value based on Stirling Index has limitations. Interdisciplinary researchers described the aggregated diversity value by the Stirling index as "a black box". It seems that this index hides or neglects: (a) how each diversity property influence this overall value with no subjective interfere; (b) the insights that can be provided by the separate calculations of each property, and (d) the opportunity to have different diversity indicators that can satisfy different criteria in different contexts[59], [101].

2.2.3 Discussion

This research accepts the views that diversity is not one thing (i.e. should be conceptualised by more than one concept) and that *diversity can be characterised* based on variety, balance and disparity.

This research recognises that there are *two broad diversity perspectives* - individuals (e.g. users) and domain by which diversity can be explored and measured. Individuals diversity should be explored with variety of user attributes, such as visible (surface-level attributes) and non-visible (deep-level attributes). Also, national culture is an interesting attribute to explore for modelling diversity, especially that there are models to support this. The GLOBE model seems to be most suitable as it facilitates comparing individuals and groups. Therefore, for this research, the *diversity taxonomy* from management and organisational diversity research that classify these attribute is adopted (discussed in section 2.2.1.1). This is to enable the exploration of users' different attributes including culture. GLOBE will inform the exploration of individual diversity with the culture attribute.

The Stirling Diversity Framework provides a generic and systematic approach for measuring diversity based on the three properties of diversity – variety, balance and disparity. This framework's definitions of the properties and how to measure them have been adopted and tested in several studies for measuring diversity. This shows the framework's potential for modelling diversity. Also, the use of multi indicators for diversity is inspiring. Although Stirling Diversity Index can give an initial insight of the overall level of diversity for a given system (e.g. a social cloud), research shows that breaking down this value to its detailed components (variety, balance and disparity) can offer multi diversity indicators, which in turn can give deeper insights into diversity from different perceptions.

To summarise, the definitions of the diversity properties – variety, balance and disparity, and how to measure them as proposed by the Stirling Diversity Framework are selected to underpin the diversity representation and measurement of social clouds. The second approach for diversity measurement (multi-concept diversity) is adopted in this research, where the diversity properties are measured separately to serve as multi indicators of the level of diversity in social clouds. Each property can be used on its own or combined with the other properties to indicate the level of diversity with regards to the domain and individuals (e.g. users) diversity perspectives. The

individuals' diversity perspective will be explored with variety of attributes supported by the diversity taxonomy and GLOBE.

2.3 Diversity in Social Clouds

There is an increasing body of research that has been investigating diversity in social clouds. Few explicitly identify understanding and/or measuring diversity as the aim for their work, others can be classified into this area of research. This is discussed next.

2.3.1 Related Work on Diversity in Social Clouds

This section reviews related work on social clouds diversity. First, grouped by the social clouds' content used for diversity, highlighting work done and techniques used. Then, classified in terms of the diversity perspectives they investigated - individuals (e.g. a user) or domain diversity (as discussed in section 2.2.1).

In general, the related work exploited microblogs, mainly from Twitter and textual comments around videos for social clouds diversity with the former being the most used. All the work focussed on one of the diversity perspectives i.e. no research investigated both. Majority of studies worked on domain differences. A summary is provided in Table 2.1.

2.3.1.1 Diversity in Social Clouds from Microblogging Spaces

The work by Gao et al. [26] compared the microblogging behaviour (e.g. writing microblogs, hash tagging, sharing links) of users from China and USA on Sina Weibo and Twitter respectively. Semantic techniques and sentiment analysis were used to analyse the content of tweets. The findings were linked to the Hofstede Model to clarify the users' microblogging behaviour.

Similarly, Ilina et al. [130] analysed the microblogging behaviour on Twitter of users from different cultural backgrounds (USA, Brazil, Germany, Spain and Japan) to construct user profiles. Descriptive statistics were used for content analysis. The findings were linked to the Lewis model to identify differences between user groups. This work was based on the assumption that "personality traits" as defined in the Lewis model is reflected in the way users' blog on Twitter.

Choudhury et al. [131] conducted a comparison of tweets that had "self-disclose" mental health concerns using machine learning techniques. The tweets by users from

the United States, the United Kingdom, India and South Africa were compared based on gender and cultural differences. The authors did not use any cultural models or culture-related studies (e.g. the ones mention in section 2.2.1.1) but develop their own techniques for identifying the users' culture based on their location.

Bhatt et al. [132] used text analysis techniques to analyse tweets of group of users, which in turn was used as an indicator of diverse crowds. The tweets with regards to fantasy sports (related to soccer teams and captains) are collected and analysed to identify what related aspects are mentioned. The users with diverse tweets are identified as a diverse crowd. Their goal was to group diverse individuals to create a diverse crowd, which in turn would enhance the wisdom of crowds.

Maynard et al. [30] detected opinions on crucial social events from tweets to enrich archives of news articles, where they used sentiment analysis and semantic techniques to detect opinions' diversity within a document that includes an event. They identified interesting documents as the ones that have diverse opinions.

Park et al. [106] used user interactions in social spaces including twitter (& profiles in music website last.fm[4]) to measure musical diversity. They measured diversity based on Stirling Diversity Index, then compared it to other attributes like culture and demographics using regression analysis to identify variables that might contributed to music consumption. The diversity properties- variety, balance, and disparity from Stirling Framework were linked to music genres and subgenres to measure diversity.

Another work into the music domain is by Schedl and Hauger [107] who compared cities and countries based on the music genre they are listening to. This is identified in the users' tweets using hashtags (e.g. #nowplaying or # itunes). The aim was first to identify artist similarities within cities and countries (e.g. New York (US) and London (UK)) i.e. who are the common artists across locations (using similarity measures). Then they used this to identify cultural listening patterns e.g. identify what type of music is popular for certain individuals within a location. The comparison to identify patterns was conducted via descriptive statistics.

2.3.1.2 Diversity in Social Clouds around Videos

Despotakis et al. [133] introduced a framework (ViewS), which is underpinned by text analysis and semantic techniques, to capture the diversity of viewpoints from YouTube

[4] https://www.last.fm/

comments on a human activity related to job interviews. This was with a focus on social signals with regards to body language and emotions. The analysis of the diversity of viewpoints was based on the demographic data of the users available from YouTube - age, gender and location. These attributes were used to visualise and manually compare the users' domain differences.

Hecking et al. [109] adapted network-text analysis of learner-generated comments to capture divergence, convergence and (dis)continuity in textual commenting. The user domain differences with regards to pitching presentation skills were analysed via visualisations of the networks. This is to identify and characterise learners' engagements while watching learning videos.

Kleanthous et al. [134] conducted an exploratory study to investigate individual differences in terms of behaviour, perception, and cognition while watching music videos for learning. Descriptive statistics and text analysis techniques were used to analyse the users' generated content including comments to compare users based on their status (musicians and amateurs) and gender.

The related work discussed above can be classified based on the two main streams of understanding diversity (discussed in section 2.2.1) as follows.

Individuals' diversity. Gao et al. [26], Ilina et al. [130] and Kleanthous [134] used the textual content of the social clouds (microblogs or comments) and visible user attributes to compare individuals. However, they did not compare domain differences that are possibly captured in the textual content.

Although Gao et al. investigated what topics are mentioned in the microblogs focussing on three concepts - organisations, people and location, these were used to compare the individuals (e.g. they identified that Chinese users tend to mention locations more than American users). They identified a term that is related to one of the three concepts (e.g. Paris is location), but did not differentiate the mentioned aspects that are related to these concepts neither within a group (e.g. between Chinese users) nor across groups (Chinese vs. American). Same applies on the work by Kleanthous [134], authors used the comments to detect differences between groups (males vs. females and musicians vs. novices). For example, who wrote more comments, who was more confident when commenting about videos' content? Although authors put aspects related to music shown in videos (e.g. melody, equipment, technique and imagination) in AVW-Space for participants to select when commenting, authors did not use these aspects to identify domain difference. They

used these aspects to identify differences in terms of frequency – e.g. how many times an aspect (e.g. melody) was selected by females. Similarly, Ilina et al. looked into the microblogs to identify difference in terms of microblogging behaviour, but with no domain of interest. They identify three features for analysis: content, activity and social networks, these are concerned with e.g. number of hashtags, number of tweets on weekends, and number of friend or followers respectively. They linked the findings to culture to identify factors that might contributed to the identified behaviour.

Domain diversity. Despotakis et al [133] and Hecking et al. [109] investigated differences in comments with regards to a domain of interest (body language and emotions for job interviews and presentation skills respectively). The former used the user visible attributes to compare domain differences, while the later used the domain differences to characterise types of learner behaviour when engaging with videos. Although both studies had access to user profiles, they did not investigate who the users who interacted with the videos in terms of diversity.

Bhatt et al. [132], Maynard et al. [30] and Schedl & Hauger [107] looked into domain differences captured in microblogs. However, they did not link these differences to the users' attributes and did not looked into users' diversity. Bhatt et al. quantified microblogs' diversity to identify the diverse tweets with regards to fantasy sports. The users are identified via their profiles in two social spaces including Twitter (& related website for fantasy sports (FPL)[5]), but none is done to identify who the users are or how diverse they are. Maynard et al. investigated opinions' diversity with regards to critical events in news articles. It would have been more interesting to see who the users who gave these opinions are and linked to their opinions e.g. whether certain user groups are mentioning something others are missing. Same can be said about the work by Schedl & Hauger, who compared cities and countries with regards to the music they are listening to. Investigations of who the users are, e.g. in terms of cultural backgrounds could have supported the findings.

Choudhury et al. [131] and Park et al. [106] used tweets to measure domain diversity, they did not measure user diversity, but used attributes to compare the domain differences. The former was interested in cultural and gender differences in terms of the way the perceived mental health issues and therapy. The later investigated music diversity based on genre and subgenre. Authors did not measure users' diversity. They

[5] https://fantasy.premierleague.com/

used users' attributes including demographics to identify factors that might contributed to their musical diversity e.g. whether users who lived in a location with mixed ethnicities listen to diverse music. Same argument can be concluded for both studies. Further investigations for the user differences would have brought more insights.

Table 2.1 Summary of research on social clouds diversity differentiated based on: (a) social cloud content, (b) source of social cloud collection, (d) diversity perspective analysed and (e) techniques used for analysing social cloud content.

Social cloud content	Social cloud source	Authors	Individuals' diversity	Domain diversity	Techniques
Microblogs	Twitter & Sina Weibo	Gao et al. [26]	√	X	*Semantic techniques* & sentiment analysis
	Twitter	Ilina et al. [130]	√	X	Descriptive statistics
		Choudhury et al. [131]	X	√	Machine learning techniques
		Bhatt et al. [132]	X	√	Text analysis techniques
		Maynard et al. [30]	X	√	*Semantic techniques* & sentiment analysis
		Park et al. [106]	X	√	Stirling Index& regression analysis
		Schedl & Hauger [106]	X	√	Descriptive statistics
Comments	YouTube & a closed social space	Despotakis et al [133]	X	√	*Semantic techniques* & text analysis
	AVW-Space [135]	Hecking et al. [109]	X	√	Network-text analysis
		Kleanthous et al. [134]	√	X	Descriptive statistics & text analysis techniques

2.3.2 Discussion

There has been an increasing body of research on understanding and/or measuring diversity in social clouds. However, these efforts have been limited, especially that few have explicitly catered for diversity.

Building on the discussions from above, majority of studies mentioned above lack automation of diversity measurements. For example, visualisations can reveal interesting patterns, but their adoption for automated diversity measurements is not feasible.

Most of the work is tailored to a certain content of the social clouds like microblogs which might hinder their applicability and replicability with other social clouds.

The inspection of users (individuals) diversity and domain diversity has been mostly disconnected, which hindered gaining further insights into social clouds diversity. Some work focussed on how different the users are, others focussed on the domain differences. Few had attempted linking the domain differences with user differences, where they mostly used user attributes to compare the domain differences.

Also, most of the research focussed on surface-level attributes of users, such as culture, age and gender either to investigate user difference or to compare domain differences. Although this is inspiring and promising, it would be more interesting to explore user diversity rigorously i.e. reveal who are the crowds (users) who interacted with the digital objects in social spaces. This should be conducted with variety of user attributes including the non-visible ones (deep-level attributes).

There is a need for a computational model that pave the way to automatically measure user and domain diversity. A model that is applicable within different domains and consider variety of user attributes. A model that can facilitate: the comparisons of users, comparisons of their perspectives with regards to a domain of interest, and connection of users with their domain knowledge, which in turn help to investigate and detect any related diversity patterns. This is the aim of this research.

2.4 Semantic Web Techniques for Analysing Social Clouds

Semantic Web (Linked Data or Web 3.0 [136]) refers to “a set of best practices for publishing and connecting structured data on the Web”, where the data is “machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets”[137]. The Semantic Web is “an ecosystem of data, where value is created by the integration of structured data from many sources” [21] resulting in the Web of Data, which is “a global data space containing billions of assertions” [137]. The Semantic Web relies on ontologies to structure the data for “comprehensive and transportable machine understanding”[138]. Ontological underpinning is much matured where existing ontologies are publicly available, and several methods are available to construct new ontologies.

This section discusses ontologies highlight their structure, usage, and how to construct them. It focusses on ontology-based semantic annotations and distance that enable

diversity representation and measurements of social clouds. It concludes with a focus on related work that used semantic techniques for social clouds diversity

2.4.1 Ontologies Underpinning

This section discusses: what is an ontology and what is its structure, existing ontologies that cover different domains and represent users, and how to construct an ontology when one is not existing. These feed to ontological underpinning for the proposed computational model discussed later in chapter 3.

2.4.1.1 Ontology Definition and Structure

An ontology, which is a term borrowed from philosophy, has several definitions to mean the same thing. One of the widely used definition is by Gruber (1993), who defined the ontology as “an explicit specification of a conceptualization”, where it is used to represent a body of knowledge in an abstract and simplified view of the domain represented. An ontology defines terms to represent Knowledge[139]. He also defines an ontology as a computer model of the world representing main concepts and the links between them [21]. This is similar to: the definition by Synak et al. (2009) who explains that “an ontology formally describes concepts and relationships which can exist between them” [140], and the definition adopted by Artificial Intelligence and Web researchers, who define an ontology as a document or file that formally defines terms and the relations among them[141].

A concept or a term in an ontology can reflect any aspect in the world. It can be a person, a building, an activity or an abstract concept like time. A relationship in an ontology represents how two concepts can be connected to each other, where this connection can represent things like characteristics of objects (e.g. Fruits are Juicy) or an activity (e.g. Mums are Busy) [140]. The concepts in the ontology can be[139] [141] [140]:

- A class, which is a concept that represent general qualities and properties to represent a group of objects.
- A subclass, which represents a part of the group of objects with some characteristics that are not common for the whole group.

- An instance (or individual), which is a single item in the world that can be a concrete object like people and animals or more abstract like words and numbers.
- A property, which represents relationships in ontologies. It is used to describe the characteristics of a class.

Classes define the main concepts in an ontology (e.g. body motion in an ontology that represent body language). Classes can have subclasses (e.g. facial expression) and instances (e.g. smile). The characteristics of a class apply to its instances. Every ontology has a class hierarchy containing and connecting all classes via the subsumption (i.e. subclass of) relationship, where characteristics of a class are inherited to its subclasses. The classes in the class hierarchy can be either (a) a root class i.e. a super/top/parent class of all classes in the ontology, (b) category class i.e. any class in the ontology except the root class that has one or more subclasses, or (c) a leaf class that has no subclasses. An instance is a type of a class. Instances belong to at least on class. An ontology that have instances belonging to more than one class is a lattice. Properties or relationships have labels that allow to define links between classes and instances. A common and important type of a relationship in ontologies is the subsumption relationship, which identifies the subclass and superclass relationships in the class hierarch. There are other domain-specific relationship types that are based on the domain represented. Classes, subclasses and instances are usually referred to as entities. The taxonomic hierarchy of entities forms a tree structure.

2.4.1.2 Using Existing Ontologies

Hundreds of ontologies are published and available online covering well-defined domain (e.g. Gene Ontology[6] and Music Ontology[7]) and ill-defined domains (e.g. emotions covered in WordNet-Affect [170] and culture aspects captured in AMOn[8] [171]). Search engines can be exploited for the retrieval and selection of suitable ontologies. The Linked Data facilitated the rise of new applications that explores the Semantic Web resources. In particular, Linked Data search engines. The search engines crawl the Web of Data by following links between data sources and provide

[6] http://www.geneontology.org/
[7] http://musicontology.com/
[8] http://imash.leeds.ac.uk/ontologies/amon/

expressive query capabilities over aggregated data[137], [142]. Example search engines are Watson[143] and Swoogle[9][144]. The World Wide Web Consortium (W3C) provides a list of examples of “good” ontologies[10] providing some criteria for good ontologies. The work in [136] lists criteria for ontology selection including: usage and uptake; maintenance and governance; coverage and Expressivity.

Beside subject domain ontologies, one of the main interests in the Semantic Web has been defining ontologies that facilitate capturing user demographics, interests and preferences (ontological user models). For example, the General User Model Ontology (GUMO) that captures the statistic aspects of the user to simplify the exchange of user models between different user-adaptive systems [145], and the User Profile Ontology with Situation-Dependent Preferences Support (UPOS), which describes the dynamic aspects of the user to be used to provide context-aware adaptations for mobile communications and information services [146]. Other ontologies built on top of these ontologies such as, the Cultural User Modelling Ontology (CUMO) [147] that extends GUMO to capture the cultural profiles of the user based on the Hofstede’s Cultural Model to overcome the bootstrapping issue in user adapted systems. Another example, is an ontology that extends GUMO and UPOS to capture static, dynamic and context profiles of the user for personalisation and adaptations for mobile applications to help people in need like people with Dementia[148]. Friend-Of-a-Friend (FOAF)[11] provides a template for user profiling by defining and describing agents (Person, Project, Group and Organisation) using contact information and demographics attributes. The Social Web User Model (SWUM) cover attributes needed to model a user in the social web. SWUM extended GUMO and FOAF by including attributes that are loosely defined in them including interests, goals and knowledge[149].

2.4.1.3 Ontology Engineering

Ontology engineering as defined by Gruber (2008), is “concerned with making representational choices that capture the relevant distinctions of a domain at the highest level of abstraction while still being as clear as possible about the meanings of terms” [21]. Ontologies are developed for several purposes, such as to enable reuse of domain knowledge, make domain assumptions explicit and analyse domain

[9] http://swoogle.umbc.edu/2006/
[10] https://www.w3.org/wiki/Good_Ontologies
[11] http://www.foaf-project.org/

knowledge. Initially, there were many formalisms (ontology languages or schema languages) used to describe ontologies in the Semantic Web research. This hindered the “interoperability” of semantic solutions[140], [150]. Then, the Web Ontology Language (OWL) became a W3C standard recommendation in 2004 [139]. It explicitly represents meaning of ontology terms and the relationships between those terms. It has three types - OWL Lite, OWL DL (DL stands for Description Logic), and OWL Full. The difference between the three is the restrictions placed on the OWL Lite and OWL DL. The main restriction is that only in OWL Full classes can be individuals at the same time. The list of all restrictions are discussed in [151].

Several methodologies were proposed to develop ontologies in the Semantic Web research. The NeOn Methodology framework[152] is one of the widely used methodologies for ontology developments. It was proposed to overcome the limitation of methodologies that build single ontologies i.e. “an ontology that has not got any type of relationship (domain dependent or independent) with other ontologies”. The continuous evolution of the ontologies in the Semantic Web required a framework that supports distributed teams to collaboratively build ontologies by reusing and re-engineering knowledge resources. The framework defines nine scenarios for building ontologies and ontology networks, such as reusing and re-engineering non-ontological resources, reusing and re-engineering ontological resources and reusing and merging ontological resources. The framework describes activities to be conducted by developers for each scenario. For example, in the reusing ontological resources scenario, there are three approaches for reusing ontological resources: (a) use ontologies as a whole; (b) use only one part or module of the ontologies; and (c) use ontology statements (i.e. subject, predicate, and object). For a comprehensive list of scenarios and activities with regards to the scenarios refer to the framework in [152].

The Stanford University team, developers of the ontology editor protégé[12], discussed some fundamental rules in ontology design in their technical report regarding how to develop your first ontology[151]. These rules are as follows: (a) there is no one correct way to model a domain, (b) developing an ontology is an iterative process, and (c) concepts in the ontology are mostly nouns (objects) and verbs (relationships) in sentences that describe the domain of interest. The team described a “simple

[12] https://protege.stanford.edu/

knowledge-engineering methodology" that consists of 7 phases. These are briefly discussed as follows:

1. Determine the ontology domain and scope. This can be by creating a set of competency questions, which are questions the ontology can answer. These questions can be used for evaluating the ontology once created, but not sufficient on its own.
2. Reuse existing ontologies (if any). Instead of starting from scratch, it is beneficial to reuse and extend available ontologies. Sometimes it is a requirement to use existing ontologies.
3. Enumerate important terms. This step involves creating a vocabulary of terms from the domain of interest.
4. Define classes and class hierarchy. This step and the next step are linked. Classes and their properties are created before moving to the branch of next classes. Some possible approaches of developing a class hierarchy are:
    - Top-down approach, which starts with creating definition of the most general concepts and then their specialization (creating subclasses); the process is recursive for every class until it we reach the most specific definitions.
    - Bottom-up approach, which is the opposite of the above i.e. first define the most specific concepts (i.e. the leave classes of the hierarchy) and then group them into more general concepts (their root/superclass).
    - Combination of both approaches, which involves defining the more "salient" concepts first and then generalise and specialise them appropriately i.e. start with few general (or top-level) concepts and few specific (or bottom-level) concepts and identify the concepts of middle levels accordingly.
5. Define class properties (or slots). Properties describe the internal structure of concepts. A property can be (a) an "intrinsic" property that describes objects' physical characteristics; (b) an "extrinsic" property that describe abstract concepts; (c) object parts if the object is structured and (d) a relationship between individuals. All subclasses of a class inherit this class properties, hence properties should be defined on the most general class.
6. Define properties restrictions (or "slot facets"). Restrictions specify the value type, allowed values, the number of the values, and other features of the values.

Types of restrictions available depend on the ontology language used. The common restrictions are:

- Property cardinality specifies how many values a property can have. Some systems specify a minimum and maximum cardinality to precisely describe the number of property values.
- Property value-type defines what kind of values a property can have. Common value types are: String, Number (or more specific types such as Integer or Float), Boolean, Enumerated (a list of specific allowed values) and instance-type value, which allow definition of relationships between individuals and define a list of allowed classes from which the instances can come.
- Domain and range of properties. A list of allowed classes of instance-type are called a range of a property. A list of classes a property is attached to is called a domain.

7. Create instances. This is the last step and it involves (a) creating a class, (b) create an individual of the chosen class and (c) fill in property values.

There are many ontology development tools, such as ontology editors (e.g. protégé) and merging and mapping tools (e.g. SMART/PROMPT [153]) that can be used to develop ontologies and ontologies network. Ontologies when implemented, each entity (class, sub class, or instance) is given a unique (URI). After defining an initial version of the ontology, there are several ways to evaluate it. For example, evaluate it by using it in applications or by discussing it with domain experts, or a combination of both. Usually this process results in a refining or revising the initial ontology, which is likely to continue throughout the development cycle of the ontology. The created ontology facilitates knowledge querying and reasoning around a domain. Several tools have been developed for this purpose and the widely used ones are Jena API[13] and SPARQL query[14].

Ontologies facilitate the exploration of the unstructured social content available on online social spaces where concepts from these ontologies are used to add semantics (i.e. meaning) to a selected text as discussed next.

[13] https://jena.apache.org/
[14] https://www.w3.org/TR/rdf-sparql-query/

2.4.2 Ontology-based Semantic Annotation

Semantic annotation (also referred to as augmentation or tagging) is defined as "the process of tying semantic models and natural language together". It is the process of attaching semantics in the form of ontological concepts to a text to assist automatic interpretation of the meaning conveyed by this text [154].

Semantic annotations can be conducted manually i.e. by humans, usually domain experts; automatically, where a computer software conduct the annotations; or semi-automatically, where the annotations are done automatically, then human validate the annotations[14].

Due to the size of the UGC, automatic semantic annotation methods are more suitable and preferable. Automatic semantic annotation can be performed with Ontology-based Information Extraction (OBIE) [155]. OBIE involves natural language processing (NLP) of text to extract particular types of information (information extraction) related to a domain. This information is then connected with entities and properties from one or more ontologies which represents knowledge about the domain [38]. OBIE systems take as input text (e.g. a document) and at least one ontology and produce links (URIs) between the words in the document and ontology concepts. The input text is firstly processed with Natural Language Processing (NLP) techniques to extract linguistic information, such as sentences and phrases (e.g. verbs and nouns). To conduct this processing a set of regular expressions based rules can be utilised in the General Architecture for Text Engineering (GATE) [156] or in NLP text parses based on grammar rules such as the Stanford parser[15]. The processing output then is linked with input ontology with textual label matching (e.g. nouns). Extensive review of different types and purposes of semantic annotations is reviewed in this survey[14]. There are several tools that provide services for semantic annotations, such as DBpedia Spotlight[16] and OpenCalais[17].

2.4.3 Ontology-based Semantic Similarity Measures

Semantic similarity computes the likeness between concepts, it is understood as the degree of taxonomical proximity[157], [158]. The distance between two concepts is a

[15] https://nlp.stanford.edu/software/lex-parser.shtml
[16] https://www.dbpedia-spotlight.org/demo/
[17] http://www.opencalais.com/

numerical representation of how far apart two concepts are from one another in some geometric space and can be considered the inverse of semantic similarity[159], [160].

Several measures exploited the taxonomic parameters extracted from the “is a” taxonomy. The widely used approaches are the following[158], [161], [162]: Path-based measures, which assess similarity based on the number of taxonomic links and the minimum path length between two concepts present in a given ontology (e.g. Rada et al., 1989[163]); Path and depth based measures is similar to the former but it uses depth of concepts measured as well(e.g. Wu and Palmer 1994 [164]); Information content-based measures: quantifies the similarity between concepts as a function of the information content (IC) that both concepts have in common in a given ontology (e.g. Resnik, 1998 [165]); Feature-based measures: estimate similarity according to the weighted sum of the number of common and non-common features (e.g. Tversky, 1977 [166]) and Hybrid measures, which a combination of the measures mentioned above (e.g. Zhou et al, 2008 [167]). An extensive review of these measures can be found in [159], [160].

Some measures rely heavily on the information available in the ontology the concepts that are compared belong to. For example, feature-based measures consider taxonomic and non-taxonomic information modelled in ontology, but mostly rely on non-taxonomic features that are rarely found in ontologies which impacts the weighting parameters. When there is a risk of having minimum or no information about the concepts in the given ontology measures such as path-based are suitable and guaranteed to work, such as the measure proposed by Rada (shortest-path measure as distance between concepts). Rada [163] defined the conceptual distance between two concepts in the “is-a” hierarchy relationships as the length of the shortest path connecting the two concepts. In this measure the semantic distance is computed by counting the number of edges between concepts in the taxonomy. Research suggests “a robustness of the semantic distance approaches thus far” [159].

Several tools and libraries have been proposed for these measurements. For example, in W3C, a list of these tools is available citing the most popular ones including the Semantic Measures Library & ToolKit (SML)[18], which is an open source Java library

[18] http://www.semantic-measures-library.org/sml/

and tools for the study and computation of semantic measures It is described as the “most complete and versatile” software library reported in the literature[160].

2.4.4 Using Semantic Techniques for Social Clouds Diversity

This section reviews related work that used semantic techniques for the analysis of social clouds to understand and/or measure diversity. The work has been mentioned in a previous section in this thesis, namely section 2.3.1 (see Table 2.1), but they are mentioned here with a focus on the semantic web techniques they used for understanding diversity.

Microblogging behaviour modelling. Gao et. al [26] used and extended their Twitter-based User Modelling Framework[168] to compare the microblogging behaviour of users from different cultures (US and China), who are using different microblogging services - Twitter and Sina Weibo respectively. The comparison included semantic content analysis. This was to understand what concepts and topics the users of those platforms mentioned in their tweets. This was with regards to three concepts - person, locations and organisations. The framework uses semantic enrichment and linkage to enhance the construction of user profiles. For semantic enrichments, they utilised OpenCalais. After that, a linkage process is conducted to further enrich the semantics of tweets by linking the identified entities to external Web resources, news articles. Finally, based on the semantic enrichment and linkage, the framework facilitates user modelling for generating hashtag-based, entity-based, and topic-based profiles.

Opinion modelling. The work by Maynard et al. [30], which is part of the ARCOMEM project[19] that uses social media content to enrich the archiving of media-related Web archives and political archives. The goal was to understand the sentiment of opinions and their dynamics on crucial social events. Their sentiment analysis application was developed in GATE, where it was used for extracting named entities, terms and events and to detect opinions about them (as described in their work [169]). They aggregated opinions for a given document and ranked its “interestingness” based on the diversity of the opinions it contains. The more diverse the opinions, the higher the interestingness score.

[19] ARCOMEM EU Project: https://cordis.europa.eu/project/rcn/97303_en.html

Viewpoints modelling. Despotakis [133] developed ViewS for exploring and extracting the diversity of individuals' viewpoints from semantic annotations and enrichments of user comments around YouTube videos. The framework was proposed to analyse viewpoints reflected in user comments with regards to interpersonal communications in terms of emotions (represented with WordNet-Affect[170] and body language (represented with Body Language Ontology which is part of Amon Ontology[171]). The framework facilitates text processing (underpinned by Stanford parser), semantic enrichment and annotations with the aforementioned ontologies. Users' viewpoints were compared (based on their age, gender and location) via visualisation of the ontology entities associated with annotated comments against the used ontologies. This was used to inspect the users' domain differences as captured in their comments.

Although Semantic Web techniques have not been extensively utilised, they show a great potential for diversity modelling. For example, they showed potential for assisting the process of sense making of user digital traces including the user comments[47]. Linked data showed potential for enriching the user modelling interactions when modelling the cultural awareness of users[172].

2.4.5 Discussion

Diversity is an ill-defined domain that requires a whole picture (range of possible values) to compare to, an ontology (with regards to domain and user) can provide this picture. Also, diversity could require qualitative explanations on top of the quantitative findings, hence Semantic Web techniques are preferable in this matter and they are chosen to be the underpinning techniques for diversity modelling.

Ontologies serve as the backbone for semantic web techniques. Several domain ontologies are available, and more are created continuously. However, there are limitations in the ontologies that model users. There is a lack of consistency in terminology between the existing ontologies. Also, they have limitations in terms of structure and hierarchy that hinder their applicability for diversity explorations. More in this point is discussed later in section 3.5.2.

The frameworks reviewed in section 2.4.4 above shows how semantic techniques assisted the process of understanding and measuring diversity. However, they encounter limitations. Building on top of discussion in section 2.3.2, the closest work to this research is the one proposed by Despotakis. His framework does not

automatically measure diversity. The framework maps the comments to the ontological entities with a focus on domain differences, but it relies on an analyst to visualise the spread and nature of distribution of ontology entities from annotations against used ontologies. This visualisation is the approach for identifying the domain differences. Also, the research neglected the user diversity with regards to their attributes. This research aims at filling in these gaps.

2.5 Summary

This section summarises the discussions throughout this chapter highlighting potential inspiring this research and limitations to fulfil.

Social clouds exhibit rich and diverse source of information about the domain discussed (textual comments) and users who made the comments (user profiles), which facilitate the representation and measurement of diversity.

This research accepts the definition of diversity as proposed by Stirling Diversity Framework as three properties - variety, balance and disparity, but not as aggregated as research shows the potential of having more than one diversity indicator rather than one overall (black box) value.

The diversity research in the management literature and social science provides a rich theoretical background for the exploration of user diversity perspective. The Diversity Taxonomy's (discussed in section 2.2.1.1) three levels, surface-level attributes, knowledge attributes and deep-level attributes, inform the proposal of an ontology for the user diversity attributes as discussed later in section 3.5.2. This ontology is intended to underpin the exploration and measurements of user diversity.

The maturity of Semantic Web techniques, especially ontologies enforces and supports the applicability of the proposed approach for modelling diversity in this research. Tools and domain ontologies are available to be exploited. This research focusses on the use of ontologies for semantic annotations of user comments and user profiles attributes. Ontologies for this research also contributes to and underpins the measurements and analysis of diversity in a social cloud. This is discussed later in chapter 3 section 3.2.

Related research on social clouds diversity is inspiring, though exhibit limitations, especially the lack of automation of diversity measurements. This research aims at

representing and automatically measuring social clouds diversity based on two diversity perspectives - domain and user. The domain diversity perspective is a subject domain that is captured in user comments and can be represented and annotated by a related domain ontology for domain diversity measurements. The user diversity perspective is based on user diversity attributes captured in the user profiles. This is for user diversity measurements. This work adds on the work on social clouds diversity captured in textual comments as little has been done compared to the microblogs, and it aims at resolving the discussed limitations with regards to diversity measurements.

To conclude, underpinned by semantic techniques and informed by diversity definitions from the Stirling Diversity Framework, this research proposes a computational model that facilitates a systematic representation and an automatic measurement of two diversity perspectives in a social cloud - domain and user. The proposed model and approach to apply it for measuring diversity are presented in next chapter.

Chapter 3 : The Proposed Semantic Approach

This chapter sets the foundations for the diversity model. It proposes the semantic approach for modelling diversity, upon which indices are developed to measure diversity in a collection of contributions within in a social cloud. A formal model is defined to represent components of a social cloud and for the application of semantic techniques in measuring diversity against a selected ontology. Stirling's Diversity Framework [113], in which three properties - variety, balance and disparity - are defined, is used as the basis for the development of the diversity indices for measurement. This enables further analysis that can be conducted by using different combination of indices to extract diversity patterns. Furthermore, the chapter discusses the ontologies to be used and describes a **Se**mantic **D**riven **D**iversity **A**nalytics **T**ool (**SeDDAT**), which operationalises the model and enables an analyst to obtain diversity profiles for a social cloud with regards to the domain and user diversity perspectives.

3.1 An Overview of the Approach

Figure 3.1 shows an overview of the semantic driven approach to represent and measure diversity in a social cloud. The initial step in the approach is the **social cloud collection**, which consists of user comments, profiles of users who made the comments, and metadata about the digital objects (e.g. title and URI). Based on the chosen diversity perspectives - domain or user - the user comments or selected attributes from the user profile are fed to the next step, **semantic annotations**. This step uses relevant underpinning ontologies (representing the domain or the user attributes) for semantic tagging: ontological entities (entities URIs) are linked to the associated words in comments or to selected user attribute. With the knowledge of which comments covering which ontology entities, the next step is **diversity profiling** where diversity indices are calculated for any entry point (selected entity) in the underpinning ontology. The output of this step captures diversity properties indicating the level of diversity in the social cloud with regards to the selected perspective (i.e. domain diversity or user diversity). Finally, **diversity analysis** covers interpretation of the indices (individually or in combination), select another perspective, and /or select another entry point in the ontology for further profiling. This will provide a deeper understanding of the diversity of the selected social cloud.

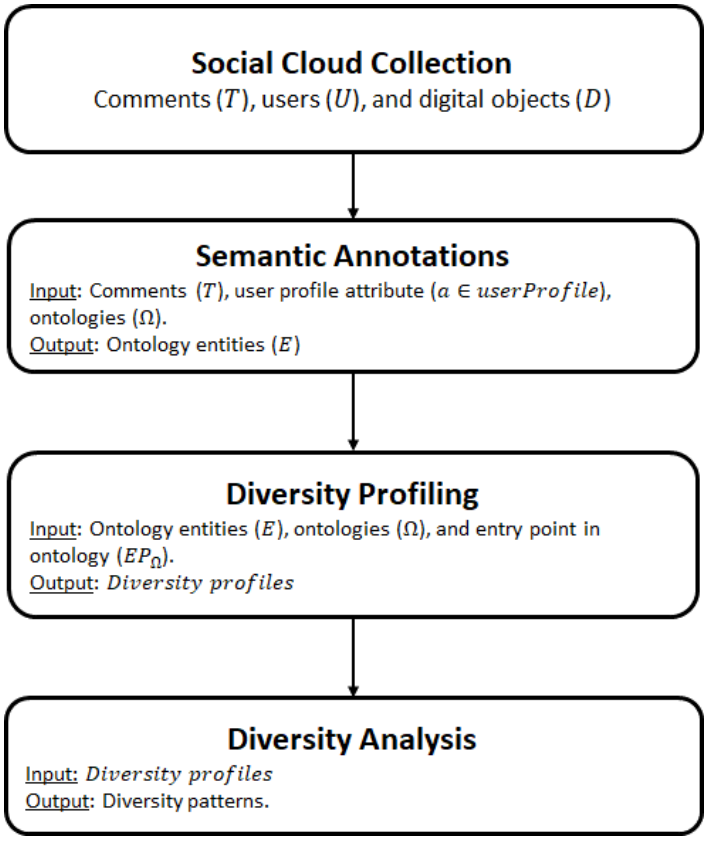

Figure 3.1 Main steps in the semantic driven approach to represent and measure diversity in a social cloud.

3.2 Formal Model for Diversity in a Social Cloud

This section describes a formal model that defines (a) the components and elements to be used in the first two steps of the approach as outlined in the above section, and (b) the formulae used for calculating diversity indices for the diversity profiling step.

3.2.1 Components in the Model

Social Cloud. Social cloud consist of a set $\boldsymbol{T} = \{t_1, t_2, \dots, t_n\}$ of textual comments t_i, $1 \leq i \leq n$ which have been created by a set $\boldsymbol{U} = \{u_1, u_2, \dots, u_m\}$ of users u_i, $1 \leq i \leq m$ while interacting with a set $\boldsymbol{D} = \{d_1, d_2, \dots, d_k\}$ of digital objects d_i, $1 \leq i \leq k$.

Social cloud is generated in social spaces and captures digital traces of users U who interact with digital objects D (e.g. videos, images or online articles). Users can express their views around these digital objects in different formats, but commonly textual comments T. An example of a social space that generates a social cloud is YouTube where users can make comments on the videos they watch. In a similar way,

social clouds are created around news articles in online newspapers collecting readers' opinions on the articles, or around TED videos where the audience indicates what they find interesting in the videos.

Social cloud components. There are three main components in a social cloud – digital objects, comments and users.

Every **digital object** $d \in D$ has a set of users U_d who commented on it; i.e. every user $u \in U$ has written at least one comment $t \in T$ on d. A user u can comment on more than one digital object d.

Every **comment** $t \in T$ is associated with a user $u_t \in U$ and a digital object $d_t \in D$ where u_t has made the comment t while interacting with the digital object d_t in a social space. The set of textual comments created by a user $u \in U$ are denoted with the set T_u, where $T_u \subseteq T$ and $T_u \neq \emptyset$. Similarly, the textual comments associated with a digital object $d \in D$ are denoted with the set T_d, where $T_d \subseteq T$ and $T_d \neq \emptyset$.

Every **user** $u \in U$ is associated with a **user profile** $userProfile(u) = \langle a_1, a_2, \dots, a_{n_u} \rangle$ which includes attributes that characterise the user u (e.g. age, gender, culture, or expertise), where n_u is the number of attributes that are associated with this user u. Users who are grouped based on an attribute $a_i \in userProfile$ (e.g. gender or age groups) are denoted as U_{a_i}, and $U_{a_i} \subseteq \mathrm{U}$. More than one profile attribute $a \in userProfile$ can be used to create user groups.

Ontological underpinning. Diversity only has meaning when it is compared against a whole picture (i.e. the whole range of possible values). In this work, we adopt an ontology-driven approach where ontology is assumed to provide the whole picture with regard to the domain or the user attributes. In the context of social clouds, an ontology may be selected to represent a specific perspective. We consider two **perspectives**: (a) domain perspective - how diverse is the social cloud with regard to the domain that is covered by the comments, and (b) user perspectives – how diverse is the social cloud with regard to its users' attributes. For any perspective, the selected ontology Ω to represent it is structured as $\Omega =< E_\Omega, H_\Omega >$, where:

E_Ω is a set of ontology entities: $E_\Omega = C_\Omega \cup I_\Omega$; C_Ω is a set of classes that represent the main categories of the ontology (e.g. domain or user) and I_Ω is a set of instances representing the individuals which belong to the classes; $C_\Omega \cap I_\Omega = \emptyset$.

H_Ω is a set of hierarchical relationships between ontology entities: $\mathrm{H}_\Omega = \{\mathrm{subClassOf}, \mathrm{instanceOf}\}$, where $subClassOf(e_i, e_j),\ e_i, e_j \in C_\Omega,\ e_i \neq e_j$ defines that e_i is a subclass of e_j; and $instanceOf(e_i, e_j),\ e_i \in I_\Omega, e_j \in C_\Omega$ defines that e_i is an instance of class e_j.

The minimum requirement for defining the selected ontology Ω is to provide the set of ontology entities and the taxonomy. This is the main input for the diversity indices (section 3.2.3). The taxonomy is in the form of a tree where every entity e has one parent class in C_Ω. This excludes lattice. If an entity e in E has more than one parent class, in the measurements one of e parent classes will be randomly selected and used for the corresponding diversity measurements.

Semantic annotations are obtained by parsing the comments and tagging any relevant word with at least one entity of the selected ontology Ω. The resulting set of entities used for annotation is denoted by $E = \{e_1, e_2, \dots, e_n\}$, where n is the number of entities used for semantic annotations and $E \subseteq E_\Omega$.

3.2.2 Two Perspectives of Diversity

The main focus of domain diversity is to map out how diverse or otherwise the collected comments have covered the domain concepts, classes and instances (i.e. entities) in the selected domain ontology. For the **domain diversity perspective** represented by a domain ontology Ω_{domain}, every comment $t \in T$ written by a user u_t on a digital object d_t is tagged with a set of entities $E_t = \{e_1, e_2, \dots, e_{n_t}\}$, where n_t is the number of entities associated with the comment t and $E_t \subseteq E_{\Omega_{domain}}$.

The following sets of entities are distinguished to assist the diversity calculations:

- The set of ontology entities associated with all comments in T is denoted as $E_{domain} = \bigcup_{i=1}^{n} E_{t_i}$, where n is number of comments in the social cloud.
- The set of entities from annotating all the comments T_d on a digital object d is denoted as $E_d = \bigcup_{i=1}^{n_d} E_{t_i}$, where $n_d$is the number of comments on the digital object d and $E_d \subseteq E_{domain}$.
- The set of entities from annotating the comments T_u by a user u is denoted as $E_u = \bigcup_{i=1}^{n_u} E_{t_i}$, where n_u is the number of comments made by the user u.

- The set of entities that belong to a category $c \in \mathrm{C}_{\Omega}$ from annotating a user's comments T_u is denoted as E_{c_u} and $E_{c_u} \subseteq E_u$. Frequency of entities per user $u \in U$ for a category $c \in \mathrm{C}_{\Omega}$ is denoted as $f_{c_u} = |E_{c_u}|$.
- The set of entities from annotating comments of a user group $U_a \in U$ is denoted as $E_{c_U} = \bigcup_{i=1}^{m} E_{u_i}$, where m is the number of users in the social cloud who have the profile attribute $a \in userProfile$ (e.g. gender, age group or cultural cluster). Frequency of entities per user group U_a for a category $c \in \mathrm{C}_{\Omega}$ is denoted as $f_{c_U} = |E_{c_U}|$.

The main focus of user diversity is to indicate how diverse or otherwise the users in the social cloud with regards to at least one user attribute. For the **user diversity perspective** represented by a user ontology Ω_{user}, a user attribute $a \in userProfile$ (e.g. age) is tagged with at least one entity e_u, where $E_a = \{e_1, e_2, \dots, e_{n_a}\}$, n_a is the number of entities associated with the user attribute a and $E_a \subseteq E_{\Omega_{user}}$. The set of ontology entities associated with the attribute a for all users is denoted with $E_{user} = \bigcup_{i=1}^{m} E_{a_m}$, where m is the number of users in the social cloud.

3.2.3 Diversity Indices

The proposed model adopts the diversity definitions from the Stirling Diversity Framework (as discussed in 2.2.3), which includes three properties - **variety, balance and disparity** - to characterise diversity, as "each is a necessary but insufficient property of diversity"[113]. This research adds another diversity property, *coverage,* that was inspired by limitations discovered during an early part of the investigation (i.e. in proportions measurements for balance). As proposed by the Stirling Diversity Framework (discussed in 2.2.2.3), to measure diversity of a system (in this case, a social cloud), it is required to identify **system elements** and **main categories**[20] of these elements[31].In this case, the system elements are relevant ontological concepts that are represented by the set of entities (i.e. E) associated with (a) the user comments T and (b) user characteristics captured by the $userProfile$ of those users who made the comments (i.e. E_{domain} and E_{user}). The elements' main categories are their classes in the ontology Ω that represent a diversity perspective. By applying

[20] A category is defined in the Cambridge Dictionary as a group of people or things that have similar features. Here a category refers to a class in a given ontology. A class is defined in section 2.4.1.1.

Stirling's Framework, diversity in a social cloud can be measured for two perspectives - domain and user. These mean: how diverse are the comments with regards to a subject domain for discussion (domain diversity) or how diverse are the users who made the comments with regards to a specific user population (user diversity). For example, to explore user diversity perspective, the system elements can be user age and the categories for those elements are the different age groups that are captured in a given ontology (classes), such as teenager and young adult or 12-17 and18-21. This means identify who and how diverse are the age groups who wrote the comments.

To measure diversity, an entry point EP_Ω in the given ontology Ω is selected, where EP_Ω is a class $c \in C_\Omega$ in the ontology Ω that has a number n of sub-classes/sub-categories and $n \geq 1$. All subsequent diversity measurement will be conducted from the chosen class.

Given an ontology Ω representing the chosen perspective, a set of entities E from Ω linked to user comments T or user attributes from $userProfile$, and a class in the ontology taxonomy providing the entry point category EP_Ω for which diversity will be calculated, the diversity properties and associated indices (discussed in section 2.2.2.1) are defined as follows.

3.2.3.1 Variety Index (v)

As the main categories for the diversity perspectives are predefined (i.e. classes in the given ontology Ω), variety is the number of sub-categories of EP_Ω which have at least one entity e from E. In other words, variety is the number of *non-empty* sub-categories of EP_Ω, hence variety is a positive integer number. The higher the number, the higher the diversity.

$$v\,(\Omega, E, EP) = |K| \qquad (1)$$

where K is set of sub-categories of EP_Ωthat has at least one entity in E and $0 \leq v \leq n$, where n is the number of sub-categories of EP_Ω.

Variety $v = 0$ indicates that none of the entities e in E are from sub-categories of EP_Ω i.e. this branch of ontology Ω is not presented in the comments or user attributes within the given social cloud.

Consequently, variety for the domain and user perspectives, respectively, is defined as follows:

Domain variety is the number of domain sub-categories of $EP_{\Omega_{domain}}$ which have been mentioned at least once in the user comments i.e. have at least one domain entity from annotating the comment in E_{domain}. High domain variety indicates that the entities from annotating the user comments in E_{domain} covered most or all the high-level aspects of the domain under EP_{Ω}.

User variety is the number of user sub-categories of $EP_{\Omega_{user}}$ (e.g. age groups or culture groups) which have been linked via semantic annotations to one of the user profile attributes $a \in userProfile$. High user variety indicates that the commenters in the selected social cloud are coming from most or all the user categories under $EP_{\Omega_{user}}$.

For example, Figure 3.2 illustrates a branch of an ontology that has been used for semantic annotations. This ontology branch is under the entry point EP_{Ω}which has 4 sub-categories (n=4) c_1, c_2, c_3 and c_4. The light-blue circles are the *distinct* entities (subclasses and/or instances) that are used for the semantic annotations. One entity under a sub-category is enough for this category to be identified for variety (e.g. c_2). An entity can be used more than once to semantically tag a record related to a perspective. For this branch, variety is 3 as the sub-category c_1 is not covered, hence excluded from diversity measurements.

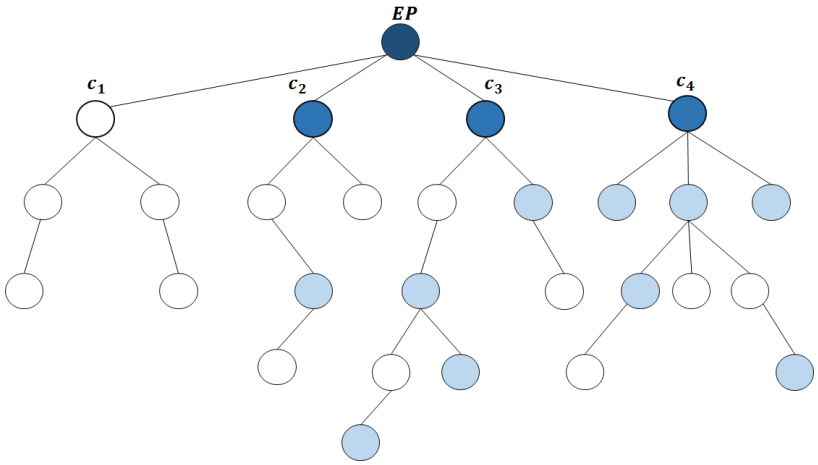

Figure 3.2 An example of an ontology branch used for semantic annotations of a selected social cloud content around a digital object *d*. Shading a node in light blue indicates its occurrence in the comments. The branch is headed by the entry point $\boldsymbol{EP_{\Omega}}$, where 3 sub-categories have been identified via semantic annotations.

If the ontology branch above is a domain ontology Ω_{domain}, it means that 3 domain categories are covered in the user comments and identified for domain variety. Similarly, if it is a user ontology Ω_{user}, it indicates that three categories of a user attribute (e.g. age groups) are identified for user variety.

As this is a top-down approach where categories are predefined, identification of variety v assists the measurements of the other diversity properties - balance, coverage and disparity. The diversity measurements of these indices are conducted on the tree under each category identified for variety i.e. excluding the category itself from calculations. In particular, the category is excluded from the count of available entities in that branch for measuring balance and coverage. This is to have generic indices for both diversity perspectives. An inclusion of the category can result in coverage being lower than it actually is for the user diversity perspective due to the nature of the user attributes and their categories.

3.2.3.2 Balance Index (b)

Balance shows whether the distribution of entities in E is evenly spread across their sub-categories under EP_Ω. Balance requires calculating the extent of each sub-category c_i of EP_Ω (i.e. proportion p_{c_i}) covered in the set of entities E. The proportions of the categories are aggregated to examine evenness in entities distribution across EP_Ω sub-categories. Only the number of distinct ontology entities is used to calculate p_{c_i}. The formula is based on the widely used index for balance (as discussed in 2.2.2.1), Shannon Entropy Index [123]. The more even the spread, the higher the value for the balance, and the higher diversity[113].

$$\boldsymbol{b}\ (\Omega, E, EP) = -\sum_{i=1}^{n} \boldsymbol{p_{c_i}} ln(\boldsymbol{p_{c_i}}) \qquad (2)$$

where n is number of sub-categories of EP_Ω and $\boldsymbol{p_{c_i}} = \frac{|E_{c_i}|}{|E|}$ is the proportion of distinct entities in E that belong to the ontology tree headed by the sub-category c_i against the total number of entities in E. The proportions are fractions that sum to unity[119].

Empty sub-categories of EP_Ω (i.e. not included in v) are excluded from calculations to avoid undefined values due to the logarithm ln used in the Shannon formula. This is following the definitions of Shannon Entropy for zero probabilities [123]. When only one sub-category is covered (i.e. $v = 1$), balance $b = 0$ [123].

Consequently, balance for both diversity perspectives is calculated as follows:

Domain balance shows whether the domain entities in E_{domain} associated with user comments are evenly distributed in their domain sub-categories of $EP_{\Omega_{domain}}$. High domain balance indicates that the domain categories identified for domain variety have even distribution by the comments on the digital object(s).

User balance shows whether users who commented on a digital object are distributed evenly across their user categories of $EP_{\Omega_{user}}$ based on E_{user} from annotating their profile attribute $a \in userProfile$. High user balance indicates that users' categories identified for user variety are represented evenly in the social cloud around the digital object(s).

For example, as shown in Figure 3.2, there are 10 entities covered from annotations i.e. $E = 10$. The sub-category c_1 is an empty category i.e. $p_{c_i} = 0$, hence it is not covered in diversity measurements. The sub-category c_2 has 4 entities in total, one of which has been covered in the semantic annotations, hence the proportion p_{c_2} of the category c_2 is $1/10$ i.e. p_{c_2}=0.1. Likewise, $p_{c_3} = 0.4$ and $p_{c_4} = 0.5$. These proportions sum to unity. This gives the balance index b (formula 2) value of 0.94.

Figure 3.3 shows two different entity distributions for the same branch (i.e. the same ontology branch is used to annotate two different social clouds). The branch to the right shows more even distribution than the one to the left. Although the sub-category c_4 is covered slightly better in the left branch, it is the only category that is covered best in this branch, which led to uneven overall distribution.

In the left branch: The total number of entities covered $E = 11$. Therefore, proportions are p_{c_1} and $p_{c_2} = 0.09$; p_{c_3} =0.18 and $p_{c_4} = 0.64$. This makes balance b =1.03.

In the right branch: $E = 17$ and the proportions are as follows: p_{c_1} and $p_{c_2} = 0.18$; p_{c_3}=0.45 and p_{c_4}=0.55. This gives b =1.30. Therefore, the right branch has more even distribution of E across the covered categories.

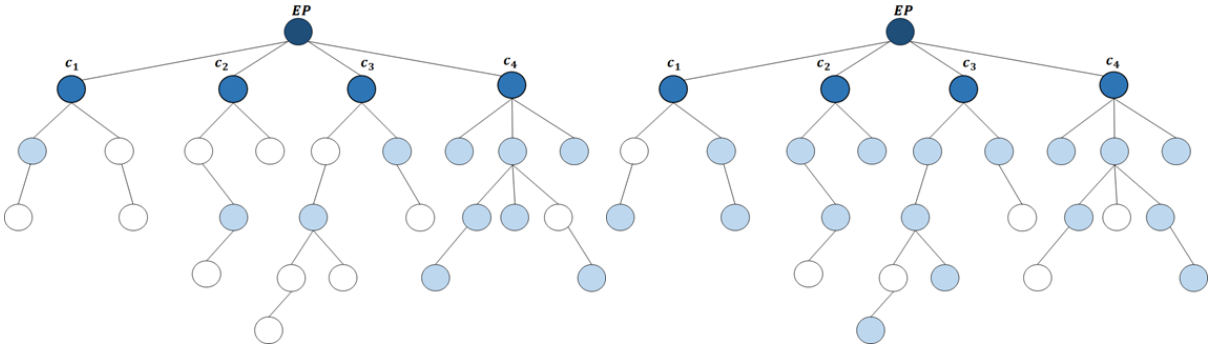

Figure 3.3 Two examples of the same ontology branch. The branch to the right has even entities distribution compared to that to the left.

3.2.3.3 Coverage Index (r)

Although the Shannon Entropy index (formula 2) for balance indicates the level of evenness across the categories, it only considers the spread of entities E covered by the annotations. This discards the density of coverage in each category against the ontology. For example, in the Figure 3.2 and Figure 3.3, entities not covered by the annotations are shown (uncoloured circles). In order to measure the density of the coverage, a separate coverage index, inspired by the proportion's definitions for balance and Shannon Entropy Index calculations, is proposed.

As the main categories and associated entities are predefined in the given ontology Ω, coverage shows how much the sub-categories of EP_Ω are represented in E. It calculates the representational proportion rep_{c_i} for each sub-category of EP_Ω in E against all the available entities in this sub-category. High representational proportion rep_{c_i} of a category c_i, indicates a good coverage of this category in E. Overall coverage r is calculated as the mean representational proportion of sub-categories identified for variety v. The higher the coverage, the stronger the diversity.

$$\boldsymbol{r}(\Omega, E, EP, v) = \frac{1}{v}\sum_{i=1}^{n} \boldsymbol{rep_{c_i}} \qquad (3)$$

where $\boldsymbol{rep_{c_i}} = \frac{|E_{c_i}|}{|c_i|}$ is the proportion of distinct entities in E that belong to the ontology tree headed by the sub-category c_i against the total number of entities in *this sub-category.*

The value of sub-category proportion is $0 \leq rep_{c_i} \leq 1$. A proportion rep_{c_i} of value 1 indicates a full coverage of the category c_i . A proportion of value zero indicates an empty category i.e. not covered in the set of entities E. When proportions of all the identified categories equal to 1 (i.e. full coverage), coverage r equals the number of

these categories (i.e. equals to variety v), which indicates that $0 \leq r \leq v$. The proportion $\boldsymbol{rep}_{c_i}$ for coverage r is labelled as *representational proportion* to distinguish it from the proportion $\boldsymbol{p}_{c_i}$ for the balance index b.

Consequently, coverage for both diversity perspectives is defined as follows:

Domain coverage indicates how well the domain is covered by the user comments. The higher the value of r, the better the domain coverage and diversity.

User coverage shows how well represented the user groups are in the social cloud. The higher the value of r, the better the user coverage and diversity. For example, when the coverage of age groups (main categories) is high (i.e. users in the social cloud have different age), it indicates that the users represent different generations ensuring age diversity.

For example, see the left ontology branch in Figure 3.3. The category c_1 has only one entity covered out of 4 i.e. $rep_{c_1} = 0.25$. Likewise, rep_{c_2} =0.25; rep_{c_3} =0.29 and rep_{c_4} =0.88. This gives average coverage r =0.42.

Similarly, in the right branch: rep_{c_1}, rep_{c_2} and rep_{c_4}=0.75 and rep_{c_3} =0.71. This gives coverage r =0.74. Therefore, the coverage in the right branch is higher than that on the left. If this is a domain ontology, the user comments associated with the right branch have better coverage of the domain. Similarly, if this is with regards to the user perspective, the user categories are better represented in the social cloud associated with the right ontology branch.

3.2.3.4 Within Disparity Index (d_w)

This research considers two types of disparity - *within and across disparity. Within disparity* is the manner and degree to which the *system elements* (entities in E) may be distinguished from the others. *Across disparity* is the manner and degree to which the *main categories* (sub-categories of EP_Ω) may be distinguished from the others. In both types, *distance* is what is being measured whether it is between system elements or main categories.

Within disparity is calculated by using each sub-category's dispersion i.e. how scattered the entities are in E under each sub-category c_i of EP_Ω. To calculate dispersion, each sub-category of EP_Ω is treated as a cluster, where each sub-category

with its entities in the ontology Ω are one independent cluster. The formula adopts Hall-Ball internal cluster validation index[173], which gives the mean dispersion across all the sub-categories under EP_Ω that are identified for variety v. The higher the number, the more dispersed the coverage.

$$\boldsymbol{d_w}(\Omega, E, EP, v) = \frac{1}{v}\sum_{i=1}^{n} dis(c_i) \quad (4)$$

where $dis(c_i) = \frac{1}{|E_{c_i}|}\sum_{j=1}^{|E_{c_i}|}\left(\min_{\forall \boldsymbol{p}}(path_p(e_j, m_{c_i}))\right)^2$ is the dispersion calculated based on the shortest path between each of the entities in E that belong to sub-category c_i and the medoid[21] m_{c_i} of the sub-category c_i. A medoid $m_{c_i}$was selected instead of a centroid to ensure that the selected entity is within the cluster (i.e. sub-category).

The shortest-path measure by [163] is selected to measure the semantic distances between the entities within their category. This is for two reasons: (a) to have a generic metric that can measure distance between entities within any selected ontology (refer to discussion in section 2.4.3); and (b) it complies with the within disparity definition i.e. how scattered/dispersed the entities are within their categories.

Empty sub-categories are given dispersion value of zero. As within disparity considers distance within sub-categories, overall dispersion is counted only for categories identified for variety to eliminate the empty categories.

When there is only one category c_i covered (i.e. variety v =1), within disparity is equal to this category's dispersion d_w= $dis(c_i)$.

The medoid m_{c_i} must be identified first to measure dispersion $dis(c_i)$ and over all within disparity d_w. The medoid is identified as the entity with the minimal average distances to all the other entities within their category. When there is only one entity e in E covered for a category c_i, there is no need to identify the medoid m_{c_i} and dispersion $dis(c_i) = 0$, as semantic distance (shortest path) between an entity and itself is zero. When there are only two entities in E covered for a category c_i, this category's medoid m_{c_i} can be either. Similarly, when the average distances between the entities in E covered for a category c_i are equal either entity can be the medoid m_{c_i}.

Within disparity for both perspectives is as follows:

[21] A medoid is the most centrally located item in a cluster that has minimal average distances to all the other items in the cluster[181].

Domain within disparity distinguishes the domain entities E_{domain} within their domain sub-categories of EP_{domain} from each other. High domain dispersion indicates that the comments and associated domain entities are scattered across their categories i.e. comments covered different aspects of the domain.

User within disparity distinguishes the users from each other based on their associated profile entities E_{user} and within their sub-categories of $EP_{\Omega_{user}}$ .The higher the dispersion, the higher the within disparity, the more different the users.

Using the example in Figure 3.2, within disparity is calculated for the three categories identified for variety. To calculate dispersion, each sub-category of EP_{Ω} is treated as a cluster. For instance, as shown in Figure 3.4, the sub-category c_4 with its entities are one independent cluster. In each cluster (i.e. sub-category) the distance between the entities from annotations is measured based on the shortest path between them and the medoid - the entity with the minimal average distances to all the entities in a category. The mean dispersion of all the clusters (sub-categories) is the *within disparity* d_w. The farther (dissimilar) the entities are from each other, the higher the dispersion and the higher the within disparity.

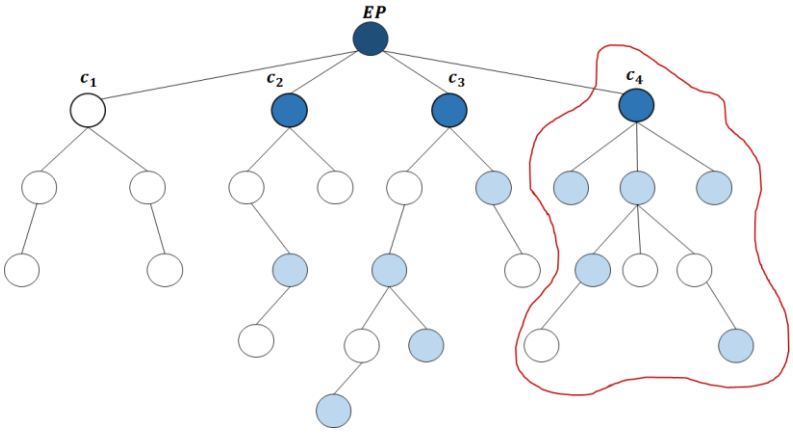

Figure 3.4 A sub-category is treated as a cluster for dispersion calculations. Within disparity is the mean of dispersion of all clusters (sub-categories).

Figure 3.5 shows the same ontology branch used with two different social clouds, where the one on the right is more dispersed compared to the one on the left. The entities in the right branch are more scattered within their categories.

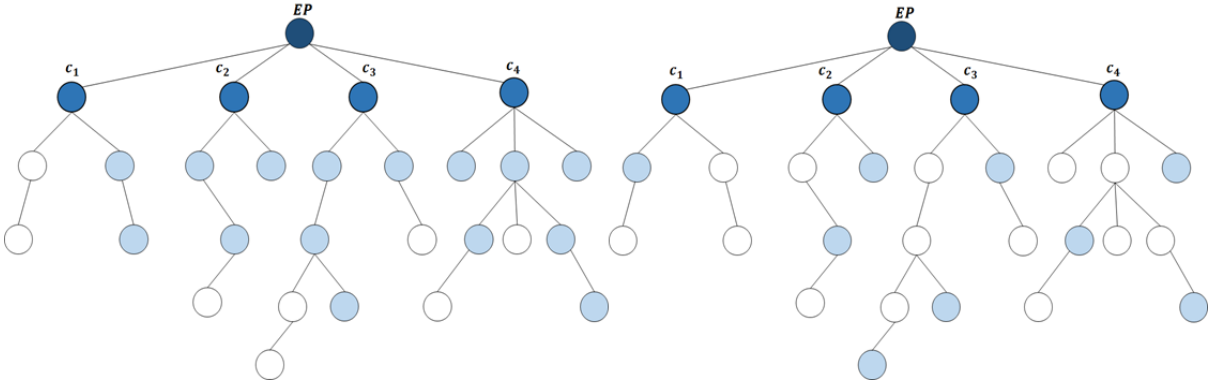

Figure 3.5 Same ontology branch where the one to the right is more dispersed (entities are scattered within their categories) compared to the one to the left.

To compare the branches shown in Figure 3.5 in terms of within disparity, medoids must be identified for the two branches (their medoids are highlighted in Figure 3.6). Consequently, in the right branch, the category c_1 has only one entity, hence dispersion for this category $dis(c_1) = 0$. For the category c_2, its dispersion $dis(c_2)$ =4.5. Similarly, dispersion $dis(c_3)$ =16.3 and $dis(c_4)$ =12. Therefore, within disparity d_w =8.21.

In the left branch, categories dispersions for the 4 covered categories are: $dis(c_1)$ =0.5 $dis(c_2)$ =3; $dis(c_3)$ =6.25 and $dis(c_4)$ =10.67.Therefore, within disparity for this branch is d_w =5.11.

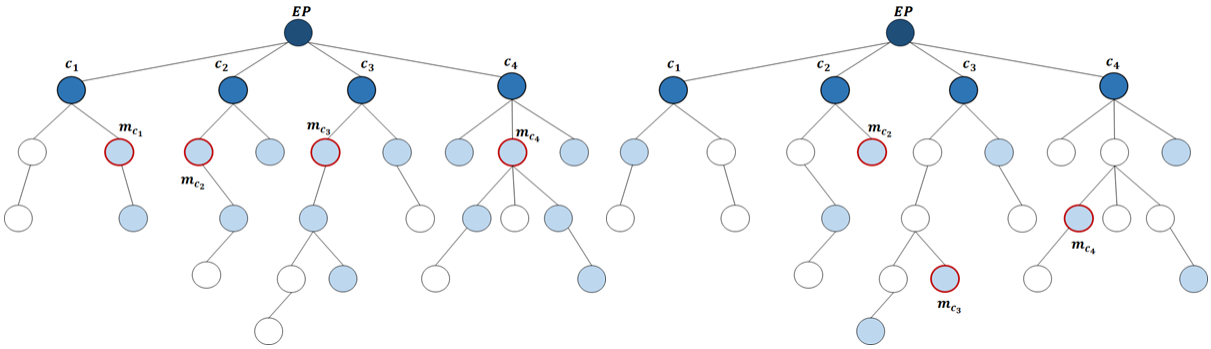

Figure 3.6 The same branch with two social clouds showing the medoid for each sub-category for dispersion and within disparity measurements.

3.2.3.5 Across Disparity Index (d_a)

This involves distinguishing the non-empty sub-categories (i.e. identified for variety) of EP_Ω from each other. That is, how distinctive the categories are from each other by measuring (pair-wise) distance between the identified categories in terms of frequency (i.e. number of distinct times a category is mentioned in the comments by a user or user group). Inspired by the work in [59], [115], [174], this is achieved by:

i. Select the main and secondary diversity perspectives. The categories to be distinguished from each other are those from the main perspective (e.g. domain categories); and within each of these categories the secondary perspective (e.g. distinct domain entities mentioned by each user or user group) is brought in.

ii. Calculate frequency of each sub-category $c_i$of EP_Ω in E. Category frequency $f_{c_{iuj}}$ is the number of times a category c_i is mentioned by an individual user u_j or a user group f_{c_U}. It is calculated by counting the number of distinct entities of a category in E.

iii. Construct a frequency vector $freq_{c_i}$ per sub-category c_i in the form of $<f_{c_{iu_1}}, f_{c_{iu_2}}, \dots, f_{c_{iu_m}}>$ for all individuals who mention this category or $<f_{c_{iU_1}}, f_{c_{iU_2}}, \dots, f_{c_{iU_l}}>$ for all user groups, where m is the number of users who interacted with the digital object d, and l is the number of user groups (e.g. age groups) who interacted with the digital object d.

iv. Calculate the Cosine Similarity Index of every pair of frequency vectors with domain categories c_i and c_j:

$$\cos(freq_{c_i}, freq_{c_j}) = \frac{\sum_{k=1}^{n} A_k\, B_k}{\sqrt{\sum_{k=1}^{n} A_k^2}\ \sqrt{\sum_{k=1}^{n} B_k^2}}$$

where A_k and B_k are *k-th* elements in the vectors $freq_{c_i}$ and $freq_{c_j}$ respectively and $0 \leq \cos \leq 1$.

v. Calculate dissimilarity d_{ij} of every pair of categories c_i and c_j. Dissimilarity $d_{ij} = 1 - \cos(freq_{c_i}, freq_{c_j})$, where $0 \leq d_{ij} \leq 1$ and d_{ij} =0 when $\cos$ =1 and d_{ij} =1 when $\cos$ =0.

vi. Finally, calculate across disparity index, which is the average dissimilarity provided by the following formula:

$$\boldsymbol{d_a}(\Omega, E, EP, v) = \frac{1}{v(v-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}, i \neq j \qquad (4)$$

For the case studies, the domain diversity perspective is selected to be the main perspective and the user is the secondary one as the domain is richer due to the limited nature of the user profiles. This involves mapping the users associated distinct domain entities in E_{domain} to their domain sub-categories of $EP_{\Omega_{domain}}$, which facilitates

creating frequency vectors per category $freq_{c_i}$. A high value in across disparity indicates that the difference in the frequencies of the entities being mentioned by different users or user groups is very high.

Across disparity can be measured on two levels- **Individual** (per user) or **group** (per user group). Table 3.1 shows the mapping of the domain entities per user across the domain sub-categories of $\boldsymbol{EP_{\Omega_{domain}}}$ that are identified for domain variety.

Table 3.2 shows the mapping of the domain entities of a user group $\boldsymbol{U_{a_i}}$ across the domain categories identified for domain variety. As more than one profile attribute $\boldsymbol{a} \in \boldsymbol{userProfile}$ can be used to group users' entities. Across disparity $\boldsymbol{d_a}$can be measured more than once based on the entities of different user groups.

Each frequency column in Table 3.1 and Table 3.2 is a frequency vector $\boldsymbol{freq_c}$ for a category $\boldsymbol{c_i}$ of $\boldsymbol{EP_{\Omega}}$ (as can be seen in Table 3.1), where the Cosine Similarity Index compares the categories pair-wise.

Table 3.1 Mapping of user entities across categories of entry point.

	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	…	$\boldsymbol{c_n}$
$\boldsymbol{u_1}$	$f_{c_1u_1}$	$f_{c_2u_1}$	…	$f_{c_nu_1}$
$\boldsymbol{u_2}$	$f_{c_1u_2}$	$f_{c_2u_2}$	…	$f_{c_nu_2}$
⋮				
$\boldsymbol{u_m}$	$f_{c_1u_m}$	$f_{c_2u_m}$	…	$f_{c_nu_m}$
				$freq_{c_n}$

Table 3.2 Mapping of user groups' entities across categories identified for variety.

	$\boldsymbol{c_1}$	$\boldsymbol{c_2}$	…	$\boldsymbol{c_n}$
$\boldsymbol{U_{1a_i}}$	$f_{c_1U_1}$	$f_{c_2U_1}$	…	$f_{c_nU_1}$
$\boldsymbol{U_{2a_i}}$	$f_{c_1U_2}$	$f_{c_2U_2}$	…	$f_{c_nU_2}$
⋮				
$\boldsymbol{U_{la_i}}$	$f_{c_1U_l}$	$f_{c_2U_m}$	…	$f_{c_nU_l}$

For example, given digital object d, the set of entities under category c_1, identified by annotating comments T_{u_1} from user $u_1 \in U_d$, is $E_{c_1u_1} = \{e_1, e_2, e_3\}$. e_1,e_2 and e_3 are distinct and they are classes or instances of c_i, hence $f_{c_1u_1} = \left|E_{c_1u_1}\right|$ =3. This is the first cell in the Table 3.3. For the same category c_1, user u_2 mentioned 4 distinct entities, hence $f_{c_1u_2}$ =4. Similar steps are conducted for all the remaining users U_d who interacted with this digital object d to construct the frequency vector $freq_{c_1}$ of this category c_1 (i.e. the first column in Table 3.3) and all the other categories that are identified for variety.

Assume that Table 3.3 is the final frequency vectors across three domain categories:

The frequency vector for c_1 is first column $freq_{c_1} =< 3,4,1 >$. Similarly, $freq_{c_2} =< 10,4,2 >$ and $freq_{c_3} =< 1,0,10 >$. Therefore, the cosine similarities between the 3 categories are as follows: $\cos(freq_{c_1}, freq_{c_2})$ =0.86, $\cos(freq_{c_1}, freq_{c_3})$ =0.25 and $\cos(freq_{c_2}, freq_{c_3})$ = 0.27. This indicates that the categories c_1 and c_2 are the most similar in terms of frequency i.e. users mentioned c_1 as frequent as c_2. The pair-wise distances for the categories c_1 and c_2 is d_{12} = 0.14. Similarly, d_{13} = 0.75 and d_{23} = 0.73. The across disparity for this digital object d is d_a =0.27.

Table 3.3 Example of frequency vectors to calculate across disparity of three categories.

	c_1	c_2	c_3
u_1	3	10	1
u_2	4	4	0
u_3	1	2	10

The same steps are conduced to calculate across disparity on the group-level.

The diversity indices (variety, balance, coverage and disparity) and associated metrics (number of sub-categories of a given entry point, proportions, dispersions, and frequencies) identified above are captured in a diversity profile generated for each selected perspective, as defined next.

3.3 Diversity Profiling

Diversity profiling is an act to characterise a given social cloud in terms of its diversity regarding the selected perspective(s) – domain or user.

3.3.1 Profiling Output

Given a social cloud, its diversity profile consists of the diversity indices introduced above and the associated metrics. Therefore, given an ontology Ω, an entry point EP_Ω within this ontology and a list of entities from annotations E, a complete diversity profile for a chosen perspective (domain or user) is the following:

$$\boldsymbol{Diversity\ profile}(\Omega, E, EP_\Omega) = < \boldsymbol{v}\ (n, K, S_{c_i}, I_{c_i}, E_{c_i}), \boldsymbol{b}\ (p_{c_i}), \boldsymbol{r}\ (rep_{c_i}), \boldsymbol{d_w}\ (dis(c_i)) >$$

where:

- $\boldsymbol{v}$: Variety, number of non-empty sub-categories of a given entry point EP_Ω and associated metrics to measure v:
  - n, number of sub-categories of a given entry point EP_Ω in the given ontology Ω.
  - K, set of non-empty sub-categories of entry point EP_Ω (i.e. covered in E).
  - S_{c_i}, set of subclasses of each category c_i in K.
  - I_{c_i}, set of instances of each category c_i in K.
  - E_{c_i}, set of entities covered in the social cloud for each category in K.

Variety v and associated metrics facilitate the calculations of the other diversity indices as they are calculated for every category $c_i \in K$, where $i = 1\ to\ |K|$ and $|K|$ is the number of non-empty sub-categories of EP_Ω (i.e. identified for variety v).

- $\boldsymbol{b}$: Balance, the distribution of entities in their sub-categories in K and associated metric to calculate b:
  - p_{c_i}, proportion of each category c_i in K against set of entities E covered in the social cloud.
- $\boldsymbol{r}$: Coverage of categories of K in E against the covered branch of the given ontology Ω with associated metric:
  - rep_{c_i}, representational proportion of each category c_i in K against all entities (subclasses and instances) in this category c_i.
- $\boldsymbol{d_w}$: Within disparity, the distinctiveness between entities in E within their categories of K with associated metric:
  - $dis(c_i)$, Dispersion, distances between entities in their categories of K.

When more than one perspective is given (e.g. domain and user), those can be linked to measure across disparity d_a. This extends the profile with:

- $\boldsymbol{d_a}$: Across disparity, the distinctiveness between the non-empty sub-categories of an entry point EP_Ω with associated metrics:
  - $freq_{c_i}$, frequency vectors of categories in K.
  - $cos(freq_{c_i}, freq_{c_j})$, cosine similarities between categories in K.
  - d_{ij}, pair-wise distances between categories in K.

Where $c_i \in K$ and $i = 1\ to\ |K|$, where $|K|$ is the number of non-empty sub-categories of EP_Ω (i.e. identified for variety v).

In summary, when more than one perspective is given, variety, balance, coverage and within disparity are measured for each perspective separately, then the perspectives are linked to measure across disparity.

3.3.2 Profiling Levels

For a given perspective (either domain or user), the profiling can be performed at multi levels- *overview* and *zoom-in*. It is achieved by slicing the entities in E in different ways. For example, profiling can be conducted based on all the entities associated with all comments from all digital objects in the social cloud. This gives an overview of the diversity of the selected social cloud. Zooming into a collection of entities associated with comments from individual digital objects, or, from individual users allows profiling with a specific focus. A list of examples are as follows.

3.3.2.1 Diversity Overview

Diversity overview profiling is based on *all distinct entities* resulting from users interacting with *all digital objects* in the social cloud, where diversity properties are calculated as an overview of a selected perspective. The diversity overview for domain and user perspectives are as follows:

Domain diversity overview for all digital objects D, this is based on the domain ontology Ω_{domain}, a selected entry point within this ontology $EP_{\Omega_{domain}}$, and a set of domain entities E_{domain} resulting from annotating all the user comments T associated with all the digital objects D in the social cloud.

User diversity overview for all users U *with regards to an attribute* $a \in userProfile$, this is based on the user ontology Ω_{user}, selected entry point within this ontology

$EP_{\Omega_{user}}$, and a set of user entities E_{user} resulting from annotating the attribute a from the $userProfile$ of all users U in the social cloud.

Diversity overview can be useful to get an insight of the level of diversity with regards to the domain or users. It is in particular useful for comparing two or more social clouds.

3.3.2.2 Diversity Zoom-in

Diversity zoom-in profiling is based on *distinct entities per a digital object, user or a user group* in the social cloud, where diversity properties are calculated to inspect closely the diversity perspectives. For examples:

Domain diversity zoom-in per digital object $d \in D$, this is based on the set of entities E_d from annotating the comments written around this digital object T_d

This facilitates diversity-based ranking of digital objects in a pool where comparisons between digital objects can be conducted (e.g. identifying videos with maximum and minimum diversity indices). It is possible to compare digital objects across social clouds if the same set of digital objects are used.

Domain diversity zoom-in per user $u \in U$, this is based on distinct entities E_u from annotating all the comments T_u written by this user around all the digital objects D in the social cloud.

The users are ranked based on the diversity of their comments. This is useful when list of users or user groups are compared based on the diversity of their domain coverage.

3.3.3 Profiling Special Cases

Profiling special cases are the extreme cases during diversity profiling that require careful considerations. For example, selecting another entry point or extending the required ontology branch under the selected entry point. These are listed below and summarised in Table 3.4:

1. If $\boldsymbol{v} = 0 \rightarrow$ $Diversity\ profile =< \boldsymbol{b} = 0,\ \boldsymbol{r} = 0, \boldsymbol{d}_{\boldsymbol{w}} = 0,\ \boldsymbol{d}_{\boldsymbol{a}} = 0 >$.

When all sub-categories of a given entry point EP_{Ω} are empty (no coverage in E), all diversity indices and associated metrics will take the value of zero.

2. If $\boldsymbol{v} = 1 \rightarrow$ $Diversity\ profile =< \boldsymbol{b} = 0,\ 0 < \boldsymbol{r} \leq 1,\ \boldsymbol{d}_{\boldsymbol{w}} = dis(c),\ \boldsymbol{d}_{\boldsymbol{a}} = 0 >$.

When there is only one sub-category c of the entry point EP_{Ω} is covered, although proportion $p_c = 1$, balance will be zero based on the Shannon Entropy formula (2). Coverage will take a vale that is greater than zero and less or equal to 1, if this category is fully covered (i.e. if rep_c=1). Within disparity will take the value of the covered category's dispersion $dis(c)$. Across disparity is meaningless as it is a pair-wise comparison between categories, hence it will be zero.

3. If $n = 1$ (with no children) →

$$Diversity\ profile =< \boldsymbol{v} = 1, \boldsymbol{b} = 0, \boldsymbol{r} = 0, \boldsymbol{d}_{\boldsymbol{w}} = 0, \boldsymbol{d}_{\boldsymbol{a}} = 0 >$$

When the entry point has only one sub-category that is covered in E with no children (i.e. no subclasses and instances for this category), variety will be 1 and all other indices take the value of zero. If $n = 1$ and has children, the case 2 above applies.

Table 3.4 Summary of the diversity profiling special cases.

Case	Variety v	Balance b	Coverage r	Within disparity d_w	Across disparity d_a
$\boldsymbol{v = 0}$	$v = 0$	$b = 0$	$r = 0$	$d_w = 0$	$d_a = 0$
$\boldsymbol{v = 1}$	$v = 1$	$b = 0$	$0 < r \leq 1$	$d_{\ w} = dis(c)$	$d_a = 0$
$\boldsymbol{n = 1}$	$v = 1$	$b = 0$	$r = 0$	$d_{\ w} = 0$	$d_a = 0$

Diversity indices captured in the diversity profiles serve as multi indicators of diversity in the social cloud with regards to the selected perspective. These indices with associated metrics facilitate the detection of diversity patterns as discussed next.

3.4 Diversity Patterns

A diversity pattern is an interpretation of any relationships: (a) between different indices within a perspective or (b) between different perspectives. The diversity profiles from the case studies have been analysed for patterns detection. The diversity properties have been combined systematically for visual inspection to detect patterns. Also, drilling down to the associated metrics of the properties enabled patterns detection. This section presents three possible ways to detect such patterns.

3.4.1 Balance Combined with Coverage

As both indices are measured based on different ways of calculating proportions of entities in E, together they can characterise the coverage of these entities. For example, low balance indicates uneven distribution of the covered entities, but it does

not necessarily mean low coverage of these entities. Combined with coverage, it can indicate the extent of the entities being mentioned in the comments against the given ontology. This analysis can be conducted for each perspective separately and combined, which gives 4 possible patterns for the former and 16 for the latter. These patterns are for the categories identified for variety v. One way to identify these patterns is to sort the list of profiled videos by balance then coverage, excluding all videos that meet the three extreme cases (Table 3.4). The patterns for each perspective are as follows.

3.4.1.1 Patterns for the Domain

There are 4 patterns (see Table 3.5) that can be detected when combining domain balance (b) and domain coverage (r) (→ denotes indicate, ↑ for high value and ↓ for low value):

- If ($\downarrow b \wedge \downarrow r$) → *niche or poor coverage*.
- If ($\uparrow b \wedge \uparrow r$) → *diverse coverage*.
- If ($\uparrow b \wedge \downarrow r$) → *lack of focus*.
- If($\downarrow b \wedge \uparrow r$) → *focus*.

Table 3.5 Possible interpretations of a combination of domain balance and coverage.

	Low coverage ($\downarrow \boldsymbol{r}$)	**High coverage** ($\uparrow \boldsymbol{r}$)
Low balance ($\downarrow \boldsymbol{b}$)	*Niche* (A)	*Focus* (B)
High balance ($\uparrow \boldsymbol{b}$)	*Lack of focus* (C)	*diverse coverage* (D)

3.4.1.2 Patterns for the User

Similarly, 4 patterns can be detected when combining user balance (b) and user coverage (r) (see Table 3.6 for a summary):

- If ($\downarrow b \wedge \downarrow r$) → *comments were made by users with very similar profiles*.
- If ($\uparrow b \wedge \uparrow r$) → *comments were made by users with very different profiles*.
- If ($\uparrow b \wedge \downarrow r$) → *comments were made by identifiable groups of similar users.*

  As fewer users are proportionately distributed across the user categories, this is a shallow or minimal representation of more than one user group.

- If ($\downarrow b \wedge \uparrow r$) → *comments were mainly made by at least one dominant user group.*

As it is an indication that there is at least one dominant user group, a good representation of views from one specific user group may be located.

Table 3.6 Possible interpretations of a combination of user balance and coverage.

	Low coverage ($\downarrow r$)	**High coverage** ($\uparrow r$)
Low balance ($\downarrow b$)	*Non-diverse users* (W)	*Dominant user group(s)* (X)
High balance ($\uparrow b$)	*Reps of user groups* (Y)	*diverse users* (Z)

Using these patterns in conjunction with the other diversity indices deepens the insights. For example, for the pattern (D), the higher the domain variety and within disparity, the more diverse are the comments made on the subject matters at all levels.

Figure 3.7 illustrates how the ontology branch would look like for each of the patterns discussed above. Patterns are applicable for the domain and user perspectives.

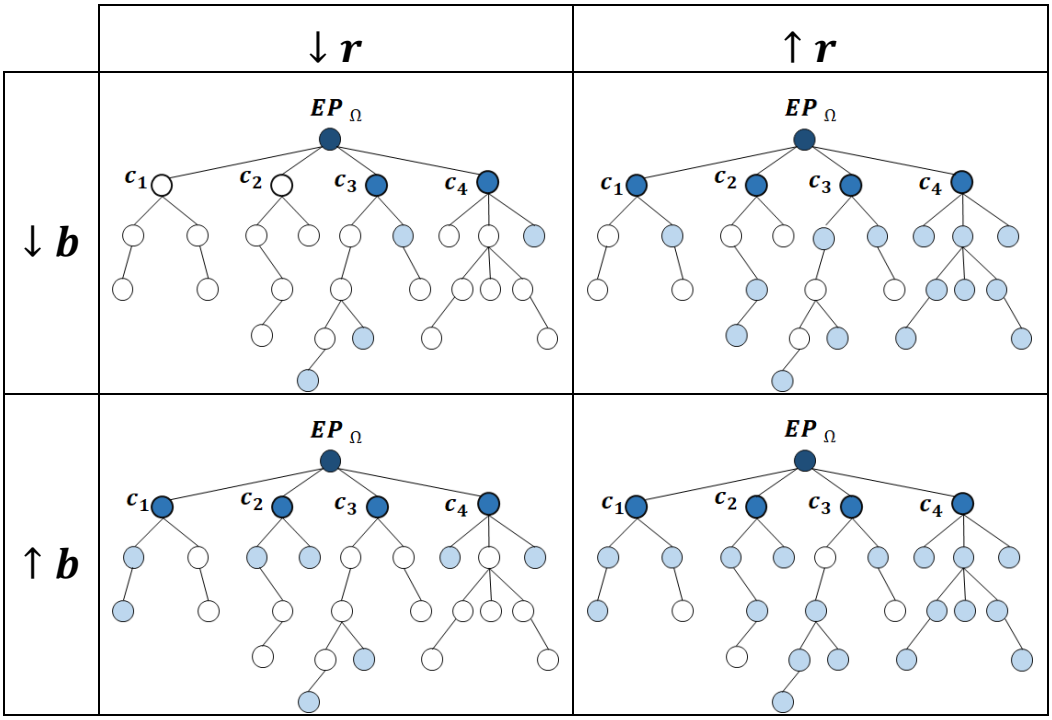

Figure 3.7 Possible patterns from combining coverage with balance.

3.4.1.3 Patterns for the Domain Combined with User

Building on the patterns from above, when the domain balance and coverage combined with the user balance and coverage, there are 16 (4*4) possible combinations that can be used for further patterns detection in a social cloud. To achieve this, the domain and users' diversity profiles are grouped and sorted based on one of the perspectives (e.g. domain). This combination does not require any linking between the diversity perspectives as done for across disparity, and it requires to measure user diversity for only the users who made domain-related comments. 3

patterns were detected using this combination in one case study, illustrating potentials for further meaningful patterns in different context. These patterns are as follows:

- $A \wedge W \rightarrow$ *Non-diverse users with niche or poor domain coverage.*
- $D \wedge Z \rightarrow$ *Diverse users with diverse domain coverage.*
- $\mathrm{C} \wedge X \rightarrow$ *Dominant user group with a lack of focus on domain aspects.*

These are possible indications and zoom-in diversity profiling per user could deepen the insights with the listed patterns above.

3.4.2 Domain Linked with User for Across Disparity

The linking of domain and user perspectives facilitates the detection of 3 patterns via the entities' frequencies (f_{c_u} and f_{c_U}). These patterns are as follows:

3.4.2.1 Dominant Domain Category

This is a sub-category c of $EP_{\Omega_{domain}}$ that is mentioned in user comments more frequently across all users - individuals or user groups (e.g. a cultural group). This can be per digital object d or for all digital objects D in the social cloud. It is identified based on the sum of all the frequencies across the users $\sum_{i=1}^{m} f_{c_{u_m}}$ or user groups $\sum_{i=1}^{l} f_{c_{U_i}}$. The category with the highest frequencies across users or user groups is the dominant category. This can be identified when there is a domain focus pattern ($\downarrow \boldsymbol{b} \wedge \uparrow \boldsymbol{r}$). This answers the question: *What is the popular topic in the domain that was discussed by users or user groups around the selected digital object(s)?*

It is worth noting that the same approach yields which categories are *under-represented*. These are categories with low to minimum aggregated frequencies.

3.4.2.2 Dominant User

This is a user $u \in U$ who have the highest frequency for one or more categories on a digital object d or all digital objects D in the social cloud. This is done via an inspection of the aggregated frequencies of each user across all the categories $\sum_{i=1}^{n} f_{c_{i_u}}$ (i.e. the summation of the values per a row in Table 3.1 in section 3.2.3.5). The user with the highest frequencies for a category is the dominant user for this category. Identification of such users can benefit others, for instance, users can learn vicariously from the comments of dominant users about the domain.

3.4.2.3 Dominant User Group

It is the user group (e.g. females) who have mentioned one or more domain categories the most in their comments on a digital object d or all digital objects D in the social cloud. Similarly, this is done via aggregating the frequencies of the user groups across the domain categories $\sum_{i=1}^{n} f_{c_{i_U}}$(i.e. the summation of the rows of Table 3.2 in section 3.2.3.5). The user group with the highest frequencies is the dominant user group for this category. This answers the question: *which is the user group (e.g. culture group) that drives the domain-related discussion on a digital object or a pool of digital objects?*

Identification of this pattern can help for example to select a video for a learner with a cold start where not much is known about the user but his/her user group. In this case, videos with dominant user groups same as the learner's group can be recommended to the learner.

3.4.3 Domain Perspective per User

Domain diversity zoom-in profiling per user $u \in U$ on a digital object d or a pool of digital objects D ranks users based on the level of diversity in their comments. This helps to identify two types of users as follows:

3.4.3.1 Domain Diversified User

This is a user $u \in U$ who scored high to maximum for the domain diversity indices among other users interacting with the same digital object d or pool D. The high values for the domain indices indicate a good exposure of the domain. Identification of such user can be useful, for instance, the comments of this user can help others to learn vicariously about this domain. The user profile can be linked to the diversity profile, where similar users (e.g. same age group or same culture) can be identified, for example, to help them overcome the domain filter bubble (i.e. users who are too focussed on one or more aspects of the domain ignoring the other aspects).

3.4.3.2 Domain Narrowed User

This is considered to be the opposite of the diversified user i.e. scored low to minimum across the domain diversity indices compared to the other users in the same social cloud. The low values of the indices can be an indication that this user has a shallow knowledge of the domain-related aspects. Identification of such user can be useful, for

example, for personalisation and adaption. Also, this user can benefit from the domain diversified user by having exposure to his/her comments on the same digital object or the pool.

One way to identify the two types of those users is to sort the domain diversity profiles of all users U in the social cloud maximum to minimum e.g. domain variety first, then coverage, balance, and finally within disparity. The user who scored maximum (top of the list) can be considered the domain diversified user in this social cloud, and the one who scored minimum (bottom of the list) is the domain narrowed user in this social cloud. The sorted list can be divided to quartiles, where users at the top quartile can be considered to be the users with the highest domain exposure, and users at the bottom quartile to be the users with the minimum exposure.

There might be a case where the users at the top quartile do not hold high values for the domain indices or users at the bottom quartile do not hold low values for these indices (e.g. a user can have the highest values for domain indices compared to others, yet it is not as high against the given entry point within the domain ontology). In such cases, these patterns apply relative to the social cloud they are in, e.g. a domain diversified user is relative to the users in the same social cloud.

3.5 Ontological Underpinning

The proposed approach for modelling diversity in a social cloud is underpinned by ontologies. An ontology of main concepts and their *taxonomical relations* representing the perspectives - domain or user is utilised for semantic tagging and diversity profiling. Selection of ontologies depend on the perspectives needed to be analysed. This section presents a discussion on the domain and user ontologies used in the two case studies of this research (chapters 4 & 5), where a User Diversity Ontology for the user diversity perspective is proposed to complement the proposed model.

3.5.1 Domain Ontology

The domain diversity perspective requires an ontology that represents the subject domain, underpins semantic annotation of the comments and assists the domain diversity profiling. Building on the discussion in section 2.4.1.2, it is assumed for this research that such ontology is available, this because there is a high number of available ontologies to cover different domains.

For this research, two domain ontologies are utilised for the two case studies in chapters 4&5. Chapter 4 used the Body Language Ontology, which is publicly available, to underpin the diversity profiling for the domain of body language in job interviews. Chapter 5 extended and implemented an available taxonomy for presentation skills and proposed the Presentation Skills Ontology (PreSOn). This ontology underpinned the diversity profiling of the presentation skills domain. Both ontologies are presented in more details in the relative chapters.

3.5.2 User Diversity Ontology

Currently, there is no adequate representation (in the form of an ontology) of the attributes of user diversity discussed in section 2.2.1.1. Some ontologies are proposed to capture the user profiles for personalisation and adaptations for different systems and platforms as was discussed in section 2.4.1.2. While these ontologies capture different aspects of the users, they have limitations that hinder their applicability for diversity measurements and explorations. These ontologies are application-driven and have limitations in terms of hierarchy and structure. For example, GUMO, which is widely recognised in the semantic community, is shallow and lacks proper classifications and grouping of its concepts, which restricts its usability for diversity exploration. Also, the consistency of the terminology used across these ontologies hinder their applicability in terms of reusing and re-engineering.

An ontology that captures the attributes of user diversity is proposed and implemented following ontology engineering methods (discussed in section 2.4.1.3). This ontology, the User Diversity Ontology[22], is intended to complement the proposed model for the annotation of user attributes as well as user diversity profiling. This research builds on the work by Thatcher [95] that extend the Diversity Taxonomy (discussed in 2.2.1.1) and classifies the user diversity as three different levels as follows:

The surface-level attributes, also referred to as the social category or demographical attributes. This level includes the visible and physical characteristics of individuals such as age, gender, and race or ethnicity.

The knowledge attributes, also referred to as Informational attributes. This level includes the functional background, experience, and education of individuals.

[22] The (extended) User Diversity Ontology is accessible via: https://doi.org/10.5518/560

The deep-level attributes, also referred to as value-based or “attitudal” attributes. This level represents the non-visible characteristics of individuals like personality, beliefs, values, and attitudes.

This research proposes these three levels as the top-level categories (i.e. top super classes) of the User Diversity Ontology as follows: ***SurfaceLevelAttribute****,* ***DeepLevelAttribute****, and* ***KnowledgeAttribute*** (see Figure 3.8). Each level is expandable as appropriate and when needed by other available ontologies, models and theories (as shown later in sections 4.2.3 and 5.2.3).

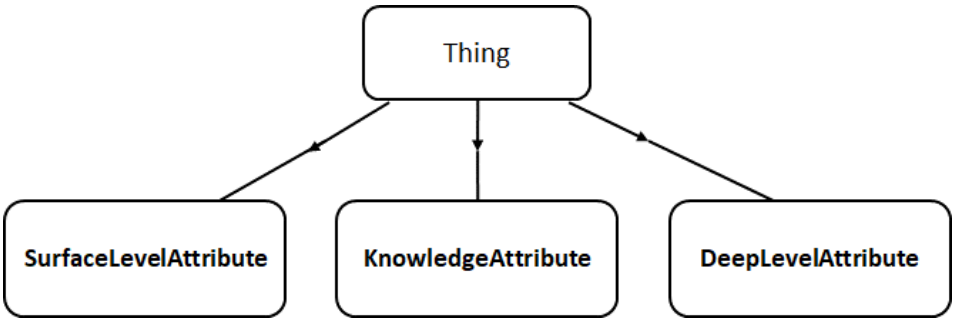

Figure 3.8 The top categories (top super classes) of User Diversity Ontology.

Protégé was used to implement this ontology in the Web Ontology Language (OWL) format. Chapters 4 and 5 show the User Diversity Ontology’s utility for annotating the user attributes (captured in their profiles) and assisting the diversity measurements in two social clouds. Further extensions to this ontology are introduced, which illustrates possible applications in case this ontology is used to categorise user attributes for diversity explorations using the proposed model. This should validate the ontology by being fit for purpose i.e. covers and enables user diversity attributes and their exploration for diversity measurements.

3.6 Semantic Driven Diversity Analytics Tool –SeDDAT

Underpinned by semantic techniques, the formal model, introduced in section 3.2, is operationalised using Java, Jena APIs and SPARQL queries. The diversity indices (discussed in section 3.2.3) are transformed into algorithms to calculate variety; balance; coverage; within and across disparity. This resulted in a semantic driven diversity analytics tool - SeDDAT. This tool is a collection of Java classes that consists of these algorithms. This section presents the input preparation and implementation of SeDDAT as well as decision taken during this process as follows.

Input preparation for SeDDAT. Some input preparation (denoted as **1** in Figure 3.9) is conducted prior to diversity profiling. SeDDAT requires the following input:

i. An ontology Ω selected to represent the perspective of interest- domain or user i.e. a domain ontology for domain diversity profiling or user ontology for user diversity profiling. SeDDAT requires the ontology URI and OWL file to conduct associated processes (e.g. querying, semantic distances) for diversity measurements.

ii. An entry point EP_{Ω}, which is a class URI within each selected ontology Ω. This $EP_{\Omega_{domain}}$ in the domain ontology and $EP_{\Omega_{user}}$ in the User Diversity Ontology. Initially the entry point can be *Thing*, here SeDDAT takes the ontology URI as an input. Other entry points can be specified for diversity measurements as required.

iii. An annotated content, which is file(s) that have been linked to the ontology entities with semantic annotation tools. This is the target dataset for diversity measurements and analysis. Expected content is: (a) digital objects metadata (e.g. Id, title, URI); (b) user profiles (e.g. Id, age, gender, cultural cluster, expertise); (c) annotated comments with associated domain entities URIs and (d) annotated user profile attributes with associated entities URIs.

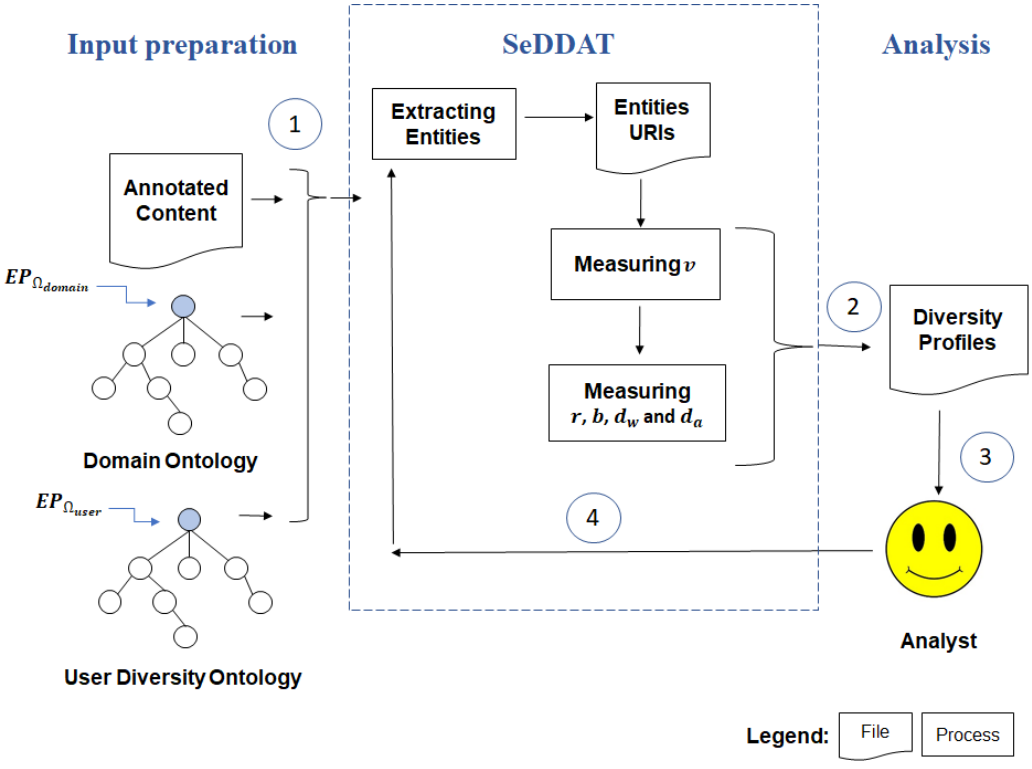

Figure 3.9 The semantic approach for diversity profiling with SeDDAT.

The input preparation stage is illustrated in the next two chapters (section 4.2 and section 5.2) showing the steps required if SeDDAT is to be used in similar scenarios for diversity profiling.

SeDDAT Execution. The initial step for diversity profiling is to calculate variety v as balance, coverage and disparity (within and across) require identification of the categories covered. For that the algorithms are called sequentially starting with an extraction algorithm. This is an algorithm for retrieving the set of entities E (E_{domain} or E_{user}) from the input file for diversity profiling based on the chosen entry point EP_{Ω}. Once the entities E are retrieved, they are passed to the variety v algorithm where entities' super classes are retrieved from the ontology. The distinct number of those super classes is considered to be variety v. These super classes with associated entities are passed to the other indices' algorithms to generate the diversity profiles.

The Semantic Measures Library and Toolkit-SML (discussed in section 2.4.3) is used within SeDDAT to calculate within disparity (d_w) - discussed in section 3.2.3.4, in particular the shortest path measure is transformed into an algorithm based on this library. Moreover, entities from annotations (in E) can be classes and instances. Semantic distances are measured between classes only, hence whenever an instance in E is used for calculating distance, it is substituted with its immediate parent class prior to calling the semantic distance algorithm. The path between the parent class and the other class is increased by one to cover the distance between the instance and the other class (see appendix A for shortest path algorithm).

SeDDAT consists of 8 java classes, 2 of which are responsible for calculating the diversity properties, while the rest include algorithms for retrieving the entities' URIs (i.e. E) from the input files and querying the given ontologies (e.g. count number of subclasses and instances of a given class/category to calculate coverage) for diversity profiling. SeDDAT takes an XML file as an input for the diversity profiling for variety, coverage, balance, and within disparity and a Microsoft excel file for across disparity. Across disparity is measured when there is associated content (e.g. user profiles) with both perspectives - domain and user (as discussed in section 3.3.1). Any required pre-processing for the input files is conducting prior to diversity profiling out of SeDDAT. The output is cleaned (the SML library creates large unnecessary and unavoidable text associated with parsing the given ontology) and transformed manually into Microsoft Excel Spreadsheets for analysis.

The diversity profiles are generated for each diversity perspective separately as each require separate and different entry points from the given ontologies for the profiling. A different entry point can be selected (e.g. to move level down within the given ontology) for one or both perspectives for further profiling.

See appendix A for example algorithms. The full code of the tool is available from here[23] with an instruction file and sample input files to use with the tool. Part of the diversity profiles of first case study are available as well to illustrate the diversity profiles described in section 3.3.1.

Analysis. All diversity indices and associated metrics, such as proportions and frequencies are captured in diversity profiles (denoted as **2** in Figure 3.9), where an analyst (denoted as **3** in Figure 3.9) can inspect these profiles in conjunction with other visualisation or presentation tools to interpret the diversity characteristics and detect patterns. If potential interesting patterns are spotted, other entry points can be selected for more fine-tuned diversity profiles (denoted as **4** in Figure 3.9).

3.7 Discussion

This chapter presented an overview of the proposed semantic driven approach to model diversity in a social cloud captured from social spaces. A formal model is proposed to pave the way for diversity measurements when using the approach in similar contexts.

SeDDAT input preparation steps are illustrated (and applied later in chapters 4 and 5). This can help to conduct the instantiation of SeDDAT if it is to be utilised in similar scenarios.

The SML contains valuable algorithms for semantic distance measures, but it generates large and unnecessary amount of text associated with parsing the given ontologies during the diversity profiling.

The profiling can be conducted for one perspective or both, separately and linked. This shows the utility of the model to overcome limitations discussed in chapter 2.

Variety is a fundamental diversity property to define as the other three properties require the identification of the system main categories to be calculated. This is one of

[23]University of Leeds Repository at: https://doi.org/10.5518/560

the reasons of why this research accepts that each diversity property is a necessary but insufficient property of diversity.

It is important to select a suitable entry point for diversity measurements. Although diversity profiles can be generated for the special cases discussed in this chapter (section 3.3.3), it is insufficient and can be meaningless for diversity understanding and exploration. Another option is to extend the ontology branch under the selected entry point as required for the diversity profiling.

The chapter lists examples of possible diversity profiles and patterns that can be generated and detected in a social cloud with regards to domain and user diversity. These can be useful, for example, for personalisation and adaptations in the domain of learning. Other patterns might be detected in different contexts. This should be explored in future work with other case studies.

Chapter 4 : Case Study 1: An Open Social Cloud

This chapter illustrates the applicability of the proposed approach (introduced in chapter 3) to measure diversity in an open social cloud and highlights interesting diversity patterns within this cloud. An existing open social cloud from YouTube was used to experiment with this approach. Diversity profiles of variety, balance, within and across disparity are generated by SeDDAT for the identified diversity perspectives - domain and user. The subject domain is **body language in job interviews**, where a Body Language Ontology is used to underpin the comments semantic annotations and domain diversity profiling. The user diversity perspective is measured based on the **users' cultural variation** represented with a cultural model from social science. The User Diversity Ontology is extended with this cultural model to underpin the annotations and user diversity profiling.

4.1 YouTube Dataset

The YouTube dataset from the work by Despotakis [133] is deployed to explore and measure the diversity properties. 600 videos with related metadata (video Ids, video URIs and titles), annotated user comments (comments with associated ontological entities from the semantic annotations), and user profiles (user Ids and locations) are utilised for this research.

The dataset is filtered and sliced as necessary to conduct diversity profiling. As the subject domain is body language in a job interview, 17,865 domain related comments were used for the domain diversity profiling. These comments helped to filter out users and videos that have no association with the domain. There are 81,147 non-distinct entities associated with these comments, 327 of which are distinct (unique) domain entities.

All the users (28,452) who interacted with the videos (i.e. wrote at least one comment) are used for user diversity profiling. 14,443 users who wrote domain-related comments are used to generate user diversity profiles to be combined with the domain diversity profiles for patterns detection (see Table 4.1 for dataset summary).

Table 4.1 Summary of The YouTube dataset.

# Videos	#Domain comments	#Total users	#Domain users	#Non-distinct domain entities	#Distinct domain entities
600	17,865	28,452	14,443	81,147	327

The user diversity profiling is explored based on the users' cultural variations. This is based on users' nationalities. As the YouTube dataset captures only the user locations, these are used as a proxy for the users' nationality and culture in this case study.

4.2 Input Preparation for SeDDAT

The diversity profiling approach described in chapter 3 is generic. This section describes its instantiation for the YouTube dataset, which also illustrates the steps required if SeDDAT were to be used in other learning contexts and domains. The ontological underpinning is discussed including the domain ontology and the extended User Diversity Ontology.

4.2.1 Domain Ontology- The Body Language Ontology

The Body Language Ontology[24] consists of six top categories (i.e. top super classes under *Thing*) – See Figure 4.1:

- *body_language_signal_meaning*
- *body_language* (equivalent (≡) to *non_verbal_communication* & *kinesics*)
- *body_motion*
- *body_sense_function*
- *body_position,*
- and *object*

The ontology was built as part of the work by Despotakis [133] to understand body language cues in job interviews captured in user comments around videos, where the comments were semantically annotated using this ontology.

For this research, the top domain categories and associated sub-categories (subclasses) and instances (or individuals) were used for domain diversity profiling and analysis, where *Thing* was the entry point EP_{Ω}. See Table 4.2 for a summary.

[24] http://imash.leeds.ac.uk/ontology/amon/BodyLanguage.owl

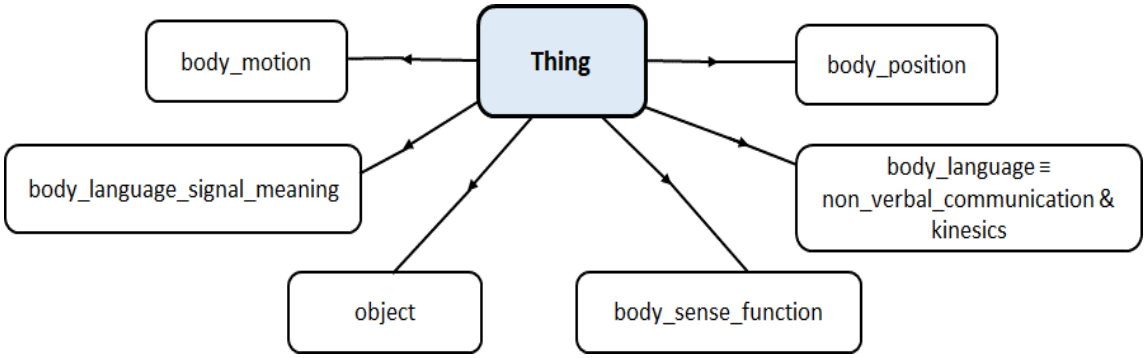

Figure 4.1 The top categories of the Body Language Ontology under "*Thing*".

Table 4.2 Number (#) of levels, classes and instances of the top domain categories in the Body Language Ontology.

Domain category	#Levels	#Classes	#Instances	Total (Sum of classes &instances)
body_language_signal_meaning	10	64	143	207
body_language	3	8	137	145
object	5	14	65	79
body_motion	5	34	42	76
body_sense_function	1	0	5	5
body_position	2	2	21	23

4.2.2 Semantic Annotations of Comments

Every domain-related comment $t \in T$ is tagged with related ontological entities E_t from the Body Language Ontology. See Table 4.3 for example comments from the dataset with associated entities from the semantic annotations, where the entities per comment are grouped by their domain categories.

The annotations are assumed to be sound (annotation validations are conducted during the work by Despotakis [133]) and used for the profiling of the domain diversity perspective.

Table 4.3 Example annotated comments with the Body Language Ontology (annotations are underlined).

Comment	Domain Category: [Entities used for annotations]
“Lol one time in an interview, I accidentally tried to shake hands with the person with my wrong hand so it was just awkward”	body_motion: [shake_hands] object: [hand]
“im a job coach , please please please dont talk or sit like that guy in the mock interview…”	body_position: [sitting]
“… The way to have good eye contact is to actually listen to what the person you're talking to is saying. Nothing beats honesty.”	body_language: [eye_contact_when_listening, eye_contact_when_speaking] body_language_signal_meaning: [honesty]
“… one thing that is a big no no in an interview - using your hands and arms too much looks threatening”	object: [hands, arms] body_language_signal_meaning: [threatening]
“… some intervieweers MIGHT feel uncomfortable if your hands are under the desk all time. Show them with open gestures (but don't overdo it!)”	object: [hands] body_language_signal_meaning: [openness, exaggeration] body_motion: [gesture]
“…while you are sitting, lean forward while you are listening to the interviewer because it makes you seem interested in what they are saying. Don't cross your arms and do not use alot of hand gestures”	body_position; [sitting] body_sense_function: [listening] body_language: [crossed_arms] body_motion: [hand_gesture]

4.2.3 User Diversity Ontology- Extension with GLOBE

This section discusses the rationale behind choosing culture, in particular national culture to explore user diversity. It also presents the extension of the User Diversity Ontology (discussed in section 3.5.2) with a national cultural model from social science. This extended ontology will be used in the user diversity profiling and patterns detection.

Understanding diversity based on cultural variations. In this case study, culture and cultural variations among individuals and groups are one of ways to explore user diversity. As with other ill-defined domains, culture research suggested different approaches for understanding cultural variations. Researchers in social science proposed that culture can be represented by the groups’ and individuals’ nationalities, while others argue that culture is beyond nationality and there are cultural groups or subcultures within nations[175]. Hsu et al supported the former and stated that “culture has centuries-old roots“ which implies that any changes to this culture occurs slowly, consequently, a culture is needed to be studied on a national level[176]. The cultural models based on nationality have been widely used in the computing community to adapt for users’ cultural variations as discussed in section 22.2.1.1. This research is

exploring culture by nationality for the understanding of diversity following the views by [176] and the computing community. It utilises the Global Leadership and Organizational Behaviour Effectiveness (GLOBE) to extend the User Diversity Ontology for the exploration and measurement of the user diversity by national culture. GLOBE is selected because it is one of the widely used models that facilitates the representation and comparison of groups and individuals based on national clusters.

GLOBE [66] (discussed in section 2.2.1.1) consists of ten cultural or societal clusters. These clusters gather similar cultures by nationality as the examples show below:

- *Nordic Europe*: (e.g. Denmark, Finland, and Sweden) ·
- *Anglo*: (e.g. Canada, The United Kingdom, and Australia) ·
- *Germanic Europe*: (e.g. Austria, The Netherlands, and Germany) ·
- *Latin Europe*: (e.g. Italy, Spain, and France) ·
- *Sub Saharan Africa*: (e.g. Zimbabwe, Namibia, and Nigeria) ·
- *Eastern Europe*: (e.g. Greece, Poland, and Russia) ·
- *Middle East*: (e.g. Turkey, Kuwait, and Egypt) ·
- *Confucian Asia*: (e.g. Singapore, Hong Kong, and China) ·
- *Southern Asia*: (e.g. Malaysia, India, and Iran) ·
- *Latin America*: (e.g. Ecuador, El Salvador, and Brail)

As GLOBE presents culture by nationality, the category/class *GLOBECulturalCluster* is injected as a sub-category of the top category *SurfaceLevelAttribute* (see Figure 4.2). The clusters (e.g. Anglo) are inserted as sub-categories (e.g. *AngloCluster*) under the category *GLOBECulturalCluster*, where country names or individuals' nationalities can be used to further extend this branch.

User locations in the YouTube dataset are used as a proxy of users' nationalities. YouTube uses the country codes to refer to the user location based on Alpha-2 code (ISO) [25] - see examples in Table 4.4 of country names, codes and associated GLOBE clusters. Therefore, each cluster category in the User Diversity Ontology is extended with the country codes based on studies by [66], [177]. This is to conduct semantic annotations and diversity profiling. The resultant branch in the **Extended User Diversity Ontology**, *GLOBECulturalCluster*, is three levels deep and each cluster category is one level deep. This branch has no instances. See Table 4.5 for the

[25] https://www.iso.org/obp/ui/#search

number of classes under each cluster category. See Figure 4.3 for the *GermanicEuropeCluster* branch and Table 4.6 for the country names associated with the country codes shown in the branch figure.

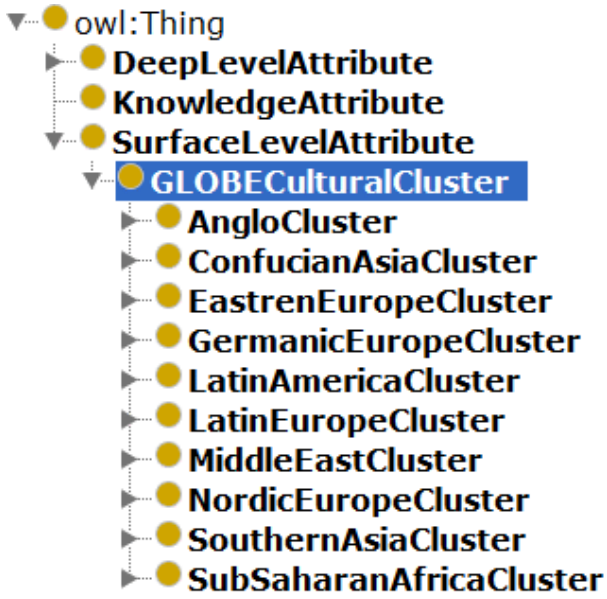

Figure 4.2 The GLOBE taxonomy as an extension of the User Diversity Ontology under the top category SurfaceLevelAttribute.

Table 4.4 Country codes with their names and associated GLOBE clusters.

Country name	Country code	Country cluster
Libya	LY	Middle East
China	CN	Confucian Asia
Canada	CA	Anglo

Table 4.5 GLOBE clusters and associated number of classes.

Cluster	#Classes	Cluster	#Classes
SouthernAsiaCluster	26	*MiddleEastCluster*	23
GermanicEuropeCluster	7	*ConfucianAsiaCluster*	9
EastrenEuropeCluster	21	*LatinEuropeanCluster*	11
AngloCluster	7	*LatinAmericaCluster*	20
SubSaharanAfricaCluster	49	*NordicEuropeCluster*	10

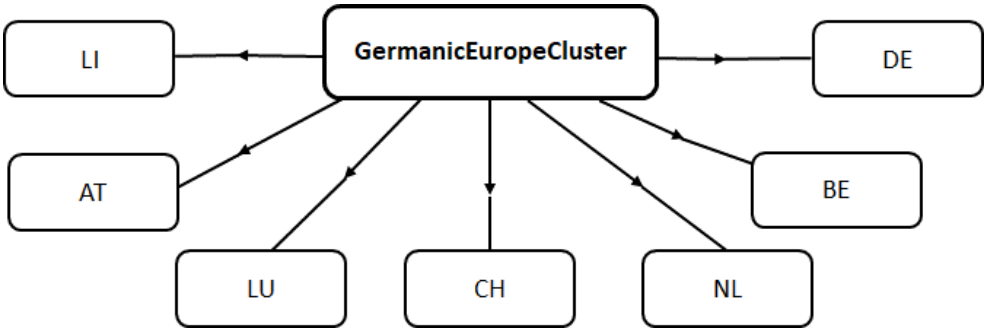

Figure 4.3 The *GermanicEuropeCluster* with country codes as its sub-categories.

Table 4.6 The *GermanicEuropeCluster* country codes and names.

Country name	Country code	Country name	Country code
Liechtenstein	LI	Netherlands	NL
Austria	AT	Belgium	BE
Luxembourg	LU	Germany	DE
Switzerland	CH		

4.2.4 Semantic Annotation of User Locations

GATE was used to conduct the automatic semantic annotations. GATE was chosen for the semantic annotation in this research as it is the biggest open-source project for textual annotation that allows the use of custom ontologies[156]. For this dataset, GATE was used for annotating the user location with the Extended User Diversity Ontology. Each user location is tagged with an associated URI from the Extended User Diversity Ontology. The output was selected to be an Inline GATE xml file, where the annotations are injected within the input xml file content. This file type was selected for easier processing and entity retrieval by SeDDAT.

There were limitations and errors with the GATE semantic annotations that required some processing, such as slicing, filtering and corrections. For example, as GATE does not handle large input files, the dataset had to be sliced to include only the video Ids, user Ids and user locations as an input xml file. In some occasions a User Id was tagged with a URI from the ontology due to the text in the Id having a partial match with a location (country code) in the ontology. It was found that GATE also occasionally tagged some locations with more than one URI for no apparent reason, which was incorrect. This required manual inspection in this case study to clean the data to ensure accuracy prior to diversity measurements.

The semantic annotation applied further filtering of users as some locations are not in the GLOBE model nor in the research that extended it (refer to section 2.2.1.1).

Therefore, these were excluded by GATE from the output file. Other users provided false locations such as “-1” and “the moon”. These are not mapped and excluded from the output file as well.

The final set of annotated locations consisted of 5,120 distinct annotated locations across all videos, 3,402 of which are annotated locations of users who made domain-related comments.

The annotated comments and locations with associated data are fed separately to SeDDAT for diversity profiling as discussed next.

4.3 Domain Diversity Profiling

This section presents the results of the diversity profiling of domain perspective - variety, coverage, balance and within disparity. These are referred to as *domain variety, domain coverage, domain balance and domain within disparity*. These diversity indices are used as diversity multi indicators separately and combined to determine the level of diversity per video (zoom-in). For each video $d \in D$, SEDDAT produced a domain diversity profile, based on the set of distinct ontology entities E_d mentioned in the user comments while watching this video d, where Ω_{domain} is the Body Language Ontology, and the ontological entry point EP_Ω for diversity profiling is *Thing*. The videos diversity profiles are analysed below (selected videos are reported in the following sections. The domain diversity profiles of all videos are available from here[26]).

4.3.1 Domain Variety (v)

Videos with high domain variety indicate that the user comments have covered most or all the high-level aspects of the domain under the selected entry point. With the entry point being “*Thing*”, the Body Language Ontology (see section 4.2.1) has 6 top super classes (i.e. 6 sub-categories). Therefore, 6 is the maximum domain variety.

Results showed that there are 34 videos that scored maximum variety ($v = 6$). Average domain variety (& standard deviation) is 3.38(1.49). 21 videos have variety $v = 0$ (i.e. special case No. 1) - i.e. these videos have no domain related comments.

[26] University of Leeds Repository at: https://doi.org/10.5518/560

58 videos have variety $v = 1$ (i.e. special case No. 2) - i.e. these videos covered only 1 domain category under *Thing*.

The initial inspection of the number of videos (# videos) that mentioned a domain category based on variety (as shown in Table 4.7) showed that *body_language_signal_meaning* seems to be the most popular topic for discussion in this dataset, followed by the category *object*. This was not the case for the category *body_position*. It does not seem that category's size is the cause of this as *body_sense_function* (5 entities) is smaller in size than *body_position (23 entities)*, but it was mentioned more in the comments (mentioned on 358 videos compared to 40).

Table 4.7 Number (#) of videos (out of 600) that had comments mentioning the six top categories of the Body Language Ontology.

	body language signal meaning	*object*	*body motion*	*body sense function*	*body language*	*body position*
#Videos	571	504	393	358	159	40

4.3.2 Domain Coverage (r)

A video with high coverage indicates that the comments linked to this video have a good representation of the high-level concepts of the domain under the selected entry point.

The maximum domain coverage (r =0.42) was scored by video 476, which reviews how some people answered political questions in a TV interview in the USA. This video covered the 6 domain categories ($v = 6$). Videos 144 (about Interview dos and don'ts) and 160 (explains how people are bias towards different things) followed closely with a score of 0.41. The former has $v = 5$ and the latter $v = 6$.

Coverage differentiates videos that have the same value for variety. This is because for variety, a category is counted even if it is mentioned just once i.e. there is only one entity from annotations from this category. For example (Table 4.8), video 156, which is about dos and don'ts in a job fair, had maximum variety ($v = 6$), but low coverage ($r = 0.06$) compared to many other videos that scored the same for variety like video 79 ($r = 0.34$),), which is about job interview tips showing examples of a good answer on how to talk about yourself.

Table 4.8 Domain categories representational proportions ($\boldsymbol{rep}_{c_i}$) for videos 79 and 156.

Video ID	*body language signal meaning*	*object*	*body motion*	*body sense function*	*body language*	*body position*
79	0.43	0.46	0.34	0.6	0.15	0.09
156	0.05	0.03	0.03	0.2	0.01	0.04

Inspecting each category's representational proportion rep_c across all videos can show which videos have a good domain coverage of a certain category. For instance, if the interest is to focus on the domain category *body_langauge_signal_meaning*, a video that scored high for this category's representational proportion can be useful, even if the other categories' proportions are low. For this dataset, below are the videos that scored top coverage for each top category of the Body Language Ontology.

Out of the 600 videos:

- 571 videos had comments that mentioned aspects related to *body_language_signal_meaning*. Video 160, which explains how people are bias towards different things, had the highest proportion for this category scoring 0.46. The comments around this video had 96 distinct entities out of the total 207 entities of this category.
- 504 videos had comments mentioning aspects related to the category *object.* Video 109, which gives advice on different situations including job interviews, scored top representational proportion for this category (0.53). This is 42 distinct entities out of the 79 classes and instances available in this category.
- 393 videos had comments mentioning *body_motion*. The top video was 476. It scored (0.37) followed by video 79 (0.34). 28 distinct entities are covered in the comments around video 476 compared to 26 by the video 79. This is out of 76 entities available in the category *body_motion.*
- 358 videos had comments that mentioned *body_sense_function*. Due to the size of this category (no sub classes and only five instances), two videos had full proportion scoring one i.e. they triggered comments that covered the 5 instances. These videos are 160 and 690 (gives advice on job interview outfits and other tips).
- 159 videos had comments that mentioned *body_language*. Videos 79 and 476 had the highest proportion with identical scores at 0.15 covering 22 entities out of total 145 entities.

- 40 videos had comments that mentioned *body_position*. The size of this branch (i.e. number of classes and instances of this category) is in total 23 entities- only two sub-classes and 21 instances. The video 476 scored top (0.22) followed by video 149 (0.17), which shows how to tie a Bow Tie. The video 476 covered 5 distinct entities of this category compared to 4 by the video 149. The rest of videos scored very low compared to the top two covering only 1 or 2 entities as follow: The following five videos after the top two ones scored the same (0.09), and the rest of videos (33 videos) scored same at 0.04.

Average (& standard deviation) coverage for this social cloud is 0.09 (0.08). It seems that the low values for this index is due to unbalanced/skewed branches (each category and its subclasses and instances are treated as a branch) of the Body Language Ontology. Some categories namely, *body_language_signal_meaning* and *body_language* are the largest branches in this ontology with 207 and 145 entities respectively (Table 4.2). This is compared to other categories like *body_sense_function* which has only 5 entities. Therefore, although for some videos some (smaller size) categories were fully represented (high rep_{c_i}), the average coverage (r) was mostly low across videos due to low representation $rep_{c_i}$of the larger ones.

4.3.3 Domain Balance (b)

Videos with a high value in balance denote that comments are distributed evenly across the high-level aspects of the domain i.e. the entities from annotations are well-proportioned across the domain categories identified for domain variety. The more even the balance, the higher the diversity. Balance can differentiate videos that have the same values for domain variety and coverage. To identify videos that had maximum domain balance, the list of profiled videos is sorted top to bottom based on domain balance.

High balance can sometimes indicate a good coverage of the domain aspects. The higher the proportions p_{c_i} (formula 2), the better the coverage against the given set of entities E_{domain}. However, as high balance signifies that all the entities around a video are well-proportioned across domain categories, some videos can be higher in proportions than other videos, yet their overall balance is lower and vice versa. For example, video 476 that scored top for coverage (as discussed above) came 10th at the sorted list for balance at 1.38. The maximum domain balance (1.58) was scored

by video 404, which is about an example of bad answer. This video and the following two top videos, 681(about phone interviews) and 679 (about the right length for your answers in interviews) at 1.46 had v =5 and had low domain coverage at 0.094, 0.07 and 0.06 respectively. An inspection of the entities for each covered category showed that video 404 covered 2 entities from each of *body_language, body_motion, object, and body_sense_function*, and it covered only 1 entity from the category *body_language_signal_meaning*. Although this is a very low coverage against the categories' actual sizes (Table 4.2), it is an even distribution of the given entities.

There are 79 videos out of 600 that scored zero. These are videos that satisfy the special profiling cases one or two shown in Table 3.4 i.e. their domain variety v is either zero or 1.

4.3.4 Domain within Disparity (d_w)

Videos with high within disparity indicate that the comments cover distinctive aspects within the domain categories - i.e. the entities from annotating the comments are widely scattered within their domain categories. The higher the distance between the entities, the higher the dispersion and overall disparity (formula (4)). Therefore, to identify videos that show most distinct domain aspects in their content, the videos can be ordered from largest to smallest according to their within disparity values.

Videos that scored top were the videos that met the special case 2 (i.e. $v = 1$). These are 16 videos out of 600. Video 65 (about how to grow your network) scored the top followed by video 716 (about how not to behave in a job interview) with the within disparity values 136.5 and 120.14 respectively. 55 videos scored the minimum within disparity value of zero. These are: (**a**) videos that met special case 1(i.e. v =0), (**b**) videos that met special case 2 (i.e. v =1) but with only one entity covered for the identified category (i.e. $|E_d|$ =1), or, (**c**) videos that had only one entity for each category identified for variety (i.e. $|E_{c_i}|$ =1). For that, semantic distance between entities were zero resulting in zero dispersion and zero overall within disparity. Average within disparity for this social cloud is 23.57 (17.95).

For this dataset, sorting videos based on within disparity shifted most videos that scored maximum based on domain variety, coverage and balance to the bottom of the list, except for the videos that scored zero for disparity.

When entities are close in distance, especially if distributed within the same level of the branch's hierarchy, the dispersion is low and hence the low value of within disparity. When the entities are distributed across the category, especially if distributed in different levels within the category's hierarchy, dispersion and within disparity are high. The videos that scored high for within disparity, had their entities E_d scattered across their categories, especially the videos that covered the larger categories. For example, video 65 had the highest within disparity score and covered only the category *body_language_signal_meaning* (v =1) with poor coverage (r = 0.02) and their entities are distributed in different levels of this category's hierarchy.

Figure 4.4 illustrates an example of how dispersion can be high when the entities are scattered (different) within their categories and low when these entities are close (similar). Within the category *object*, although the entities *lip* and *mouth* (mentioned on video 459(about interview persuasion) and *head* and *heart* (mentioned on video 737(about why executives fail in interviews)) are subclasses of *body_part (heart* is a subclass of *organ* under *body_part)*, the former are distributed closely scoring 2 for dispersion compared to the latter, which scored 4.5. Another example, *body_sense_function* (5 entities) had the lowest dispersion compared to all the other categories. Maximum dispersion for this category was 3.2 scored by video 160. As can be seen in Figure 4.5, this video had the maximum coverage (full coverage) for this category scoring 1. The entities (listening; looking; smelling; tasting and touching) are distributed closely within the same level of this branch. The snapshots are obtained using the framework ViewS[27] implemented by Despotakis[133].

High representational proportion $rep_{c_i}$for a category results in a low dispersion as the entities are distributed closely to each other. For example, the video 402 had lower representational proportion rep_{c_i}(0.25) than video 160(0.46) for the category *body_language_signal_meaning*, yet both videos had similar dispersion $dis\,(c_i)$ (36.27 and 37.77 respectively) as the 96 entities covered by video 160 are distributed closely within the *body_language_signal_meaning* similar to the 38 entities covered by video 402 (tips on how to work in a certain company) - shown in Figure 4.6. This

[27] A graph in ViewS shows the entities (classes and instances) of a domain category (super class). The coloured (e.g. green) shapes are the entities from annotating the comments on the video and the uncoloured ones are the entities not present in the user comments. Squares are classes and circles are instances. Entities' colours refer to comment content e.g. entities with red edges indicates negotiation in comment.

might be due to the shortest path measure used for measuring dispersion. See Table 4.9 for a summary of diversity profiles mentioned examples.

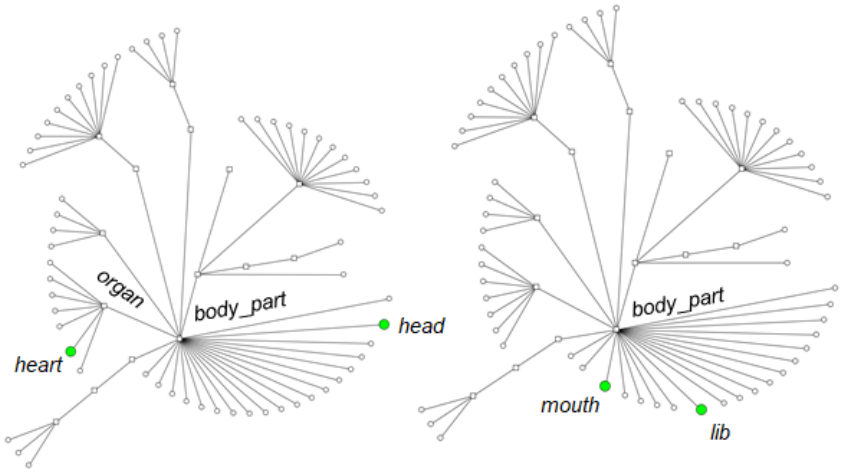

Figure 4.4 The distribution of the entities *head* and *heart* for the video 373 (left), and *lib* and *mouth* for the video 459 within the category *object*.

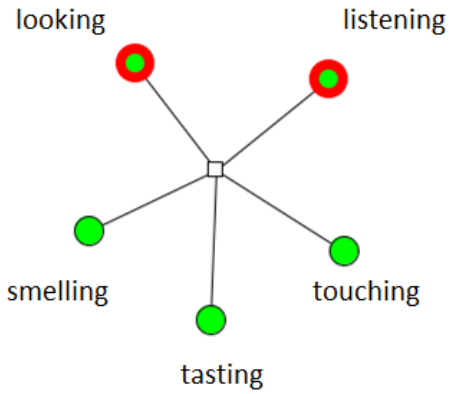

Figure 4.5 The distribution of the entities mentioned in comments around video 160 within their categories *body_sense_function*.

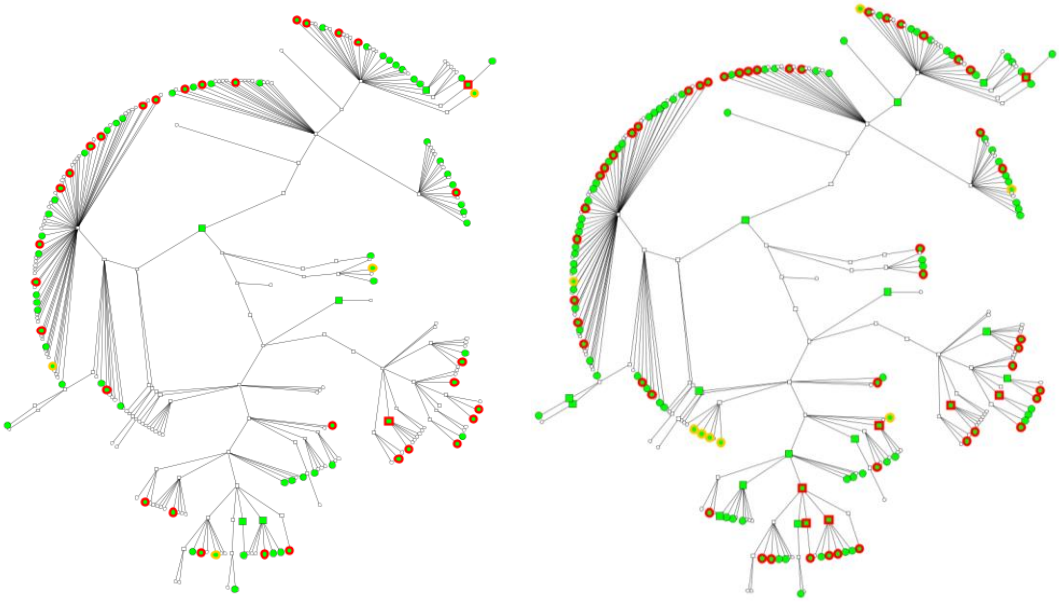

Figure 4.6 The distribution of the entities for videos 402 (left) and 160 within the category *body_language_signal_meaning*.

Table 4.9 Number of comments, distinct entities and diversity indices of example 6 videos.

Video Id	#Comments	#Distinct entities	Domain variety	Domain balance	Domain coverage	Domain within disparity
65	1	4	1	0	0.02	136.5
459	2	7	3	1.08	0.03	7.56
737	2	4	2	0.69	0.02	32.5
160	559	181	6	1.28	0.41	11.54
402	230	105	6	1.28	0.25	10.65

A high number of domain-related comments is likely to result in a high number of entities from annotations, but the important point is that the entities from annotating the comments must be: apportioned to many domain categories to be ranked high based on domain variety v; cover well the domain categories to be ranked high for domain coverage r; well-proportioned across the domain categories to be ranked high based on balance b, and widely dispersed within the domain categories to be ranked high based on within disparity d_a.

Domain diversity profiling investigates diversity based on the domain coverage reflected in user comments regardless to the users who wrote these comments. Next section investigates diversity based on the users who were triggered to comments around the videos.

4.4 User Diversity Profiling

This section explores the user diversity perspective based on the cultural variation by nationality. The Extended User Diversity Ontology using the GLOBE Model was used for the measurements and analysis (as discussed in 4.2.30). In this chapter, it is referred to as *GLOBE user diversity*, and the diversity indices are referred to as *GLOBE variety, GLOBE coverage, GLOBE balance, and GLOBE within disparity*. The *system elements* here are the entities E_{user} associated with user locations, and the *main categories* are the GLOBE clusters.

SeDDAT generates the diversity profile for each video $d_i \in D$ (zoom-in profiling) based on users E_{user} who interacted around this video d_i, where E_{user} is the list of distinct entities from annotating user locations on this video d_i. The entry point $EP_{\Omega_{user}}$ in the User Diversity Ontology is the class *GLOBECulturalCluster*. It is a sub-category of the *SurfaceLevelAttribute* in the User Diversity Ontology (see Figure 4.2).

There are two possible approaches to profile the user diversity- *regardless of a domain* or *domain focussed*. The former is useful as sometimes the interest is to know who is interested to comment on a video regardless of their views on the video content. The latter is useful when combined with the domain diversity profiles for patterns detection (as discussed in section 3.4). Therefore, firstly, the users were diversity profiled in isolation of their comments - the 28,452 users who interacted with the 600 videos. Then, the user diversity profiles were generated for those who made domain-related comments i.e. 14,443 users. Videos with no domain-related comments were filtered out, resulting in 579 out of 600. These are used later in section 4.6. The diversity indices from the first approach are reported next (The user diversity profiles of both approaches are available here[28]).

4.4.1 User Variety (v)

High GLOBE variety indicates that users who interacted with a video are coming from most or all the GLOBE clusters (see Figure 4.2). Out of 600 videos, 33 videos had commenters coming from all ten clusters i.e. scored maximum GLOBE variety (v =10), 24 videos from 9 cultural clusters, down to 190 videos from only one cluster (v =1-special case 2). Average GLOBE variety for this dataset is 3.44(2.73).

Users from the *AngloCluster* (e.g. Canada and The United Kingdom) participated widely as they commented on 582 videos out of 600. This might be due to the fact that the videos are in English. Users from *SouthernAsiaCluster* (e.g. Malaysia and India) came second and commented on around 288 videos. *EasternEuropeCluster* users (e.g. Greece and Poland) came last as they commented on only 82 videos. The other seven clusters together commented on fewer than 200 videos: *LatinAmericaCluster* users (e.g. Ecuador and El Salvador) commented on 188 videos, *ConfucianAsiaCluster* (e.g. Hong Kong and China) on 155, GermanicEuropeCluster users (e.g. Austria and Germany) on 143, *LatinEuropeCluster* (e.g. Spain and France) and *NordicEuropeCluster* (e.g. Finland and Sweden) on 138, *MiddleEastCluster* (e.g. Libya and Turkey) on 132, and *SubSaharanAfricaCluster* (e.g. Namibia and Nigeria) on 105.

[28] University of Leeds Repository at: https://doi.org/10.5518/560

4.4.2 User Coverage (r)

High GLOBE coverage indicates that the users who interacted with a video give good representation of the countries grouped in the clusters identified for GLOBE variety v.

Maximum coverage (r =0.71) for this social cloud was scored by video 354 (tips for teenagers on job interviews) with the users from one cluster, namely the *AngloCluster*. (Figure 4.7). In fact, the top 5 videos had users from this cluster, followed by 8 videos that had maximum GLOBE variety (v = 10), such as video 561 (about hair style for work) which came 6th on the list with coverage r =0.51. Average GLOBE coverage for this social cloud is 0.2(0.1).

An inspection of the clusters' representational proportions rep_{c_i} across all videos identifies which videos had a good or poor coverage (i.e. well- or under-represented) of the user cultural clusters in this dataset. For example, in the 600 videos (with content presented in English and mostly by native English speakers), 582 had commenters from the *AngloCluster* (seven countries including GB, US, and CA as shown in Figure 4.7), only 17 videos had full coverage of this cluster (i.e.rep_{c_i} = 1). Furthermore, the *AngloCluster* is the only cluster that had a full coverage in this dataset.

The *SubSaharanAfricaCluster* (e.g. Namibia and Nigeria) seems to be under-represented in this dataset. Although 105 videos had commenters from this cluster, the highest representational proportion $rep_{c_i}$for this cluster is 0.16 scored by two videos 79 and 97 (about tips on job interview question and answer). Both videos covered only 8 out of the 49 countries within this cluster.

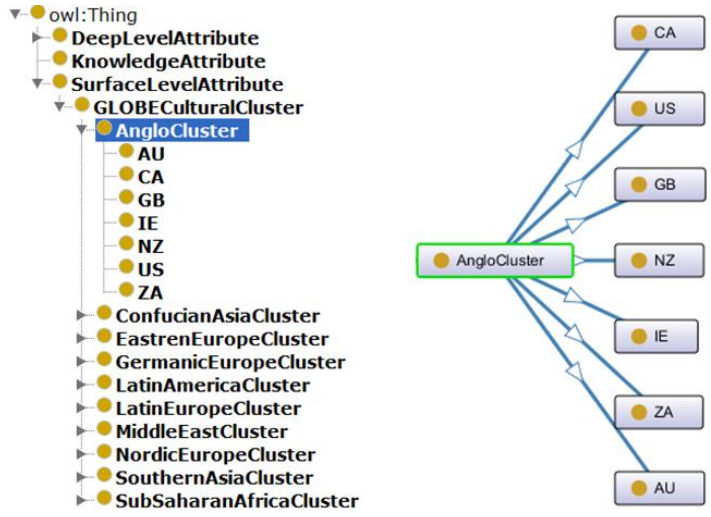

Figure 4.7 Protégé snapshot showing the countries under the *AngloCluster*.

4.4.3 User Balance (b)

High GLOBE balance indicates that the commenters on a video are distributed evenly across the GLOBE clusters identified for GLOBE variety. Many videos that scored high for GLOBE balance are the videos that scored high for GLOBE variety and coverage (the top 26 videos on the sorted list had maximum GLOBE variety). The maximum GLOBE balance was 2.26 scored by the video 109 (about general tips including job interviews).

Regardless of the number of distinct user locations on a video, the even the distribution of users across GLOBE clusters, the higher the balance. For example, as shown in Table 4.10, top video 109 had 43 distinct location ($r = 0.36$) compared to video 561 that had maximum coverage ($r = 0.51$) in this dataset, which triggered users from 64 distinct locations. However, the distribution of the users is more even for video 109 (Table 4.11). This resulted in video 109 scoring higher for GLOBE proportions (Table 4.12) and balance.

Table 4.10 Summary of the GLOBE diversity profiles for the videos 109 (top based on GLOBE balance) and 561 (top based on GLOBE coverage).

Video Id	#Distinct locations	GLOBE variety	GLOBE coverage	GLOBE balance	GLOBE within disparity
109	43	2.26	0.36	3	2.26
561	64	2.25	0.51	3.25	2.25

Table 4.11 The distribution of the distinct locations of users triggered to comment around the videos 109 and 561.

Cluster	#Countries in a cluster	**Video Id**	#Covered countries	**Video Id**	#Covered countries
SouthernAsiaCluster	26	**109**	3	**561**	7
GermanicEuropeCluster	7		5		5
EastrenEuropeCluster	21		6		9
AngloCluster	7		7		7
SubSaharanAfricaCluster	49		4		2
MiddleEastCluster	23		4		9
ConfucianAsiaCluster	9		3		6
LatinEuropeCluster	11		4		6
LatinAmericaCluster	20		3		5
NordicEuropeCluster	10		4		8

Table 4.12 Clusters proportions ($\boldsymbol{p_{c_i}}$) for videos 109 and 561.

Cluster	**Video Id**	Proportion	**Video Id**	Proportion
SouthernAsiaCluster	**109**	0.07	**561**	0.11
GermanicEuropeCluster		0.12		0.08
EastrenEuropeCluster		0.14		0.14
AngloCluster		0.16		0.11
SubSaharanAfricaCluster		0.09		0.03
MiddleEastCluster		0.09		0.14
ConfucianAsiaCluster		0.07		0.09
LatinEuropeCluster		0.09		0.09
LatinAmericaCluster		0.07		0.08
NordicEuropeCluster		0.09		0.13

4.4.4 User within Disparity (d_a)

High GLOBE within disparity indicates that the commenters on a video are scattered (coming from different countries) within a GLOBE cluster; the higher the dispersion, the higher the within disparity.

In general, GLOBE within disparity scores are low in value compared to the domain within disparity. The average GLOBE within disparity is 1.07(0.92). This is due to the GLOBE taxonomy being shallow (every cluster is 1 level deep), hence this gave small values for dispersion and overall within disparity as the entities are distributed closely (similar).

Most of the highest values for GLOBE within disparity were scored by videos for appropriate appearance for a job interview. Maximum within disparity was 3.27 scored by two videos 578 (it's a tutorial for ladies' hairstyle) and 79. The videos 561 (about ladies' hairstyle for a job interview) and 768 (about job interview outfit) followed closely scoring 3.25. It seems that appearance for an interview is a common interest in different cultures as these videos scored high to maximum for the other diversity indices.

Minimum within disparity is zero scored by 180 videos. These are videos that have covered one country within each cluster identified for GLOBE variety. For example, the video 347(about job interview tips of a good answer) covered 6 GLOBE clusters with one distinct country within each cluster, hence overall within disparity is zero.

4.5 Across Disparity: Linking Domain and User Diversity Perspectives

The two identified diversity perspectives, domain and user, are linked to measure the across disparity. The linking involved using and mapping attributes across perspectives, where one perspective was the main/anchor and the other was the secondary/selection. The domain and user diversity perspectives were linked based on two levels - individual and group levels considering domain as the main perspective and user as the secondary perspective. Only users who made domain-related comments were used in measuring across disparity.

Across disparity indicates distinctiveness between domain categories in terms of frequency. Frequency here refers to the number of times a category (its subclasses and entities) is mentioned by a user or user group in their comments. Frequency vectors are established to measure distance between them (pairwise) based on the Cosine Similarity Index. These vectors are established for the users (i.e. individual level) and their GLOBE clusters (i.e. group level) and across disparity is calculated separately for each level. The higher the similarity between categories (i.e. the cosine similarity closer to one), the lower the across disparity. The lower the similarity (i.e. cosine similarity closer to zero), the higher the across disparity; indicating dominance in one or more categories among the users.

Across disparity for this dataset based on both levels was generated by SeDDAT and is reported next.

4.5.1 Individual-level Across Disparity

Videos with high values for this property indicate that the ways that user comments cover the domain categories are distinctive. It is an indication that one or more categories are dominant categories (popular topics) among users. Hereafter, the term *individual across disparity* is used to refer to the individual-level across disparity.

In this dataset, the average individual across disparity is 1.52(1.75). Video 479 (about most important questions to ask in a job interview) scored the highest (Table 4.13), followed by Video 97 (about questions and answers in job interviews) and 79 (about an example of a good answer to the question tell me about yourself) with the values 8.65, 8.26 and 7.96 respectively. These videos triggered users to mention the domain categories in a dissimilar manner in terms of frequency.

Drilling down to the cosine similarities $cos(freq_{c_i}, freq_{c_j})$ across the domain categories on a video shows which categories are similar/dissimilar in terms of frequency. For the top video 479, minimum similarity was 0.04 between categories *body_langauge_signal_meaning* and *body_position*, and maximum was 0.77 between categories *body_motion* and *object*. In other words, the categories *body_langauge_signal_meaning* and *body_position* were mentioned dissimilarly in the comments, whereas *body_motion* and *object* are similar in terms of frequency in user comments.

Table 4.13 Summary of diversity profile of video 479 that scored top for individual across disparity (d_a =8.65). (V-Id ≡ Video Id).

V-Id	Domain variety	Domain balance	Domain coverage	Domain within disparity	GLOBE variety	GLOBE balance	GLOBE coverage	GLOBE within disparity
479	6	1.30	0.25	15.16	9	2.16	0.36	2.95

The minimum value for this index was zero scored by 134 videos. There are three causes of zero across disparity d_a as follows: (a) cosine similarity (discussed in section 3.2.3.5) is calculated pair-wise for categories, and when there is only one category mentioned in the comments (i.e. domain variety equals one) distance is not calculated; (b) cosine similarity does not apply on frequency vectors that include only zeros; and (c) when cosine similarities between the domain categories identified for domain variety are maximum (equal to one), distances between these categories (d_{ij}) are minimal (equal to zero), hence overall across disparity d_a is zero.

An inspection of cosine similarity values for each pair of categories across all the videos (videos were sorted top to bottom for each pair) showed that maximum similarity between every pair across all videos was 1 except for the dominant categories (*body_langauge_signal_meaning* and object) and under-represented category (*body_position*). The maximum similarity between the *body_langauge_signal_meaning* and *body_position* was 0.96 and it was 0.99 between *object* and *body_position*. These scores were on video 113 (about tips on how to “ace” a job interview). This video triggered users from only the *AngloCluster* covering 3 distinct locations (GB, CA, and US) out of 7 in this cluster ($r = 0.43$). The users wrote domain-related comments that covered 5 domain categories with a lack of focus

(relatively high domain balance ($b = 0.96$) and low coverage ($r = 0.09$). The individual across disparity for this video was low at 0.12.

4.5.2 Group-level Across Disparity

Videos with high values for this property indicate that the domain categories mentioned by GLOBE clusters around a video are different in terms of frequency. It is an indication that one or more categories are dominant for these cultural clusters. Hereafter, the term *group across disparity* is used to refer to the group-level across disparity.

In general, this index based on the group level is lower than that on the individual level (Average group across disparity is 0.28 (0.6)). Most videos scored lower on the group across disparity compared to their score on the individual level. As user entities in the *GlobeCulturalCluster* branch of Extended User Diversity Ontology are grouped to the cultural clusters they belong to, more entities are filtered out to maintain only distinct entities within each cluster. This resulted in smaller frequency vectors. This in turn seems to result in higher cosine similarities, and hence lower across disparity values compared to the individual across disparity. This also resulted in more videos scoring zero as 269 videos compared to 134 from the individual level. Moreover, unlike the individual across disparity, all maximum similarities between categories are 1 including the ones with the category *body_position*. For example, video 113 had maximum similarities for 5 domain categories and minimum group across disparity (d_a =0). This can be an indication that cultural groups within this dataset tended to see and mention domain aspects in a similar manner.

This however did not change the fact that *body_langauge_signal_meaning* was dominant. Similarly, an inspection of aggregated frequencies shows which category is the dominant domain category across GLOBE clusters per video and who is the dominant cluster per category on a video. For example, in comments on video 97, the *body_language_signal_meaning* is the dominant category across the 9 GLOBE clusters and had 976 distinct entities from this category. The *AngloCluster* is the dominant cluster across the 6 domain categories. Users from this cluster mentioned 1166 distinct entities in total - 13 about *body_language*; 815 about *body_language_signal_meaning*; 102 about *body_motion*; 40 about *body_sense_function*; 194 about *object* and 2 about *body_position*. The category

body_position is not only very underrepresented on this video, but also, it is only mentioned by one cluster, the *AngloCluster*. The *ConfucianAsiaCluster* contributed the minimum to the domain diversity profiling on this video with only 13 distinct entities in total (i.e. across the 6 domain categories).

4.6 Diversity Patterns in Case Study 1

As discussed in section 3.4, a list of possible patterns can be detected by (a) combining indices per or across perspectives, (b) linking perspectives, (c) profiling one perspective based on the other (e.g. profiling user based on the domain). From the diversity profiling in the above sections, the first two methods for patterns detection were applied. Although methods for detecting patterns in domain perspective per user are possible it is not useful with the scant user profiles in YouTube dataset. Examples of the patterns that were detected are as follows.

4.6.1 Combining Balance and Coverage

This section describes the patterns for understanding the nature of diversity in the YouTube dataset by combining balance with coverage from domain and user perspectives, individually or together, as discussed in in section 3.4.

4.6.1.1 Patterns for the Body Language Domain

Domain balance combined with domain coverage can provide deeper insights on the diversity level in the comments around a video. For example, it reveals which videos triggered focus on certain aspects of a domain and which videos offered diverse coverage of domain aspects. For a learning context, this can address different learners' needs as some require focussing while others like to explore the different aspects of a domain. Examples of the 4 patterns listed in section 3.4.1.1 are below (Table 4.14):

- An example of a niche or poor domain coverage (low balance and low coverage) is video 158 (about eliminating fillers when speaking in interviews). It covered 2 domain categories scoring 0.27 for balance and 0.04 for coverage.
- Video 476 covering the 6 domain categories with 1.38 for domain balance and 0.42 for domain coverage can be an example of a video with a diverse coverage (high balance and high coverage).

- An example of a video that triggered comments with a lack of focus on the domain (high balance and low coverage) is video 404 (shows examples of bad answers in interview). This video covered 5 categories scoring high for balance at 1.58 and low for coverage 0.09.
- Video 87 (shows an interview of a book author) illustrates domain focus (low balance and high coverage) covering 4 categories scoring relatively low for domain balance ($b = 0.86$) and relatively high for coverage ($r = 0.21$).

Table 4.14 Examples of diversity patterns with regards to the body language domain.

Video Id	#Distinct entities	Domain variety	Domain balance	Domain coverage	Domain within disparity
158	13	2	0.27	0.04	32.21
476	193	6	1.38	0.42	10.78
404	9	5	1.58	0.09	4.90
87	154	4	0.86	0.21	22.95

4.6.1.2 Patterns for the GLOBE Clusters

GLOBE balance combined with GLOBE coverage can characterise the distribution of users within their GLOBE clusters. Examples of the patterns listed in 3.4.1.2 are listed below (Table 4.15):

- An example of the pattern **non-diverse users** (low balance and low coverage) is video 284, which is about employment interview techniques. This video had three distinct user locations covering 2 GLOBE clusters – Philippine and India from *SouthernAsiaCluster* and The United States of America from the *AngloCluster*. This resulted in low balance (b =0.64) and low coverage (r =0.11).
- An example of a video that triggered comments from **diverse users** (high balance and high coverage) is video 561, which is about hair style for work. This video covered the 10 GLOBE clusters with high coverage ($r = 0.51$) and balance ($b = 2.25$).
- The pattern **representatives of user groups** (high balance and low coverage) indicates that few, but different user countries were triggered to comment on this video, such as video 235 (about tips on how to interview for a job). This video covered the 10 GLOBE clusters with only 17 distinct user locations and scored high for balance ($b = 2.15$) and low for coverage ($r = 0.15$).
- The pattern **dominant user group** (low balance and high coverage) implies that a dominant user group was triggered to comment on the video. For

instance, the video 331 (about questions not to ask in job interview) had the *AngloCluster* as a dominant user group as 4 distinct locations are from this cluster. The other user location is from the *ConfucianAsiaCluster*.

Table 4.15 Examples of diversity patterns with regards to the GLOBE clusters.

Video Id	#Distinct locations	GLOBE variety	GLOBE balance	GLOBE coverage	GLOBE within disparity
284	3	2	0.64	0.11	1
561	64	10	2.25	0.51	3.25
235	17	10	2.15	0.15	0.97
331	5	2	0.50	0.34	1.5

Identification of such videos can be very useful in the domain of learning. For example, videos that triggered focus on domain aspects can be useful for learners with the cold start scenario. A learner from the United Kingdom (GB) could benefit from video 331 (see Table 4.15) as majority of users are coming from the *AngloCluster*, especially that the countries covered include the learner's country. This could initially ensure the learner's engagement with the video since it has already triggered other users from his/her culture. Also, when there are other cultures triggered with the dominant user group on a video, the learner can benefit from the perspectives of culturally dissimilar users as well.

4.6.1.3 Patterns for Body Language Combined with GLOBE Clusters

This is conducted via combining the diversity profiles of the domain and users for the same set of videos, where the diversity indices for one perspective (e.g. domain) are used to sort the diversity profiles of both perspectives. Combining diversity indices of the domain and GLOBE helps to identify *what domain topics are covered by cultural groups and how diverse are the users and the topics they noticed*. This combination required GLOBE user diversity indices for only the users who made domain-related comments i.e. for 14,443 users (Table 4.1). The GLOBE diversity properties for this group of users can be lower in value due to the exclusion of users who did not write domain-related comments. For example, video 49 (about how to have a good job interview) covered 9 clusters (24 distinct locations) when diversity was measured regardless of the domain, 7 clusters of which (17 distinct locations out of 24) wrote domain-related comments.

Below is a list of patterns that can be detected in a social cloud. Example of the three patterns are (Table 4.16):

- Videos that trigger **diverse groups of users to notice diverse aspects of the domain**. For example, video 79 triggered users from the 10 GLOBE clusters to cover the 6 domain categories scoring high for domain balance and domain coverage indicating a diverse domain coverage. The high GLOBE balance and coverage indicate good diverse coverage of the GLOBE clusters.
- Videos that **did not trigger diverse users nor diverse topics** i.e. had low GLOBE user diversity and low domain diversity. There are 34 videos in the dataset that had low interaction of user cultural groups with poor coverage of the domain, such as video 804, which is about appropriate dress code for a job interview. This video had one domain entity *caution* from the domain category *body_language_signal_meaning* category by a user from the United Kingdom (GB).
- **Dominant with a lack of focus** was detected across this social cloud. For example, video 133 had users from *GermanicEuropeCluster* and the *AngloCluster*. The latter was the dominant cluster as more distinct locations (5) are from this cluster. The comments covered 5 sub-categories (v =5), yet with a lack of focus as coverage was low while balance was high indicating even but low coverage of the sub-categories.

Identification of such videos can be valuable for the domain of learning. It can help a tutor or a learner to select possibly useful videos to support the learning of the subject domain. For example, the identification of videos that trigger different cultural groups around them covering certain domain aspects can be useful when a learner or a tutor needs diverse perspectives on these aspects. This also can be useful with learners who are in a GLOBE cluster filter bubble i.e. learners who are interested in particular aspects of the domain and tend to watch videos that are by or trigger users from the same cultural cluster as the learner. This helps the learner to notice what other cultures see and know. Also, the learner can learn vicariously by reading the diverse perspectives of other cultural groups reflected in their comments. Another example, identifying videos that trigger comments covering most or all topics of a domain by certain cultural groups can help learners who are in a domain filter bubble i.e. learners who seem to miss the other aspects of the domain. Also, videos that trigger certain cultural groups to focus on certain aspects can be useful with the cold start issue when not much is known about the learners except for their cultural background. Videos that

did not trigger diverse users nor diverse topics i.e. had low GLOBE user diversity and low domain diversity might be avoided or reconsidered for learning purposes.

Table 4.16 Combining GLOBE (domain-focussed) and domain diversity indices showing number of distinct user locations and domain entities.

V-Id	#Distinct location	GLOBE variety	GLOBE balance	GLOBE coverage	GLOBE within disparity	#Distinct entities	Domain variety	Domain balance	Domain coverage	Domain within disparity
79	50	10	2.24	0.38	3.08	177	6	1.33	0.34	10.42
804	1	1	0	0.14	0	1	1	0	0.005	0
133	6	2	0.45	0.43	1.60	47	5	1.32	0.15	24.53

4.6.2 Domain Linked with User for Across Disparity

Based on frequencies used to calculate across disparity (section 3.4.2), 3 different patterns are detected as follows.

4.6.2.1 Dominant Domain Category per Video

Aggregated frequencies f_{c_u} of domain categories across users per video shows which categories are dominant and which are under-represented (section 3.4.2.1). For this dataset, most videos had the categories *body_language_signal_meaning* and *object* as dominant categories and *body-position* as the under-represented one. This seems to support findings from the coverage index (section 4.3.2). For example, frequencies of video 479 showed that although the users mentioned aspects related to all the domain categories with diverse coverage, they had the domain category *body_langauge_signal_meaning* as main discussion in their comments - it was more frequent than the other categories at total frequency value of 746. The domain category *object* followed at 502. The domain category *body_position* was hardly mentioned in the comments with total frequency value of 2 throughout all users on the video (504 users).

4.6.2.2 Dominant User for a Domain Category per Video

Inspecting the aggregated frequencies per user for a category can identify who is the dominant user per category on a video. This is a user who had the highest number of distinct entities from a category on a video i.e. a user who noticed and mentioned more entities of a category than any other user. For example, video 79 covered the 6 domain categories of Body Language ontology. It triggered 514 users from the 10 GLOBE clusters. The highest number of distinct domain entities from the category *body_language_signal_meaning* was 22 from 2 comments mentioned by one user.

This user is considered the dominant on this video for this category. The user profile can be linked to such findings where any available attributes, such as the cultural cluster, can contribute to gaining more insights. This user is male aged 23 and from the *MiddleEastCluster*. Other users, e.g. similar GLOBE cluster, can benefit from the comments of this user.

4.6.2.3 Dominant GLOBE Cluster for a Domain Category

This can be detected by inspecting frequencies per each cluster with users triggered to comment on a video. The cluster with the highest frequency for this category is the dominant cluster. For video 476, 9 GLOBE clusters were triggered to comment on this video covering the 6 top level categories of Body Language ontology. If the aim is to find dominant cluster for the dominant category *body_language_signal_meaning*, the *AngloCluster* is the one with the highest frequency with 2262 entities. In fact, this cluster is a dominant cluster across all domain categories for this video.

Identification of a dominant category on a video helps to identify which videos are more suitable for learning aspects related to this topic (category). Some videos are meant to give broader coverage of a domain (diverse coverage), while others are specific for certain aspects (one or more dominant topics/categories).

4.7 Discussion

This chapter introduced an application of the proposed semantic driven approach on an open social cloud from YouTube to show applicability of the proposed approach. Here the approach worked as an automatic mechanism for characterising and sorting the pool of YouTube videos based on the level of diversity in the social cloud around them. This shows utility of the approach to identify possible useful videos using the diversity profiles and patterns within an open and large social cloud.

Domain profiling showed that the more comments, the more domain exposure and coverage, hence users should be encouraged to write more comments. However, more comments do not necessarily mean high diversity. The comments might focus on few aspects of the domain.

User profiling and diversity patterns shed light on dominant user groups around videos, which is in this case was one cultural cluster, the Anglo. This could be due to the

criteria associated with the video selection. This raises a question: could we make a video in a way that could trigger diverse audience.

The diversity profiling of each perspective separately and linked assisted the identification of interesting diversity patterns, such as domain focus where a topic was the main discussion and a dominant user group where a certain user group was the one driving the discussion. These open doors for further understanding of diversity and further future work.

This research conforms with the importance of exploring diversity based on a combination of diversity properties. The order of the digital objects can change greatly based on the selected property. Each property brings more insights to social cloud diversity from a different angle, which in turn shows the complexity of understanding diversity.

The user locations were used as a proxy of users' nationalities for inclusion in GLOBE clusters, which might be inaccurate for some users (e.g. a user can be living, studying or working abroad and their YouTube account is linked to this location, which is not their nationality), but for this research it is assumed that the users have cultural exposure related to the location they are in. Further investigations of the related findings will be explored in other case studies in the future, where user nationalities or national cultures are well defined in their profiles.

High coverage (representational proportion $\boldsymbol{rep}_{c_i}$) for a category resulted in a low dispersion for this dataset when the entities are distributed closely to each other, this could be an indication to explore a different semantic distance measure to be added as another metric for within disparity. The shortest path was selected to avoid possible limitations of entities' information within their ontologies, which can hinder distance measures, but another semantic measure can be added to this property to investigate how they influence dispersion and within disparity. This is planned as a future work.

The ontology-based automatic semantic annotations using GATE created three main challenges, (a) incapacity to work with large input files, which required slicing and reducing the size of input file, (b) selection and parsing of the output file format, and (c) cleaning the annotation errors to ensure accuracy to underpin the diversity profiling. Undesired annotations can be generated based on data that is included in the input

file. Caution must be taken and any unnecessary data for annotations can be filtered out to reduce the annotations errors.

GATE also filtered out locational that are not in the extended User Diversity Ontology (with GLOBE), this highlights a limitation in this model and the study that extended it. The inclusion of the filtered locations is beyond this research as it is a research on its own and should be conducted by researchers from a different discipline (e.g. social scientists).

The YouTube dataset used for this case study facilitates openness and scale but consists of limited user profiles. Richer characteristics (e.g. deep-level characteristics, such as cognitive thinking) of the user are preferable for better understanding of diversity and patterns detection, especially for the domain of learning. This is achieved in next chapter.

Chapter 5 : Case Study 2: A Closed Social Cloud

This chapter illustrates the transferability of the proposed approach as well as the methods for detecting diversity patterns. Another social cloud was obtained from a closed social platform, Active Video Watching (AVW-Space), setup for learning pitch presentation skills from videos. In this case study, two datasets were obtained from two cohorts of learners – one was postgraduate students and the other undergraduate students. SeDDAT was then applied, involving the extension of Presentation Skills Ontology (PreSOn), as the domain ontology. This ontology was used for annotation of user comments and the domain diversity profiling. For this social cloud, domain diversity profiling was carried out for the videos and learners.

This case study also collected richer user profiles (such as cognitive and metacognitive strategies for learning) by an online survey before the learners interacted with the AVW-Space. Hence, the User Diversity Ontology was further extended with the extra user information, including the deep-level attributes, and used for annotations as well as user diversity profiling and patterns detection.

5.1 Datasets from the AVW-Space

This section introduces the AVW-Space, the two user studies conducted with it and the resultant datasets that are used for this research.

Environment. AVW-Space is a video-watching environment that supports engagement via interactive notetaking during watching a video[49], [135]. It taps into learners' familiarity with commenting on videos in social networking sites. There are two spaces: 'Personal' for individuals to make comments during watching, and 'Social' for browsing other learner's comments and reflecting on them. Presentation skills was the topic for learning and all learners went through eight selected videos from YouTube: four tutorials (T) on presentations and four examples (E) (two TED talks and two 3-minute PhD pitch presentations)[51] - see Table 5.1.

User Studies. Two studies were involved: **Study A** with 38 postgraduate students and **Study B** with 141 first-year undergraduate students. These students are classified as constructive according to the ICAP framework [178]. The studies included online surveys pre- and post-interaction with AVW-Space to collect students' profile and other attributes, and AVW-Space collects their interaction data [49]. Only the comments

made in the ‘Personal Space’ are used as a proxy for the learners’ thinking during the video watching.

Table 5.1 Titles of the four tutorials and four example presentations

Video	Title	Video	Title
T1	*How to Give an Awesome (PowerPoint) Presentation*	**E1**	*How can we make better medicines? Computer tools for chemistry*
T2	*How to open and close presentations*	**E2**	*Social media and the end of gender*
T3	*Make a presentation like Steve Jobs*	**E3**	*A Magna Carta for the web*
T4	*The five secrets of speaking with confidence*	**E4**	*Hypoxia-activated pro-drugs: a novel approach for breast cancer treatment*

Datasets Collected. The following collected data were specifically relevant to this research: (i) data about the videos (Ids and titels); (ii) the profiles of participants (users), such as demographic information, background experiences, scores used from the Motivated Strategies for Learning Questionnaire (MSLQ)[179]; and (iii) user comments. The total number of comments was 744 from Study A and 1129 from Study B.

5.2 Input Preparation for SeDDAT

This section describes SeDDAT instantiation for the AVW-Space studies. It introduces the selected domain ontology and the second extension of the User Diversity Ontology.

5.2.1 Domain Ontology- Presentation Skills Ontology (PreSOn)

To utilise SeDDAT, an ontology that represents the important concepts in the domain, i.e. **presentation skills**, is required to enable semantic annotations of learner comments. There is no published ontology for presentation skills that meets this purpose. To develop a presentation skills ontology (called hereafter PreSOn), the NeOn ontology development framework was adopted (discussed in section 2.4.1.3). Earlier work presenting an educational environment for presentation skills suggested two key aspects when developing presentation skills - presentation slides and presentation delivery[180]. These were broadened to *Delivery*, *VisualAid*, and *Structure*, providing the starting top level categories. A vocabulary of domain terms belonging to these categories was derived using a semi-automatic analysis of the conceptual knowledge answers in Study A (see description in [109]). Using the vocabulary, an initial domain taxonomy was created. The non-verbal communication

concept (under *Delivery*) was then extended using concepts from the Body Language Ontology used in earlier work on interpersonal communication [133]. Furthermore, it was noted that attribute terms were commonly used in comments to characterise delivery, visual aids and structure (e.g. clear, engaging, disorganised). Therefore, a fourth category (*PresentationAttribute*) was added to include such terms.

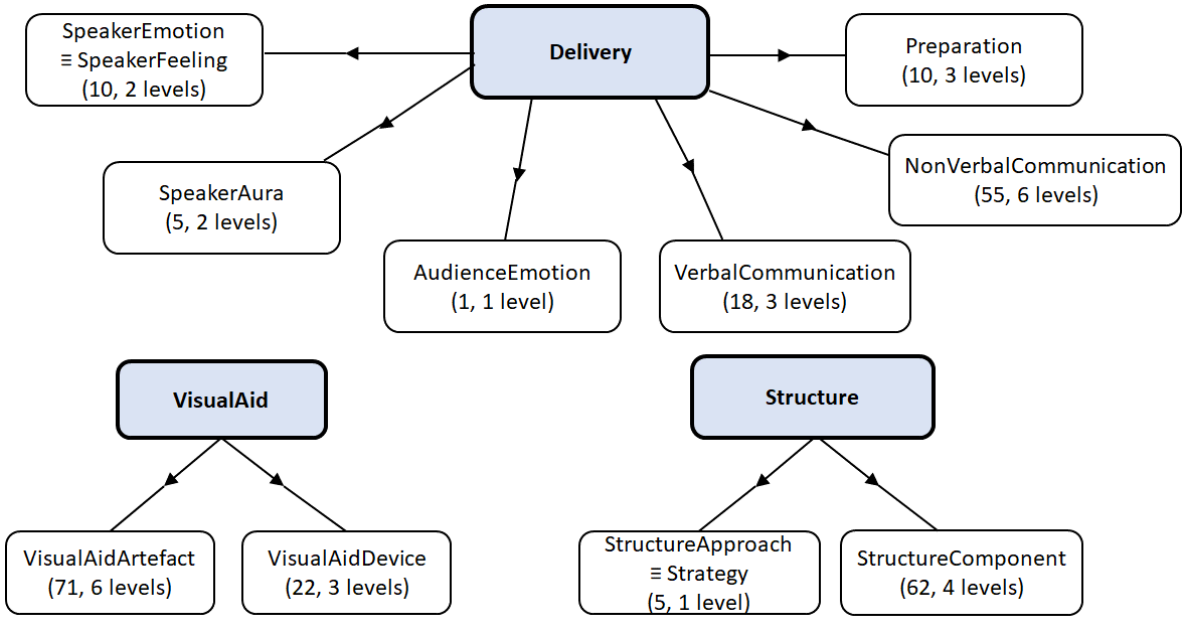

Figure 5.1 An overview of PreSOn's top three domain categories – *Delivery, VisualAid, and Structure*; the total number of sub-classes and the depth of the category's tree are given in brackets. Note that *PresentationAttribute* (27, 3 levels), is not included due to space limit (see Appendix B).

Following the NeOn framework, the derived ontology was validated with three domain experts from the institutional study skills training team at the University of Leeds. The experts were not involved in the research team nor the video watching study presented earlier. They were given the ontology taxonomy to inspect individually and asked to note any inappropriate or missing concepts. A focus group was then carried out, walking through PreSOn to review the ontology concepts and instances. The domain experts' feedback was used to refine PreSOn.

The resultant PreSOn[29] ontology consists of 299 classes and 240 instances. Figure 5.1 shows an overview of the top categories in PreSOn.

[29] https://doi.org/10.5518/560

5.2.2 Semantic Annotations of Comments

GATE was used to semantically annotate the comments using PreSOn. This process resulted in a total of 1,217 annotations with PreSOn entities for Study A and 2,070 for Study B; with 197 and 220 distinct entities, respectively. The average number of annotations and distinct entities per video covered by the comments are: Study A - 152.1 (30.1) and 66 (8.7); and Study B - 258.8 (65.0) and 78.5 (9.1). Table 5.2 shows example comments with PreSOn entities in semantic annotations.

Table 5.2 Example annotated comments with PreSOn (annotations are underlined).

Comment	Domain Category: [Entities used for annotations]
"In my opinion, this presentation is not so organized and clear. The presenter just used one slide in his whole presentation. He did not give concise theme of his presentation in the beginning."	*PresentationAttribute*: [organized, clear, concise] *Structure*: [Beginning]
"voice variation for clarity of points "	*Delivery:* [Voice] *PresentationAttribute*: [Clarity]
"… so simple and effective and doesn't distract from the spoken text."	*PresentationAttribute*: [simple, distracting] *VisualAid*: [Text]

5.2.3 User Diversity Ontology- Extension with MSLQ

This section explains the steps conducted to extend the User Diversity Ontology for the second time to profile the datasets in this case study. The extension is to utilise the extra user information, in particular the MSLQ answers, to populate the deep-level attributes in the user ontology.

Motivated Strategies for Learning Questionnaire (MSLQ). MSLQ is a self-report instrument intended to evaluate college students' motivational orientations and learning strategies to succeed in a college course. The MSLQ consists of two scales-motivation and learning strategies scales. The number of items (questions associated with the scales to be answered) in the MSLQ scales progressed over time, for example, there are 81 items on the 1991 version (scales' components and strategies are listed in Table 5.3). Both scales consist of 31 main items - The motivation items assess, for a particular course, students' goals and value beliefs for this course, their beliefs about their skill to succeed in it, and their anxiety about the course tests. The learning strategy items assess students' use of different cognitive and metacognitive strategies. The learning strategies scale has 19 additional items to evaluate students'

management of different resources. Each item on these scales consists of different number of questions that must be answered by the students as Likert scales (e.g. scale of 1 to 7). The answers on each question are then aggregated as a mean to determine the score for the item(more details are discussed in [179]). Example definitions of MSLQ items are as follows (these are used later for user diversity profiling – section 5.4): **Task value**, a value component of the motivation scales, refers to the student's perception of a task in terms of how interesting, how important, and how useful this task is. High task value results in more involvement in student's learning. **Effort regulation**, a resource management strategy of the learning strategies, refers to the student's ability to control their effort and attention in case of distraction and uninteresting task. High effort regulation ensures academic success as it shows goal commitment and continues use of learning strategies. **Organization,** a cognitive and strategy of the learning strategies**,** helps the learner to select appropriate information to be learned and make connections among them using strategies such as, clustering and outlining. High organisation indicates close involvement in the task and results in better performance.

Table 5.3 MSQL: motivation and learning strategies including the components and strategies respectively.

Motivation Scales	Learning Strategies Scales
1. Value Components	1. Cognitive and Metacognitive Strategies
a) *Intrinsic Goal Orientation*	a) *Rehearsa*l
b) *Extrinsic Goal Orientation*	b*) Elaboration*
c) *Task Value*	c) *Organization*
2. Expectancy Components	d) Critical Thinking
a) *Control Beliefs*	e) *Metacognitive Self-Regulation*
b) *Self-Efficacy for Learning and Performance*	2. Resource Management Strategies
3. Affective Components	a) Time and Study Environment
a) Test Anxiety	b) *Effort Regulation*
	c) Peer Learning
	d) Help Seeking

The User Diversity Ontology, specifically the category *DeepLevelAttribute*, was extended with concepts from the MSLQ. Although the user studies used only part of

the questionnaire (highlighted in *italic* in Table 5.3), the ontology was extended with all concepts from the MSLQ. The ontology hierarchy followed the listing of the MSLQ items i.e. the two MSLQ scales, Motivation and Learning Strategies, were used as top sub-categories of the *DeepLevelAttribute*. The three components of the motivation scales (left column in Table 5.3) were added as its sub-categories. Similarly, the two strategies of the learning strategies (right column) scales were added as its sub-categories. This resulted in a three-level deep branch for the *DeepLevelAttribute*.

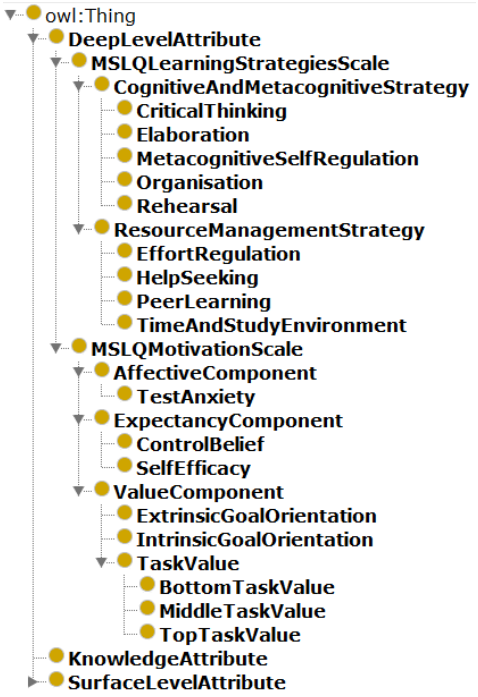

Figure 5.2 The User Diversity Ontology - MSLQ Extension showing the sub-categories of the category *TaskValue*.

As this research attempts to distinguish the extent of diversity in users, their profiles were explored against the proposed ontology via different entry points EP_{Ω}. This requires the identification of the EP_{Ω} sub-categories for diversity profiling. For this case study, a final extension was required for the leaf/bottom level concepts in the *DeepLevelAttribute* branch (e.g. *TaskValue*) to incorporate other user information in the datasets other than MLSQ. This was done by adding three other sub-categories- top, medium, and bottom (i.e. *TopTaskValue*, *MediumTaskValue*, and *BottomTaskValue- see* Figure 5.2). Each of these three sub-categories has the expected user scores as instances. The user scores, collected from the AVW-Space, were sorted top to bottom and divided to quartiles. The top and bottom sub-categories had the mean scores of the top and bottom quartiles respectively. The middle sub-

category had the scores of the two middle quartiles. This helped to classify the user scores as appropriate through semantic annotations for the diversity profiling. The resultant branch, *DeepLevelAttribute*, after final extension is five levels deep (see Figure 5.2). This final extension to the User Diversity Ontology shows a possible approach for diversity profiling using SeDDAT when the sub-categories of an interesting EP_{Ω} are not available as was the case with the MSLQ items.

5.2.4 Semantic Annotations of User MSLQ Scores

The MSLQ scores of users from both studies and the extended User Diversity Ontology were fed to GATE for tagging. The semantic annotations were done for each study at a time. Each user score is tagged with an entity URI that is the matching instance from the User Diversity Ontology. The annotations were done separately for each selected item to avoid tagging a user with an entity URI of another item that has the same score. This is due to some scores of MSLQ items being equal. In study A and study B, seven out of the nine selected MSLQ items have the same scores ranging from 1 to 5. For example, the top categories, such as *TopOrganisation,* has as instances the scores 5, 4.75, 4.5 and 4.25 for the 7 mentioned items. Because of this, the User Diversity Ontology has different versions for each selected MSLQ item that was used for annotations and diversity profiling.

5.3 Domain Diversity Profiling

This section discusses the domain diversity perspective profiling for both the videos and users. It is diversity zoom-in profiling based on the user comments from study A and study B. Comparisons are provided between both studies where applicable.

5.3.1 Domain Profiles for Videos

This section reports a zoom-in domain diversity profiling of the 8 videos listed in Table 5.1. For each video $d \in D$, SEDDAT produced a domain diversity profile based on: the set of distinct ontology entities E_d mentioned in the user comments while watching this video d; the domain ontology Ω_{domain} PreSOn and 4 entry points $EP_{\Omega_{domain}}$ in this ontology- *Thing* and the three top level sub-categories: *Delivery*, *Structure* and *VisualAid* (see Figure 5.1).

5.3.1.1 Videos' Domain Variety (v)

When the entry point was *Thing*, all videos in both studies had variety 4, meaning that for each video, the comments generated by the users referred to all top categories of PreSOn (i.e. *Delivery*, *Structure, VisualAid* and *PresentationAttribute*). This gives an indication that all videos covered at least one domain aspect related to these categories. To further investigate the videos domain coverage, each of the main presentation skills aspects (i.e. PreSOn top categories) - *Delivery*, *Structure* and *VisualAid* – is used as an entry point $EP_{\Omega_{domain}}$ for profiling. The category *PresentationAttribute* was excluded as it consists of descriptive concepts of the other three categories and it is more abstract compared to them. Table 5.4 presents and compares the videos domain variety of the three entry points for study A and study B (the diversity profiles with all properties are in Appendix C).

Table 5.4 Domain varieties of the 3 top categories of PreSOn – *Delivery*, *Structure*, and *VisualAid* against the number of their direct sub-categories (No. of sub-cat.): comparing Study A and Study B for the eight videos (E1-E4, T1-T4)

Category	**Study A**								**Study B**							
(No. of sub-cat.)	**E1**	**E2**	**E3**	**E4**	**T1**	**T2**	**T3**	**T4**	**E1**	**E2**	**E3**	**E4**	**T1**	**T2**	**T3**	**T4**
***Delivery* (6)**	4	4	3	5	2	4	4	4	5	4	4	5	3	4	5	5
***Structure* (2)**	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1
***VisualAid* (2)**	2	2	2	2	2	2	1	1	2	2	2	2	2	2	1	2

In general, domain variety based on the three entry points were similar for both studies. There were cases when the comments did not cover any entities from specific domain sub-categories. When the entry point was *Structure*, all videos, except E2 and T1 in Study A, had variety 1 because the user comments missed the sub-category *StructureApproach* (with 5 entities). Similarly, the sub-category *AudienceEmotion* (with only 1 entity) under *Delivery* was not mentioned at all in the user comments. Taking into consideration the hierarchy of the PreSOn ontology for these categories (Figure 5.1), these cases could be an indication of potential ontology deficiency – both categories would require extension.

Certain domain sub-categories were expected to be discussed in the comments, but this was not the case. When *VisualAid* was the entry point, comments on some videos missed the sub-category *VisualAidDevice* (i.e. v =1). It was in particular surprising

with video T3 as it does refer to the use of *props* (which is under *VisualAidDevice*) when delivering presentations.

Differences in the coverage by both studies, especially with regards to the entry point *Delivery* provided an indicator for further investigation. The comments in Study B has better coverage of the high-level aspects of this entry point. For example, comments in study B referred to *Preparation* (sub-category of *Delivery*) in all videos. This did not happen in Study A. Similarly, the sub-category *VerbalCommunication* (sub-category of *Delivery*) was missed in user comments in Study A on videos T1, T3 and E3.

Video E4 has the highest variety for *Delivery* in both studies – this gives an indication that this video triggered users to notice different aspects related to delivery, hence may be particularly useful for learning about presentation delivery. It is also noted that the comments in all videos covered *NonVerbalCommunication* – key skills for pitch presentation.

5.3.1.2 Videos' Domain Coverage (r)

This index shows the actual coverage of the entry points' sub-categories against the corresponding branch within PreSOn ontology. For the entry point *Thing*, Figure 5.3[30] shows the coverage r of the sub-categories of entry point *Thing* (left) and the presentational proportions $rep_{c_i}$of each category.

E2 in study A and T1 in study B had the maximum coverage ($r = 0.15$ and 0.17 respectively). E1 had the same coverage for domain categories in both studies ($r =$ 0.13). E3 and E4 covered the sub-categories equally in study A ($r = 0.11$). Similarly, in study B, E3 and T1 had similar coverage of these categories ($r =$0.17). Minimum coverage in both studies was by comments on video T4. This was in particular low in study A (0.09).

An inspection of the sub-categories' representational proportions rep_{c_i}shows the sub-categories coverage individually, which in turn shows which videos triggered good coverage for one or more categories.

[30] Coverage is rounded to three decimal places as in Tableau (https://www.tableau.com/), although some numbers when rounded will be the same, it shows them in figures as their values prior to rounding e.g. values 0.124 and 0.12 when rounded to 2 decimal places, they will be same at 0.12, yet in Tableau figure, the former number (0.124) will be shown higher in the axis.

In general, postgraduates and undergraduates tended to mention aspects related to sub-category *PresentationAttribute* which resulted in a good coverage of this category.

Delivery on the other hand was not highly covered except for video T4, which had the highest coverage for this category in both studies. This might be due to this category's size (it is the largest branch). T1 seems to be a good video for aspects related to *VisualAid* as this category had the highest coverage in both studies on this video. E3 covered the categories fairly the same in both studies. This can be an indication that this video is good for general knowledge about different presentation skills. E1 triggered postgraduates to focus on *Structure*, while undergraduates covered *PresentationAttributes*.

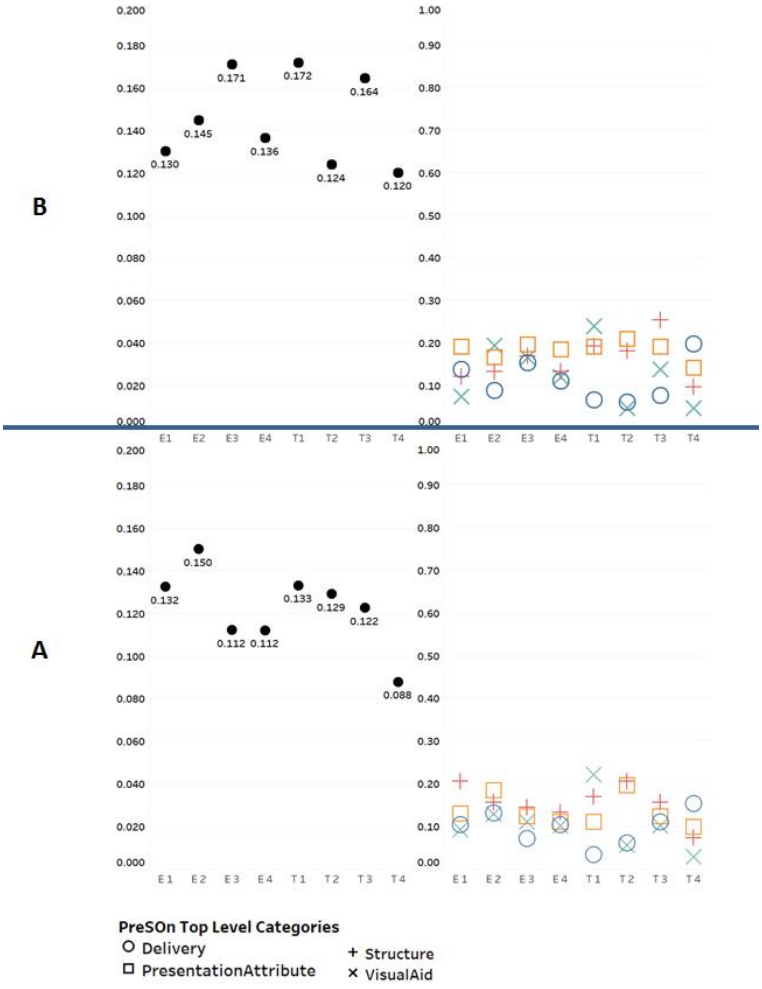

Figure 5.3 Domain coverage (left) of *Thing* in PreSOn and proportions (right) of the top 4 categories.

Zooming into the PreSOn ontology by selecting each top category- *Delivery*, *Structure* and *VisualAid*, as an entry point deepens the understanding of the domain coverage of the videos.

For entry point *Delivery,* in general both studies had very similar coverage - 0.12(0.05) and 0.12(0.03) for study A and study B respectively. T3 had maximum coverage for study A (r =0.23), while E3 was maximum for this index in study B (r =0.17). This video had the minimum coverage in study A (r =0.07). It was expected that T4 would be highest for this entry point, but it followed closely as it came second for study A and third for study B (r =0.16 and 0.15 respectively). An inspection of number of entities around these videos showed that although T4 had higher number of entities compared to the maximum videos in both studies (28 in study A and 36 in study B), entities are mostly coming from sub-category *NonVerbalCommunication* (20 out of 28 and 28 out of 36 respectively). Therefore, the overall coverage was lower. This could be an indication that this video is a good one for learning aspects related to non-verbal communication skills - a key skill for pitching presentations. Although E4 had covered most of the high-level aspects under *Delivery* (v = 5 - Table 5.4), the coverage was low (0.09). This was the case for both studies.

For entry point *Structure,* again both studies had similar coverage - 0.17(0.05) and 0.18(0.06) for study A and study B respectively. Results were similar to the findings based on the entry point *Thing*. In study A, E1 and T2 were maximum with identical coverage at 0.23. Both videos had domain variety of 1 with regards to this entry point covering sub-category *StructureComponent* with 17 distinct entities out of 75. T3 was maximum for study B at 0.28 covering the same category with 23 distinct entities. T4 was minimum for both studies covering 6 and 8 distinct entities for same sub-category resulting in r =0.08 and 0.11 for study A and study B respectively. T1 and E2 were the only videos that covered both sub-categories of this entry point in study A, they had 1 entity out of 5 from sub-category *StructureApproach*. T1 had 13 entities compared to 12 from E2 covering sub-category *StructureComponent*. This resulted in coverage being r = 0.19 and 0.18 respectively.

For entry point *VisualAid,* the coverage of this sub-category was very low for study A-0.09(0.04) compared to study B 0.12(0.06). T1 was top for both studies (0.16 for study A and 0.22 for study B respectively). It had 26 entities in total - 22 out of 84 (71 classes and 13 instances) from *VisualAidArtefact* and 2 out of 23 (22 classes and 1 instance) from *VisualAidDevice*. T4 was bottom for both studies at 0.04 and 0.05 for study A and study B respectively. This video covered only 3 entities in study A from sub-category

VisualAidArtefact and 5 in study B from the 2 sub-categories of *VisualAid*- 1 entity from *VisualAidDevice* and 4 entities from *VisualAidArtefact*.

5.3.1.3 Videos' Domain Balance (b)

Domain balance shows whether the distribution of the annotated entities under the entry point's sub-categories is even or not. Figure 5.4 (left) shows the overall balance (b) for every video in both studies. It also shows proportions (p_{c_i}) of the top 4 sub-categories of *Thing* against the set of entities E covered in the social cloud around each video d.

With entry point being *Thing*, videos scored closely for balance b in both studies except for T2 and T4 as both had low domain balance. Looking at the proportions p_{c_i} in Figure 5.4 shows the uneven distribution of the categories' coverage around these videos. In T2 (on how to close and open a presentation), the comments in both studies focused on *PresentationAttribute*, whereas in T4 (on how to speak with confidence), the comments in both studies focussed on *Delivery*. Hence the low balance for each of these two videos is not a surprise as they are aiming for a narrower range of topics. The category *PresentationAttribute* had highest proportions p_{c_i}against the set of entities E covered in the social cloud, especially in study B. This conform to finding of domain coverage discussed above. This could be an indication that the users tended to describe the other presentation aspects (*Delivery*, *Structure* and *VisualAid*) in their comments.

Moving a level lower with *Delivery* as an entry point resulted in T2 having maximum domain balance. T4 remained unbalanced for study B. this is expected as building from domain coverage based on this entry point showed that the main coverage is coming from one sub-category, *NonVerbalCommunication* i.e. entities are distributed unevenly across sub-categories of entry point *Delivery* on video T4. Video T3 did not only had maximum coverage based on this entry point in study A, but it also scored high for balance ($b = 1.31$) coming second after top video T2 ($b = 1.37$). T3 was also maximum for study B at $b = 1.51$.

As majority of videos covered only one sub-category of *Structure* – *StructureComponent*, domain balance with this entry point was zero for those videos (special case 2 - $v = 1$). This is 6 videos in study A and all videos (8) in study B. In

study A, videos E2 and T1 covered the 2 sub-categories of structure and scored 0.27 and 0.25 respectively.

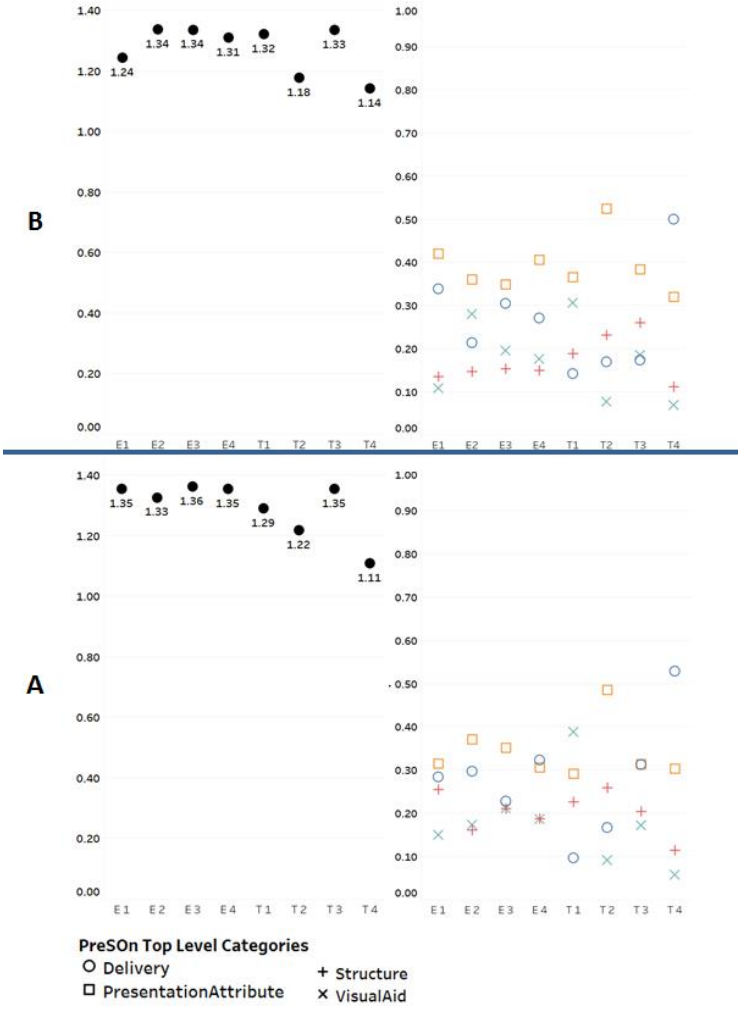

Figure 5.4 Domain balance for *Thing* (left) and the proportions of PreSOn 4 top categories.

When the entry point was *VisualAid,* E1 was top in both studies scoring 0.50 for study A and 0.56 for study B. T3 had minimum domain balance, zero, for both studies as it covered only one sub-category of this entry point. T4 scored zero for study A. Distribution of entities across 2 sub-categories of *VisualAid* on E3 was more even in study A ($b = 0.45$) compared to that in study B as this video scored lower at 0.21.

5.3.1.4 Videos' Domain within disparity (d_w)

Within Disparity indicates the spread within a category – the higher the within disparity, the broader the spread of entities in a category. Domain within disparity of both studies were close (11.73(1.83) and 11.91(1.18) for study A and study B respectively). Study A had more varied values for this index ranging from 8.78 (minimum) and 14.37 (maximum).

Figure 5.5 (left) shows the within disparity of all the videos at the level *Thing*. By inspecting the categories' dispersions (Figure 5.5 (right)), *Delivery* consistently scores the highest (except for T1), whilst *Structure* tends to be lowest. In other words, a more dispersed coverage on *Delivery* (this is the ontology branch with most entities - 183). For T2, an inspection of the entities showed that comments in Study B seems very scattered in terms of concepts on *Delivery*; and comments on *Structure* are concentrated on concepts around *StructureComponent* (opening, body, closing).

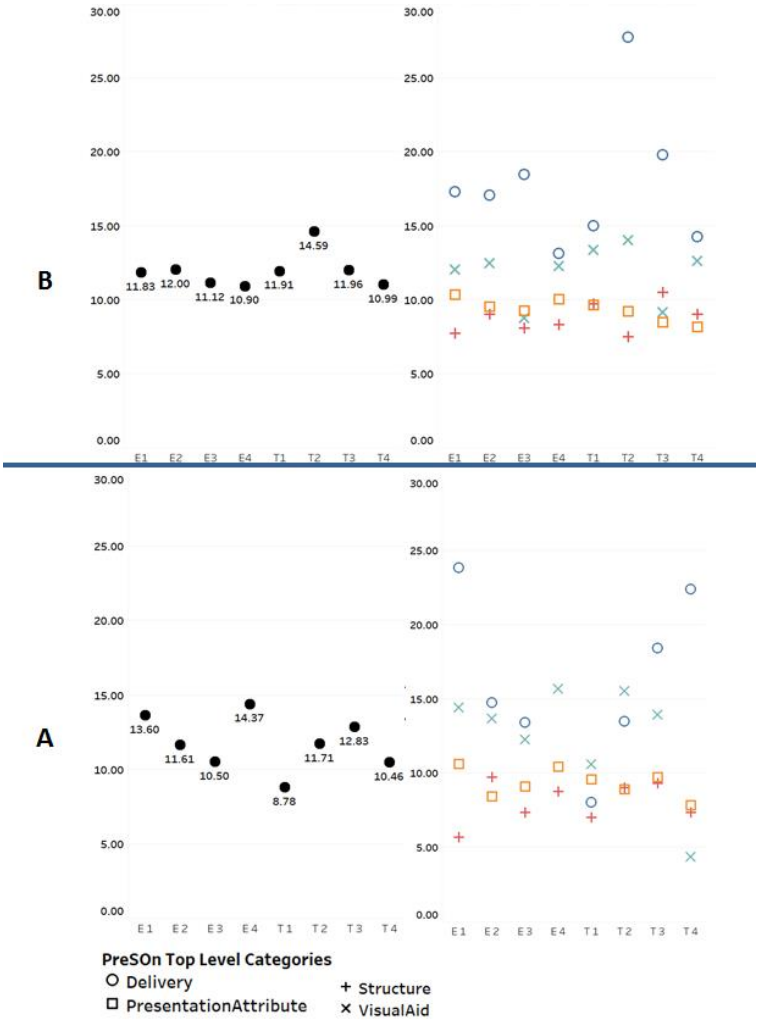

Figure 5.5 Domain within disparity for *Thing* (left) and dispersion for the PreSOn's 4 top categories for study A and study B.

Zooming one level down into PreSOn with the entry point *Delivery* gave more insight to the broader and dispersed coverage of this category. In study A, T4 had the highest within disparity (d_w =6.23), while E3 had the minimum (d_w = 2.70). In study B, T1 scored maximum d_w =6.47 and E4 scored minimum d_w =2.32.

The dispersion of entities within their categories was high when these entities are from different parts of this category's branch and low when they are close and from same

level. For example, in study A, T4 triggered 28 distinct entities which covered 4 sub-categories of *Delivery* (Table 5.4) - It covered *NonVerbalCommunication* with 20 entities out of 107 that are available in this sub-category; *Preparation* with 1 out of 18; *SpeakerEmotion* with 6 out of 18 and *VerbalCommunication with* 1 out of 22. The entities from *NonVerbalCommunication* were scattered on different parts on this branch e.g. comments covered the entities *eye_contact* and *HandGesture* which are from different levels of sub-category *bodyLanguage* under *NonVerbalCommunication*. Also, comments covered entities like *Speed* and *Volume* which are under another sub-category, *Paralanguage*. This resulted in high dispersion (dis_{c_i} =17.9) for sub-category *NonVerbalCommunication*. The other sub-categories, *Preparation* and *VerbalCommunication*, had zero dispersion and *SpeakerEmotion* had dispersion of dis_{c_i} =7. This explains the results on T4.

On the other hand, video E3, which scored minimum within disparity for study A, covered 3 sub-categories of *Delivery* with 13 distinct entities – 11 entities from *NonVerbalCommunication* and 1 entity from *Preparation* and *SpeakerEmotion*. The entities from *NonVerbalCommunication* were mostly concentrated (distributed closely) from one part of the branch under this sub-category, which is subclass *Paralanguage.* This gave dispersion value of dis_{c_i} =8.09, hence gave lower within disparity compared to that on T4 and other videos. The same explanation can be given about the videos in study B.

Switching the entry point to *Structure* showed that T3 was maximum for both studies (d_w =9.31 and 10.48 respectively). T2 was second top in study A (d_w =9). Although it was minimum in study B, it can still be considered high compared to its value in study A (d_w =7.47). T1 came second top in study B (d_w =9.69) but minimum in study A (d_w =2.81). An inspection of the entities and their dispersion on this video (T1) showed that in study A, it covered the two sub-categories – *StructureComponent* and *StructureApproach* of *Structure* with 14 distinct entities, one of which is from *StructureApproach* resulting in zero dispersion for this sub-category and 5.62 for the other as entities are distributed closely within this sub-category. In study B, T1 covered only *StructureComponent* with 16 distinct entities scattered within this category resulting in higher dispersion of 9.69.

Finally, profiling based on *VisualAid* - similarly T3 was top for both studies (9.13 for study A and 13.91 for study B). It was expected that T2 (from B) and E4 (from A) could be maximum considering results based on entry point *Thing* (left Figure 5.5), but this was not the case. T4 was low for both studies 4.75 for study A and (4.33) for study B - it was lowest for study B. E3 was minimum for study A (3.88). T1 had good and close coverage of *VisuaAid* in both studies (0.16 and 0.22), but within disparity for study A was higher (8.49) than that in study B (4.74). The reason is same as discussed above with the other entry points- it is the distribution of the entities within their categories. This could be an indication that this video triggers the users to give a good coverage of this category, however, this coverage can be different by different users.

5.3.2 Domain Profiles for Users

For each user $u \in U$, SEDDAT produced a domain diversity profile (zoom-in profiling per user). This was based on the set of distinct entities E_u mentioned in all comments made by u while watching videos. For users, the entry point $EP_{\Omega_{domain}}$ *Thing* in PreSOn was used to generate the diversity profiles for both studies. The user diversity profiles and other profile attributes (e.g. demographics, MSLQ, knowledge, etc.) are analysed below, where significant differences are reported.

Comparing study groups. To see if postgraduates (Study A) differ from the undergraduates (Study B) in their background knowledge, attitudes towards learning, and their behaviour during video watching, their personal profiles and diversity profiles were compared. Table 5.5 reports the profile items with significant difference between the two studies (see Appendix D for other profile items with associated domain properties).

There were high significant differences between the two study groups on domain variety, balance, coverage and within disparity. Postgraduates seem to have better recognition of the top domain categories of PreSOn (higher variety), tend to mention more aspects related to these categories (higher coverage) evenly (higher balance), and have dispersed domain coverage within these categories (higher within disparity) compared to undergraduates. This indicates that postgraduates were more diverse with regards to presentation skills.

Postgraduates reported significantly more training and experience on presentation skills. Postgraduates scored higher on most of MSLQ items covered in the case

studies (refer to Table 5.3 for MSQL scales). In terms of motivation (based on Value Components and Expectancy Components scales), postgraduates scored higher for Task Value and Intrinsic Goal Orientation (Intrinsic), whereas undergraduates scored higher on Extrinsic Goal Orientation (Extrinsic) and Control Beliefs. There was no difference between the study groups in Self-Efficacy for Learning and Performance (Self-Efficacy). Postgraduates scored higher for all the items covered in the Learning Strategies scales (Cognitive and Metacognitive Strategies and Resource Management Strategies) - they scored significantly higher in terms of Organisation, Effort regulation, Elaboration and Metacognitive Self-Regulation (Self-Regulation). There was marginal significant difference in terms of Rehearsal between the study groups. These figures seem to correlate to the fact that Study A was on a volunteer basis (more comments) whereas Study B was part of a course assessment (fewer comments).

Table 5.5 Significant differences between participants from Studies A and B; † denotes Likert scale was used - 1 (lowest) to 5 (highest); *** significance at $p < .001$, ** at $p < .01$ and * at $p < .05$.

User profile items	Study A (38)	Study B (141)	Significance
Domain variety	3.66 (0.78)	2.89(1.16)	U= 1667.5***
Domain balance	1.19 (0.32)	0.87 (0.46)	U= 1260.5***
Domain coverage	0.05 (0.02)	0.03(0.02)	U= 1387***
Domain within disparity	10.66 (4.58)	8.01 (5.62)	U= 1824**
Comments (no. of)	19.58 (13.19)	7.87 (9.56)	U= 1040***
Training†	2.16 (0.95)	1.7 (0.78)	U= 1905**
Experience†	2.87 (0.78)	2.34 (0.84)	U= 1735***
Task Value†	4.49 (0.39)	3.97 (0.59)	U= 1265***
Control Beliefs†	3.91 (0.46)	4.14 (0.57)	U= 1961.5**
Intrinsic†	4.05 (0.52)	3.76 (0.59)	U= 1984**
Extrinsic†	3.38 (0.83)	4.19 (0.63)	U= 1193***
Effort Regulation†	2.93 (0.44)	3.57 (0.66)	U= 1128***
Organisation†	3.84 (0.94)	3.23 (0.9)	U= 1667.5***
Elaboration†	4.13 (0.54)	3.67 (0.66)	U= 1547***
Self-Regulation†	3.61 (0.39)	3.28 (0.51)	U= 1396***
Rehearsal†	3.4(0.8)	3.17((0.72)	U= 2185*

Comparing user personal attributes. To see if the learners' personal characteristics, gender or language (native/non-native speaker), influenced diversity scores, each diversity index was compared across all users' personal attributes as follows:

Gender. The only significant difference was in Study A (Table 5.6). The domain within disparity ($U = 103$, $p < .05$) was significantly higher for female participants ($n = 26$) than male participants ($n = 12$). Study B had 82 males and 58 females- one user (ID: ari49) in study B was excluded from this statistical analysis as it was classified as other for gender.

Table 5.6 Comparing diversity indices based on gender for study A and study B.

	Study A		
	Male (12)	**Female (26)**	**Significance**
Domain within disparity	8.31(5.35)	11.75(3.67)	U=103*

Language. There was no significance difference for natives versus non-natives in study A (23 versus 15) and Study B (119 versus 22) respectively. This is surprising, as it was expected that the language attribute will have an impact on the coverage of the domain (i.e. diversity scores).

Comparing the most and least diverse users. To further understand if some of the learners' profile items contributed to the domain coverage, their diversity scores are ranked for each study - variety, then balance, then coverage and finally within disparity. The two extreme quartiles, top and bottom quartiles, were analysed for the following attributes: all MSLQ items, training for giving a presentation, experience in giving presentations, watching YouTube and using it for learning, and number of comments.

There were significant differences on the number of comments in both studies (Table 5.7) between these three subgroups of participants (all pairwise comparisons significant at $p < .001$). It surprising that there are no significant differences between the selected profile attributes. Although it is expected that more trained and experienced users have higher diverse coverage, this was not the case.

Comparing correlations. For both studies (Table 5.8 and Table 5.9), domain variety, balance, coverage, and within disparity are strongly correlated (with correlations ranging from .51 to .94). Also, the number of comments is strongly correlated with domain variety, balance, coverage and within disparity in both studies (with correlations ranging from .47 to .84). This indicates that users should be triggered to write more comments, as the more they write the more they notice with regards to the domain while watching the videos.

Table 5.7 Comparing quartiles defined on domain variety, balance, coverage and within disparity for study A and study B.

	Study A			**Study B**		
	Top quartile (10)	Bottom quartile (10)	Significance	Top quartile (35)	Bottom quartile (35)	Significance
Comments (no. of)	24.8(10.62)	8.4 (7.14)	U=8***	16.94(14.52)	1.86(1.4)	U=15***

Table 5.8 Correlation between diversity properties.

	Variety & Balance	Variety & coverage	Varity & Within disparity	Balance & coverage	Balance & Within disparity	Coverage & Within Disparity
Study A	0.71***	0.53***	0.51***	0.66***	0.72***	0.57***
Study B	0.94***	0.70***	0.54***	0.70***	0.52***	0.70***

Table 5.9 Correlation between the number of comments and diversity properties.

	Domain variety	Domain balance	Domain coverage	Domain within disparity
Comments - Study A	0.58***	0.61***	0.84***	0.47**
Comments - Study B	0.82***	0.79***	0.76***	0.65***

5.4 User Diversity Profiling

Diversity profiling for this perspective involves the users U who wrote comments on the eight videos (tutorials and examples). This is conducted based on their deep-level attributes. SeDDAT used the distinct entities E_a from annotating users' MSLQ scores for diversity profiling, i.e. overview profiling.

Initial analysis has been conducted across the established quartiles of the MSLQ scores from both studies. This is to identify possible interesting items for diversity profiling. Three items have been selected as they show interesting results to be reported. The rest of the items calculated for study A and B showed no significance difference as the scores are similarly distributed across the quartiles i.e. will produce very close, mostly the same scores for variety, coverage, balance, and within disparity in both studies. The selected items for user diversity profiling are, Task value, Effort regulation, and Organisation. Their definitions (discussed in section 5.2.3) should help to interpret the user diversity profiles.

Therefore, three diversity profiles are generated for postgraduates (study A) and undergraduates (study B) - six profiles in total, where the ontology Ω_{user} is the Extended User Diversity Ontology. For each profile, set of distinct entities E_a from annotating one of the selected MSLQ items was used with associated entry point

$EP_{\Omega_{user}}$ within the *DeepLevelAttribute* branch - *TaskValue*, *EffortRegulation* or *Organisation*. Postgraduates' and undergraduates' profiles are compared based on these items as follows.

5.4.1 Diversity Profiling of User Task Value

The *TaskValue* (*TV*) entry point has three sub-categories in the User Diversity Ontology (see Figure 5.2) – *TopTaskValue* (4 entities), *MiddleTaskValue* (5 entities) and *BottomTaskValue* (4 entities).

For postgraduates, their scores (5 distinct entities in total) covered only *TopTaskValue* and *MiddleTaskValue* (i.e. task value variety for study A was $v = 2$). All distinct entities except for one are in the *TopTaskValue* sub-category, which resulted in full coverage of this sub-category (rep_{c_i} =1). Having one entity in the *MiddleTaskValue* sub-category resulted in low proportions for balance and coverage p_{c_i} and $rep_{c_i} = 0.2$ and minimum dispersion $dis_{c_i} = 0$. This in turn resulted in low overall user coverage (r =0.6), unbalanced distribution of user entities ($b = 0.5$) and fairly low within disparity (d_w =1.5).

The undergraduates, on the other hand, had higher diversity indices with regards to task value compared to the postgraduates. Their scores (11 distinct entities) covered the three sub-categories of *TaskValue* i.e. undergraduates had different task value levels ranging from bottom to top. They had full coverage of the *MiddleTaskValue and TopTaskValue* sub-categories (i.e. their representational proportion rep_{c_i} =1). The *BottomTaskValue* sub-category had two entities (out of 4) covered by users from study B (rep_{c_i} =0.5), hence the overall coverage is higher (r =0.83). The user distribution across the sub-categories is more even resulting in higher balance (b =1.04). Against this shallow branch (2 levels deep), within disparity can be considered high for undergraduates (d_w =2.73) and it is higher than postgraduates.

Overall, although undergraduates had higher and diverse coverage of task value levels (sub-categories), postgraduates can be considered to have higher motivation in terms of Task value. This confirms findings in Table 5.5 and seems to correlate with the fact that study A with postgraduates was on a voluntary basis, whereas study B was part of a course work for undergraduates. In such case having higher diverse

coverage, does not necessarily mean better performance in terms of an item within MSLQ. See diversity profiles of both studies in Table 5.10 and Table 5.11.

Table 5.10 *TaskValue* (*TV*) diversity indices - variety, balance, coverage and within disparity for study A and study B.

	User (*TV*) **variety**	User (*TV*) **balance**	User (*TV*) **coverage**	User (*TV*) **within disparity**
Study A	2	0.50	0.6	1.5
Study B	3	1.04	0.83	2.73

Table 5.11 Proportions for balance $\boldsymbol{p_{c_i}}$ and coverage $\boldsymbol{rep_{c_i}}$with dispersions $\boldsymbol{dis_{c_i}}$of the *TaskValue*'s sub-categories - *TopTaskValue* (Top), *MiddleTaskValue* (Middle) and *BottomTaskValue* (Bottom).

	User (*TV*) $\boldsymbol{p_{c_i}}$			User (*TV*) $\boldsymbol{rep_{c_i}}$			User (*TV*) $\boldsymbol{dis_{c_i}}$		
	Bottom	*Middle*	*Top*	*Bottom*	*Middle*	*Top*	*Bottom*	*Middle*	*Top*
Study A	0.0	0.2	0.8	0.0	0.2	1.0	0.0	0.0	3.0
Study B	0.18	0.45	0.36	0.5	1.0	1.0	2.0	3.2	3.0

5.4.2 Diversity Profiling of User Effort Regulation

The *EffortRegulation* (*ER*) entry point had three levels (sub-categories) - *TopEffortRegulation* (with 4 entities), *MiddleEffortRegulation* (with 9 entities), and *BottomEffortRegulation* (with 4 entities).

Postgraduates (9 distinct entities) covered only the *MiddleEffortRegulation* sub-category i.e. $v = 1$. The *MiddleEffortRegulation* is fully covered, hence proportions for balance and coverage p_{c_i} and rep_{c_i} =1. This meet the special case 2 (i.e. when v =1). Therefore, user coverage $r = 1$, balance $b = 0$, and within disparity equals the covered sub-category dispersion (i.e. $d_w = dis_{c_i} = 3.56$).

This was not the same for undergraduates (13 distinct entities). All three sub categories/levels are covered i.e. maximum variety $v =$ 3. All entities of *TopEffortRegulation* are covered (p_{c_i} =0.31 and rep_{c_i} =). The *MiddleEffortRegulation* category is covered except for one entity i.e. 8 entities out of 13 are from this sub-category ($p_{c_i} =$ 0.62 and $rep_{c_i} =$ 0.89). Only one entity is covered from the *BottomEffortRegulation* ($p_{c_i} =$ 0.08 and $rep_{c_i} =$ 0.25). This resulted in high overall coverage ($r =$ 0.71) and slightly low balance ($b =$ 0.86) due to the uneven distribution of users (majority are in *MiddleEffortRegulation*). As the *MiddleEffortRegulation* sub-

category for postgraduates was fully covered, this resulted in higher dispersion and higher overall disparity (3.56) compared to that for undergraduates (2.17). The low dispersion ($dis_{c_i} = 0$) of the *BottomEffortRegulation* category influenced the overall with disparity for the undergraduates. See Table 5.12 and Table 5.13.

Overall, based on the distribution of the user scores across the *EffortRegulation* levels (sub-categories), it seems that undergraduates had higher level of resource management strategies in terms of effort regulation especially that most students' scores are in the top and middle categories. Again, this confirm findings in Table 5.5.

Table 5.12 *EffortRegulation* (*ER*) diversity indices - variety, balance, coverage and within disparity for study A and study B.

	User (*ER*) **variety**	User (*ER*) **balance**	User (*ER*) **coverage**	User (*ER*) **within disparity**
Study A	1	0.0	1.0	3.56
Study B	3	0.86	0.71	2.17

Table 5.13 Proportions for balance $\boldsymbol{p_{c_i}}$ and coverage $\boldsymbol{rep_{c_i}}$with dispersions $\boldsymbol{dis_{c_i}}$of the *EffortRegulation*'s sub-categories - *TopEffortRegulation* (Top)*, MiddleEffortRegulation* (Middle) and *BottomEffortRegulation* (Bottom).

	User (*ER*) $\boldsymbol{p_{c_i}}$			User (*ER*) $\boldsymbol{rep_{c_i}}$			User (*ER*) $\boldsymbol{dis_{c_i}}$		
	Bottom	*Middle*	*Top*	*Bottom*	*Middle*	*Top*	*Bottom*	*Middle*	*Top*
Study A	0.0	1.0	0.0	0.0	1.0	0.0	0.0	3.56	0.0
Study B	0.08	0.62	0.31	0.25	0.89	1.0	0.0	3.5	3.0

5.4.3 Diversity Profiling of User Organisation

The entry point *Organisation* (*O*) had three levels (sub-categories) - *TopOrganisation* (with 4 entities), *MiddleOrganisation* (with 9 entities), and *BottomOrganisation* (with 4 entities). Users from both studies had similar profiles with regards to this entry point. Postgraduates covered 8 entities from this branch compared to 9 from undergraduates. The entities for the postgraduates are distributes as follows: 2 entities from the *TopOrganisation* sub-category, 5 entities from *MiddleOrganisation*, and 1 entity from *BottomOrganisation*. Undergraduates had same distribution for *TopOrganisation* and *MiddleOrganisation* sub-categories and had 2 entities from the *BottomOrganisation*. For that, diversity indices were close in value where undergraduates had slightly higher values.

Both studies had same scores for *Organisation* variety $v = 3$ i.e. students in both studies had all different levels of organisation strategies ranging bottom to top. Overall balance is slightly higher for undergraduates ($b = 1$ compared to 0.90) due to the more even distribution of users (for *TopOrganisation* and *BottomOrganisation*). This influenced the within disparity scores, which was higher for undergraduates (d_w =1.73 and 2.40 for study A and study B respectively). As only one entity was covered within the *BottomOrganisation* category for study A, this resulted in its dispersion dis_{c_i} being zero. See Table 5.14 and Table 5.15.

Overall, postgraduates and undergraduates showed almost the same level of cognitive and metacognitive strategies in terms of organisation, which conform to results in Table 5.5.

Table 5.14 *Organisation* (*O*) diversity indices - variety, balance, coverage and within disparity for study A and study B.

	User (*O*) **variety**	User (*O*) **balance**	User (*O*) **coverage**	User (*O*) **within disparity**
Study A	3	0.90	0.44	1.73
Study B	3	1	0.52	2.40

Table 5.15 Proportions for balance $\boldsymbol{p_{c_i}}$ and coverage $\boldsymbol{rep_{c_i}}$ with dispersions $\boldsymbol{dis_{c_i}}$ of the *Organisation*'s sub-categories- *TopOrganisation* (Top)*, MiddleOrganisation* (Middle) and *BottomOrganisation* (Bottom).

	User (*O*) $\boldsymbol{p_{c_i}}$			User (*O*) $\boldsymbol{rep_{c_i}}$			User (*O*) $\boldsymbol{dis_{c_i}}$		
	Bottom	*Middle*	*Top*	*Bottom*	*Middle*	*Top*	*Bottom*	*Middle*	*Top*
Study A	0.13	0.63	0.25	0.25	0.56	0.5	0.0	3.2	2.0
Study B	0.22	0.56	0.22	0.5	0.56	0.5	2.0	3.2	2.0

5.5 Across Disparity: Linking Domain and User Diversity Perspectives

As discussed and illustrated in chapter 4, linking the two diversity perspectives facilitates the measurement of across disparity. Across disparity distinguishes between the domain categories that are mentioned together in the comments via their entities frequencies i.e. how many (distinct) times a user or user group mentioned this category in the comments. Some categories are mentioned together in the comments, yet one is more/less frequent.

Across disparity was measured per video for both studies. The entry point *Thing* was used to calculate across disparity on the individual level (per user) and group level (per user groups: gender (males and females) and language (native and non-native speakers). The personal attributes (i.e. gender and language) in the user profile $userProfile$ are used to categories and group the user entities on a video. Frequency vectors across PreSOn's top 4 categories are constructed based on the entities of male and female users and native and non-native users. Across disparity is measured based on gender and language for the 8 videos in both studies i.e. each video has 4 values for the across disparity property. This facilitates comparing user groups within a study and across both studies.

Group across disparity in this chapter illustrates how user diversity attributes that lack proper categorisations against ontologies (i.e. the User Diversity Ontology) for diversity profiling can be utilised for across disparity measurements and diversity analysis. For example, the attribute gender cannot be extended any further for diversity measurements, specifically for balance and disparity (only variety can be identified). Results are reported next.

5.5.1 Individual-level Across Disparity

Undergraduates (study B) scored significantly higher (U=14, p<.5) than postgraduates (study A) at 2.48(0.52) and 2.15(0.35) respectively. Maximum across disparity (d_a) for study A was for the video T4 (The five secrets of speaking with confidence) at 2.75, whereas T1 (How to Give an Awesome (PowerPoint) Presentation) at 2.97 scored top for study B. Minimum across disparity for study A was scored by video E2 (1.60). T3 scored minimum for study B at 1.36. This indicates that E2 and T3 triggered the users to mention the categories similarly in both studies. T3 was minimum for study B but scored high (top 3) for study A at 2.23 (Figure 5.6).

Drilling down the cosine similarity values across the four domain categories of PreSOn reveals which categories are similar/dissimilar in terms of frequency per video. The higher the cosine similarity (closer to 1), the closer the categories in terms of frequency. The lower the cosine similarity (closer to zero), the more disparate the categories. For this social cloud and based on the individual level, there were no zero cosine similarities between PreSOn categories i.e. the categories are mentioned differently in the user comments. However, most of pair-wise cosine similarities are

relatively low - see Figure 5.7 for the cosine similarities across the 8 videos for this level.

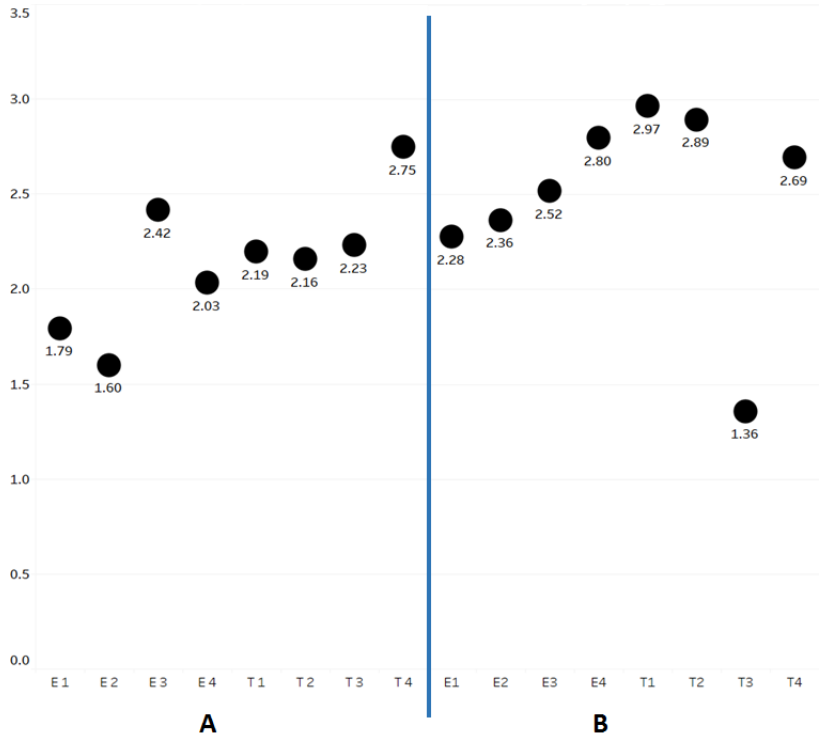

Figure 5.6 Across disparity based on the individual level (per user) for study A and study B.

The maximum cosine similarity for study B was scored by video T3 (Make a presentation like Steve Jobs) between the categories *Structure* and *PresentationAttribute* at 0.79. Although the video T3 scored high (top 3) across disparity (d_a) for study A, it scored high for cosine similarity for the same categories (*Structure* and *PresentationAttribute*) at 0.81. This is an indication that this video triggered the users to mention aspects related to these categories similarly in terms of frequency. The maximum cosine similarity for study A was scored by video E2 (Social media and the end of gender) between the categories *Delivery* and *PresentationAttribute* at 0.86.

As the category *PresentationAttribute* consists of descriptive concepts (e.g. engaging, informative, organised, etc.) of the other aspects of presentation skills (*Structure*, *Delivery* and *VisualAid*), this could explain the reason behind this category having similar frequencies with the other categories. As mentioned before, it seems that users tended to describe these categories when they mention them. It seems that E2 was mostly high (in frequency) for this category compared with the other categories.

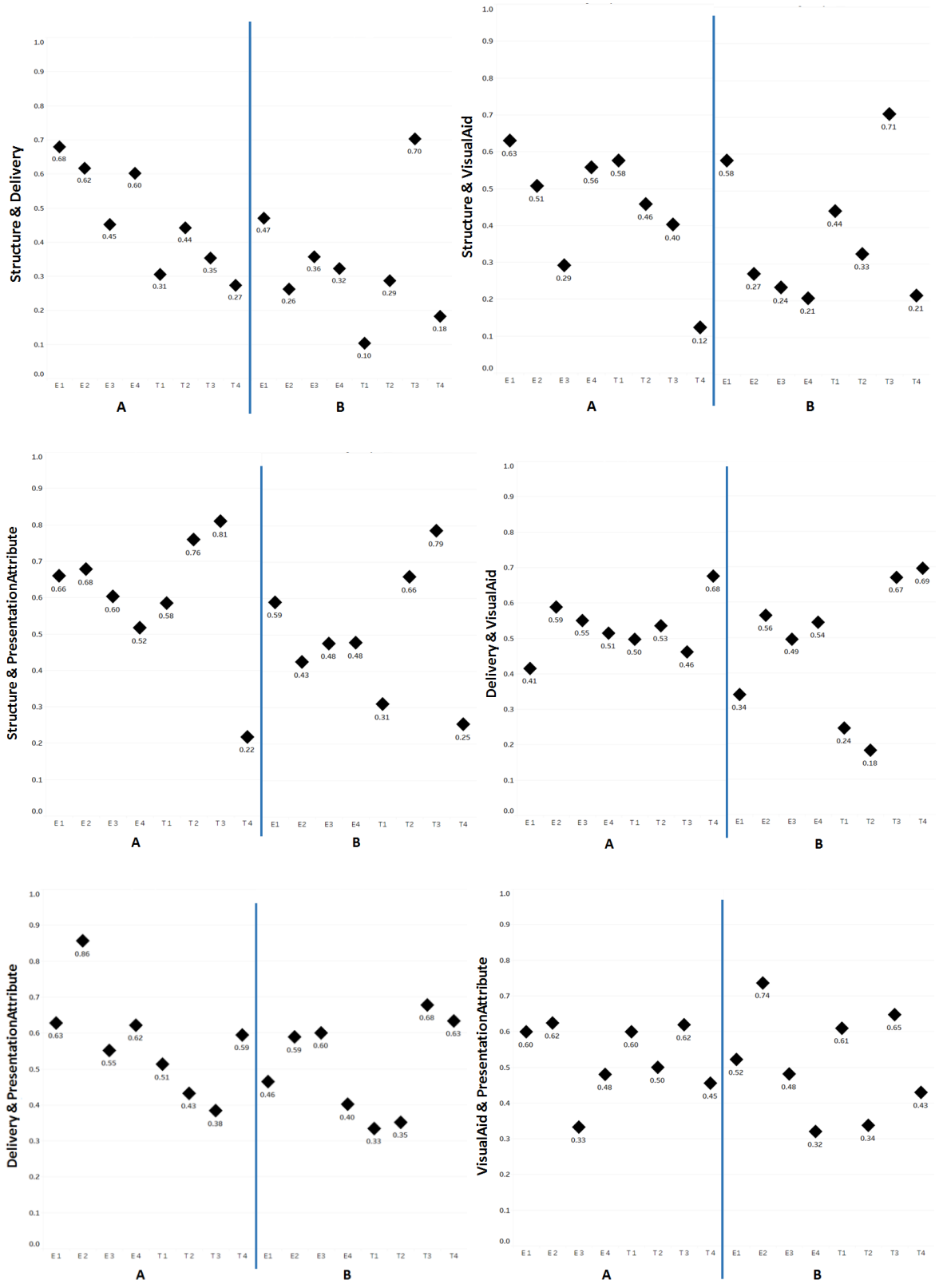

Figure 5.7 Cosine similarities between PreSOn's top level 4 categories for study A and B on the individual-level.

The minimum cosine similarity for study A was between categories *Structure* and *VisualAid* at 0.12 scored on the top video for this study, T4. Study B had minimum

cosine similarity between *Structure* and *Delivery* at 0.10 scored by the top video T1. As the categories *Structure* and *VisualAid* were not covered well in the comments (as was discussed in section 5.3.1), hence the frequencies were lower resulting in low overall across disparity, especially for the category *Structure*.

Building from findings on videos domain profiling can explain why some videos triggered high/low frequencies for each pair of categories. For example, T4 had very low cosine similarity (0.27) with *Structure* (see top left of Figure 5.7) as this video had high domain coverage of the category *Delivery* and low coverage for *Structure* (see section 5.3.1), hence the low similarity is not a surprise as the focus was on *Delivery*. On the other hand, E1 had highest coverage for *Structure* and a good coverage for *Delivery*, here it seems that the users mentioned both categories as regular resulting in high similarity.

5.5.2 Group-level Across Disparity

The scores for this level based on users' gender and language are reported next.

5.5.2.1 Across Disparity Based on Gender

Male and female users are compared within a study and across both studies in terms of how frequent they mention a category. The categories are distinguished across males and females. This is to identify which categories are noticed more often and which categories are usually mentioned together with regards to gender.

In general (Figure 5.8), there was no significant difference between the user groups for both studies based on gender. Users in study B had very similar results for this index (2.47(0.62) and 2.44(0.49) for males and females respectively). The top three videos for males and females in this study are videos T2, T1 and E4 at: (3.05, 2.95 and 2.76) for males and (3.07, 2.98 and 2.80) for females respectively. Video T3 scored minimum for males (1.11), and it scored low as well for females (1.89), but T4 was minimum (1.77).

Results from user groups in study A were more different compared to that in study B (1.72 (0.64) for males and 2.05 (0.28). Video T2 was top for males (d_a =2.50), whereas video T1 was top for females (d_a =2.39). Video E2 was top two for male postgraduates. Video E4 had low across disparity for both genders (0.77 and 1.65 for males and females respectively).

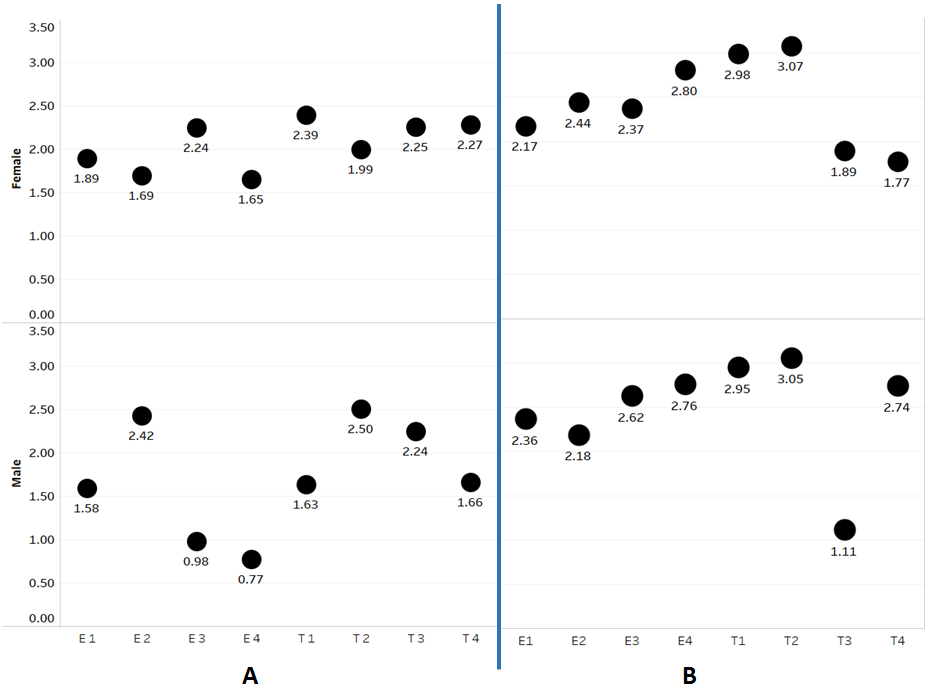

Figure 5.8 Across disparity for male and female users from study A and study B.

Drilling down to cosine similarities between categories across genders for both studies can deepen findings from the individual level and domain coverage in general. For example, an inspection of cosine similarities for both genders on study A showed that male users missed completely the category *Delivery* on video E4 and *Structure* on video E3 (d_a =0.98) i.e. the domain coverage of these categories on both videos was by female postgraduates. Although T4 had relatively high cosine similarities between *Structure* and *Delivery* by female undergraduates (study B) and male postgraduates (study A) - 0.78 and 0.54 respectively, the opposite gender in both studies had low similarities (0.39 by female postgraduates and 0.13 by male undergraduates). This seems to be the cause of the overall low similarity on the individual level (see section above). E1 on the other hand had fairly high cosine similarities for *Structure* and *Delivery* by both genders, which in turn seems to result in the high similarity based on individual level. It was interesting to see that T3 (Make a presentation like Steve Jobs) had completely different cosine similarities for the same gender, males - it scored zero cosine similarity between *Structure* and *PresentationAttribute* by male postgraduates (Study A) and relatively high by male undergraduates (0.79).

5.5.2.2 Across Disparity Based on Language

There were significance differences between native (English speakers) and non-native users in both studies- in study A - 1.84(0.27) and 1.09(0.36) respectively, where U=0,

p < 0.001, and in study B – 2.5(0.57) and 1.87(0.38) respectively, where U=12, p < .05.

In both studies, maximum video for non-natives was minimum for natives. In study A (22 natives versus 15 non-natives), E2 scored top for this index at 1.39 by non-natives, whereas it was bottom for natives at 1.46 (although it was higher than non-natives for natives). In study B (119 natives versus 23 non-natives), T3 was top for non-natives at 2.49 and bottom for natives at 1.23. See Figure 5.9.

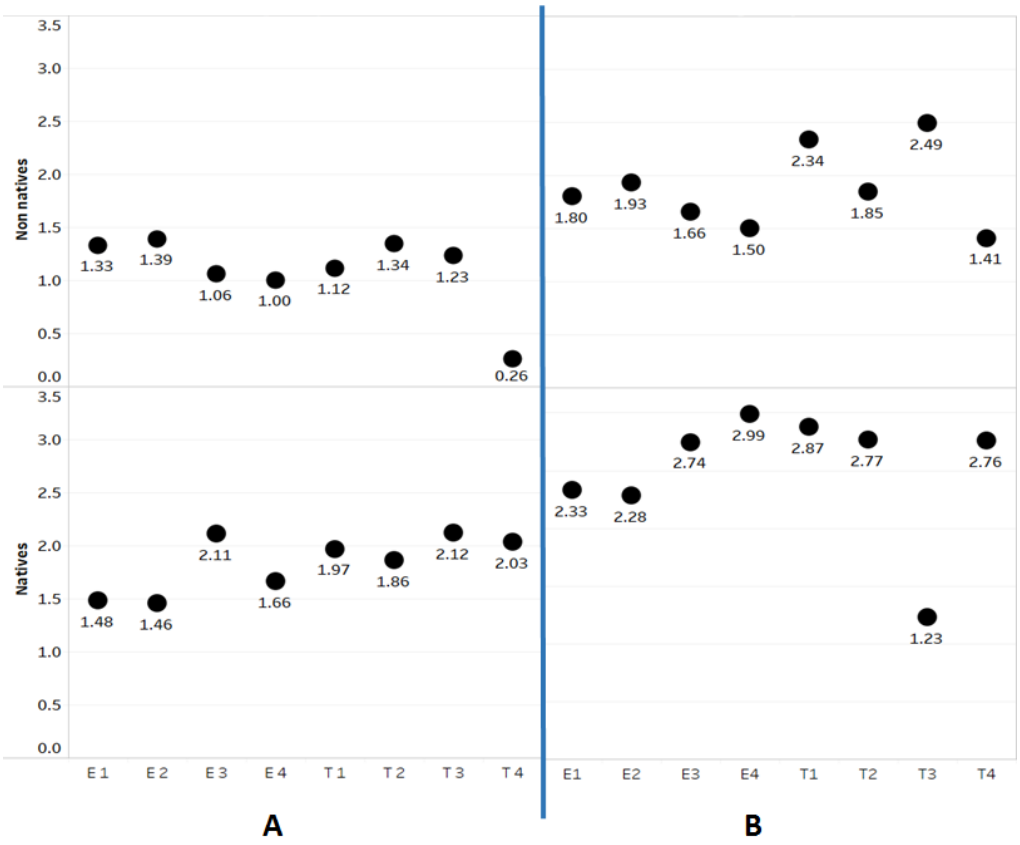

Figure 5.9 Across disparity for native speakers versus non-native speakers for study A and study B.

All videos in study A for non-natives had zero cosine similarities between categories: *Structure* and *Delivery*, *Structure* and *VisualAid*, *Structure* and *PresentationAttribute*. T4 had zero cosine similarities even for *Delivery* and *VisualAid* and *Delivery* and *PresentationAttribute*. This is due to very low or zero frequencies across the categories. For example, non-natives did not cover the category *Structure* at all on T4 (11 non-native speaker commented on this video). Category *Delivery* was mentioned by only one user on T1 (14 non-native speaker commented on this video). For same reasons, in study B for non-natives, T1, T2 and E2 had zero cosine similarities

between *Structure* and *Delivery*, and T2 had zero cosine similarity between *Delivery* and *VisualAid* as well.

These findings are interesting as there were no differences in terms of domain coverage between user groups based on language for the other diversity indices (see section 5.3.2). This is due to the fact that the level of profiling was different. In section 5.3.2, domain diversity is profiled per user (zoom-in per user) based on the entities he/she covered on all 8 videos, whereas here the entities are grouped by certain user group (e.g. natives) per video (zoom-in per video). A user (e.g. a non-native speaker) could miss one domain aspect on a video but mention it on another. This can give an indication that the video did not trigger this user to see this aspect, but the user is aware about it. This shows the utility and importance of diversity different profiling levels and different indices as both broaden the insights for the nature of diversity in a social cloud.

5.6 Diversity Patterns in Case Study 2

This section reports on the patterns that were detected in a closed social cloud.

5.6.1 Combining balance and coverage

This section reports patterns for the presentation skills domain as the user diversity perspective was measured on the overview level of users where an overview diversity profile is produced for all users U in the social cloud.

Looking at the tutorial videos (refer to the left side of Figure 5.3 and Figure 5.4), several observations can be made. In study A, E2 seems to have fairly **diverse coverage** (high balance and high coverage). In study B, E3 had a diverse coverage, while it had a **lack of focus** in study A (high balance and low coverage). T4 had **non-diverse coverage** (low balance and low coverage) for both studies. Although coverage on this video focussed on *Delivery*, it is low as this category is large (largest branch in PreSOn). T4 in study B had 73 distinct entities, 36 of which are from category *Delivery.* In study A, this video had 53 entities, 28 of which are from *Delivery*. This out of 183 entities available in this category. T1 in study A had a **focus** on category *VisualAid* (low balance and high coverage).

5.6.2 Domain Perspective per User

On way to identify the two types of users- **domain diversified** and **domain narrowed user**, the domain diversity profiles per user are sorted maximum to minimum based on domain variety first, then balance, then coverage, and finally within disparity. Then one can divide the user diversity profiles to quartiles to inspect top quartile for the former type and the bottom for the latter. In study A, the user UCW12 can be considered most domain diversified user for this study. This is a native male user who wrote 29 comments in total. The comments covered the 4 top level categories of PreSOn with an even distribution (high balance at 1.39), fairly high coverage (it was maximum coverage in this study) and high within disparity 15.93. In study B, the user mfu36 can be considered the domain diversified user as it covered the 4 top categories of PreSon with even balance ($b = 1.37$) and high domain coverage ($r =$0.13) and within disparity ($d_w =$10.73).

For the domain narrowed user, in study A, a male non-native speaker (UCT23) wrote two comments that had only one domain entity covering one domain category with zero balance and within disparity and very low coverage. There are other 2 users in the bottom quartile with similar profiles. Similarly, in study B, a male native user (rpd24) wrote 2 comments in total. These comments were not domain related (i.e. had no aspects related to presentation skills domain). This resulted in zero values for all diversity indices. In the bottom quartile, there are 12 users who had low to minimum values for diversity indices. Majority (10 users) wrote only one comment and the other 2 users wrote 4 comments. All resulted in only one domain entity from one domain category with zero balance and within disparity and very low coverage. In both studies, all mentioned above users are considered domain narrowed users.

5.6.3 Domain Linked with User for Across Disparity

Identifying **dominant domain category per video** can be done per user or user groups. For example, per natives and non-natives - for non-natives, *PresentationAttribute* was a dominant category on 5 (out of 8) videos. Maximum frequency (30 entities) was on video T2, which had comments from 11 non-native users. This category was dominant on 3 videos for native speakers, where maximum frequency for this category (38 entities) was on video E2. VideoT4 had maximum frequencies for the category *Delivery* by natives and non-natives (58 and 32

respectively). An inspection of frequencies per user on this category from both studies revealed that in study A, a female native speaker (LstudyID599) had 11 entities from *Delivery*, while in study B, a male native speaker (mbw49) had 17 entities from this category**.** Both users are **dominant users** for this category. Native users can be considered the **dominant user group** for the category *Delivery* on T4 as they mentioned 58 entities from this category compared to 32 mentioned by non-natives.

5.7 Discussion

This chapter presented the application of the proposed semantic driven approach in a closed social cloud. An instantiation of this approach is illustrated for (a) profiling domain diversity in user comments when watching videos to gain deeper insights into the user perspectives of the domain in the context of informal learning of soft skills, where the ability to see different domain perspectives is crucial. (b) Profiling user diversity based on attributes from their profiles to understand who the users are and how diverse they are. Example of diversity patterns based on the diversity indices are illustrated and discussed with possible applications.

This case study conforms with the findings discussed in the previous chapter. Initially, it was surprising that the native language did not impact on the domain diversity, which initially indicated that cognitive understanding of presentation skills was orthogonal to language. However, drilling down to frequencies for across disparity showed some differences of this coverage between natives and non-natives. This shows the utility and importance of different diversity profiling levels and different indices as both broaden the insights for the nature of diversity in a social cloud.

Diversity indices were significantly positively correlated with the number of comments; this can be an indication for intelligent learning environments designers to implement suitable prompts that nudge learners to write comments while watching videos for learning. Again, more comments do not mean high diversity.

Although the user diversity can be detected by comparing the quartiles due to sample size, the profiling facilitated the automation of diversity measurements and showed the applicability and potential of the proposed approach for exploring diversity even on a small scale. This is beside showing the utility of the diversity profiles for comparing user groups and observing interesting patterns about using videos for learning. This

can be beneficial, for example, for educators, where they can conduct multi group comparisons based on these profiles to identify potential or limitations.

The User Diversity Ontology was extended to include MSLQ items, which show further flexibility and utility of this ontology, but the final extension with the MSLQ scores are tailored to the users' scores from the used social clouds (e.g. instances under *TopTaskValue*). This means that for future use, this extension must be updated if different scores are obtained. Although a general extension can be conducted based on the work in MSLQ, but researchers tend to adopt this questionnaire to their work (as for the case studies from AVW-Space here). Therefore, this is left open to be tailored for future work with other case studies.

Chapter 6 : Conclusion

This chapter summarises the research conducted during this PhD highlighting potential contributions, limitations and future direction.

6.1 Summary

Adopting a well-established Diversity Framework and underpinned by Semantic Web techniques, diversity in social clouds has been represented and measured automatically based on two diversity perspectives (domain and user) and by four diversity indices (variety, balance, coverage and disparity (within and across). This is captured in diversity profiles.

For the domain diversity perspective - variety refers to breadth of domain coverage, i.e. how many top-level categories are covered via associated domain concepts mentioned in the comments. This is useful for learning to gather the learners' overview of the domain. Balance goes further and captures the evenness of domain coverage, i.e. the distribution of concepts in the top categories covered by the pool of comments. This is useful to see the degree of consistency in the level of understanding across domain categories. Coverage measures how much of the domain is covered in the comments. Coverage complements balance, which enables an understanding on whether the learners' attention is focused on specific areas or is dispersed across the domain. Within disparity refers to the distinctiveness of domain coverage i.e. measures how scattered the domain concepts mentioned in the comments within the top categories.

Similarly, for the user perspective, with regards to a selected user attribute captured in the user profile: variety measures how many user categories are covered in the social cloud. It helps to identify main user categories who were triggered to contribute to the social cloud content. Balance captures user distribution across identified user categories. It indicates whether the users are evenly distributed across their categories. Coverage shows how much of the user categories is captured in the social cloud. Within disparity distinguish the users from each other within the user categories i.e. how different are the users from each other.

Across disparity links the diversity perspectives showing possible deeper connections between individuals or groups and their domain differences, which assists identifying

in terms of frequency whether the domain categories mentioned in the user comments are similar or not.

Linking diversity perspectives and combining diversity properties showed further utility of the diversity profiles. They highlighted several possible patterns like dominant domain categories to identify popular topics among users that were discussed in their comments. Another example is combination of coverage with balance which can further the insights like identify whether there are dominant user groups in the social cloud. These patterns deepen the understanding of diversity and open doors for further utilities of the diversity profiles, such as personalisation and adaptations in the domain of learning.

The proposed model was operationalised to generate a Semantic Diversity Analytics Tool (SeDDAT) to be applied for social clouds diversity profiling. Two case studies capturing the two types of social clouds, open and closed (discussed in chapter 2), were selected. They show applicability and transferability of the proposed model with different social clouds and ontologies underpinning. The first case study with YouTube explored the body language domain and user cultural variations for the domain and user perspectives respectively. The second case study explored presentation skills domain and users' hidden attributes for domain and user perspectives. The steps of the proposed semantic approach (discussed in chapter 3) are illustrated for both case studies showing the instantiation of SeDDAT if to be used in similar context.

Two ontologies were implemented for this research - the User Diversity Ontology was proposed to complement the model for the profiling of the user diversity perspective, and PreSOn which was extended and implemented for domain diversity profiling for the second case study.

The proposed semantic approach underpinned by the model pave the way for further work in similar context. It is applicable in similar scenarios, such as in the social clouds around MOOCs, reviews (e.g. products or hotels), and news articles.

6.2 Contributions

This work contributes to the ongoing diversity research, facilitates possible personalisation and adaptations and shows and supports the potential of the Semantic Web techniques with the following:

Computational diversity model. This contributes to the diversity research on understanding and measuring diversity by proposing a model and an approach to use it for modelling diversity in social clouds. This is with regards to two main diversity perspectives found in the diversity literature: individuals and their perspectives on a domain of interest. This is achieved using four diversity properties, one of which was proposed in this research to refine the findings related to actual coverage against ontologies. The computational model facilitates automatic measurements of diversity in social clouds with variety of domains and individuals' attributes. The model profiles diversity perspectives separately and connected showing variety of utilities of the approach for diversity profiling. The model is applicable and can be extended to work with other social clouds.

Diversity profiles and patterns. This contributes to personalisation and adaptations for learning. The diversity profiles generated for the identified perspectives indicate the level of diversity and enable the detection of possible diversity patterns. The patterns deepen the understanding for both diversity perspectives – the nature of domain coverage and users who contributed to this coverage. This can, for example, help educators to understand diversity of their learners and identify their domain knowledge, which in turn enforces possible educational methods for identifying and enhancing any potential or limitations. A list of these patterns, how to detect them and possible utilities in learning are discussed in this thesis with examples.

Semantic Driven Diversity Analytics Tool- SeDDAT. This shows applications of semantic technologies for modelling diversity. SeDDAT was built based on the proposed model for diversity profiling, where its instantiation in the case studies illustrates the proposed approach step by step. The approach and tool show and support the potential provided by the semantic technologies for understanding and measuring diversity, which adds to the vision of utilising these techniques for harvesting and exploiting the collective intelligence of the crowds generated on online social spaces. Moreover, two ontologies, the User Diversity Ontology and Presentation Skills Ontology (PreSOn) are made publicly available for researchers to exploit for their work. This research illustrates, using available methods for building ontologies even when there are challenges or limitations with the resources intended to be used for extending or re-engineering.

6.3 Limitations

While the diversity indices are generic and independent on the quality of the ontology, the findings when applying these indices are dependent on the size and shape of the ontology. This is similar to the dependency of a data sample on the population it represents.

As in social science, where diversity indices can highlight under-representation of certain quarter(s), the unexpected low diversity in a social cloud may indicate a need for revision or extension of used ontology branches.

Automatic ontology-based semantic annotations of the social content (e.g. comments) is prone to errors (e.g. miss a word/term (in a comment) or an attribute (of a user)), which can impact the findings. Manual inspection of the resultant annotations prior to the diversity profiling enables terminating any errors.

The format of input files for SeDDAT might restrict its use. The tool will be extended to accept further format.

6.4 Future Work

There are two main streams of future work- immediate and long-term as follows.

6.4.1 Immediate work

Diversity patterns validations. The diversity patterns (discussed in section 3.4) showed potential to gain more insights into diversity of social clouds with the two case studies. Validations of these patterns and their usefulness in the domain of learning is planned to be the next immediate work. Further case studies to be utilised to validate the patterns and extend/update the list of patterns when required. Then, the usefulness of these patterns is planned to be evaluated by presenting these patterns to educators and investigate how they can assist them in real life applications. Findings from these two steps should lead the way for further expansion and validations as appropriate. The actual utilisation and implementation of these patterns to assist the process of education require further work, which might be a research of its own.

Further measures for semantic distances. The property disparity is arguably the trickiest diversity property to be understood and measured. For within disparity d_w in this research, a shortest path measure was selected to measure the semantic

distances between entities from annotations. This measure was used to facilitate generic measurement of this property regardless of the information available about the entities within the given ontology. Other taxonomic and non-taxonomic semantic distance indices could be explored, such as path and depth (discussed in Chapter 2 section 2.4.3) to investigate their impact on distinguishing elements (entities) within their categories. These can serve as further indicators of within disparity.

Generalising to other input file format. Currently SeDDAT takes an XML and a Microsoft excel files as input. SeDDAT will be extended to accept more common file formats.

6.4.2 Long-term Work

There are several directions for future work building on findings from this research as follows:

Generalising to other social clouds' content. The proposed semantic approach is planned to be extended to include any social cloud content (e.g. microblogs and Q&A forums' textual content). This will require an extension of the steps conducted within this approach (discussed in section 3.1) to include further pre-processing, semantic enrichments (e.g. for tweets), and other methods for diversity analysis. The diversity model (discussed in section 3.2) will be extended as appropriate.

Contributing to intelligent learning environments. Modelling diversity is especially valuable in soft skills learning, where contextual awareness and understanding of different perspectives is crucial. Hence, this work can be extended to contribute to future intelligent learning environments that address pressing training needs of the modern society (e.g. unconscious bias, cultural awareness), which would require automated ways to capture and compare different domain perspectives and user backgrounds.

One example application is extending SeDDAT to serve as a built-in tool in educational social spaces (e.g. MOOCs). A tool that can provide a feedback to educators in terms of diversity of the content (e.g. comments) created by the learners, and who are the learners interacting with the content uploaded by the educators. For instance, there has been an increasing research on identifying reasons of MOOCs dropouts (learners starting a course and leaving it before it is finished). Understanding who are the learners interacting with the educational content can help identify some patterns

associated with those learners and what they contribute, which in turn could point to limitations or protentional (e.g. within the educational content) to hedge against dropping out the courses.

Another application is extending SeDDAT to serve as a recommender system where educators can use it to identify suitable educational mediums (e.g. videos) for their learners.

Exploring other fields. It is also intended to apply the approach and model in different fields from learnings. For example, in the news domain. The diversity profiles and patterns can shed light on interesting findings with regards to this domain, especially during crucial events. Also, in the health and safety domain, where diversity profiles can help to identify critical health and safety issues (e.g. metal health issues) exposed via social content (e.g. in the user microblogs or comments).

Journalists can benefit from this work, for instance, people nowadays read news online, where major newspapers are uploading articles, images and videos for readers' consumption. Many newspapers (e.g. the Guardian) link to other social media spaces (e.g. Facebook) to allow the readers to interact with the news and leave their opinions. Diversity profiling of this social cloud can help the journalists to enhance the quality of their articles, for example, diversify their articles content to trigger diverse readers via understanding who is reading their news and what they think.

This work can contribute to the growing body of research concerned with understanding human behaviour and wellbeing from user interactions in online social spaces. For instances, help exploring online discussions about student stress during exams or self-inclusion of terms that indicate mental health issues. Diversity profiles can first identify terms associated with such issues and second compare and identify user groups that are more prone to these issues.

References

[1] A. M. Konrad, P. Prasad, and J. K. Pringle, *The handbook of workplace diversity*. Sage Publications, 2006.

[2] A. Stirling, “On the economics and analysis of diversity,” SPRU Electronic Working Paper Number 28. University of Sussex, 1998.

[3] A. G. Greenwald *et al.*, “Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes Implicit Social Cognition: Introduction and Overview,” 1995.

[4] J. Surowiecki, *The Wisdom of Crowds*. Random House, 2004.

[5] K. W. Phillips, M. Duguid, M. Thomas-Hunt, and J. Uparna, “Diversity as Knowledge Exchange: The Roles of Information Processing, Expertise, and Status,” in *The Oxford Handbook of Diversity and Work*, vol. 1, Q. M. Roberson, Ed. Oxford University Press, 2013.

[6] L. Hong and S. E. Page, “Groups of diverse problem solvers can outperform groups of high-ability problem solvers.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 46, pp. 16385–9, Nov. 2004.

[7] S. E. Page, *The difference : how the power of diversity creates better groups, firms, schools, and societies*. .

[8] A. Ascolese *et al.*, “ImREAL Final Publishable Summary Report, Deliverable 1.5 Project,” 2013.

[9] J. Vassileva, “Toward Social Learning Environments,” *IEEE Trans. Learn. Technol.*, vol. 1, no. 4, pp. 199–214, Oct. 2008.

[10] L. Gómez Chova, A. P. Lopes, and F. Soares, “The potential benefits of using video in higher education,” in *6th International Conference on Education and New Learning Technologies*, 2014, pp. 750–756.

[11] E. Tan, “Informal learning on YouTube : exploring digital literacy in independent online learning,” *Learn. Media Technol.*, vol. 38, no. 4, pp. 463–477, 2013.

[12] A. Mohamed, F. Yousef, M. A. Chatti, and U. Schroeder, “The State of Video-Based Learning: A Review and Future Perspectives,” *Int. J. Adv. Life Sci.*, vol. 6, no. 3&4, 2014.

[13] R. K. Nielsen, “Where do people get their news?,” *Journalism Studies*, 25-Jan-2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1461670X.2016.1171163.

[14] K. Bontcheva and D. Rout, “Making sense of social media streams through semantics: A survey,” *Semant. Web*, vol. 5, no. 0, pp. 373–403, 2014.

[15] D. Davis, I. Jivet, R. F. Kizilcec, G. Chen, C. Hauff, and G.-J. Houben, “Follow the Successful Crowd: Raising MOOC Completion Rates through Social Comparison at Scale,” in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017.

[16] K. Oyibo, R. Orji, and J. Vassileva, “Investigation of the Social Predictors of Competitive Behavior and the Moderating Effect of Culture,” in *UMAP-WPPG: Fifty Shades of Personalization - Workshop on Personalization in Serious and Persuasive Games and Gameful Interaction*, 2017.

[17] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horiz.*, vol. 53, p. 59—68, 2010.

[18] H. Khang, E. J. Ki, and L. Ye, “Social media research in advertising, communication, marketing, and public relations, 1997-2010,” *Journal. Mass Commun. Q.*, 2012.

[19] M.-F. Moens, J. Li, and T.-S. Chua, *Mining User Generated Content*. Chapman and Hall/CRC, 2014.

[20] D. C. Brabham, *Crowdsourcing*. The MIT Press Essential Knowledge Series , 2013.

[21] T. Gruber, “Collective Knowledge Systems: Where the Social Web meets the Semantic Web,” *Semant. Web Web 2.0*, vol. 6, no. 1, pp. 4–13, 2008.

[22] P. Lévy, Pierre/Translator-Bononno, and Robert, *Collective intelligence : mankind’s emerging world in cyberspace*. Plenum Trade, 1997.

[23] M. Sabou, K. Bontcheva, and A. Scharl, “Crowdsourcing research opportunities: lessons from natural language processing,” *Proc. 12th Int. Conf. Knowl. Manag. Knowl. Technol.*, pp. 1–8, 2012.

[24] J. Kim, “Learnersourcing: Improving Learning with Collective Learner Activity,” Massachusetts Institute of Technology, 2015.

[25] K. Bontcheva and D. Rout, “Making sense of social media streams through semantics: A survey,” *Semant. Web*, vol. 5, no. 5, pp. 373–403, Jan. 2014.

[26] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu, “A Comparative Study of Users’ Microblogging Behavior on Sina Weibo and Twitter,” *User Model. Adapt. Pers.*, vol. 7379, pp. 88–101, 2012.

[27] D. Maynard and A. Funk, “Automatic Detection of Political Opinions in Tweets,” in *The Semantic Web: ESWC 2011 Selected Workshop Papers*, 2011.

[28] D. Maynard, K. Bontcheva, and D. Rout, “Challenges in developing opinion mining tools for social media,” *Lr. 2012 Work. @NLP can u tag #usergeneratedcontent*, p. 8, 2012.

[29] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers, 2012.

[30] D. Maynard, G. Gossen, A. Funk, and M. Fisichella, “Should I Care about Your Opinion? Detection of Opinion Interestingness and Dynamics in Social Media,” *Futur. Internet*, vol. 6, pp. 457–481, 2014.

[31] D. Eilander, P. Trambauer, J. Wagemaker, and A. Van Loenen, “Harvesting Social Media for Generation of Near Real-time Flood Maps,” *Procedia Eng.*, vol. 154, no. 0, pp. 176–183, 2016.

[32] F. Abel, C. Hauff, and R. Stronkman, “Twitcident : Fighting Fire with Information from Social Web Streams,” pp. 305–308, 2012.

[33] C. Musto, G. Semeraro, de M. Gemmis, and P. Lops, “Modeling Community Behavior through Semantic Analysis of Social Data: the Italian Hate Map Experience,” in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP ’16*, 2016.

[34] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo,

"Measuring Urban Social Diversity Using Interconnected Geo-Social Networks," *Proc. 25th Int. Conf. World Wide Web - WWW '16*, pp. 21–30, 2016.

[35] F. Abel, Q. Gao, G. J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6643 LNCS, no. PART 2, pp. 375–389.

[36] G. Piao and J. G. Breslin, "Analyzing Aggregated Semantics-enabled User Modeling on Google+ and Twitter for Personalized Link Recommendations," in *in Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016.

[37] Y. Kiprov, P. Gencheva, and I. Koychev, "Generating Labeled Datasets of Twitter Users," in *Proceedings of the 2017 Conference on User Modeling Adaptation and Personalization - UMAP '17*, 2017.

[38] J. Herzig, G. Feigenblat, M. Shmueli-Scheuer, D. Konopnicki, and A. Rafaeli, "Predicting Customer Satisfaction in Customer Support Conversations in Social Media Using Affective Features," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016, pp. 115–119.

[39] M. Al-Ghossein, T. Abdessalem, and A. Barré, "Exploiting Contextual and External Data for Hotel Recommendation," in *UMAP '18 Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, 2018.

[40] R. Dong and B. Smyth, "From More-Like-This to Better-Than-This: Hotel Recommendations from User Generated Reviews," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016.

[41] S. Chelaru, C. Orellana-Rodriguez, and I. S. Altingovde, "How useful is social feedback for learning to rank YouTube videos?," *World Wide Web*, vol. 17, no. 5. pp. 997–1025, 2014.

[42] K. Filippova and K. B. Hall, "Improved video categorization from text metadata and user comments," *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*. p. 835, 2011.

[43] M. Galli, D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti, "Analysis of user-generated content for improving youtube video recommendation," in *CEUR Workshop Proceedings*, 2015, vol. 1441.

[44] A. Serbanoiu and T. Rebedea, "Relevance-based ranking of video comments on youtube," *Proceedings - 19th International Conference on Control Systems and Computer Science, CSCS 2013*. pp. 225–231, 2013.

[45] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, "How Useful are Your Comments? - Analyzing and Predicting YouTube Comments and Comment Ratings," in *Proceedings of the 19th international conference on World Wide Web*, 2010, vol. 15, pp. 891–900.

[46] A. Ammari, L. Lau, and V. Dimitrova, "Deriving group profiles from social

media to facilitate the design of simulated environments for learning," *Proc. 2nd Int. Conf. Learn. Anal. Knowl. - LAK '12*, no. May, p. 198, 2012.

[47] D. Thakker, V. Dimitrova, and L. Lau, "I-CAW : Intelligent Data Browser for Informal Learning Using Semantic Nudges," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7603, pp. 434–437, 2012.

[48] G. Chen, D. Davis, C. Hauff, and G.-J. Houben, "On the Impact of Personality in Massive Open Online Learning," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016.

[49] A. Mitrovic, V. Dimitrova, L. Lau, A. Weerasinghe, and M. Mathews, "Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies," Springer, Cham, 2017, pp. 224–237.

[50] B. Sjödén, V. Dimitrova, and A. Mitrovic, "Using Thematic Analysis to Understand Students' Learning of Soft Skills from Videos," in *European Conference on Technology Enhanced Learning*, 2018.

[51] V. Dimitrova, A. Mitrovic, A. Piotrkowicz, L. Lau, and A. Weerasinghe, "Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17*, 2017, pp. 22–31.

[52] I. Adaji and J. Vassileva, "Modelling User Collaboration in Social Networks Using Edits and Comments," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016, pp. 111–114.

[53] A. Joorabchi, M. English, and A. E. Mahdi, "Text mining stackoverflow: an insight into challenges and subject-related difficulties faced by computer science learners," *J. Enterp. Inf. Manag.*, 2016.

[54] O. Odiete, T. Jain, I. Adaji, J. Vassileva, and R. Deters, "Recommending Programming Languages by Identifying Skill Gaps Using Analysis of Experts. A Study of Stack Overflow," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17*, 2017, pp. 159–164.

[55] M. Polignano, P. Basile, G. Rossiello, M. de Gemmis, and G. Semeraro, "User's Social Media Profile as Predictor of Empathy," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17*, 2017, pp. 386–390.

[56] G. Piao and J. G. Breslin, "Analyzing MOOC Entries of Professionals on LinkedIn for User Modeling and Personalized MOOC Recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP '16*, 2016.

[57] J. R. Lambert and M. P. Bell, "Diverse Forms of Difference," in *The Oxford Handbook of Diversity and Work*, vol. 1, Q. M. Roberson, Ed. Oxford University Press, 2013.

[58] D. Harrison and K. J. Klein, "What's the difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations," *Acad. Manag. Rev.*, vol. 32, no. 4, pp. 1199–1228, 2007.

[59] I. Rafols, L. Leydesdorff, A. O'Hare, P. Nightingale, and A. Stirling, "How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management," *Research Policy*, vol. 41, no. 7. pp. 1262–1282, 2012.

[60] D. Steel, S. Fazelpour, K. Gillette, B. Crewe, and M. Burgess, "Multiple diversity concepts and their ethical-epistemic implications," *Eur. J. Philos. Sci.*, vol. 8, no. 3, pp. 761–780, 2018.

[61] J. E. McGrath, J. L. Berdahl, and H. Arrow, "Traits, expectations, culture, and clout: The dynamics of diversity in work groups," in *American Psychological Association*, vol. 10099, Washington: American Psychological Association, 1995, pp. 17–45.

[62] S. Fletcher-Watson, H. De Jaegher, J. van Dijk, C. Frauenberger, M. Magnée, and J. Ye, "Diversity Computing," *Interactions*, vol. 25, no. 5, pp. 28–33, 29-Aug-2018.

[63] M. Aktas, M. J. Gelfand, and P. J. Hanges, "Cultural Tightness–Looseness and Perceptions of Effective Leadership," *J. Cross. Cult. Psychol.*, vol. 47, no. 2, pp. 294–309, 2016.

[64] M. Minkov and G. Hofstede, "The evolution of Hofstede's doctrine," *Cross Cult. Manag. An Int. J.*, vol. 18, no. 1, pp. 10–20, 2011.

[65] M. Minkov and G. Hofstede, "Is National Culture a Meaningful Concept?: Cultural Values Delineate Homogeneous National Clusters of In-Country Regions," *Cross-Cultural Res.*, vol. 46, no. 2, pp. 133–159, Nov. 2011.

[66] R. J. House, P. J. Hanges, M. Javidan, P. W. Dorfman, and V. Gupta, *Culture, Leadership and organizations: The Globe study of 62 societies*. Sage publications, Thousand Oaks, CA., 2004.

[67] M. Javidan and a. Dastmalchian, "Managerial implications of the GLOBE project: A study of 62 societies," *Asia Pacific J. Hum. Resour.*, vol. 47, no. 1, pp. 41–58, Apr. 2009.

[68] E. Blanchard, "Adaptation-oriented culturally-aware tutoring systems: When adaptive instructional technologies meet intercultural education," in *Handbook of Research on Human Performance and Instructional Technology*, & T. K. H. Song, Ed. Hershey, PA: Information Science Reference, 2009, pp. 413–430.

[69] A. Ogan, "Supporting Learner Social Relationship with Encultured Pedadogical Agents," Carnegie Mellon University, 2011.

[70] Y. M. Mensah and H.-Y. Chen, "Global Clustering of Countries by Culture – An Extension of the GLOBE Study," *SSRN Electron. J.*, Apr. 2012.

[71] R. Lewis, *When Cultures Colide*, 3rd ed. Nicholas Brealey, 2006.

[72] M. J. Gates, R. Lewis, I. P. Bairatchnyi, and M. Brown, "Use of the Lewis Model to Analyse Multicultural Teams and Improve Performance by the World Bank: A Case Study," *ijm.cgpublisher.com*, vol. 8, no. 12, 2009.

[73] V. Dimitrova, R. Denaux, L. Lau, D. Thakker, P. Brna, and C. M. Steiner, "Utilising Linked Data for Interactive Learner Modelling of Culture-related Aspects," *Int. J. Artif. Intell. Educ. Util.*, 2010.

[74] W. Johnson, "Serious use of a serious game for language learning," *Front.*

Artif. Intell. Appl., vol. 20, pp. 175–195, 2007.

[75] F. Kistler, E. André, and S. Mascarenhas, “Traveller: An Interactive Cultural Training System Controlled by User-Defined Body Gestures,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8120, no. PART 4, pp. 697–704, 2013.

[76] K. Reinecke, P. Minder, and A. Bernstein, “MOCCA - a system that learns and recommends visual preferences based on cultural similarity,” *Proc. 15th Int. Conf. Intell. user interfaces - IUI ’11*, p. 453, 2011.

[77] E. Kapros, “International Large-Scale Assessments and Culture: Implications for Designing Educational Technology,” in *UMAP’17: HAAPIE: Human Aspects in Adaptive and Personalized Interactive Environments* , 2017.

[78] E. Zangerle, M. Pichl, and M. Schedl, “Culture-Aware Music Recommendation,” *Proc. 26th Conf. User Model. Adapt. Pers. - UMAP ’18*, pp. 357–358, 2018.

[79] Kiemute Oyibo, “Designing Culture-based Persuasive Technology to Promote Physical Activity among University Students,” in *UMAP*, 2016.

[80] J. Yang, M. R. Morris, J. Teevan, L. A. Adamic, and M. S. Ackerman, “Culture Matters: A Survey Study of Social Q&A Behavior,” 2011.

[81] K. Oyibo, Y. S. Ali, and J. Vassileva, “Gender Difference in the Credibility Perception of Mobile Websites: A Mixed Method Approach,” in *UMAP* , 2016.

[82] M. Rokicki, T. Kusmierczyk, and C. Trattner, “Plate and Prejudice: Gender Differences in Online Cooking,” in *UMAP*, 2016.

[83] A. K. Vail, J. F. Grafsgaard, K. E. Boyer, E. N. Wiebe, and J. C. Lester, “Gender Differences in Facial Expressions of Affect During Learning,” in *24th Conference on User Modeling, Adaptation and Personalization UMAP 2016*, 2016, pp. 65–73.

[84] I. Arroyo, B. P. Woolf, D. G. Cooper, W. Burleson, and K. Muldner, “The Impact of Animated Pedagogical Agents on Girls’ and Boys’ Emotions, Attitudes, Behaviors and Learning,” in *2011 IEEE 11th International Conference on Advanced Learning Technologies*, 2011, pp. 506–510.

[85] B. McLaren, R. Farzan, D. Adams, R. Mayer, and J. Forlizzi, “Uncovering Gender and Problem Difficulty Effects in Learning with an Educational Game,” Springer, Cham, 2017, pp. 540–543.

[86] B. Ferwerda, A. Vall, M. Tkalcic, and M. Schedl, “Exploring Music Diversity Needs Across Countries,” in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization - UMAP ’16*, 2016, pp. 287–288.

[87] A. M. Abdullahi, R. Orji, and K. Oyibo, “Personalizing Persuasive Technologies: Do Gender and Age Affect Susceptibility to Persuasive Strategies?,” in *UMAP*, 2018.

[88] K. Oyibo, R. Orji, and J. Vassileva, “The Influence of Culture in the Effect of Age and Gender on Social Influence in Persuasive Technology,” in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP ’17*, 2017, pp. 47–52.

[89] I. Adaji and J. Vassileva, “Personalizing Social Influence Strategies in a

Q&A Social Network," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization - UMAP '17*, 2017, pp. 215–220.

[90] A. M. Abdullahi, K. Oyibo, and R. Orji, "The influence of cognitive ability on the susceptibility to persuasive strategies," in *CEUR Workshop Proceedings*, 2018, vol. 2089, pp. 22–33.

[91] A. Piotrkowicz, V. Dimitrova, A. Mitrovic, and L. Lau, "Using the Explicit User Profile to Predict User Engagement in Active Video Watching," *Proc. 26th Conf. User Model. Adapt. Pers. - UMAP '18*, pp. 365–366, 2018.

[92] S. Lallé, M. Taub, N. V. Mudrick, C. Conati, and R. Azevedo, "The Impact of Student Individual Differences and Visual Attention to Pedagogical Agents During Learning with MetaTutor," Springer, Cham, 2017, pp. 149–161.

[93] Q. Roberson, *The Oxford Handbook of Diversity and Work*. 2012.

[94] P. Harrison, David A; Price, Kenneth, H; Bell, Myrtle, "Beyond Rational Demography: Time and the Effects on Surface and Deep-level Diversity on Work Group Cohesion," *Acad. Manag. J.*, vol. Vol. 41, no. 1, pp. 96–107, 1998.

[95] S. M. B. Thatcher, "Moving Beyond a Categorical Approach to Diversity: The Role of Demographic Faultlines," in *The Oxford Handbook of Diversity and Work*, vol. 1, Q. M. Roberson, Ed. Oxford University Press, 2013.

[96] E. Salas, M. R. Salazar, and M. J. Gelfand, "Understanding Diversity as Culture," in *The Oxford Handbook of Diversity and Work*, vol. 1, Q. M. Roberson, Ed. Oxford University Press, 2013.

[97] N. Wixon, S. Schultz, K. Muldner, and et al., "Internal &External Attributions for Emotions Within an ITS," in *UMAP*, 2016.

[98] J. Peperkamp and B. Berendt, "Diversity checker: Toward recommendations for improving journalism with respect to diversity," *UMAP 2018 - Adjun. Publ. 26th Conf. User Model. Adapt. Pers.*, pp. 35–41, 2018.

[99] N. Tintarev, E. Sullivan, D. Guldin, S. Qiu, and D. Odjik, "Same, Same, but Different," in *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization - UMAP '18*, 2018, pp. 7–13.

[100] A. Funk, Y. Li, H. Saggion, K. Bontcheva, and C. Leibold, "Opinion analysis for business intelligence applications," in *Proceedings of the first international workshop on Ontology-supported business intelligence - OBI '08*, 2008, pp. 1–9.

[101] D. G. Mcdonald and J. Dimmick, "The Conceptualization and Measurement of Diversity," *Communic. Res.*, vol. 30, no. 1, pp. 60–79, 2003.

[102] UNESCO Institute for Statistics (UIS), *MEASURING THE DIVERSITY OF CULTURAL EXPRESSIONS: Applying the Stitling Model of Diversity in Culture*, no. 6. 2011.

[103] F. Benhamou and S. Peltier, "How should cultural diversity be measured? An application using the French publishing industry," *J. Cult. Econ.*, vol. 31, pp. 85–107, 2007.

[104] F. Benhamou and S. Peltier, "APPLICATION OF THE STIRLING MODEL TO ASSESS DIVERSITY USING UIS CINEMA DATA," *UNESCO Inst. Stat.*, pp.

1–73, 2010.

[105] J. Farchy and H. Ranaivoson, “Do Public Television Channels Provide More Diversity than Private Ones,” *J. Cult. Manag. Policy*, 2011.

[106] M. Park, I. Weber, M. Naaman, and S. Vieweg, “Understanding Musical Diversity via Online Social Media,” in *Ninth International AAAI Conference on Web and Social Media*, 2015, no. Wikipedia, pp. 308–317.

[107] M. Schedl and D. Hauger, “Mining Microblogs to Infer Music Artist Similarity and Cultural Listening Patterns,” in *Proceedings of the 21st international conference companion on World Wide Web - WWW ’12*, 2012.

[108] D. Despotakis, D. Thakker, V. Dimitrova, and L. Lau, “Diversity of user viewpoints on social signals: A study with youtube content,” in *CEUR Workshop Proceedings*, 2012, vol. 872.

[109] T. HECKING, V. DIMITROVA, A. MITROVIC, and & H. U. HOPPE, “Using Network-Text Analysis to Characterise Learner Engagement in Active Video Watching,” in *Proceedings of the 25th International Conference on Computers in Education*, 2017.

[110] D. Schleuter, M. Daufresne, F. Massol, and A. C. Argillier, “A user’s guide to functional diversity indices,” *Ecol. Monogr.*, vol. 80, no. 3, pp. 469–484, 2010.

[111] H. Ranaivoson, “Measuring cultural diversity: a review of existing definitions,” *concept Pap. UNESCO Inst. Stat.*, no. September, pp. 1–31, 2007.

[112] J. Skea, “Valuing diversity in energy supply,” *Energy Policy*, vol. 38, no. 7, pp. 3608–3621, 2010.

[113] A. Stirling, “A general framework for analysing diversity in science, technology and society.,” *J. R. Soc. Interface*, vol. 4, no. February, pp. 707–719, 2007.

[114] C. Ricotta and L. Szeidl, “Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao’s quadratic index,” *Theor. Popul. Biol.*, vol. 70, pp. 237–243, 2006.

[115] I. Rafols and M. Meyer, “Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience,” *Scientometrics*, 2008.

[116] R. H. MacArthur, “Patterns of Species Diversity,” *Geogr. Ecol. Patterns Distrib. Species*, 1965.

[117] L. Zhang, R. Rousseau, and W. Glänzel, “Diversity of References as an Indicator of the Interdisciplinarity of Journals: Taking Similarity Between Subject Fields Into Account,” *Int. Rev. Res. Open Distance Learn.*, vol. 14, no. 4, pp. 90–103, 2013.

[118] R. C. Guiasu and S. Guiasu, “Weighted Gini-Simpson Quadratic Index of Biodiversity for Interdependent Species,” *Nat. Sci.*, vol. 06, no. 07, pp. 455–466, Apr. 2014.

[119] R. R. Laxton, “The Measure of Diversity,” 1978.

[120] E. C. Pielou, “An introduction to mathematical ecology.,” *An Introd. to Math. Ecol.*, 1969.

[121] M. O. Hill, “Diversity and Evenness: A Unifying Notation and Its Consequences,” *Ecology*, vol. 54, no. 2, pp. 427–432, Mar. 1973.

[122] S. Finkelstein, S. Scherer, and A. Ogan, “Investigating the influence of virtual peers as dialect models on students’ prosodic inventory.,” *WOCCI*, 2012.

[123] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. July 1928, pp. 379–423, 1948.

[124] C. Gini, *Variabilita e mutabilita*. Studi economico-giuridici pubblicati per cura della Facoltà di Giurisprudenza della Regia, 1912.

[125] E. H. SIMPSON, “Measurement of Diversity,” *Nature*, vol. 163, no. 4148, pp. 688–688, Apr. 1949.

[126] C. Ricotta, “A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species,” *Divers. Distrib.*, vol. 10, no. 2, pp. 143–146, Feb. 2004.

[127] L. Leydesdorff, D. Kushnir, and I. Rafols, “Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC),” *Scientometrics*, vol. 98, no. 3, 2014.

[128] C. R. Rao, “Diversity and dissimilarity coefficients: A unified approach,” *Theor. Popul. Biol.*, vol. 21, no. 1, pp. 24–43, Feb. 1982.

[129] J. H. Kang and K. Lerman, “Leveraging user diversity to harvest knowledge on the social web,” in *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 2011.

[130] E. Ilina, F. Abel, and G. Houben, “Mining Twitter for Cultural Patterns.,” in *Mensch & Computer Workshopband*, 2012, pp. 1–9.

[131] M. De Choudhury, S. S. Sharma, T. Logar, W. Eekhout, and R. C. Nielsen, “Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW ’17*, 2017, pp. 353–369.

[132] S. Bhatt, B. Minnery, S. Nadella, B. Bullemer, V. Shalin, and A. Sheth, “Enhancing Crowd Wisdom Using Measures of Diversity Computed from Social Media Data,” *Proc. Int. Conf. Web Intell.*, 2017.

[133] D. Despotakis, “Modelling viewpoints in user generated content,” University of Leeds, 2013.

[134] S. Kleanthous, G. Michael, and G. Samara, “Individual Differences in Music Video Interaction: An exploratory Analysis,” in *UMAP*, 2017.

[135] A. Mitrovic, V. Dimitrova, A. Weerasinghe, and L. Lau, “Reflective Experiential Learning: Using Active Video Watching for Soft Skills Training,” in *Proceedings of the 24th International Conference on Computers in Education*, 2016.

[136] T. Heath and C. Bizer, “Linked Data: Evolving the Web into a Global Data Space,” *Synth. Lect. Semant. Web Theory Technol.*, vol. 1, pp. 1–1, 2011.

[137] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far,” *Int. J. Semant. Web Inf. Syst.*, vol. 5, pp. 1–22, 2009.

[138] A. Maedche and S. Staab, “Ontology Learning for the Semantic Web,” *IEEE INTELLIGENT SYSTEMS*, 2001.

[139] T. Gruber, “A Translation Approach to Portable Ontology Specifications,” *Knowl. Acquis.*, vol. 5, no. April, pp. 199–220, 1993.

[140] S. R. Kruk, M. Synak, and M. Dabrowski, “Semantic Web and Ontologies,” in *Semantic Digital Libraries*, Springer, 2009, pp. 41–54.

[141] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284, no. 5. pp. 34–43, 2001.

[142] M. d’Aquin, L. Ding, and E. Motta, “Semantic Web Search Engines,” in *Handbook of Semantic Web Technologies*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 659–700.

[143] P. Cimiano, U. Bielefeld, L. Hollink, and E. Motta, “Watson, more than a Semantic Web search engine,” *Semant. Web*, vol. 2, no. 1, pp. 55–63, 2011.

[144] L. Ding *et al.*, “Swoogle: A Search and Metadata Engine for the Semantic Web,” *Proc. Thirteen. ACM Conf. Inf. Knowl. Manag.* , Nov. 2004.

[145] A. Kleinsmith, P. De Silva, N. Bianchi-Berthouze, L. Ardissono, P. Brna, and A. Mitrovic, “Gumo – The General User Model Ontology,” *User Model. 2005*, vol. 3538, no. February 2016, p. 148, 2005.

[146] M. Sutterer, O. Droegehorn, and K. David, “UPOS: User profile ontology with situation-dependent preferences support,” *Proc. 1st Int. Conf. Adv. Comput. Interact. ACHI 2008*, no. March 2008, pp. 230–235, 2008.

[147] K. Reinecke, G. Reif, and A. Bernstein, “Cultural User Modeling With CUMO: An Approach to Overcome the Personalization Bootstrapping Problem,” *Proc. First Int. Work. Cult. Herit. Semant. Web 6th Int. Semant. Web Conf.*, 2007.

[148] K. Skillen, Liming Chen, C. D. Nugent, M. P. Donnelly, and I. Solheim, “A user profile ontology based approach for assisting people with dementia in mobile environments,” *2012 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, no. December 2013, pp. 6390–6393, 2012.

[149] T. Plumbaum, S. Wu, E. W. De Luca, and S. Albayrak, “User Modeling for the Social Semantic Web,” in *2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, in conjunction with ISWC 2011*, 2011.

[150] A. Gomez-Perez and O. Corcho, “Ontology languages for the Semantic Web,” *IEEE Intell. Syst.*, vol. 17, no. 1, pp. 54–60, Jan. 2002.

[151] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology,” 2001.

[152] A. Gómez-Pérez and M. Carmen Suárez-Figueroa, “NeOn Methodology for Building Ontology Networks: a Scenario-based Methodology,” in *Proceedings of the International Conference on Software, Services & Semantic Technology*, 2009, pp. 28–29.

[153] N. F. Noy and M. A. Musen, “PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment,” in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI’00)*, 2000.

[154] S. Karanasios, D. Thakker, L. Lau, D. Allen, V. Dimitrova, and A. Norman, “Making Sense of Digital Traces : An Activity Theory Driven Ontological Approach,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 12, pp. 2452–2467,

2013.

[155] Y. Li and K. Bontcheva, “Hierarchical, Perceptron-like Learning for Ontology-Based Information Extraction,” in *Proceedings of the 16th international conference on World Wide Web*, 2007.

[156] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “GATE: an Architecture for Development of Robust HLT Applications,” in *Proceeding of the 40th Annual Meeting of the Association for Computaional Linguistics (ACL)*, 2002.

[157] M. Batet, D. Sánchez, and A. Valls, “An ontology-based measure to compute semantic similarity in biomedicine,” *J. Biomed. Inform.*, vol. 44, pp. 118–125, 2011.

[158] M. A. Hadj Taieb, M. Ben Aouicha, and A. Ben Hamadou, “Ontology-based approach for measuring semantic similarity,” *Eng. Appl. Artif. Intell.*, vol. 36, pp. 238–261, 2014.

[159] W.-N. Lee, N. Shah, K. Sundlass, and M. Musen, “Comparison of ontology-based semantic-similarity measures.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2008, pp. 384–8, Nov. 2008.

[160] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, “HESML: A scalable ontology-based semantic similarity measures library with a set of reproducible experiments and a replication dataset,” *Inf. Syst.*, vol. 66, pp. 97–118, Jun. 2017.

[161] W.-N. Lee, N. Shah, K. Sundlass, and M. Musen, “Comparison of ontology-based semantic-similarity measures.,” *AMIA Annu. Symp. Proc.*, pp. 384–388, 2008.

[162] K. Saruladha, G. Aghila, and S. Raj, “A Survey of Semantic Similarity Methods for Ontology Based Information Retrieval,” *2010 Second Int. Conf. Mach. Learn. Comput.*, pp. 297–301, 2010.

[163] R. Rada, H. Mili, E. Bicknell, and M. Blettner, “Development and Application of a Metric on Semantic Nets,” *IEEE Trans. Syst. Man Cybern.*, vol. 19, no. 1, pp. 17–30, 1989.

[164] Z. Wu and M. Palmer, “Verb Semantics and Lexical Selection,” in *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, 1994, pp. 133–138.

[165] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language,” *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, Jul. 1999.

[166] A. Tversky, “Features of similarity.,” *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.

[167] Z. Zhou, Y. Wang, and J. Gu, “New model of semantic similarity measuring in wordnet,” in *2008 3rd International Conference on Intelligent System and Knowledge Engineering*, 2008, pp. 256–261.

[168] F. Abel, Q. Gao, G. Houben, and K. Tao, “Analyzing User Modeling on Twitter for Personalized News Recommendations,” in *In 19th International conference, UMAP 2011*, 2011, pp. 1–12.

[169] S. Dietze *et al.*, “Preservation of Social Web Content based on Entity Extraction and Consolidation,” in *Proceedings of 2nd International Workshop on Semantic Digital Archives (SDA)*, 2012.

[170] A. Valitutti, C. Strapparava, and O. Stock, “Developing affective lexical resources,” *PsychNology J.*, 2004.

[171] D. Thakker, V. Dimitrova, L. Lau, R. Denaux, and F. Yang-turner, “A Priori Ontology Modularisation in Ill-defined Domains,” in *I-Semantics ’11 Proceedings of the 7th International Conference on Semantic Systems*, 2011, pp. 167–170.

[172] R. Denaux, V. Dimitrova, L. Lau, P. Brna, D. Thakker, and C. Steiner, “Employing Linked Data and Dialogue for Modelling Cultural Awareness of a User,” in *IUI ’14 Proceedings of the 19th international conference on Intelligent User Interfaces*, 2014, pp. 241–246.

[173] G. H. Ball and D. J. Hall, “ISODATA, a novel method of data analysis and pattern classification,” *Menlo Park Stanford Res. Inst.*, no. AD699616, 1965.

[174] L. Leydesdorff and I. Rafols, “A global map of science based on the ISI subject categories,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 2, pp. 348–362, 2009.

[175] M. Rehm, F. Gruneberg, Y. Nakano, A. A. Lipi, Y. Yamaoka, and H.-H. Huang, “Creating a Standardized Corpus of Multimodal Interactions for Enculturating Conversational Interfaces,” in *Proceedings of the IUI-Workshop on Enculturating Conversational Interfaces*, 2008.

[176] S.-Y. Hsu, a. G. Woodside, and R. Marshall, “Critical Tests of Multiple Theories of Cultures’ Consequences: Comparing the Usefulness of Models by Hofstede, Inglehart and Baker, Schwartz, Steenkamp, as well as GDP and Distance for Explaining Overseas Tourism Behavior,” *J. Travel Res.*, vol. 52, no. 6, pp. 679–704, Feb. 2013.

[177] B. L. Hallen, C. B. Bingham, C. Hill, N. Carolina, and S. L. Cohen, “GLOBAL Clustering of Countries By Culture-An Extension of the GLOBE Study,” *SSRN Electron. J.*, vol. 4, 2012.

[178] M. T. H. Chi and R. Wylie, “The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes,” *Educ. Psychol.*, vol. 49, no. 4, pp. 219–243, Oct. 2014.

[179] E. V De Groot and P. R. Pintrich, “Motivational and Self-Regulated Learning Components of Classroom Academic Performance,” *J. Educ. Psychol.*, vol. 82, no. 1, pp. 33–40, 1990.

[180] K. Seta and M. Ikeda, “Design of an environment for developing presentation skills,” *Front. Artif. Intell. Appl.*, vol. 151, no. 29, 2006.

[181] J. E. Gentle, L. Kaufman, and P. J. Rousseuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley-Interscience, 1990.

Appendix A: Example Algorithms from SeDDAT

A.1 Retrieval of Subclasses/Sub-categories of Entry Point *Thing*

```
package DiversityPackage;
import java.util.ArrayList;
import org.apache.jena.ontology.Individual;
import org.apache.jena.ontology.OntClass;
import org.apache.jena.ontology.OntModel;
import org.apache.jena.rdf.model.ModelFactory;
import org.apache.jena.rdf.model.Resource;
import org.apache.jena.util.iterator.ExtendedIterator;

public class RetrieveSuperclassesJena {

public static String RetriveSuperClass(String Source, String URI) {
                ArrayList IndividualSuperclasses = new ArrayList();
                ArrayList ThingDirectSubClasses = new ArrayList();
                ArrayList ClassSuperClass = new ArrayList();
                ArrayList individuals = new ArrayList();
                String SuprerClass = null;
                OntModel model = ModelFactory.createOntologyModel();
                model.read(Source, "RDF/XML");
                // Retrieve Thing direct subclasses
                for (ExtendedIterator<OntClass> l = model.listHierarchyRootClasses(); l.hasNext();) {
                        OntClass ThingSubclasses = l.next();
                        ThingDirectSubClasses.add(ThingSubclasses);
                }
                // Remove Equivalent classes-Thing Subclasses
                for (int i = 0; i < ThingDirectSubClasses.size(); i++) {
                        for (int j = 0; j < ThingDirectSubClasses.size(); j++) {

                        if (((OntClass) ThingDirectSubClasses.get(i))
                                                .hasEquivalentClass((OntClass)
                                                ThingDirectSubClasses.get(j))) {

                                        ThingDirectSubClasses.remove(ThingDirectSubClasses.get(j));
                                }
                        }
                }
                // check if the URI is individual
                if (model.getIndividual(URI).isIndividual()) {
                        Individual d = model.getIndividual(URI);
                        for (ExtendedIterator<OntClass> i = d.listOntClasses(false); i.hasNext();) {
                                OntClass cls = i.next();
                                if (cls.isHierarchyRoot())
```

```
                                {
                                        IndividualSuperclasses.add(cls);
                                }
                        }
                        // retrieve only the top super class
                        for (int o = 0; o < IndividualSuperclasses.size(); o++) {
                                if (ThingDirectSubClasses.contains(IndividualSuperclasses.get(o))) {

                                        SuprerClass = IndividualSuperclasses.get(o).toString();
                                }
                        }
                }
                // URI is for a class
                else {
                        Resource r = model.getResource(URI);
                        OntClass Concept = r.as(OntClass.class);
                        // check if the class is the top super class
                        if (Concept.isHierarchyRoot()) {
                                SuprerClass = Concept.toString();
                        } else {
                for (ExtendedIterator<OntClass> j = Concept.listSuperClasses(false); j.hasNext();) {
                                        OntClass cls = j.next();
                                        if (cls.isHierarchyRoot()) {
                                                ClassSuperClass.add(cls);
                                        }
                                }
                                // retrieve only the top super class
                                for (int o = 0; o < ClassSuperClass.size(); o++) {
                                        if (ThingDirectSubClasses.contains(ClassSuperClass.get(o))) {

                                                SuprerClass = ClassSuperClass.get(o).toString();
                                        }
                                }
                        }
                }
                return SuprerClass;
        }
}
```

A.2 Shortest Path for Within Disparity d_w

```
package DiversityPackage;
import java.util.ArrayList;
import java.util.Enumeration;
import java.util.Hashtable;
import java.util.Iterator;
import java.util.List;
```

```
import java.util.Map;
import org.apache.jena.ontology.Individual;
import org.apache.jena.ontology.OntClass;
import org.apache.jena.ontology.OntModel;
import org.apache.jena.ontology.OntResource;
import org.apache.jena.rdf.model.ModelFactory;
import org.apache.jena.rdf.model.Resource;
import org.apache.jena.util.iterator.ExtendedIterator;
import org.openrdf.model.URI;
import slib.graph.algo.utils.GAction;
import slib.graph.algo.utils.GActionType;
import slib.graph.io.conf.GDataConf;
import slib.graph.io.conf.GraphConf;
import slib.graph.io.loader.GraphLoaderGeneric;
import slib.graph.io.util.GFormat;
import slib.graph.model.graph.G;
import slib.graph.model.impl.graph.memory.GraphMemory;
import slib.graph.model.impl.repo.URIFactoryMemory;
import slib.graph.model.repo.URIFactory;
import slib.sml.sm.core.engine.SM_Engine;
import slib.sml.sm.core.utils.SMConstants;
import slib.sml.sm.core.utils.SMconf;
import slib.utils.ex.SLIB_Exception;


public class ShortestPathFromSML {
        URIFactory factory = URIFactoryMemory.getSingleton();
        final String ONTO_FILE = "PresentationOntologyV4.2Validated.owl";
public Hashtable<Object, Float> computeShortestPath(Hashtable<String, ArrayList>
EntitiesPerSuperClass, String source) throws SLIB_Exception {
                double Disparity = 0;
                // prepare for SML semantic distances calculations
                URI graphURI = factory.getURI(source);
                G g = new GraphMemory(graphURI);
                GDataConf dataConf = new GDataConf(GFormat.RDF_XML, ONTO_FILE);
                GAction actionRerootConf = new GAction(GActionType.REROOTING);
                GraphConf gConf = new GraphConf();
                gConf.addGDataConf(dataConf);
                gConf.addGAction(actionRerootConf);
                GraphLoaderGeneric.load(gConf, g);
                SMconf smConf = new SMconf(SMConstants.FLAG_SIM_PAIRWISE_DAG_EDGE_RADA_1989);
                SM_Engine engine = new SM_Engine(g);
                OntModel model = ModelFactory.createOntologyModel();
                model.read(source, "RDF/XML");
                // Access each super class and it is entities
                Hashtable<Object, Float> superClassDispersion = new Hashtable<Object, Float>();
                Enumeration en = EntitiesPerSuperClass.keys();
```

```
		while (en.hasMoreElements()) {
			Object Superclasses = en.nextElement();
			ArrayList EntitiesofSuperClasses = EntitiesPerSuperClass.get(Superclasses);
List<List<Integer>> pathMatrix = new ArrayList<List<Integer>>(EntitiesofSuperClasses.size());
Hashtable<Object, List<Integer>> PathMatrix = new Hashtable<Object, List<Integer>>(
					EntitiesofSuperClasses.size());
			for (int i = 0; i < EntitiesofSuperClasses.size(); i++) {
				int path = 0;
				List<Integer> pathLenght = new ArrayList<Integer>();
				Object u1 = EntitiesofSuperClasses.get(i);
				Object temp = u1;
				int pathCounter1 = 0;
				for (int j = 0; j < EntitiesofSuperClasses.size(); j++) {
					Object u2 = EntitiesofSuperClasses.get(j);
					int pathCounter2 = 0;
					if (u1.equals(u2)) {
					path = 0;
					} else {
					// u1 is an instance
					if (model.getIndividual(u1.toString()).isIndividual()) {
						u1 = FindParentClass(u1, model, Superclasses);
							pathCounter1++;
						}
					// u2 is an instance
					if (model.getIndividual(u2.toString()).isIndividual()) {
						u2 = FindParentClass(u2, model, Superclasses);
							pathCounter2++;
						}
				double score = engine.compare(smConf, factory.getURI(u1.toString()),
								factory.getURI(u2.toString()));
						path = ((int) (1 / score)) - 1;
						if (pathCounter1 != 0 || pathCounter2 != 0) {
							path = path + pathCounter1 + pathCounter2;
						}
					}
					pathLenght.add(path);
					pathCounter1 = 0;
					u1 = temp;
				}
				// Add each entity with its distances from all other entities
				PathMatrix.put(u1, pathLenght);
				// add the list of short paths to the overall list
				pathMatrix.add(pathLenght);

			}
```

```
                // calculate (dispersion) Ball-Hall Internal clustering validation for each
                // super class
                superClassDispersion.put(Superclasses, FindMedoid(PathMatrix));
        }
        return superClassDispersion;
}
public double CalculateDisparity(Hashtable<Object, Float> superClassDispersion) {
        float meanSum = 0;
        Enumeration en1 = superClassDispersion.keys();
        while (en1.hasMoreElements()) {
                Object superclass1 = en1.nextElement();
                float dispersion1 = superClassDispersion.get(superclass1);
                meanSum = meanSum + dispersion1;
        }
        double disparityBH = meanSum / superClassDispersion.size();
        return disparityBH;
}
private float FindMedoid(Hashtable<Object, List<Integer>> PathMatrix) {
        Object medoid = null;
        float min;
        float disparity = 0;
        Hashtable<Object, Float> average = new Hashtable<Object, Float>();
        // calculate average (midoid is the element with the minimal average distances//
        // the closest entity to all other entities
        Enumeration e = PathMatrix.keys();
        while (e.hasMoreElements()) {
                int sum = 0;
                float avg = 0;
                Object tag = e.nextElement();
                List<Integer> shortestPath = PathMatrix.get(tag);
                for (int i = 0; i < shortestPath.size(); i++) {
                        sum = sum + shortestPath.get(i);
                }
                avg = (float) sum / shortestPath.size();
        // put uri and average distances (average of shortest path for this entity with
        // all other entities)
                average.put(tag, avg);
        }
        // Find medoid, minimum average
        Enumeration e1 = average.keys();
        Object tag1 = e1.nextElement();
        min = average.get(tag1);
        while (e1.hasMoreElements()) {

                Object tag2 = e1.nextElement();
```

```
                float ave = average.get(tag2);
                if (average.get(tag2) < min) {
                        min = average.get(tag2);
                }
        }
        for (Map.Entry<Object, Float> entry : average.entrySet()) {
                if (entry.getValue() == min) {
                        medoid = entry.getKey();
                }
        }
        if (medoid.equals(null)) {
                System.out.println("Oops no mediod is found");
        } else {
        }
        // Now, identify which element is the medoid and save its distances
        Enumeration e3 = PathMatrix.keys();
        while (e3.hasMoreElements()) {
                Object tag4 = e3.nextElement();
                List<Integer> distanceValues = PathMatrix.get(tag4);
                if (tag4.equals(medoid)) {
                        disparity = BallHallDespersion(distanceValues, PathMatrix.size());
                }
        }
        return disparity;
}
private float BallHallDespersion(List<Integer> distanceValues, int superclassSize) {
        int sum = 0;
        float sum1 = 0;
        float dispersion = 0;
        // find mean (dispersion) from Ball-Hall for each cluster/super class
        for (int i = 0; i < distanceValues.size(); i++) {
                sum = sum + (distanceValues.get(i) * distanceValues.get(i));
        }
        dispersion = (float) sum / superclassSize;

        return dispersion;
}
public static Object FindParentClass(Object c, OntModel model, Object superclass) {
        Individual d = model.getIndividual(c.toString());
        OntClass individualParentClass = d.getOntClass();
        Resource Superclass = model.getResource(superclass.toString());
        if (individualParentClass.hasSuperClass(Superclass)) {
                return individualParentClass;
        } else {
                return CheckParentClass(d, model, superclass);
```

```
		}
	}
	public static Object CheckParentClass(Individual d, OntModel model, Object superclass) {
		OntClass SC = null;
		OntClass Sclass = model.getOntClass(superclass.toString());
	for (ExtendedIterator<? extends OntResource> j = Sclass.listInstances(); j.hasNext();) {
			Object instance1 = j.next();
			if (instance1.equals(d)) {
				SC = Sclass;
				break;
			}
		}
		if (SC == null) {
		for (ExtendedIterator<OntClass> k = Sclass.listSubClasses(false); k.hasNext();) {
				OntClass cls = k.next();
				for (Iterator i = cls.listInstances(true); i.hasNext();) {
					Object instance2 = i.next();
					if (instance2.equals(d)) {
						SC = cls;
						break;
					}
				}
			}
		}
		if (SC.equals(null)) {
			System.out.println("I am an instance cannot find my parent class");
		}
		return SC;
	}
}
```

Appendix B: *PresentationAttribute* Category in PreSOn

The figure illustrates the *PresentationAttribute* category in the PreSOn (mentioned in chapter 5 section 5.2.1) showing the sub-categories (subclasses). This category has 27 subclasses (excluding equivalent classes (≡)) and 136 instances, such as *organised, overwhelming, easy_to_follow, wordy*, and *eye_catching.*

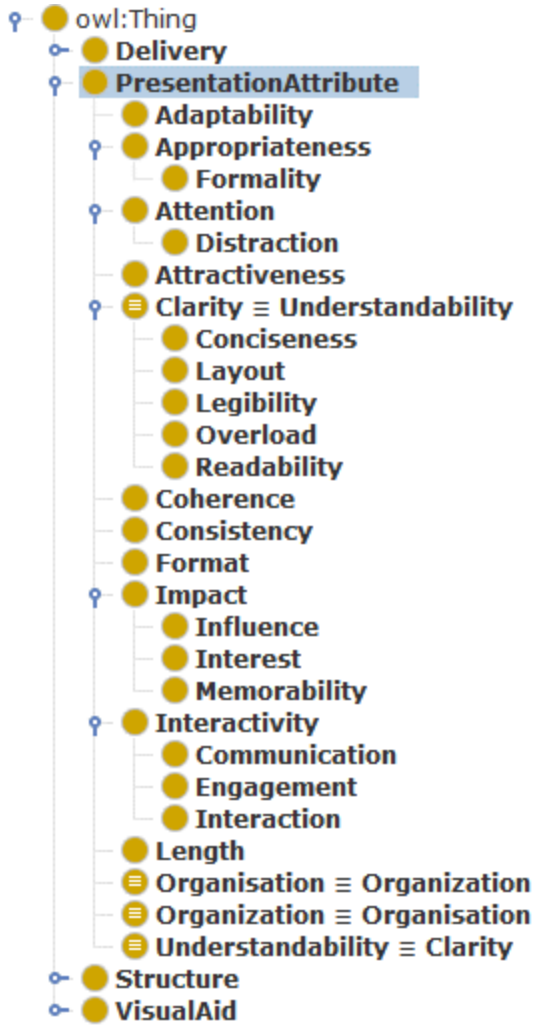

Figure B.1 *PresentationAttribute* category of PreSOn.

Appendix C: Domain Profiles of Videos with PreSOn's Entry Points

The following is part of the domain diversity profiles of the 8 videos in case study 2 (section 5.3.1) with the three entry points within PreSOn: *Delivery*, *Structure and VisualAid*. The videos are sorted based on their Ids alphabetically.

Table C.1 Domain diversity profiles for study A with *Delivery*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	4	0.83	0.11	5.9
E 2	4	0.73	0.11	4.6
E 3	3	0.54	0.07	2.7
E 4	5	0.93	0.09	4.47
T 1	2	0.5	0.12	4.13
T 2	4	1.37	0.09	5.83
T 3	4	1.31	0.23	5.61
T 4	4	0.81	0.16	6.22

Table C.2 Domain diversity profiles for study B with *Delivery*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	5	1.23	0.15	3.37
E 2	4	0.95	0.09	4.49
E 3	4	0.81	0.17	4.8
E 4	5	0.9	0.09	2.32
T 1	3	1.07	0.12	6.47
T 2	4	1.16	0.08	3.88
T 3	5	1.51	0.12	2.52
T 4	5	0.77	0.15	3.58

Table C.3 Domain diversity profiles for study A with *Structure*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	1	0	0.23	5.65
E 2	2	0.27	0.18	4.67
E 3	1	0	0.16	7.33
E 4	1	0	0.15	8.73
T 1	2	0.26	0.19	2.81
T 2	1	0	0.23	9
T 3	1	0	0.17	9.31
T 4	1	0	0.08	7.33

Table C.4 Domain diversity profiles for Study B with *Structure*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	1	0	0.13	7.7
E 2	1	0	0.15	9
E 3	1	0	0.19	8.07
E 4	1	0	0.15	8.27
T 1	1	0	0.21	9.69
T 2	1	0	0.2	7.47
T 3	1	0	0.28	10.48
T 4	1	0	0.11	9

Table C.5 Domain diversity profiles for study A with *VisualAid*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	2	0.5	0.09	8.69
E 2	2	0.26	0.1	5.46
E 3	2	0.45	0.1	5.85
E 4	2	0.3	0.08	6.95
T 1	2	0.17	0.16	4.74
T 2	2	0.45	0.05	6.8
T 3	1	0	0.13	13.91
T 4	1	0	0.04	4.33

Table C.6 Domain diversity profiles for study B with *VisualAid*.

Video Id	Domain variety	Domain balance	Domain coverage	Domain within disparity
E 1	2	0.56	0.08	6.83
E 2	2	0.41	0.17	7.03
E 3	2	0.21	0.12	3.88
E 4	2	0.27	0.09	5.58
T 1	2	0.43	0.22	8.49
T 2	2	0.5	0.05	5.63
T 3	1	0	0.18	9.13
T 4	2	0.5	0.05	4.75

Appendix D: User Profiles for User Domain Diversity

This appendix is related to findings in section 5.3.2 Domain Profiles for Users. These are the user attributes with *non-significant* findings when compared across diversity properties.

Gender:

Table D.1 Average (&standard deviation) of diversity properties based on gender for study A and study B. *Within disparity was significant for study A*.

	Study A		**Study B**	
	Male (12)	*Female* (26)	*Male* (82)	*Female* (58)
Domain variety	3.17(1.14)	3.88(0.33)	2.77(1.22)	3.05(1.07)
Domain balance	0.97(0.48)	1.28(0.11)	0.82(0.49)	0.94(0.42)
Domain coverage	0.04(0.02)	0.05(0.02)	0.03(0.02)	0.03(0.02)
Domain within disparity			7.57(6.02)	8.59(5.03)

Language:

Table D.2 Average (&standard deviation) of diversity properties based on gender for study A and study B.

	Study A		**Study B**	
	Native (23)	*Non-native* (15)	*Native* (119)	*Non-native* (22)
Domain variety	3.65(0.88)	3.67(0.62)	2.9(1.18)	2.86(1.08)
Domain balance	1.18(0.39)	1.19(0.19)	0.87(0.47)	0.89(0.46)
Domain coverage	0.05(0.03)	0.04(0.02)	0.03(0.02)	0.03(0.02)
Domain within disparity	11.38(4.33)	9.57(4.89)	7.76(5.67)	9.34(5.25)

The most and least diverse learners:

Table D.3 Average (&standard deviation) of top and bottom quartiles of study A and study B. YT4L refers to using YouTube for Learning.

	Study A		**Study B**	
	Top quartile(10)	*Bottom quartile*(10)	*Top quartile*(35)	*Bottom quartile*(35)
Training	2.5(1.18)	1.8(0.79)	1.8(0.9)	1.6(0.77)
Experience	2.8(1.03)	2.7(0.67)	2.37(0.77)	2.11(0.8)
YouTube	3.2(1.32)	3.7(1.06)	4.09(0.98)	4(1.16)
YT4L	2.3(0.95)	3.1(1.2)	3.14(1.09)	2.86(1.14)