INFLIBNET
ETD 2023
NDLTD
Gandhinagar, Gujarat, India
26-28 October 2023

VIRGINIA TECH

Case Study of Analyzing the Variety of ETD Layouts

Sung Hee Park[1,3], Bipasha Banerjee[1,2], William A. Ingram[1,2], and Edward A. Fox[2]

1 University Libraries, Virginia Tech, USA
2 Department of Computer Science,Virginia Tech, USA
3 Department of Library and Information Science, Hannam University, South Korea

Content

1. Introduction
2. Research Objectives
3. Data Preparation
4. Layout Variety of Numbering Styles
5. Results
6. Conclusion

Introduction

- Existing ETDs can be easily reused by AI for functionalities like:
  - Searching, browsing, summarizing, and topic modeling
- Yet, automatic and accurate segmentation of ETDs into chapters is challenging:
  - In dealing with various ETD layouts from different majors, disciplines, and universities.
- To overcome this challenge,
  - We need to understand the variation in document templates across various disciplines and universities.
  - Thus, it is imperative that we analyze the ETD layouts to determine the variation among them.
- This study can help us create better models to perform automatic ETD segmentation more accurately.

Research Objectives

(1) We identify numbering styles that characterize the hierarchical structure of a document.

(2) We identify if elements other than the dependent ones affect dependent variables such as the ETD layout and numbering style, which will be used as independent variables in this study.

- Examples include universities, departments, STEM/non-STEM, majors, and year of publication.
- We want to specifically identify variations in layouts among documents from STEM and non-STEM fields and what makes those variables.

Data Preparation (1)

- Sampling source
  - Over 500,000 ETDs from U.S. research institutions (Uddin et al., 2021)
- Universities
  - STEM(5): Ohio State University(OSU), Caltech(CalTech), Virginia Tech(VT), Georgia Tech(GT), University of Texas at Austin(UTAustin)
  - Non-STEM(5): Ohio State University(OSU), Virginia Tech(VT), University of Texas at Austin(UTAustin), Penn State University(PSU), North Carolina State University(NCSU)
- Discipline/departments
  - STEM(5): Computer science(CS), Biology(Bio), physics(Phy), Mechanical engineering(ME), Civil engineering(CE)
  - Non-STEM (4) : Psychology(Psy), Education(Edu), English(Eng), Business(Bus)

Data Preparation (2)

- Publication time periods (5): 1995~2000, 2001~2005, 2006~2010, 2011~2015, 2016~2020
- And then, we conducted purposive sampling to ensure even distribution across universities, discipline fields, and time periods.
- Finally, we sampled 20 non-STEM ETDs and 27 to 33 STEM ETDs from each university.
- Overall, 100 non-STEM ETD samples were drawn and 140 STEM ETD samples were drawn, for a total of 240 samples.

Data Preparation (3)

- Sample statistics

Table 1 : Dataset Statistics

University	Category	1991~2000	2001~2005	2006~2010	2011~2015	2015~	SUM
OSU	STEM	5	5	5	5	7	27
	Non STEM	4	4	4	4	4	20
VT	STEM	5	5	5	5	11	31
	Non STEM	5	5	4	3	3	20
UTA	STEM	1	9	5	5	10	30
	Non STEM	4	4	4	4	4	20
Caltech	STEM	5	5	5	5	7	27
GaTech	STEM	5	5	5	5	5	25
PSU	Non STEM	3	5	4	4	4	20
NCSU	Non STEM	4	7	4	4	1	20
SUM		41	54	45	44	56	240

Overall Observations

- Learning from Observations:
  - The layout tends to be similar within a university, but with minor variety by the departments.
  - The layouts tend to vary significantly across different universities.
  - This is likely to occur as each university library or graduate school typically provides an ETD template.

Defining Chapter/section Numbering Styles

- The five styles indicate how deeply ETDs use numbering in the chapter/section headings.

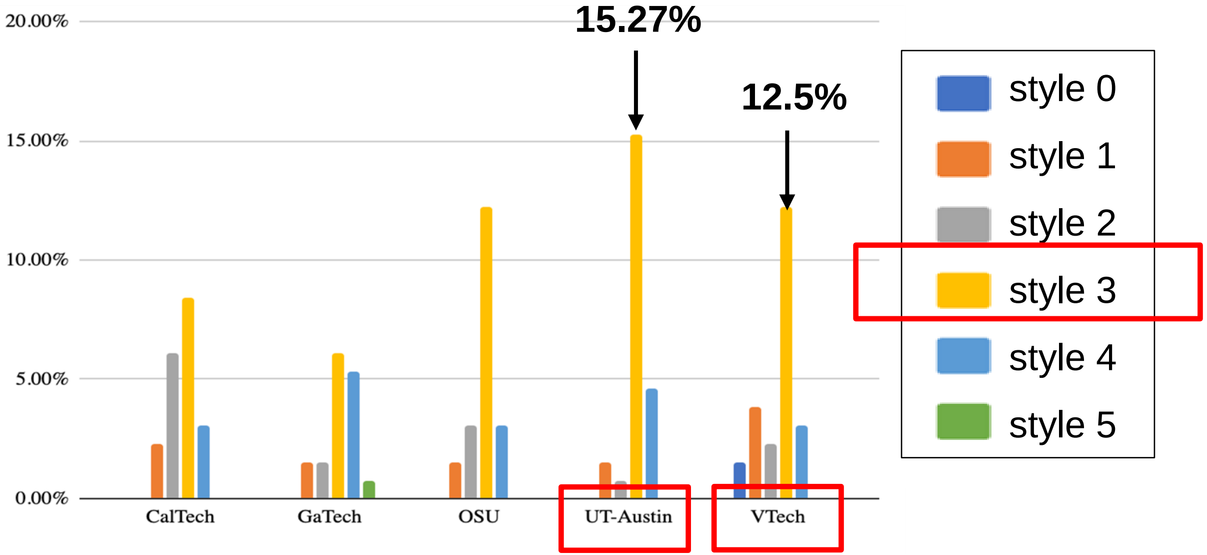
Table 4 : Chapter/Section numbering Styles for in document text

Numbering style	Description	Example	source
0	Even chapter level numbers don’t exist.	Introduction	ETD id: 247230 from English, UT Austin
1	Only chapter level numbers exist but section level numbers do not exist.	CHAPTER 1 INTRODUCTION	ETD id: 98948 from Education, OSU
2	Chapter and section level numbers exist but subsection level numbers do not exist.	1.1 Introduction	ETD id: 116377 from Business, OSU
3	Even subsection level numbers exist	4.3.1 User algorithms	ETD id: 42990 from CS, GaTech
4	Even subsubsection level numbers exist	4.1.3.1 Probability of Collision	ETD id: 63305 from Mechanical Engineering, CalTech

Result – STEM ETDs across Universities

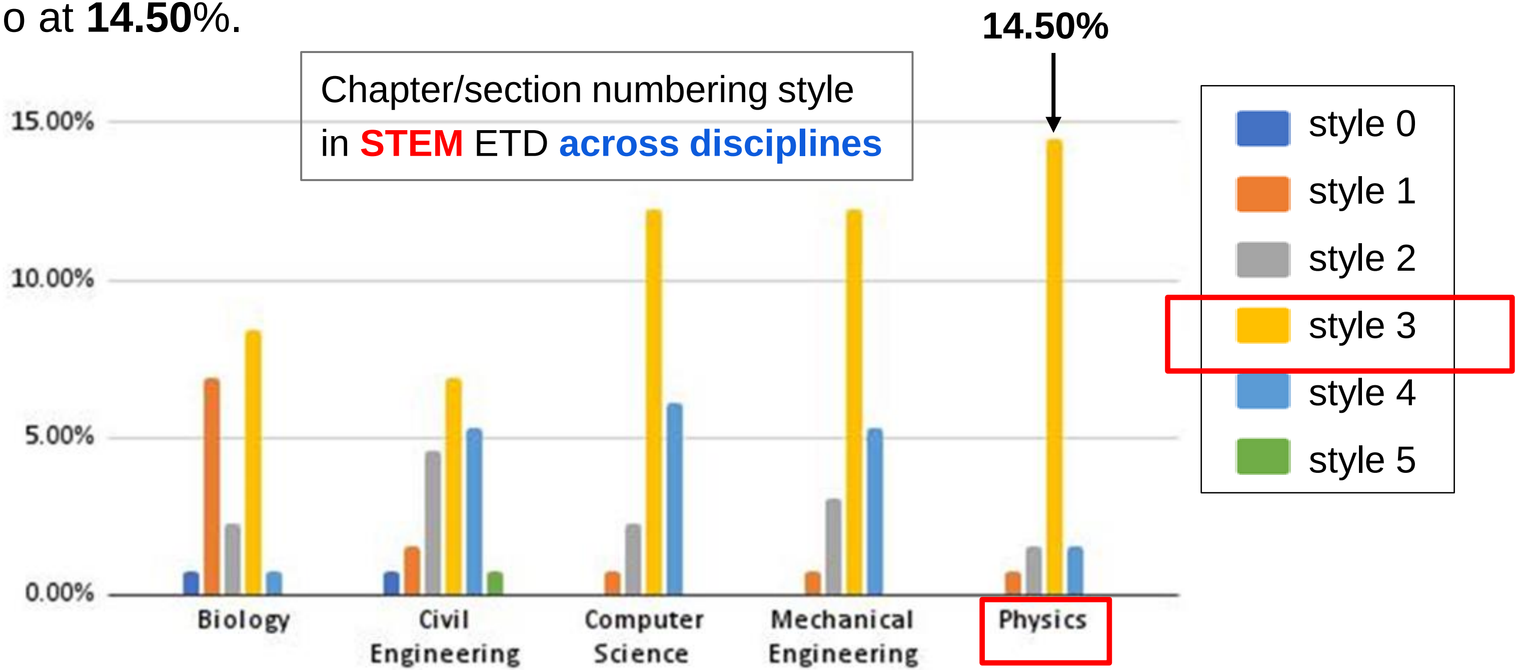
Among universities, the **University of Texas at Austin** showed the highest rates of **15.27%** for **the style 3** and **the OSU and Virginia Tech** showed the next high rate of **12.5%** for the **style 3**, respectively,

Chapter/section numbering style in **STEM** ETD **across Universities**



Result – STEM ETDs across Disciplines

Among disciplines, style 3 had the highest ratio at 54.20%, and physics showed the highest ratio at **14.50**%.

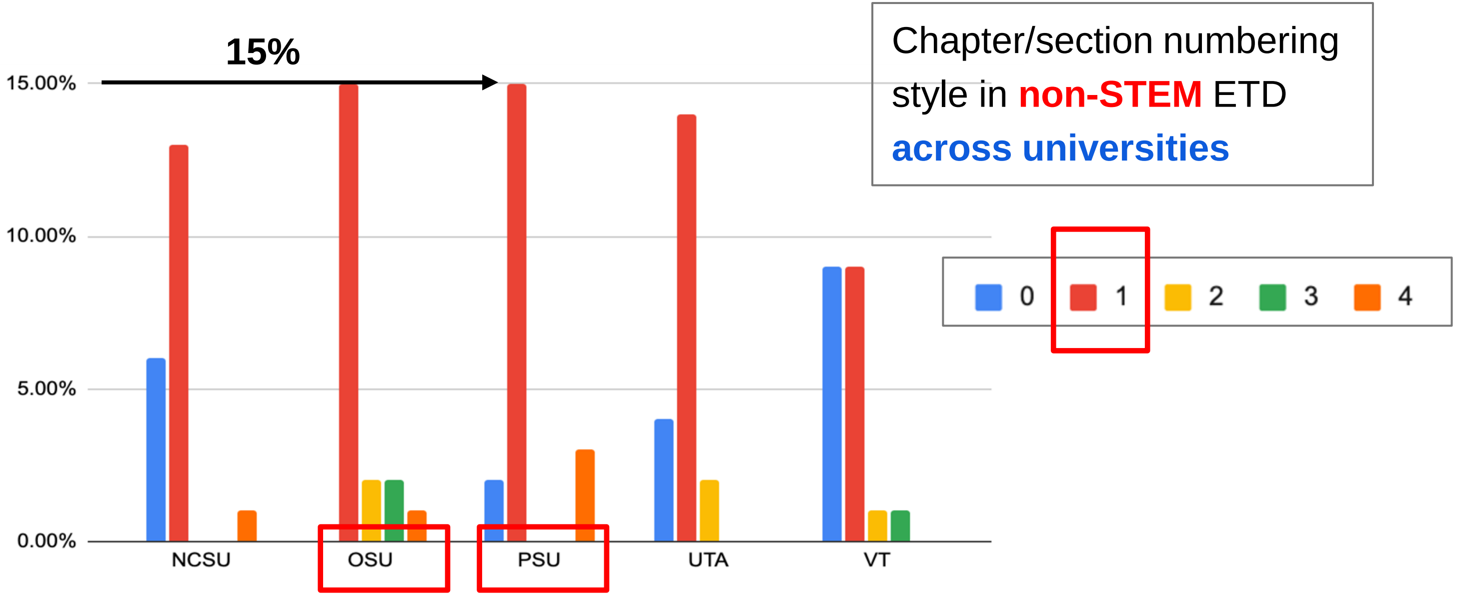


Chapter/section numbering style in **STEM** ETD **across disciplines**

14.50%

Results - Non-STEM across Universities

- Among universities, **OSU and PSU** showed **equally high rates** of style 1 at 15% and 15%, respectively, and **UTAustin (14%) and NCSU (13%)** showed **no statistically significant difference** for each style among schools.

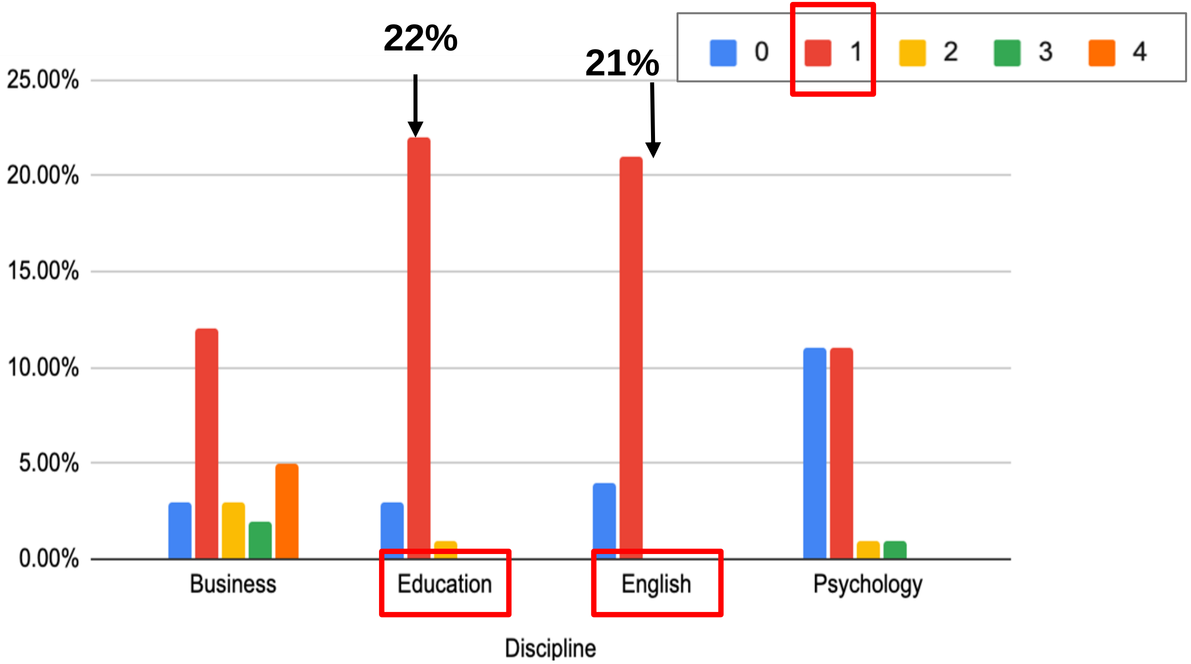


Chapter/section numbering style in **non-STEM** ETD **across universities**

Results of Non-STEM ETDs across Universities

- Among non-STEM sub areas, **education**, and **English** majors showed **the highest usage** of **style 1** at **22%** and **21%**, respectively.

Chapter/section numbering style in **non STEM** ETD **across disciplines**



Learning from Frequency Analysis

- We learned from frequency analysis of chapter/section numbering style
  - **STEM** fields,
    - **style 3** had **the highest ratio** at 54.20% in different disciplines, and **physics** showed the highest ratio at 14.50%.
    - Among universities, **UTAustin** showed the **highest ratio** at 15.27% for **style 3**
  - **non-STEM** fields,
    - **education, and English** majors showed **the highest usage** of **style 1** at 22% and 21%, respectively
    - Among universities, **OSU and PSU** showed **equally high rates** of **style 1** at 15% and 15%, respectively, and
    - UTAustin (14%) and NCSU (13%) showed **no statistically significant difference** for each style among schools
- That is, in **STEM** fields, the numbering is introduced up to the subsubsection, and in **non-STEM**, the numbering is introduced only up to the chapters.

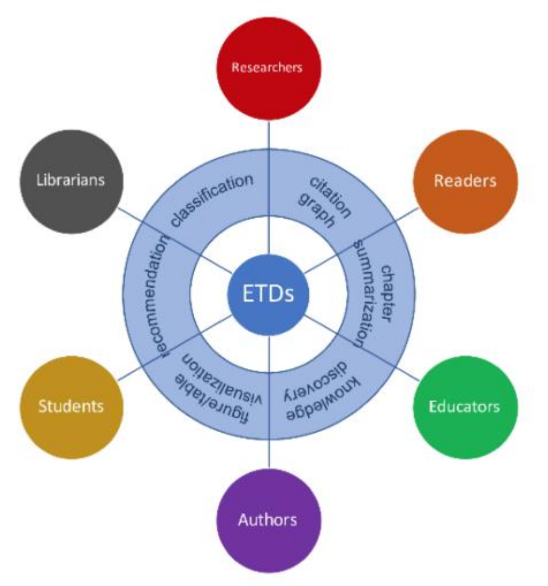
χ2 Test of Chapter/section Numbering Style Variation

- Chi-square(χ2) independence test
  - In order to check if the independence of STEM/non-STEM fields between the numbering styles exists
  - The p-value of the Chi-square independency test is <0.001
  - Thus, we reject the null hypothesis that STEM/non-STEM fields are independent of the variety of numbering styles.
  - The numbering style in chapter/section heading is dependent on STEM/non-STEM areas

Conclusion and Future Work

- Summary:
  - Data Sampling : 240 STEM/non-STEM ETDs
  - Defining 6 types of chapter/section numbering style
  - Observation of overall ETD layouts variety
  - Frequency analysis across different universities, disciplines
  - Chi-Square independence test
- We conclude:
  - **The discipline information** of the ETD **affects the variation of its layout**, particularly, **the numbering style of chapter/section titles**.
- Future Work
  - To analyze the layout variations in terms of **other layout elements** of ETD (e.g. **title page, table of contents, citation styles and reference styles, page numbering style**, figures and tables with/without captions, equations, and algorithms).

ETD Research Team from VT & ODU



- Research Project
  - Title: “Opening Books and the National Corpus of Graduate Research”
  - Period: 2019 – 2022 (3 years)
  - IMLS funded research grant project
  - Led by the University Libraries at Virginia Tech
- Research Team

William A. Ingram,
Principle Investigator,
Assistant Dean and Director of IT, University Libraries, Virginia Tech

Dr. Edward A. Fox,
Co-PI,
Professor, Computer Science, Virginia Tech

Dr. Jian Wu,
Co-PI,
Professor, Computer Science, Old Dominion University

Bipasha Banerjee,
Graduate Assistant,
Ph.D. Candidate,
Computer Science,
Virginia Tech

Muntabir Choudhury,
Graduate Research Assistant
Ph.D. Candidate
Computer Science, Old Dominion University

Q&A

Thank you!

Question: Email to shpark@vt.edu