Spring 2020

Parsimonious Sociology Theory Construction: From a Computational Framework to Semantic-Based Parsimony Analysis

Mingzhe Du

Recommended Citation

Du, M.(2020). *Parsimonious Sociology Theory Construction: From a Computational Framework to Semantic-Based Parsimony Analysis.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/5938

Parsimonious Sociology Theory Construction: From A Computational Framework to Semantic-based Parsimony Analysis

by

Mingzhe Du

Bachelor of Software Engineering
Harbin University of Science and Technology, 2007

Master of Software Engineering
University of South Carolina, 2012

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2020

Accepted by:

Jose M. Vidal, Major Professor

Jijun Tang, Committee Member

John R. Rose, Committee Member

Manton M. Matthews, Committee Member

Barry Markovsky, Committee Member

Cheryl L. Addy, Vice Provost and Dean of Graduate Studies

Acknowledgments

First and foremost, I would like to express thanks to my advisor Dr. Jose Vidal. Dr. Vidal is a tremendous mentor for me. He has supported my research for many years and sincerely encouraged me to grow as a research scientist. I also would like to express my sincerest gratitude to Dr. Markovsky for his support and guidance throughout my graduate studies. I cannot be here without all your constant feedback and unlimited support. I appreciate all your contributions of time, ideas, and critical thinkings to make my Ph.D. experience productive and stimulating.

To my doctoral dissertation committee: Dr. Jijun Tang, Dr. John Rose, and Dr. Manton Matthews, I appreciate the time that you have spent on reading my dissertation and invaluable advice for my research. It is an honor to have you as my committee members.

I would like to thank my colleagues and friends for their continued support. This dissertation would not have been possible without the contributions of Zaid Alibadi, Dazhou Guo, Bing Feng, Jing Wang, Jake Frederick, and Nicolas Harder.

Thanks are also due to the (NSF) National Science Foundation (under Grants SES-1123040 and IIS-1551458). I otherwise would not have been able to develop my scientific discoveries.

Last but not least, I would like to express my deepest gratitude to my family and friends. This dissertation would not have been possible without their warm love, continued patience, and endless support.

Abstract

In the social sciences, theories are used to explain and predict observed phenomena in the natural world. Theory construction is the research process of building testable scientific theories to explain and predict observed phenomena in the natural world. Conceptual new ideas and meanings of theories are conveyed through carefully chosen definitions and terms.

The principle of parsimony, an important criterion for evaluating the quality of theories (e.g., as exemplified by Occam's Razor), mandates that we minimize the number of definitions (terms) used in a given theory.

Conventional methods for theory construction and parsimony analysis are based on heuristic approaches. However, it is not always easy for young researchers to fully understand the theoretical work in a given area because of the problem of "tacit knowledge", which often makes results lack coherence and logical integrity. In this research, we propose to help with this problem in three parts.

In the first part of this study, we present Wikitheoria, a generic knowledge aggregation framework, to facilitate the parsimonious approach of theory construction with a cloud-based theory modularization platform and semantic-based algorithms to minimize the number of definitions. The proposed approach is demonstrated and evaluated using the modularized theories from the database and sociological definitions retrieved from the system lexicon and sociological literature. This study proves the effectiveness of using a cloud-based knowledge aggregation system and semantic analysis models for promoting the parsimonious sociology theory construction.

In the second part, our study is focused on semantic-based parsimony analysis. We introduce an embedding-based approach using machine learning models to reduce the semantically similar sociological definitions, where definitions are encoded with word embeddings and sentence embeddings. Given several types of embeddings exist, we compare the definition's encodings with the goal of understanding what embeddings are more suitable for knowledge representation, and what classifiers are more capable of capturing semantic similarity in the task of parsimonious theory construction.

In the final part of this study, we propose SOREC, a novel semantic content-based recommendation system (CBRS) with the supervised machine learning model for theoretical parsimony evaluation by checking the semantic consistency of definitions while constructing theories. Specifically, we evaluate the XGBoost tree-based classifier with the combination of low-level features and high-level features on our dataset. The proposed CBRS substantially outperforms conventional matrix factorization-based CBRS in suggesting semantically related sociological definitions. In this study, we provide a solid baseline for future studies in the research area of sociological definition semantic similarity computation. Moreover, theory construction is a common research process in a lot of human science-related disciplines such as psychology, criminology, and other social sciences. The results of this study can be further applied to the theory construction in these disciplines.

Contents

Acknowledgments iii

Abstract iv

List of Tables viii

List of Figures ix

Chapter 1 Introduction 1

1.1 Scope of the Research 2

1.2 Dataset 8

1.3 The Structure of the Thesis 9

Chapter 2 Modularized Theory Construction and Parsimony Analysis with Wikitheoria 10

2.1 Related Works 10

2.2 Method 14

2.3 Results and Discussion 24

2.4 Conclusion 25

Chapter 3 Towards Parsimonious Theory Construction with Neural Embeddings and Semantic Measurement . . . 27

3.1 Related Works 27

3.2 Method 28

3.3 Results and Discussion 35

3.4 Conclusion 38

Chapter 4 Semantic Content-based Recommendation with SOREC 40

4.1 Related Works 40

4.2 Method 43

4.3 Results and Discussion 51

4.4 Conclusion 57

Chapter 5 Conclusion 58

5.1 Summary 58

5.2 Future Works 59

Bibliography 60

List of Tables

Table 2.1 The experiment results of k-nearest neighbors with different values of k. 26

Table 2.2 The results of k-nearest neighbors by category when k is equal to 5. 26

Table 3.1 The word-level embedding methods and their corresponding training corpora. 29

Table 3.2 The sociological definition with the sentence-level embedding methods and their corresponding training corpora. 32

Table 3.3 The word-level sociological definition embedding results on test data with 10-fold cross validation. 36

Table 3.4 The sentence-level sociological definition embedding results on test data with 10-fold cross validation. 37

Table 4.1 The comparison results of prediction on sociological definition test dataset with 10-fold cross validation. 54

Table 4.2 The comparison results of prediction with Word2Vec and Transformer. 54

List of Figures

Figure 1.1 An illustration of Wikitheoria user interface. The designed framework was deployed and served on Google App Engine. 3

Figure 1.2 An illustration of Wikitheoria theory construction with the process of theory modularization and parsimony analysis. 4

Figure 1.3 An illustration of annotation with sociological definitions and scores to train the semantic classifiers for binary classification. 8

Figure 2.1 An illustration of published formal theory on Wikitheoria platform. 11

Figure 2.2 An schematic illustration of two simple theory modules with two propositions and a derivation. 13

Figure 2.3 An illustration of Wikitheoria infrastructure with devices and Google cloud. 15

Figure 2.4 An illustration of Wikitheoria guide page for the newly registered users. 16

Figure 2.5 An illustration of Wikitheoria library page for experienced users to quickly access the published theory modules. 17

Figure 2.6 An illustration of a theory that are under construction on Wikitheoria. 18

Figure 2.7 An illustration of how the administrators manage different roles of users with the system. 19

Figure 2.8 An illustration of peer review process for suggesting modification on the published theory module. 20

Figure 2.9 An illustration of how the administrators manage the proposed contributions for the peer review system. 21

Figure 2.10 An illustration of how the administrators manage the lexicon and terms that associate with the published theory modules. . . . 22

Figure 2.11 An illustration of how the administrators manage the binary files in blobstore. 23

Figure 2.12 An illustration of published modularizrd theory module on Wikitheoria platform. 24

Figure 3.1 An illustration of semantic-based parsimony analysis workflow with encoding and training process. 29

Figure 4.1 An illustration of SOREC recomendation workflow with offline training and online prediction process. 44

Figure 4.2 An illustration of three stage recommendation with sociological terms and definitions on Wikitheroia platform. 45

Figure 4.3 An illustration of sociological terms and definitions recommendation user interface on Wikitheoria. 46

Figure 4.4 An illustration of SOREC prediction workflow with three feature sets and definition transformation. 51

Figure 4.5 The results with descriptive features in normalized confusion matrix. 55

Figure 4.6 The results with descriptive features and tokenized features. . . . 55

Figure 4.7 The results of prediction when descriptive features, tokenized features and embedding features are concatenated together. . . . 56

Figure 4.8 The results of conventional content-based recommendation system prediction accuracy with TF-IDF, Word2Vec and SVD. . . . 56

Chapter 1

Introduction

In the social sciences, theory construction is the research process of formulating scientific theories with references to explicit logical and semantic criteria, where theories are defined as a set of explicit, abstract, general, logically related statements designed to explain observed phenomena[42]. A successful theory is one that, when applied to specific empirical cases, describes relationships among phenomena, and explains and predicts the occurrence of certain events. Good scientific theories include four essential components: terms, statements, arguments, and scope conditions. Terms are used to build statements; statements are used to build arguments; arguments apply under a set of scope conditions[42]. Terms are carefully chosen by the theorist to convey new concepts or ideas in theory, and their meanings are implicated in the definitions[42, 43].

The principle of parsimony, an important criterion for evaluating the quality of theories (e.g., as exemplified by Occam's Razor) mandates that we minimize the number of definitions (terms) used in a given theory[1, 32]. For example, consider the following two definitions for the term "mechanical solidarity" extracted from the Blackwell Encyclopedia of Sociology. Definition 1: According to Emile Durkheim, mechanical solidarity refers to the factors that hold primitive societies together, mostly through family and kinship ties and a collective consciousness shared by all members of the community. Definition 2: Durkheim's term for the unity (a shared consciousness) that people feel as a result of performing the same or similar tasks. In the process of theory construction, if definition 1 were in theory already and the sociologist plans to add a new definition 2, he should determine whether the new definition 2 is already in theory or is similar to the pre-defined definition 1. If either is the case, only definition 1 should be used instead of creating a new definition.

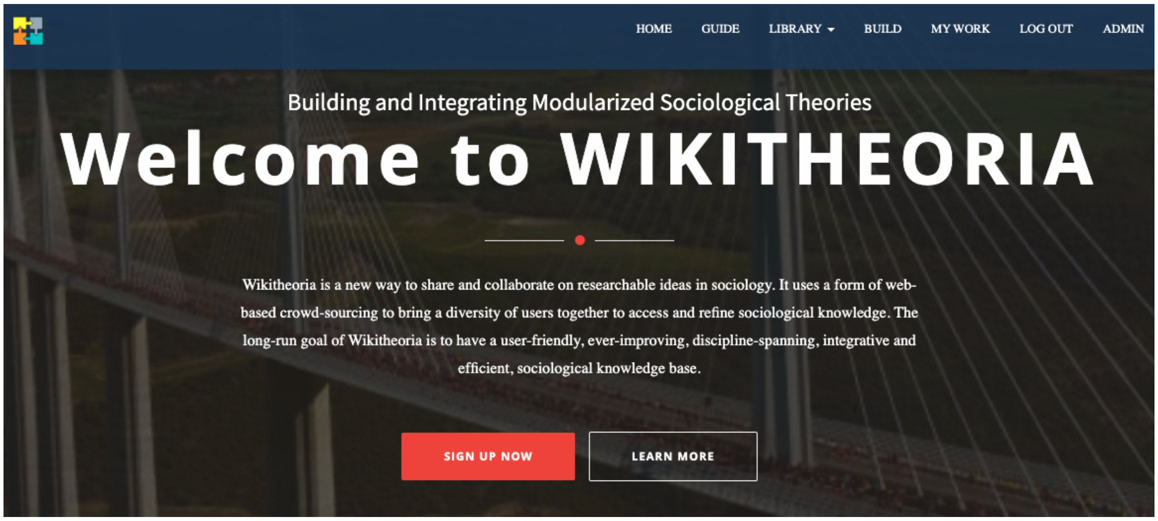
The conventional methods for parsimony analysis in social science theory construction are based on the heuristic approaches, which are determined by the human[43, 32]. However, it is not always easy for young researchers to fully understand the theoretical work in a given area – although they are trained by mentors who are familiar with accepted views in that area. To try to acquire a sense of understanding in another theoretical area can be difficult because of the problem of "tacit knowledge" – essentially inside information about how to interpret certain vague or ambiguous terminologies, which often makes results lack coherence and logical integrity[1, 21, 42, 43]. In this research, we propose to help with this problem in the following studies.

1.1 Scope of the Research

1.1.1 Modularized Theory Construction and Parsimony Analysis with Cloud-based Wikitheoria

For the first part of this study, we present Wikitheoria, a fully functional knowledge aggregation web framework we built for modularized theory construction in social sciences. As shown in Figure 1.1, it corporates various sub-systems for managing a lexicon, a library of theory modules, a set of registered users, a peer review process, automated checks to rule out definitional circularizes and non-causal propositions, plus various administrative operations. The lexicon of Wikitheoria preserves the pre-defined terms and definitions and helps to encourage more parsimonious theory construction. In turn, this facilitates communication along with logical and empirical analysis of the theory[29, 16, 53].

As illustrated in Figure 1.2, the proposed approach consists of three major steps: (1) Sociology theories were modularized and constructed with the cloud-based tools



Building and Integrating Modularized Sociological Theories

Welcome to WIKITHEORIA

Wikitheoria is a new way to share and collaborate on researchable ideas in sociology. It uses a form of web-based crowd-sourcing to bring a diversity of users together to access and refine sociological knowledge. The long-run goal of Wikitheoria is to have a user-friendly, ever-improving, discipline-spanning, integrative and efficient, sociological knowledge base.

Figure 1.1: An illustration of Wikitheoria user interface. The designed framework was deployed and served on Google App Engine.

provided by our platform. (2) Definitions in theories were pre-processed, then encoded with Transformer-based Universal Sentence Encoder. (3) Cosine similarity and K-Nearest Neighbors (KNN) algorithms were employed to calculate the semantic similarity of definitions and further reduce the redundant definitions for theory construction. To the best of our knowledge, our work is the first to systematically apply cloud-based modularized theory construction with semantic-based parsimony analysis by using neural embedding and machine learning model.

The main contributions of this research are as follows:

- We develop a computational framework for theory modularization and theory construction.

- We prove the effectiveness of using embedding models on the semantic similarities of sociological definitions.

- Additionally, we experiment with textual similarity measurement (cosine similarity) and similarity prediction (KNN) in which the result achieves an accuracy of 81.69%.

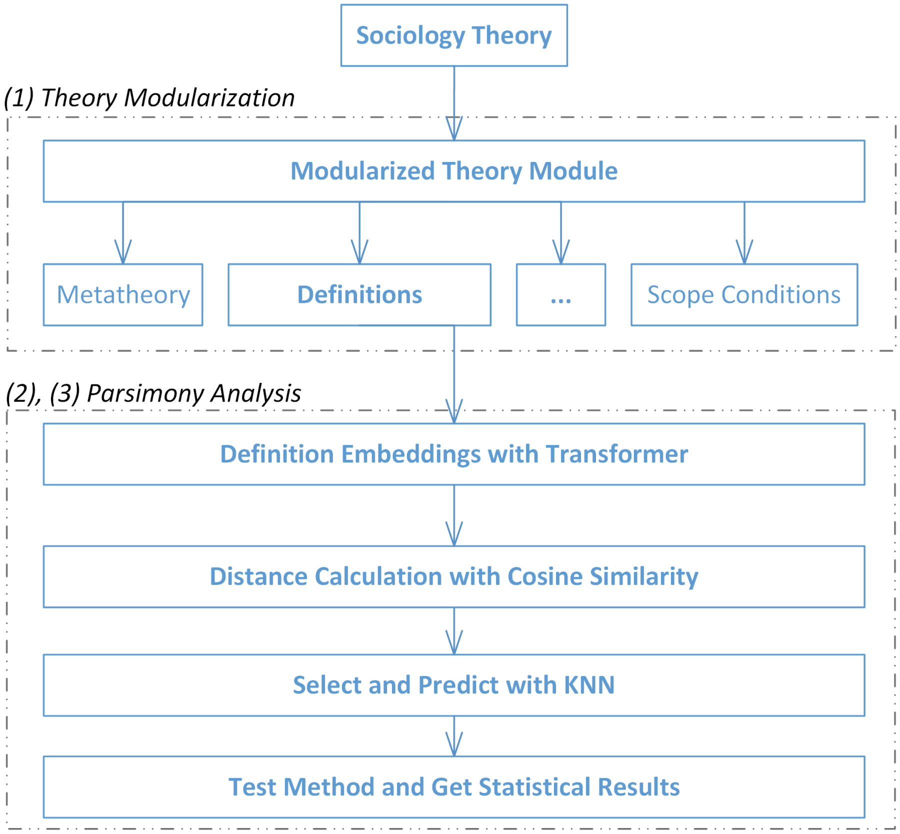

Sociology Theory
(1) Theory Modularization
Modularized Theory Module
Metatheory
Definitions
...
Scope Conditions
(2), (3) Parsimony Analysis
Definition Embeddings with Transformer
Distance Calculation with Cosine Similarity
Select and Predict with KNN
Test Method and Get Statistical Results

Figure 1.2: An illustration of Wikitheoria theory construction with the process of theory modularization and parsimony analysis.

1.1.2 Towards Parsimonious Theory Construction with Neural Emebeddings and Semantic Measurement

In the second part, our study is focused on semantic-based parsimony analysis. We introduce an automatic approach using the distributed semantic embeddings and machine learning models to evaluate the semantic consistency of sociological definitions and reduce the semantically similar sociological definitions for theory construction, which ensures the similar definitions for different terms could converge into one definition for a single term [15][16]. Therefore, it helps safeguard against redundancy and fosters the more parsimonious theories.

The developed approach consists of three components: data pre-processing, feature extraction, and definition similarity prediction. For data pre-processing, the definitions are extracted from a collection of sociological books, then annotated by sociologists. The sociological definitions are tokenized with the removal of stop words, then converted to the lower case. Considering the excellent performance of neural embeddings on capturing the semantic similarities, we experiment with four word-level embeddings [46, 51, 52, 10] and four sentence-level embeddings [17, 20, 25] with the goal of gaining insights of which embeddings are most suitable for sociological definitions. For feature extraction, we exploit 11 features with the embedding-based similarity metrics. For definition training and model prediction, we adopt four different types of supervised classifiers on the feature representations to predict the sociological definition semantic similarity [50]. Our work is the first to apply recent state-of-art pre-trained word-level embeddings and sentence-level embeddings models to measure the semantic similarities of sociological definitions, and the first one that evaluates four different types of classifiers on promoting the parsimony for sociology theory construction.

The main contributions of this study are as follows:

- We develop an embedding-based approach using the supervised machine learning model to reduce the semantically similar sociological definitions, where definitions are encoded with word embeddings and sentence embeddings.

- We study eight sociological definition embedding methods to understand what representations are more suitable for our task.

- Additionally, we experiment with four different supervised models to gain insights into classifiers that generalize well from embeddings.

Our experimental results showed that the Transformer outperforms other seven embedding methods when employed with supervised machine learning models. The proposed approach achieves the best accuracy of 84.82%, comparing with Word2Vec

(81.7%), GloVe (82.14%), ELMo (64.73%), fastText (79.91%), InferSent (75%), USE-DAN (83.48%) and BERT (55.8%).

1.1.3 Semantic Content-based Recommendation with SOREC

In the final part of this study, we propose SOciology RECommendation system (SOREC), a novel semantic content-based recommendation system with the supervised machine learning model for theoretical parsimony evaluation by checking the semantic consistency of definitions while constructing theories.

With the explosion of big data and model-based content-based recommendation system [55, 30, 3, 36], there are multiple approaches to tackle this problem. One of them is a semantic ontology-based approach, which uses of WordNet[4, 45, 14] in enhancing semantic-based analysis where hierarchies of concepts are built to capture conceptual relations between words and sentences. This approach performs well on the general domain, but many sociological terms are utilized and/or defined differently from generic English.

Another promising approach is based on Latent Semantic Analysis, which originates from the principle that words used in the same context tend to have similar meanings. Semantic relatedness is discovered through matrix factorization[54]. This approach performs well on various CBRS[3, 36], and is considered as the baseline method.

In recent years, word embeddings and sentence embeddings have produced high-quality representations for words and sentences on a broad spectrum of natural language understanding applications[17, 46, 20]. Especially, deep neural language models have demonstrated the efficacy by using pre-trained language models followed by fine-tuning dataset and achieved state-of-the-art results in semantic similarity related tasks[17]. Considering the excellent performance on representing the semantic similarities[27], the definitions in our study are embedded with Transformer in [17].

The proposed SOREC is designed to check the semantic consistency of sociological definitions, which consists of three components: data pre-processing, feature extraction, and definition recommendation. For data pre-processing, the definitions were extracted from a collection of sociological books, then annotated by sociologists. Prior to feature extraction, definitions were tokenized, stop words were removed, and all words were converted to the lower case. For feature extraction, we exploit 15 low-level features from the basic properties of definitions and the edit distance, 11 high-level features from the embedding-based distance metrics. For definition training and recommendation, we adopt XGBoost[18] on the feature representations extracted from 26 features to predict the definition similarity.

To sum up, the main contributions of this research are:

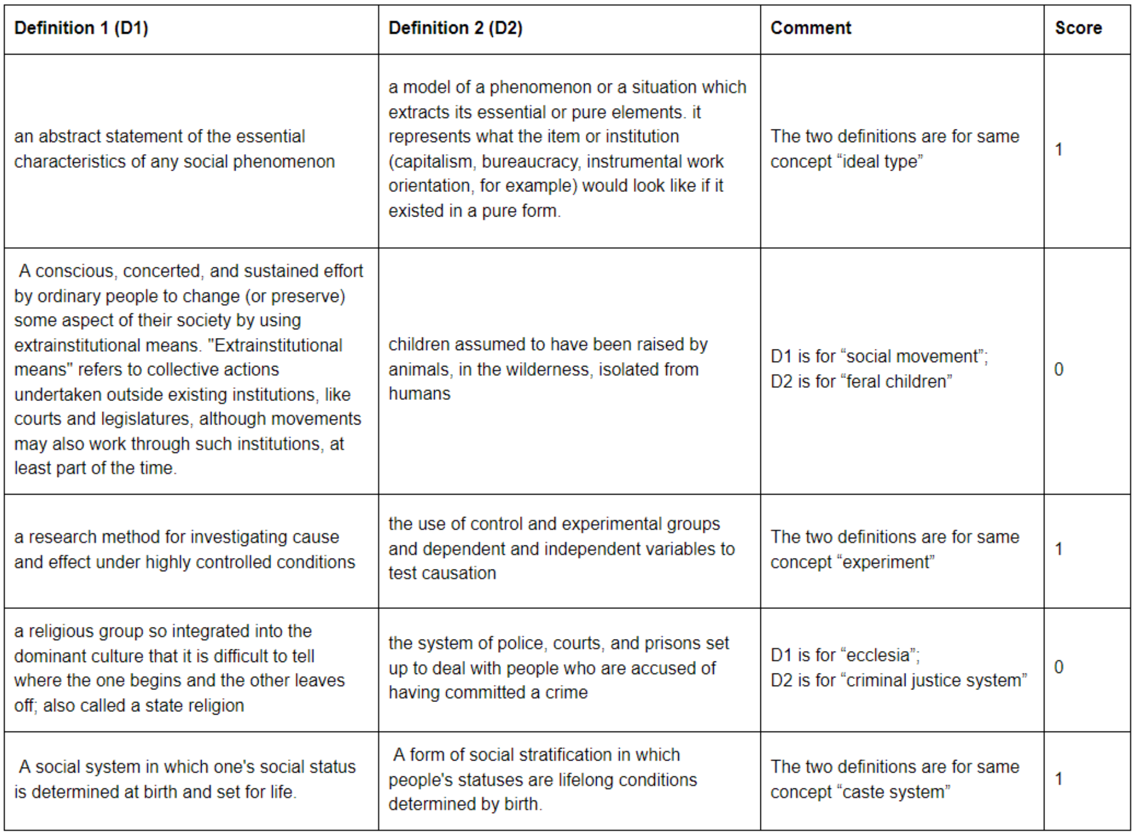
- We present a novel semantic CBRS which adapted specifically for our domain of interest.
- We compare the importance of the feature sets, analyze the impact of each feature set.
- Additionally, we present a manual annotation benchmark dataset for training and evaluation in future research in this area.

The experiment results showed that the proposed system achieves 86.16% accuracy, 84.42% F-measure, and 86% precision in suggesting semantically related sociological definitions. The proposed CBRS outperforms conventional matrix factorization-based CBRS by 18.75% in overall accuracy. In this study, we provide a solid baseline for future studies in the research area of sociological definition semantic similarity computation.

1.2 Dataset

As far as we know, there is neither a similar dataset we could use nor the published research on pairwise sociological definition semantic similarity computation. We, therefore, created a benchmark dataset for this purpose.

To build this corpus, we collected a corpus with over 4000 sociological definitions from Wikitheoria system lexicon, online resources, and the glossaries of a broad range of sociological books. These sociological definitions are used to generate pairwise comparisons of definitions offered for single terms, then evaluated by sociologists who judge the similarity with scores of 0 (different concept) and 1 (same concept). The annotated dataset includes 2235 definition pairs, including 959 positive samples and 1276 negative samples, as presented in Figure 1.3.



Definition 1 (D1)	Definition 2 (D2)	Comment	Score
an abstract statement of the essential characteristics of any social phenomenon	a model of a phenomenon or a situation which extracts its essential or pure elements. it represents what the item or institution (capitalism, bureaucracy, instrumental work orientation, for example) would look like if it existed in a pure form.	The two definitions are for same concept “ideal type”	1
A conscious, concerted, and sustained effort by ordinary people to change (or preserve) some aspect of their society by using extrainstitutional means. "Extrainstitutional means" refers to collective actions undertaken outside existing institutions, like courts and legislatures, although movements may also work through such institutions, at least part of the time.	children assumed to have been raised by animals, in the wilderness, isolated from humans	D1 is for “social movement”; D2 is for “feral children”	0
a research method for investigating cause and effect under highly controlled conditions	the use of control and experimental groups and dependent and independent variables to test causation	The two definitions are for same concept “experiment”	1
a religious group so integrated into the dominant culture that it is difficult to tell where the one begins and the other leaves off; also called a state religion	the system of police, courts, and prisons set up to deal with people who are accused of having committed a crime	D1 is for “ecclesia”; D2 is for “criminal justice system”	0
A social system in which one's social status is determined at birth and set for life.	A form of social stratification in which people's statuses are lifelong conditions determined by birth.	The two definitions are for same concept “caste system”	1

Figure 1.3: An illustration of annotation with sociological definitions and scores to train the semantic classifiers for binary classification.

1.3 The Structure of the Thesis

The remainder of the thesis is organized as follows. In Chapter 2, 3, 4, we overview the related work and relevant knowledge, detail the research method, and discuss their associated experimental results, respectively. The conclusion and future research are discussed in Chapter 5.

Chapter 2

Modularized Theory Construction and Parsimony Analysis with Wikitheoria

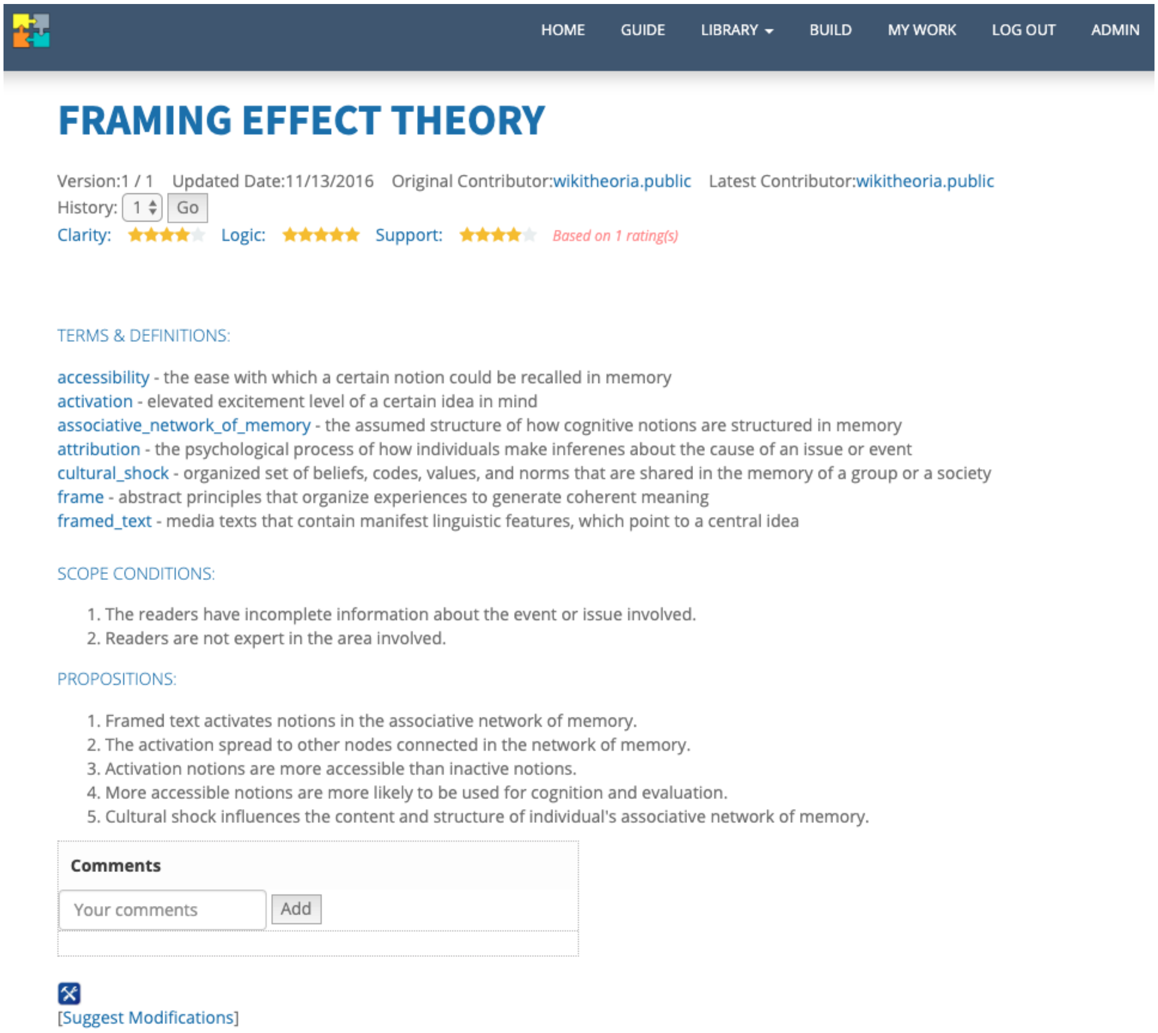
2.1 Related Works

2.1.1 Formal Theory

To formalize a theory means to express its statements using a formal language and explicit logic. Various branches of mathematics, as well as computer simulation programming, have provided language elements and logical frameworks for some of our theories. However—and this is crucial—not all logics are mathematical, and so formal theories need not be mathematical. As illustrated in Figure 2.1, what distinguishes formal theories from informal theories is the clear identification of the following components:

- Basic and defined terms. Basic terms are the foundational expressions whose meanings are presumed to be already well-understood by an intended audience. Defined terms employ previously defined terms and/or basic terms as definitions. Together the basic and defined terms provide the entire terminological system of a theory.

- Propositions and derivations. These are the core assertions of the theory. Propositions also may be called assumptions, axioms, premises, postulates, etc. Derivations are statements that have been logically deduced from previously stated propositions. They are sometimes called deductions, theorems, or con-



FRAMING EFFECT THEORY

Version:1 / 1 Updated Date:11/13/2016 Original Contributor:wikitheoria.public Latest Contributor:wikitheoria.public

History: 1 Go

Clarity: ★★★★☆ Logic: ★★★★★ Support: ★★★★☆ *Based on 1 rating(s)*

TERMS & DEFINITIONS:

accessibility - the ease with which a certain notion could be recalled in memory
activation - elevated excitement level of a certain idea in mind
associative_network_of_memory - the assumed structure of how cognitive notions are structured in memory
attribution - the psychological process of how individuals make inferenes about the cause of an issue or event
cultural_shock - organized set of beliefs, codes, values, and norms that are shared in the memory of a group or a society
frame - abstract principles that organize experiences to generate coherent meaning
framed_text - media texts that contain manifest linguistic features, which point to a central idea

SCOPE CONDITIONS:

1. The readers have incomplete information about the event or issue involved.
2. Readers are not expert in the area involved.

PROPOSITIONS:

1. Framed text activates notions in the associative network of memory.
2. The activation spread to other nodes connected in the network of memory.
3. Activation notions are more accessible than inactive notions.
4. More accessible notions are more likely to be used for cognition and evaluation.
5. Cultural shock influences the content and structure of individual's associative network of memory.

Comments

Your comments Add

[Suggest Modifications]

Figure 2.1: An illustration of published formal theory on Wikitheoria platform.

clusions. In general, derivations may be thought of as the claims that a theorist seeks to explain via the propositions used to derive them. Implicit in these definitions is the employment of a logical calculus—a set of rules governing the transformation of propositions into derivations.

- Scope conditions. These are the general conditions that must be satisfied for the appropriate application of the theory. As with every other theoretical component, they are provisional and, ideally, relaxed as progress is made, and the theory broadens in its scope.

2.1.2 Theory Modularization and Integration

A theory uses clearly defined terms in propositions that are amenable to the application of logical calculi [42, 31, 23]. Terms can be from natural language, and logical forms can be simple as “If x, then y.” Various branches of mathematics, logic, and simulation programming [48] have successfully provided operators and frameworks for some of our theories.

The concept of modularization is critical to theory construction. Generally, a module is a self-contained assemblage of elements that behave as a unit within a larger system. Cornforth and Green [21] described the nature and benefits of modularization, from genetics to social networks to manufacturing. These ideas apply readily to theories [22, 26, 44, 43, 42]. Figure 2.2 is a schematic illustration of two simple theory modules, each with two propositions (e.g., “The greater the A, then the greater the B.”) and a logical derivation. The modules intersect at B. Logically conjoining the intersecting statements integrates Module 1 and Module 2, yielding A -> Y, a derived proposition unavailable from either module alone. The ability to facilitate integrations is central to Wikitheoria. Building a new and more specialized theory tying A to Y only would have increased the complexity of the knowledge base without actually contributing anything new.

Theory modularization offers a novel approach to guiding empirical applications and to solving complex real-world problems: A user is able to withdraw modules from the Wikitheoria library on an as-needed basis, integrate them for the purpose at hand, and thus build a customized applied theory. The transparency of the method makes evident when modules are candidates for integrating into more comprehensive explanations or more detailed road-maps for applications and interventions.

In fact, modules ultimately must demonstrate their utility through useful integrations. Having a searchable library with a large number of small modules, rather than a small number of large theories, presents more opportunities for integrations and ap-

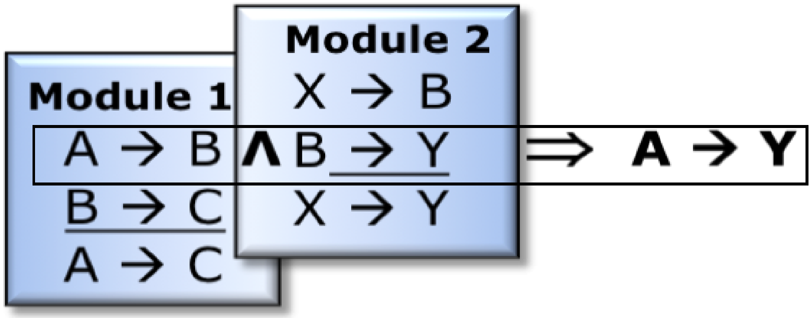


Figure 2.2: An schematic illustration of two simple theory modules with two propositions and a derivation.

plications. In the long-run, building modular theories should encourage a broadening of the range of potential integrations and a more nimble and fault-tolerant system.

2.1.3 Parsimony Analysis

When applied to specific empirical cases, a successful theory is one that describes relationships among phenomena, explains, and predicts the occurrence of certain events. Terms, statements, arguments, and scope conditions are the four fundamental components in any good scientific theories. Terms are used to build statements; statements are used to build arguments; arguments apply under a set of scope conditions. [43, 42, 1].

In a theory, terms are carefully chosen by the theorist to convey ideas or concepts, and their meanings are clearly implicated in the definitions[43, 42, 32]. Parsimony favors the use of relatively few definitions (terms), rather than creating new ones when the user goes to add the new definitions. Considering the following definitions for the term "denomination". D1: a church, independent of the state, that recognizes religious pluralism. D2: a brand name within a major religion; for example, Methodist or Baptist. If D1 were in theory already and the sociologist plans to add a new definition D2, he should determine whether D2 is in theory or is similar to D1. If either is the case, only D1 should be used instead of adding D2.

In recent years, many word embeddings and sentence embeddings have demonstrated the outstanding performances for language models on a broad spectrum of natural language understanding applications[16, 7, 6, 5, 4]. Especially, deep neural language models have demonstrated the efficacy by training with large corpora, such as Wikipedia, Google News, and 1 Billion Word Benchmark followed by fine-tuning dataset and achieved state-of-the-art results in semantic similarity related tasks[45, 14, 17]. Considering the excellent performance on representing the semantic similarities of textual snippets, the sociological definitions in our study are embedded with Transformer-based Universal Sentence Encoder in [17].

2.2 Method

In this section, we specify the details of cloud-based modularized theory construction, and we outline the details of the methodology used to perform our parsimony analysis. The system architecture of our proposed system is illustrated in Figure 2.3. To promote the theoretically-driven research, Wikitheoria was built with various subsystems such as modularized theory construction system, user management system, email system, peer-review system, ratings and incentive system, etc. In the following subsections, our focuses are on the system user interface, modularized theory construction, and parsimony analysis.

2.2.1 User Interface Development

Much attention was paid to making the system simple, engaging, functional, and expandable. Once a new user is registered and logged in, the system will lead the user to the guide page, as shown in Figure 2.4, which provides a comprehensive guide to system introduction, glossary, perspective, and tutorials, etc.

For the experienced user, the banner spans the top of the home page and provides some of the most frequent functions. LIBRARY links to the published theory modules

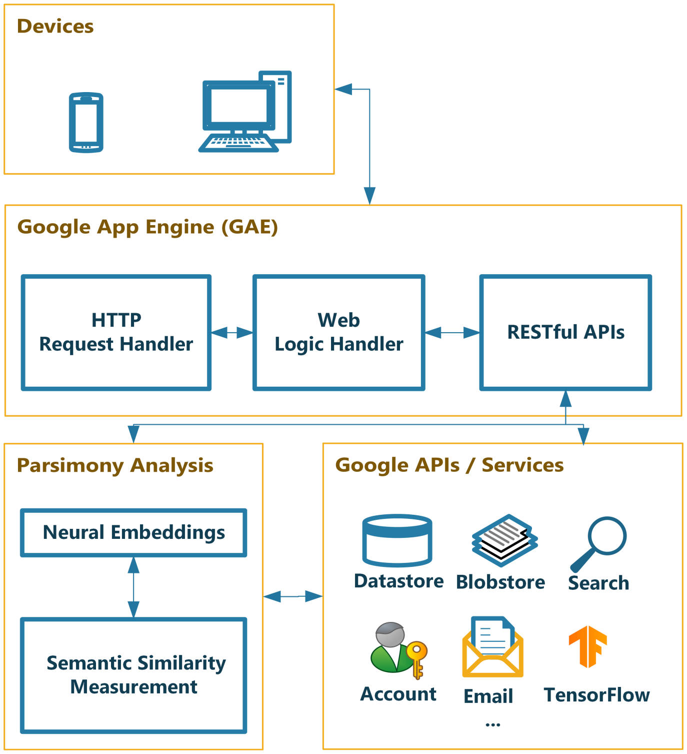
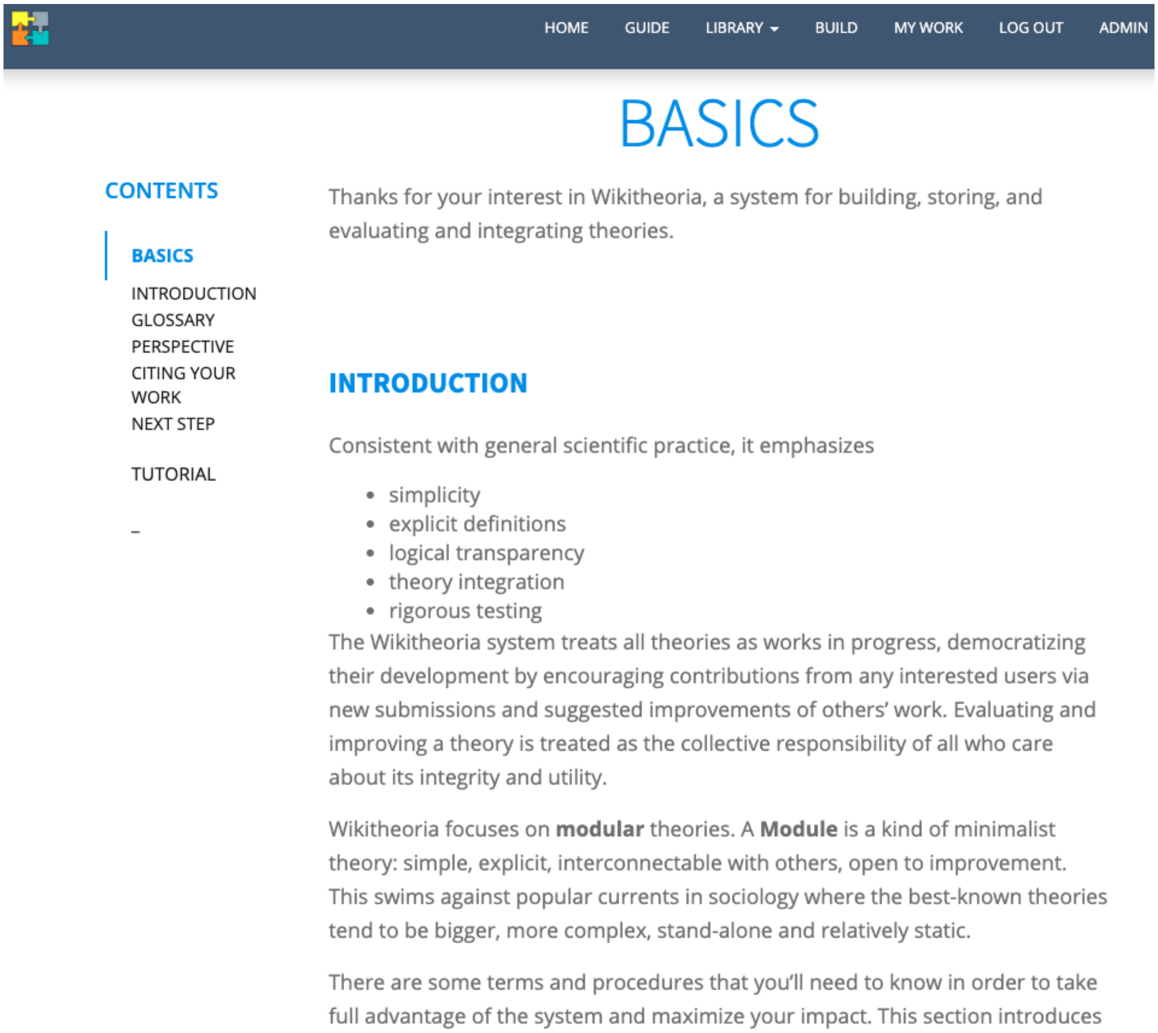


Figure 2.3: An illustration of Wikitheoria infrastructure with devices and Google cloud.

and system lexicon (shown in Figure 2.5). For each of the theory module, the user could browse the details of the theory, rate the quality of the theory, add comments, and make suggestions to improve the quality of a published theory further. Within the library, the system supports both word to word in-table quick search and the Google snippet style full-text search on theory title, meta-theory, terms, definitions, propositions, scopes, and derivations, etc. The full-text search is very similar to the Elastic Search, and the implementations of search function are based on the Google Search API, which maintains a reverse index on the specific columns of the text fields.

MY WORK lets users manage and edit previously saved work. BUILD takes users to the system for submitting or editing a new module (shown in Figure 2.6). At the bottom, CONTACT US and ADMINISTRATORS links provide, respectively, a brief sharable description of Wikitheoria, information about the research team and credit to NSF for support, and a portable to the administrators' interface, described shortly.



BASICS

CONTENTS

BASICS
INTRODUCTION
GLOSSARY
PERSPECTIVE
CITING YOUR WORK
NEXT STEP

TUTORIAL

–

Thanks for your interest in Wikitheoria, a system for building, storing, and evaluating and integrating theories.

INTRODUCTION

Consistent with general scientific practice, it emphasizes

- simplicity
- explicit definitions
- logical transparency
- theory integration
- rigorous testing

The Wikitheoria system treats all theories as works in progress, democratizing their development by encouraging contributions from any interested users via new submissions and suggested improvements of others' work. Evaluating and improving a theory is treated as the collective responsibility of all who care about its integrity and utility.

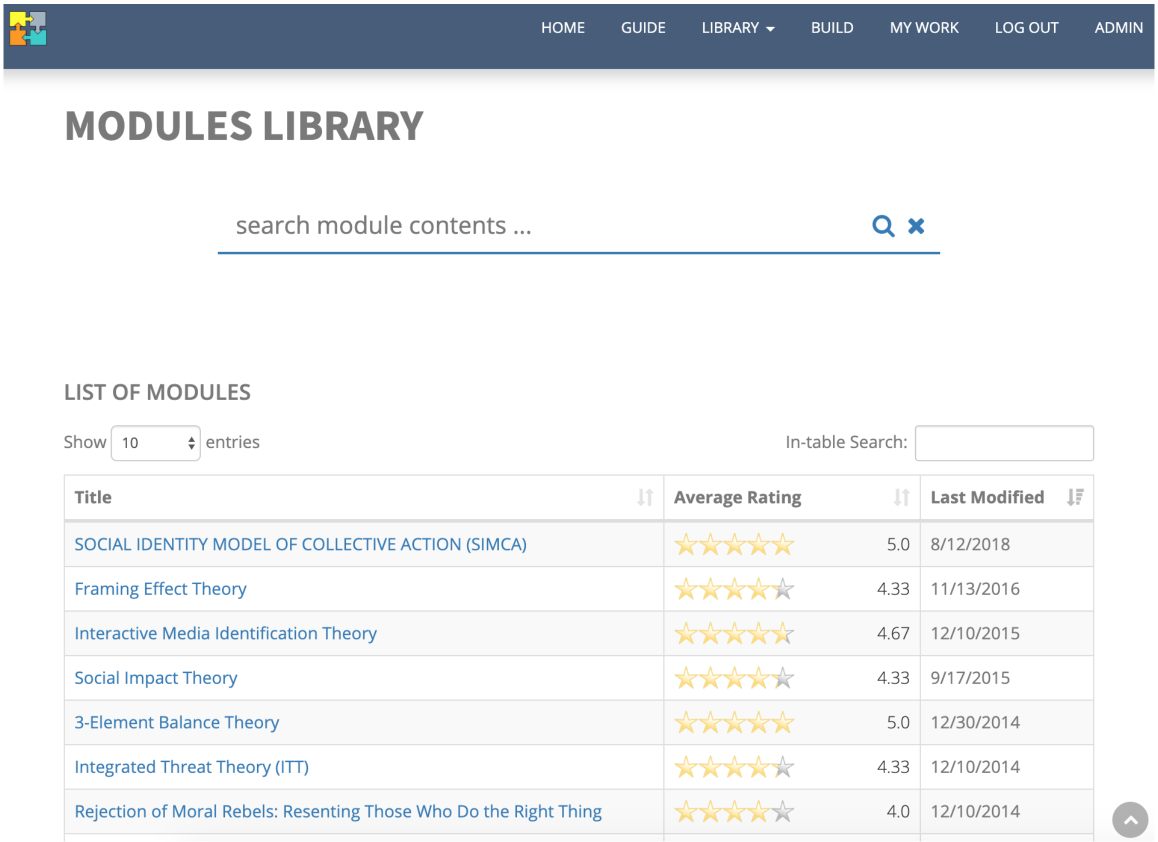
Wikitheoria focuses on **modular** theories. A **Module** is a kind of minimalist theory: simple, explicit, interconnectable with others, open to improvement. This swims against popular currents in sociology where the best-known theories tend to be bigger, more complex, stand-alone and relatively static.

There are some terms and procedures that you'll need to know in order to take full advantage of the system and maximize your impact. This section introduces

Figure 2.4: An illustration of Wikitheoria guide page for the newly registered users.

ADMINISTRATORS link is inaccessible to regular users, and a significant effort of our design and programming work has been devoted to the design and development of administrative functions. Figure 2.7 displays the tasks of managing the different roles for different users. Across the top is a series of links, each of which opens a unique page. Although masking a great deal of what we have completed, for the sake of brevity, we provide only brief descriptions of these functions.

As a critical part of the peer review system, the administration system manages email communication between administrators and users, generating various automated alerts and emailing contents of review forms back to appropriate authors. As



MODULES LIBRARY

LIST OF MODULES

Show 10 entries In-table Search:

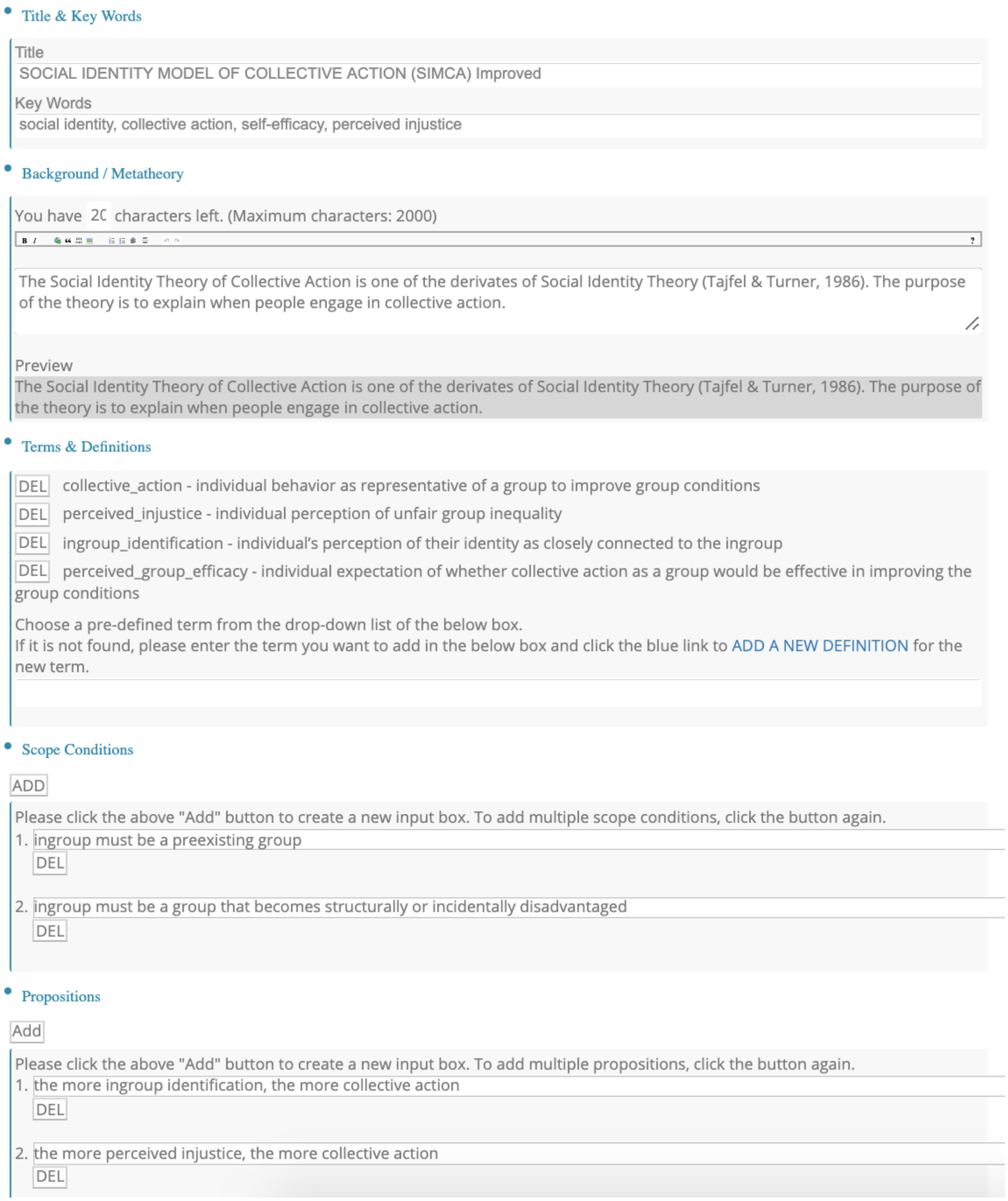
Title	Average Rating	Last Modified
SOCIAL IDENTITY MODEL OF COLLECTIVE ACTION (SIMCA)	5.0	8/12/2018
Framing Effect Theory	4.33	11/13/2016
Interactive Media Identification Theory	4.67	12/10/2015
Social Impact Theory	4.33	9/17/2015
3-Element Balance Theory	5.0	12/30/2014
Integrated Threat Theory (ITT)	4.33	12/10/2014
Rejection of Moral Rebels: Resenting Those Who Do the Right Thing	4.0	12/10/2014

Figure 2.5: An illustration of Wikitheoria library page for experienced users to quickly access the published theory modules.

illustrated in Figure 2.8 Users are automatically sent signals if they are editing a previously accepted module that is currently being written by someone else. All edits must be submitted to and approved by an editor. It is crucial for a given user to know of the possibility that someone's updated version of the module may be approved before the user has submitted her recommended changes.

Proposed Contributions page, as shown in Figure 2.9, appears upon entering the administrator system. This page displays a table with an overview of submitted modules pending administrative decisions.

Manage Modules page and Manage Terms has a comprehensive listing of modules and terms. This page allows the administrator to view and operate any term in the lexicon and any theory module in the library (shown in Figure 2.10).



- Title & Key Words

Title
SOCIAL IDENTITY MODEL OF COLLECTIVE ACTION (SIMCA) Improved

Key Words
social identity, collective action, self-efficacy, perceived injustice

- Background / Metatheory

You have 20 characters left. (Maximum characters: 2000)

The Social Identity Theory of Collective Action is one of the derivates of Social Identity Theory (Tajfel & Turner, 1986). The purpose of the theory is to explain when people engage in collective action.

Preview
The Social Identity Theory of Collective Action is one of the derivates of Social Identity Theory (Tajfel & Turner, 1986). The purpose of the theory is to explain when people engage in collective action.

- Terms & Definitions

DEL collective_action - individual behavior as representative of a group to improve group conditions
DEL perceived_injustice - individual perception of unfair group inequality
DEL ingroup_identification - individual's perception of their identity as closely connected to the ingroup
DEL perceived_group_efficacy - individual expectation of whether collective action as a group would be effective in improving the group conditions

Choose a pre-defined term from the drop-down list of the below box.
If it is not found, please enter the term you want to add in the below box and click the blue link to ADD A NEW DEFINITION for the new term.

- Scope Conditions

ADD

Please click the above "Add" button to create a new input box. To add multiple scope conditions, click the button again.
1. ingroup must be a preexisting group
DEL
2. ingroup must be a group that becomes structurally or incidentally disadvantaged
DEL

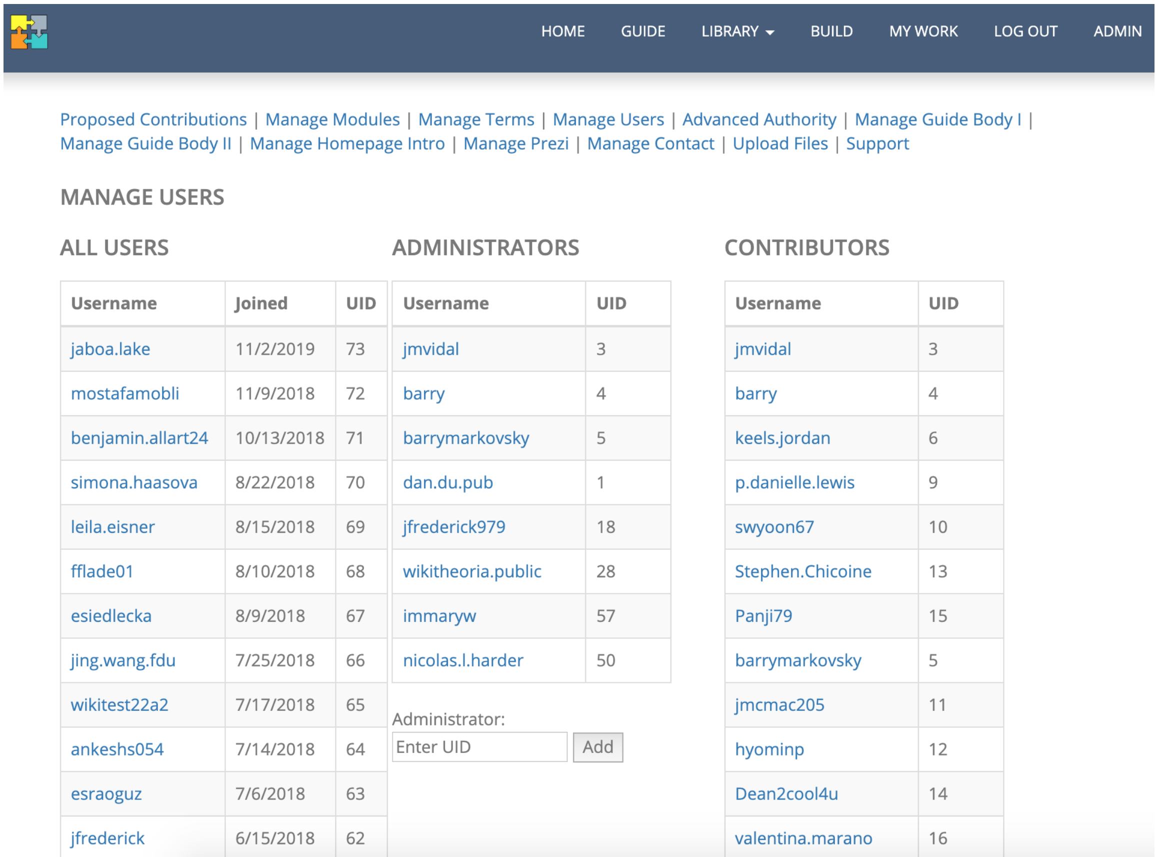
- Propositions

Add

Please click the above "Add" button to create a new input box. To add multiple propositions, click the button again.
1. the more ingroup identification, the more collective action
DEL
2. the more perceived injustice, the more collective action
DEL

Figure 2.6: An illustration of a theory that are under construction on Wikitheoria.

Upload Files allows administrators to upload and reference binary files to the server (shown in Figure 2.11). These files are used in conjunction with the links displayed on the Guide page, Tutorial page, and some area of the home page. The binary files are stored in the Google blob store, which is integral storage for the



Proposed Contributions | Manage Modules | Manage Terms | Manage Users | Advanced Authority | Manage Guide Body I | Manage Guide Body II | Manage Homepage Intro | Manage Prezi | Manage Contact | Upload Files | Support

MANAGE USERS

ALL USERS

Username	Joined	UID
jaboa.lake	11/2/2019	73
mostafamobli	11/9/2018	72
benjamin.allart24	10/13/2018	71
simona.haasova	8/22/2018	70
leila.eisner	8/15/2018	69
fflade01	8/10/2018	68
esiedlecka	8/9/2018	67
jing.wang.fdu	7/25/2018	66
wikitest22a2	7/17/2018	65
ankeshs054	7/14/2018	64
esraoguz	7/6/2018	63
jfrederick	6/15/2018	62

ADMINISTRATORS

Username	UID
jmvidal	3
barry	4
barrymarkovsky	5
dan.du.pub	1
jfrederick979	18
wikitheoria.public	28
immaryw	57
nicolas.l.harder	50

Administrator:

Enter UID

CONTRIBUTORS

Username	UID
jmvidal	3
barry	4
keels.jordan	6
p.danielle.lewis	9
swyoon67	10
Stephen.Chicoine	13
Panji79	15
barrymarkovsky	5
jmcmac205	11
hyominp	12
Dean2cool4u	14
valentina.marano	16

Figure 2.7: An illustration of how the administrators manage different roles of users with the system.

development with Google App Engine. All the binary files' related operations are controlled through this page.

2.2.2 Cloud-based Theory Modularization

Wikitheoria is a cloud-based application that could be accessed with smartphones and computers. With the help of Wikitheoria, a theory is modularized and constructed with several essential components. For example, in Figure 2.12, the theory "Social Identity Model of Collective Action (SIMCA)" is constructed using theory title, keywords, metatheory, terms and definitions, propositions, derivations, scope conditions, and evidence. As illustrated in Figure 2.3, the HTTP requests sent from devices are accepted and processed by the application. The various services such as

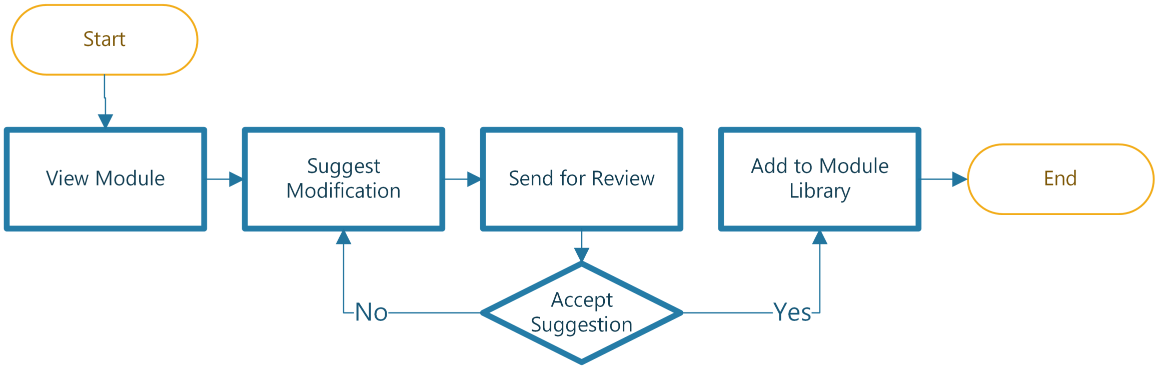
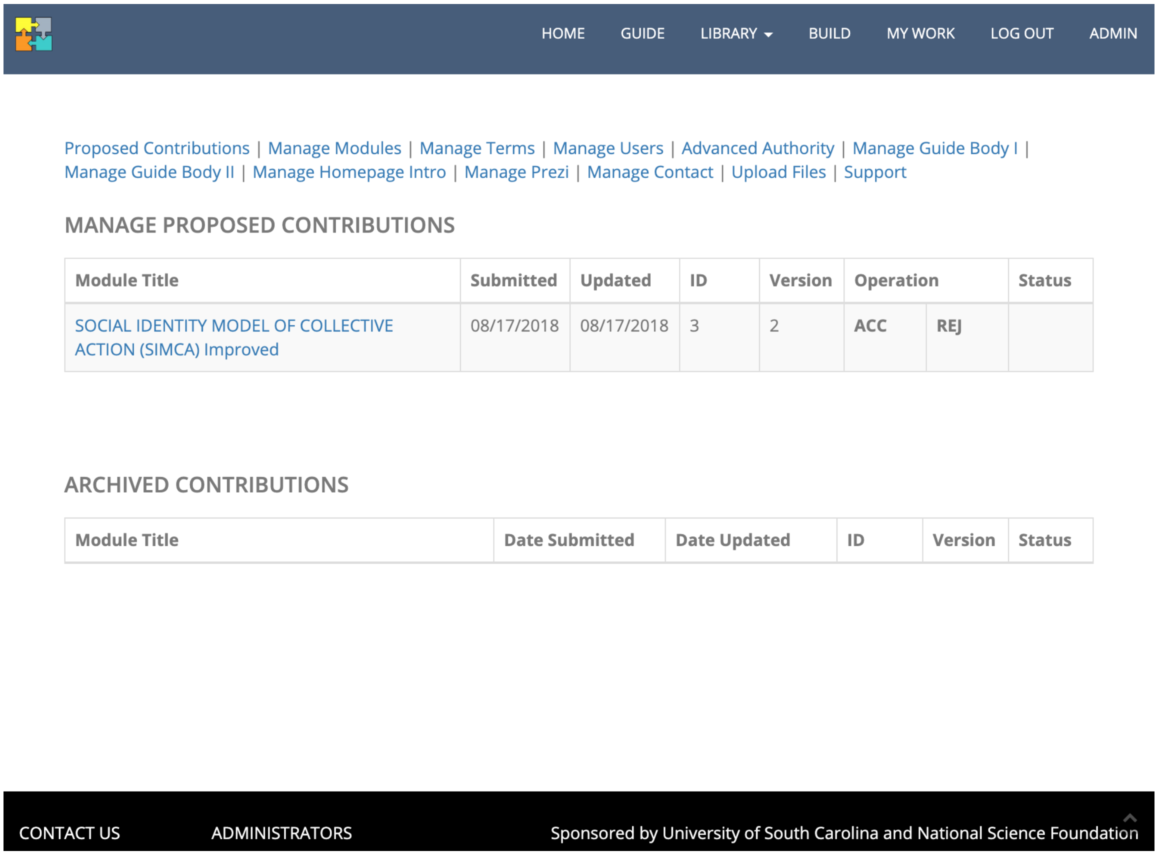


Figure 2.8: An illustration of peer review process for suggesting modification on the published theory module.

parsimony analysis, email, account, datastore, and blob store, etc. communicate with system backend through RESTful APIs. We implement the proposed framework with Python, Jinja2 framework, and web-related technologies such as HTML, JavaScript, and jQuery libraries[53] to handle the web related logics. For the parsimony analysis, we encode the sentence definitions with Tensorflow Hub and Keras[2, 33], then train the classifier with Scikit-learn[50].

The proposed application utilizes the Google App Engine (GAE), a Software-as-a-Service (SaaS) platform[53], which offers significant advantages for this application. (1) Google handles security, bandwidth, server space, certain administrative functions, and more. (2) The scale is a non-issue, as the App Engine would transparently duplicate our web application across multiple servers if usage ever exceeds capacity. (3) The App Engine provides access to the Object Relational Model (ORM) on top of Google's BigTable database implementation. The latter was designed for rapid location and fetching of documents, in contrast with relational databases optimized for complex queries, making it ideal for Wikitheoria. (4) We use Google accounts and other services, thus leveraging web functions with which most users are familiar already.


HOME
GUIDE
LIBRARY
BUILD
MY WORK
LOG OUT
ADMIN
Proposed Contributions | Manage Modules | Manage Terms | Manage Users | Advanced Authority | Manage Guide Body I | Manage Guide Body II | Manage Homepage Intro | Manage Prezi | Manage Contact | Upload Files | Support

MANAGE PROPOSED CONTRIBUTIONS

Module Title	Submitted	Updated	ID	Version	Operation		Status
SOCIAL IDENTITY MODEL OF COLLECTIVE ACTION (SIMCA) Improved	08/17/2018	08/17/2018	3	2	ACC	REJ	

ARCHIVED CONTRIBUTIONS

Module Title	Date Submitted	Date Updated	ID	Version	Status

CONTACT US
ADMINISTRATORS
Sponsored by University of South Carolina and National Science Foundation

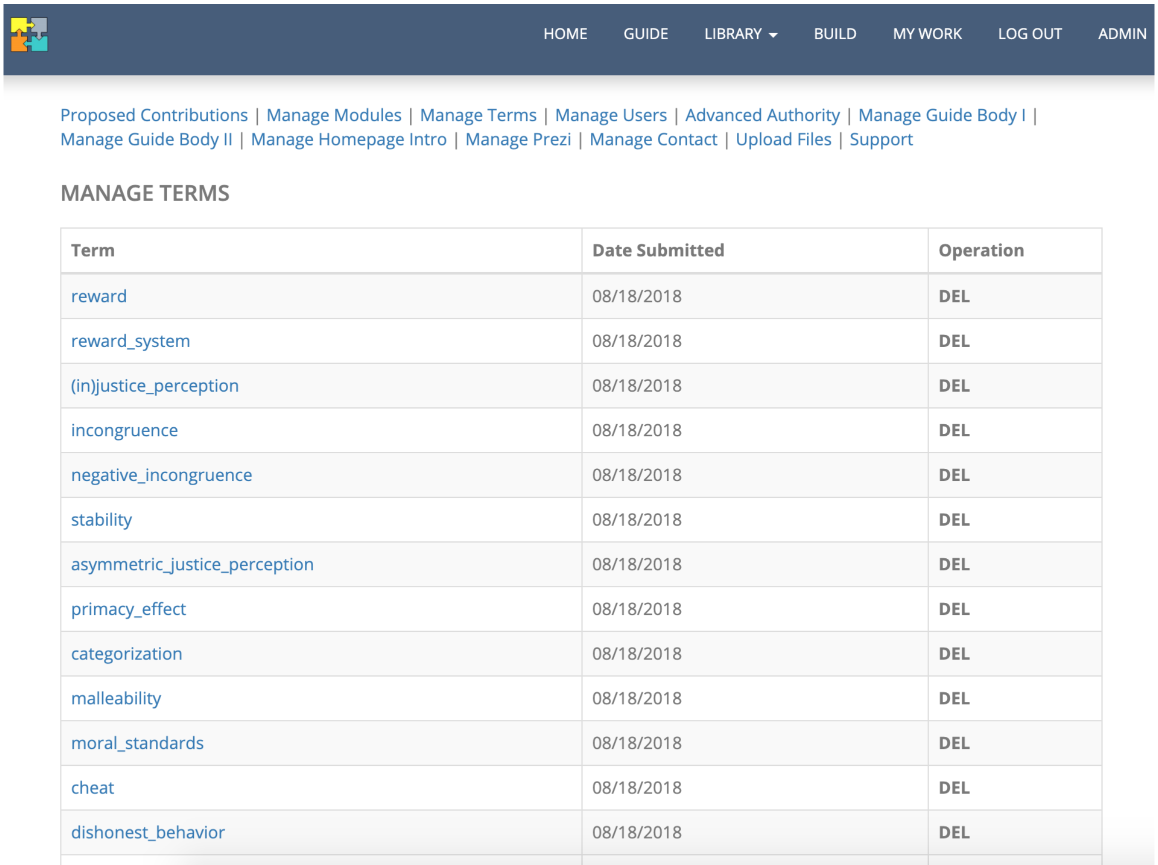
Figure 2.9: An illustration of how the administrators manage the proposed contributions for the peer review system.

2.2.3 Parsimony Analysis with Semantic Evaluation

Textual similarity metrics detect similarities between the two definition. To minimize the redundant sociological definition and optimize our model, we spent considerable effort on definition encoding and semantic similarity experiments[17, 30, 38, 46, 9, 59, 11].

Before applying the feature analysis, the pre-processing methods include definition tokenization, removing stop words, and converting all words to lowercase were applied to all the sociological definitions. The definition tokenization utilized the TreeBank tokenizer implemented in the NLTK toolkit[9].

The Universal Sentence Encoder mixed an unsupervised task using a large corpus together showed significant improvement by leveraging the attention-based Trans-


HOME
GUIDE
LIBRARY
BUILD
MY WORK
LOG OUT
ADMIN
Proposed Contributions | Manage Modules | Manage Terms | Manage Users | Advanced Authority | Manage Guide Body I | Manage Guide Body II | Manage Homepage Intro | Manage Prezi | Manage Contact | Upload Files | Support

MANAGE TERMS

Term	Date Submitted	Operation
reward	08/18/2018	DEL
reward_system	08/18/2018	DEL
(in)justice_perception	08/18/2018	DEL
incongruence	08/18/2018	DEL
negative_incongruence	08/18/2018	DEL
stability	08/18/2018	DEL
asymmetric_justice_perception	08/18/2018	DEL
primacy_effect	08/18/2018	DEL
categorization	08/18/2018	DEL
malleability	08/18/2018	DEL
moral_standards	08/18/2018	DEL
cheat	08/18/2018	DEL
dishonest_behavior	08/18/2018	DEL

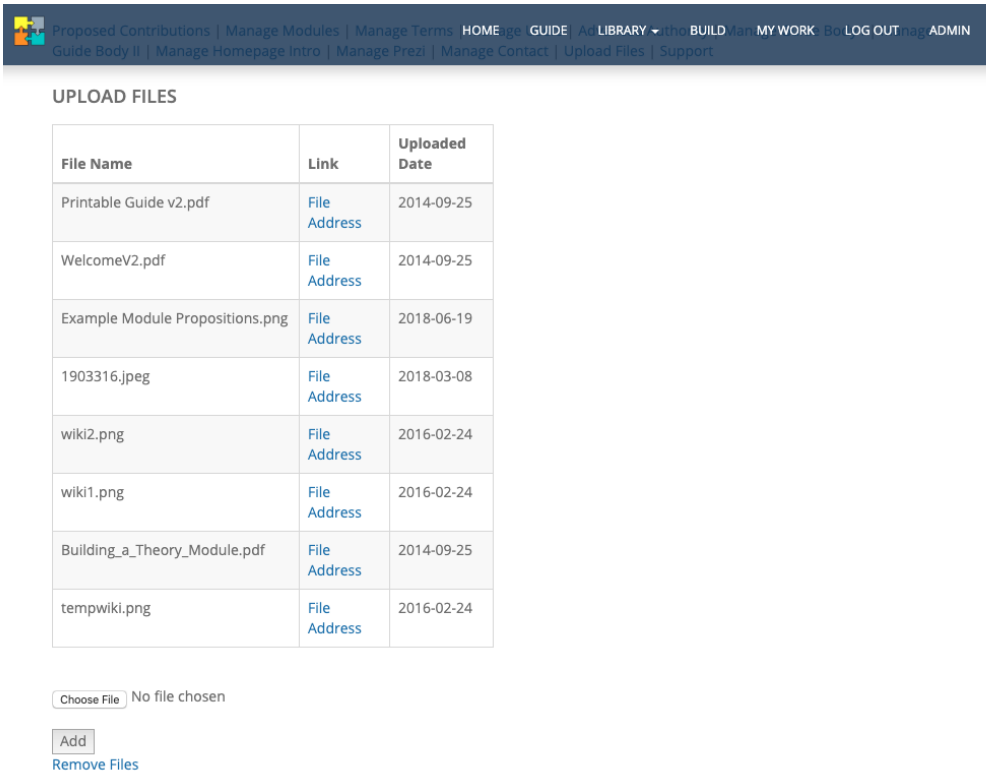
Figure 2.10: An illustration of how the administrators manage the lexicon and terms that associate with the published theory modules.

former architecture[17]. In our experiment, each definition was transformed into a 512-dimensional sentence vector. With the Transformer encoded embedding output, we computed the distance of two definition vector (u and v) with the kernel function show below and the KNN algorithm.

Cosine Distance[19] between two vectors u and v is defined as

$$sim = 1 - \frac{u \cdot v}{\|u\|_2 \, \|v\|_2} \tag{2.1}$$

To find the potential redundant definition, we employed an approach with KNN model [8]. The model was implemented with Scikit-learn[50], which computes the cosine distance from every definition in the lexicon, keeping track of the "most similar definition so far". It has a running time of O(dN) where N is the cardinality of S, and d is the dimensionality of u and v, where d equals to 512.


Proposed Contributions | Manage Modules | Manage Terms |
Guide Body II | Manage Homepage Intro | Manage Prezi | Manage Contact | Upload Files | Support
HOME
GUIDE
LIBRARY
BUILD
MY WORK
LOG OUT
ADMIN

UPLOAD FILES

File Name	Link	Uploaded Date
Printable Guide v2.pdf	File Address	2014-09-25
WelcomeV2.pdf	File Address	2014-09-25
Example Module Propositions.png	File Address	2018-06-19
1903316.jpeg	File Address	2018-03-08
wiki2.png	File Address	2016-02-24
wiki1.png	File Address	2016-02-24
Building_a_Theory_Module.pdf	File Address	2014-09-25
tempwiki.png	File Address	2016-02-24

Figure 2.11: An illustration of how the administrators manage the binary files in blobstore.

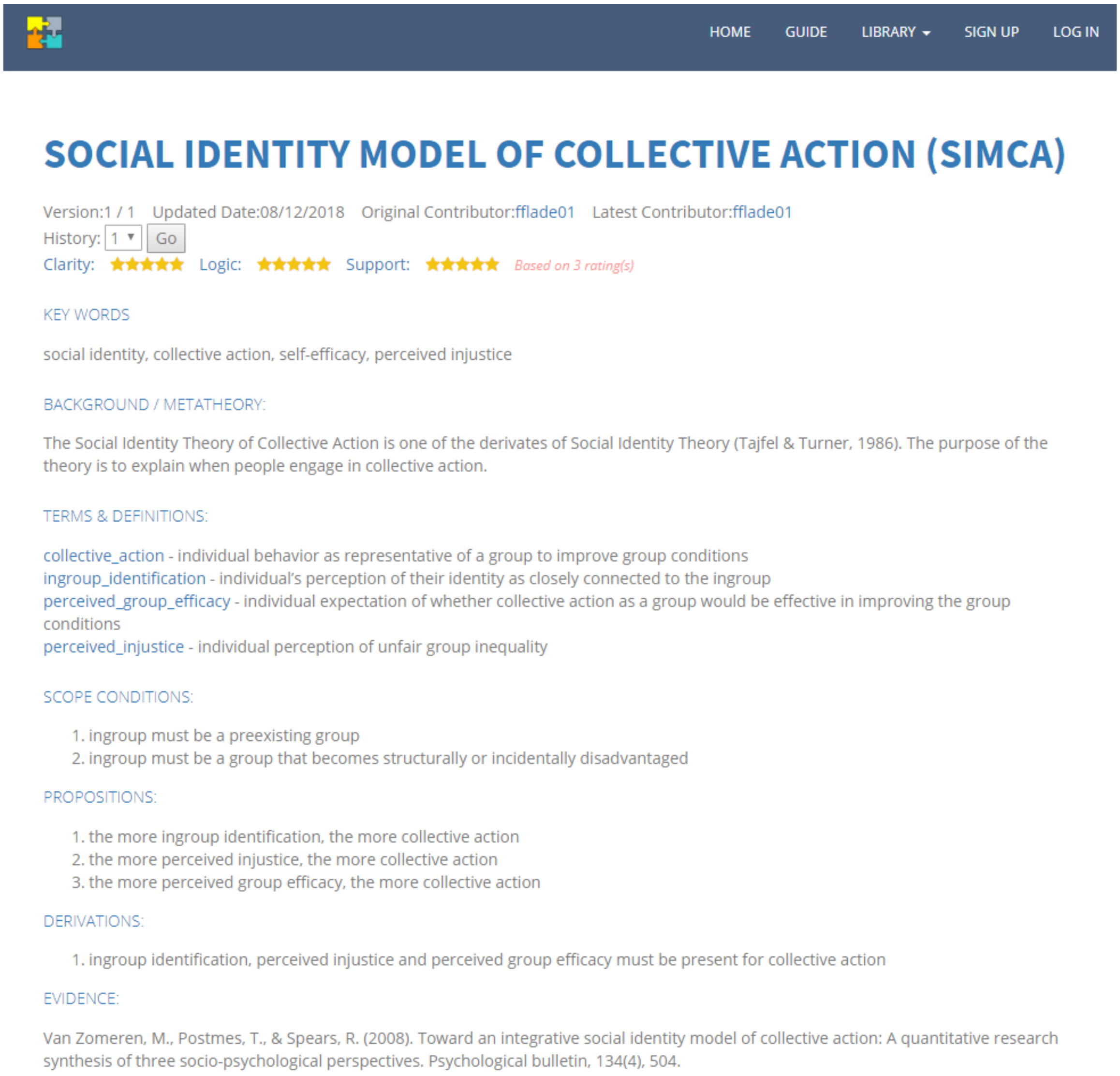
The quality and correctness of the proposed method is evaluated as 1) True positive (TP), the number of correct predictions on "same concept"; 2) True negative (TN), the number of correct predictions on "different concept"; 3) False positive (FP), the number of incorrect predictions on "same concept"; 4) False negative (FN), the number of incorrect predictions on "different concept". The precision(2), recall(3), and accuracy (5) were used to evaluate the semantic similarity measurement.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.5}$$



SOCIAL IDENTITY MODEL OF COLLECTIVE ACTION (SIMCA)

Version:1 / 1 Updated Date:08/12/2018 Original Contributor:fflade01 Latest Contributor:fflade01
History: 1 ▾ Go
Clarity: ★★★★★ Logic: ★★★★★ Support: ★★★★★ *Based on 3 rating(s)*

KEY WORDS

social identity, collective action, self-efficacy, perceived injustice

BACKGROUND / METATHEORY:

The Social Identity Theory of Collective Action is one of the derivates of Social Identity Theory (Tajfel & Turner, 1986). The purpose of the theory is to explain when people engage in collective action.

TERMS & DEFINITIONS:

collective_action - individual behavior as representative of a group to improve group conditions
ingroup_identification - individual's perception of their identity as closely connected to the ingroup
perceived_group_efficacy - individual expectation of whether collective action as a group would be effective in improving the group conditions
perceived_injustice - individual perception of unfair group inequality

SCOPE CONDITIONS:

1. ingroup must be a preexisting group
2. ingroup must be a group that becomes structurally or incidentally disadvantaged

PROPOSITIONS:

1. the more ingroup identification, the more collective action
2. the more perceived injustice, the more collective action
3. the more perceived group efficacy, the more collective action

DERIVATIONS:

1. ingroup identification, perceived injustice and perceived group efficacy must be present for collective action

EVIDENCE:

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. Psychological bulletin, 134(4), 504.

Figure 2.12: An illustration of published modularizrd theory module on Wikitheoria platform.

2.3 Results and Discussion

This study seeks to identify semantically similar sociological definitions with binary classification. To evaluate the performance of the proposed approach, we extracted over 4000 sociological definitions from system lexicon and sociological books. These definitions were paired according to their semantic similarities, then evaluated by sociologists with two categories: 1 (same concept) and 0 (different concept). The evaluation of the proposed approach was performed using 10-fold cross-validation[37, 57] over 2235 definition pairs, including 959 positive samples, and 1276 negative samples. The final result was calculated by averaging the results of each fold.

Table 2.1 presents the performance of KNN on the sociological definition data when different values of k (number of neighbors) are considered. It can be found that the value of k doesn't significantly affect the classifier's precision, recall, and accuracy. KNN model achieves the best performance with k = 5.

Comparing with both categories, the prediction performance on "same concept" is more important since it indicates whether the semantically similar sociological definitions can be detected. In Table 2.2, it shows the precision, which indicates the ratio of the number of correct predictions on "same concept" in the total number of correct and wrong predictions on "same concept" is 78%. With the recall on "same concept", 82% of the "same concept" definitions are detected from all the "same concept" definitions in the dataset. Considering the overall performance in both categories, the best prediction accuracy is 81.69%.

As shown in Table 2.1 and Table 2.2, the performance of Transformer embedded sociological definitions with the KNN model is an effective model for evaluating the semantic similarities of sociological definitions. The experiment results indicate that the proposed cloud-based theory modularization and embedding-based semantic analysis obtained a strong performance on recall, precision, and accuracy for promoting the parsimonious theory construction.

2.4 Conclusion

In this study, we proposed and implemented a generic knowledge aggregation framework to facilitate the parsimonious approach of theory construction with a cloud-based theory modularization platform and semantic-based algorithms to minimize the number of definitions. The presented approach is demonstrated and evaluated using the modularized theories from the database and sociological definitions retrieved

Table 2.1: The experiment results of k-nearest neighbors with different values of k.

No. of Neighbors	Precision	Recall	F-measure	Accuracy
k = 2	0.79	0.79	0.78	0.7857
k = 3	0.80	0.80	0.80	0.8035
k = 4	0.81	0.80	0.79	0.7991
k = 5	0.82	0.82	0.82	0.8169
k = 6	0.82	0.82	0.82	0.8169
k = 7	0.82	0.82	0.82	0.8169

Table 2.2: The results of k-nearest neighbors by category when k is equal to 5.

	Precision	Recall	F-measure
0 (different concept)	0.85	0.81	0.83
1 (same concept)	0.78	0.82	0.80
Average	0.82	0.82	0.82
Overall Accuracy	0.8169		

from the system lexicon and sociological literature to reduce the semantic redundancy.

To the best of our knowledge, our work is the first framework, which systematically applies cloud-based modularized theory construction with semantic-based parsimony analysis by using neural embedding and machine learning model. Our results demonstrated the effectiveness of using the cloud-based knowledge aggregation system and semantic analysis models for promoting the parsimonious sociology theory construction.

The proposed approach achieves the precision of 82%, recall of 82%, and accuracy of 81.69%. The proposed platform is fully implemented and publicly accessible via (https://www.wikitheoria.com). The results of this study can be further applied to the theory construction in human science-related disciplines such as psychology, criminology, and other social sciences.

Chapter 3

Towards Parsimonious Theory Construction with Neural Embeddings and Semantic Measurement

3.1 Related Works

Much effort has been expended in the development of text similarity data sets and related models. For example, the SemEval Semantic Text Similarity (STS) challenge has been organized for more than six years [7, 6, 5, 16]. These challenges have greatly accelerated the study of semantic texts. The manual annotation data set given by the STS enables various methods for semantic similarity estimation to be improved and checked[40, 56]. Many supervised learning models have proven useful for semantic evaluation. WordNet-based sentence distance calculations and distributed word embedding representations have proven to perform well in the general field of comparable semantic text similarity calculations. But in our research, many sociological terms are used and defined differently from generic English[4, 45, 14].

In recent years, text representations through text embedding and sentence embedding have shown excellent performance in capturing the semantic meanings of various tasks, and thus can be used to address the limitations of bag-of-words representation[46, 51, 52]. By using the neural network, the word embedding and sentence embedding produce high-quality word and sentence dense vectors in a wide range of natural language understanding applications. In particular, the deep neural language model demonstrates its effectiveness by using pre-trained language models, then fine-tuning on a small dataset and implementing state-of-the-art results in semantic similarity related tasks.

Words with similar meanings are close in the vector space when using neural embeddings to represent vocabulary words. Embedded vectors can well capture the similarity, especially the semantic similarity – texts that use different words but have similar semantic meanings to human. In 2013, Mikolov et al. demonstrated the effectiveness of neural embeddings for semantic similarity [46]. For word similarities, the use of Continuous Bag of Words and Skip-gram with 1.6 billion words corpus respectively achieved an accuracy of 63.7% and 65.6%.

In 2018, two sentence-level language models based on the Transformer and DAN demonstrated good performance as the general sentence encoders on the related semantic tasks [18]. Word-level and sentence-level embeddings, which pre-trained on large corpora such as news and Wikipedia, can be fine-tuned and used to encode the sociological definitions. More sophisticated pre-trained models such as sentence-level deep neural network encodings have also shown good performance on various NLP tasks[51, 52, 10, 17, 20, 25, 27].

3.2 Method

In this section, we specify the details of our data and provide the details of the methodology used to perform our analysis. The general architecture of our proposed training and prediction process is illustrated in Figure 3.1.

3.2.1 Encoding with Word-level Embeddings

Word embeddings are extensively used in state-of-the-art NLP techniques, mainly due to their ability to capture semantic and syntactic information[46]. In our experiments,

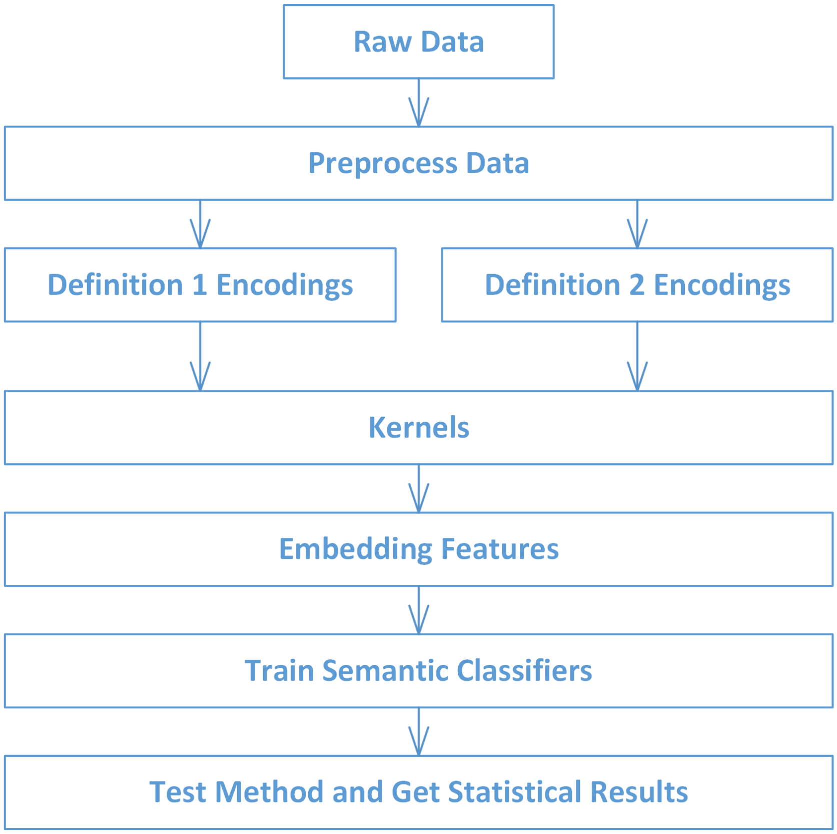


Figure 3.1: An illustration of semantic-based parsimony analysis workflow with encoding and training process.

Table 3.1: The word-level embedding methods and their corresponding training corpora.

Embedding	Corpus	Pooling	Size
Word2Vec	Google News 2013	Average	300
GloVe	English Wikipedia Feb 2017	Average	300
ELMo	1 Billion Word Benchmark	Mean	1024
fastText	English Wikipedia Feb 2017	Average	300

we encode definitions with four word-level embeddings. Their pre-trained corpora, pooling methods and embedding sizes are listed in Table 3.1.

Word2Vec: Word2Vec[46] is one of the first and the most popular approaches of word embeddings based on neural networks. It can preserve semantic relationships between words and their context, where context is modeled by nearby words using a shallow feed-forward neural network. In our experiment, a traditional average pooling was employed to produce the sentence embeddings with the obtained model [1].

Global Vectors (GloVe): GloVe[51] aims to overcome some limitations of Word2Vec, focusing on the global context for learning the representations. The global context is captured by the statistics of word co-occurrences in a corpus (count-based, as opposed to the prediction-based method as in Word2Vec), while still capturing semantic and syntactic meaning as in Word2Vec. In our experiment, an average pooling method was employed to produce the sentence embeddings with the pre-trained GloVe[2] model.

fastText: fastText[52] is a recent method for learning word embeddings for large datasets. It is an extension of Word2Vec that treats each word as a composition of character n-grams. The sub-word representation allows fastText to represent words more efficiently, enabling the estimation of rare and out-of-vocabulary words. In [16], the authors used fastText word representation combined with techniques such as the bag of n-gram features. They demonstrated that fastText obtained performance on par with deep learning methods while being faster. In our experiment, the pre-trained fastText[3] model was applied with the average pooling method.

Embedding from Language Models (ELMo): The challenges exist when learning high-quality representations from Word2Vec, Glove, and fastText: they should capture semantic and syntax and the different meanings the word can represent in different contexts. For example, a bowl (a round food container) and bowl (a stadium) should have different word vectors. To solve this problem, ELMo[10] uses representations from a bi-directional LSTM that is trained with a language model objective on a large text dataset. In ELMo[10], they use a deep representation by incorporating internal representations of the LSTM network, therefore capturing the meaning and

[1] https://code.google.com/archive/p/word2vec/

[2] https://nlp.stanford.edu/projects/glove/

[3] https://fasttext.cc/docs/en/english-vectors.html

syntactical aspects of words. In our experiment, the ELMo[4] model was pre-trained with 1 billion word benchmark corpus. A built-in mean-pooling method was applied to produce the sentence embeddings.

3.2.2 Encoding with Sentence-level Embeddings

Although the traditional bag-of-words model showed excellent performance for some tasks, it is still unclear how to accurately represent the full sentence meaning. Nowadays, there is still no consensus or studies on how to represent sociological towards the sociology theory construction. We experimented with four sentence-level embeddings with the listed pre-trained corpora and embedding sizes in Table 3.2.

InferSent: InferSent[20] proposes a supervised training for the sentence embeddings. The sentence encoders, which based on a bidirectional LSTM (BiLSTM), are trained using the Stanford Natural Language Inference (SNLI) dataset, which consists of 570k human-generated English sentence-pairs. The pre-trained InferSent[5] model was obtained and applied to definition encodings.

Universal Sentence Encoder - Transformer (USE-Transformer): While word embeddings such as Word2Vec or GloVe try to embed a single word into a high dimensional vector, USE[17] works to embed not only words but phrases, sentences, and short paragraphs. It takes variable-length English text as input and outputs a 512-dimensional vector by utilizing a stack of 6 identical layers, where each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise, fully connected feed-forward network. A residual connection around each of the two sub-layers, followed by layer normalization is employed. The Transformer[6] model was trained with various data from Google.

[4] https://tfhub.dev/google/elmo/2

[5] https://github.com/facebookresearch/InferSent

[6] http://www.tensorflow.org/hub/modules/google/universal-sentence-encoder-large/3

Table 3.2: The sociological definition with the sentence-level embedding methods and their corresponding training corpora.

Embedding	Corpus	Size
InferSent	Stanford Natural Language Inference	4096
USE-DAN	Various Data from Google	512
USE-Transformer	Various Data from Google	512
BERT	English Wikipedia Feb 2017	768

Universal Sentence Encoder - Deep Averaging Network (USE-DAN): USE-DAN[17] is based on a deep averaging network where input embeddings for words and bi-grams are first averaged together and then passed through a feedforward deep neural network to produce sentence embeddings. The main advantage of the DAN encoder over the Transformer is that compute time is linear in the length of the input sequence. Our USE-DAN[7] was trained and obtained with various data from Google.

BERT: BERT[25] starts by training a general-purpose "language modeling" (LM) on a large text corpus, and then use that model for various tasks. It applies the bidirectional training of Transformer, an attention model, to language modeling. When training language models, many models predict the next word in a sequence, a directional approach that inherently limits context learning. To overcome this challenge, BERT uses two training strategies: (1) Masked LM: before feeding word sequences into BERT, some of the words in each sequence are replaced with a mask token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. (2) Next Sentence Prediction: In the training process, the model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence. The BERT[8] model was trained with English Wikipedia Feb 2017 corpus.

[7] https://tfhub.dev/google/universal-sentence-encoder/2

[8] https://github.com/google-research/bert

3.2.3 Feature Extractions

Textual similarity metrics detect similarities between the two sociological definitions. These metrics are then used as features for our machine learning models. To optimize our model, we spent considerable effort on feature engineering, experimented with the effectiveness of embedding-based features, and obtained the best result with the 11 features described below. With the 8 embedding methods listed in Section 3.2 and Section 3.3, the pairwise sociological definitions were transformed into 8 representations of pairwise definition vectors. With the encoded embedding output, we computed the distance of two definition vectors (u and v) with the kernel functions shown below.

Cosine Distance[19] of two vectors u and v is defined as

$$sim_1 = 1 - \frac{u \cdot v}{\|u\|_2 \, \|v\|_2} \tag{3.1}$$

Manhattan Distance[19], also known as block distance, computes the distance between two vectors by summing the differences of their corresponding components in u and v, which is defined as

$$sim_2 = \sum_i |u_i - v_i| \tag{3.2}$$

Jaccard Distance[23] measures the dissimilarity between two vectors u and v. It is defined as

$$sim_3 = \frac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v} \tag{3.3}$$

Canberra Distance[38] between u and v is defined as

$$sim_4 = \sum_i \frac{|u_i - v_i|}{|u_i| + |v_i|} \tag{3.4}$$

Euclidean Distance[13] between u and v is defined as

$$sim_5 = \|u - v\|_2 \tag{3.5}$$

Minkowski Distance[19] between u and v is defined as

$$sim_6 = \|u - v\|_p = (\sum |u_i - v_i|^p))^{1-p} \tag{3.6}$$

Bray-Curtis Distance[12] between u and v is defined as

$$sim_7 = \sum |u_i - v_i| \, / \sum |u_i + v_i| \tag{3.7}$$

Skewness and Kurtosis[41] are used to measure the symmetry of definition vectors when comparing with the normal distribution. Skewness is a measure of the symmetry. If the distribution or dataset appears to be the same as the left and right sides of the center point, it is symmetrical. Kurtosis is a measure of tail thickness, i.e., distribution with high kurtosis often has a heavy tail.

3.2.4 Supervised Models

With the features we created in Section 3.4, our goal was to create a relevance model that would accurately predict if a new definition is semantically similar to an existing definition in theory. Here, we choose four representative supervised machine learning algorithms: K Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and XGBoost (XGB). For KNN, we set neighbors number to 5 because of the better performance on grid search. We included the Naive Bayes algorithm in our study as it does not require hyperparameter tuning and can be trained in linear time. For LR and XGB, we chose the default hyper-parameter settings in scikit-learn[50] and XGBoost library[18].

3.2.5 Evaluation

For this study, the model evaluation was performed using 10-fold cross-validation over 2235 sociological definition pairs, including 959 positive samples and 1276 negative samples [22]. Each fold contained 1811 definitions pairs for training, 200 pairs for validation, and 224 pairs for testing. The result for the supervised semantic similarity was calculated by averaging the results of each fold. The quality and correctness of the proposed method is evaluated as 1) True positive (TP), the number of correct predictions on “same concept”; 2) True negative (TN), the number of correct predictions on “different concept”; 3) False positive (FP), the number of wrong predictions on “same concept”; 4) False negative (FN), the number of wrong predictions on “different concept”. The precision (3.8), recall (3.9), F-measure (3.10) and accuracy (3.11) were used to evaluate the recommendation system.

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

$$F - measure = 2 \ast \frac{Precision \ast Recall}{Precision + Recall} \tag{3.10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.11}$$

3.3 Results and Discussion

Given the variety of available embedding methods, we aim to understand how the performance of semantic classifiers varies with different types of word embeddings and sentence embeddings. We analyzed the results of the experiments presented in Tables 3.3 and Table 3.4 in the following subsections.

3.3.1 Usefulness of Word-level Embeddings

By analyzing the results in Table 3.3, we can compare the performance of four types of word-level embeddings, when used with four different machine learning models. From Table 3.3, we can see that the GloVe embeddings outperform the other three word-level embedding methods. Specifically, training on the XGBoost model achieves the best result with an accuracy of 82.14%, a precision of 82%, a recall of 82%, and an F1 of 80.59%. The Word2Vec with the LR model also achieves a good result with

Table 3.3: The word-level sociological definition embedding results on test data with 10-fold cross validation.

Embedding	**Model**	**Precision**	**Recall**	**F1**	**Accuracy**
Word2Vec	KNN	0.77	0.77	0.7488	0.7723
	GNB	0.80	0.79	0.7416	0.7946
	LR	**0.82**	**0.82**	**0.7960**	**0.8170**
	XGB	0.81	0.81	0.7817	0.8080
GloVe	KNN	0.80	0.79	0.7767	0.7946
	GNB	0.78	0.77	0.6941	0.7679
	LR	0.82	0.82	0.8000	0.8170
	XGB	**0.82**	**0.82**	**0.8059**	**0.8214**
ELMo	KNN	0.63	0.63	0.5638	0.6339
	GNB	0.76	0.63	0.3025	0.6295
	LR	0.66	0.65	0.4903	0.6473
	XGB	**0.66**	**0.65**	**0.4903**	**0.6473**
fastText	KNN	0.75	0.75	0.7264	0.7545
	GNB	0.80	0.78	0.7135	0.7813
	LR	0.77	0.77	0.7437	0.7723
	XGB	**0.80**	**0.80**	**0.7716**	**0.7991**

a slightly lower accuracy of 81.7% and F1 of 79.60%. The best accuracy of ELMo is 64.73%. One possible explanation for this may be that the pre-trained “Google News” corpus and “Wikipedia” corpus are better representations for the knowledge of our domain.

Based on the best precision and recall given from Table 3.3, the ratio of correct predictions on “same concept” in the total number of correct and wrong predictions on “same concept” is 82%. With the recall on “same concept”, 82% of the ”same concept” definitions are detected from all the “same concept” definitions in the dataset. Considering the overall performance in both categories, the best prediction accuracy is 82.14%.

3.3.2 Usefulness of Sentence-level Embeddings

By comparing columns in Table 3.4, we can see that the USE-Transformer is the best sentence-level sociological definition embedding, which achieves the best accuracy of

Table 3.4: The sentence-level sociological definition embedding results on test data with 10-fold cross validation.

Embedding	**Model**	**Precision**	**Recall**	**F1**	**Accuracy**
InferSent	KNN	0.71	0.71	0.6596	0.7143
	GNB	0.75	0.74	0.6548	0.7411
	LR	0.73	0.73	0.6592	0.7277
	XGB	**0.76**	**0.75**	**0.6782**	**0.7500**
USE-DAN	KNN	0.78	0.78	0.7525	0.7768
	GNB	0.84	0.83	0.7912	0.8304
	LR	0.82	0.82	0.8019	0.8170
	XGB	**0.84**	**0.84**	**0.7978**	**0.8348**
USE-Transformer	KNN	0.83	0.83	0.8020	0.8259
	GNB	0.84	0.83	0.7978	0.8348
	LR	0.83	0.83	0.8173	0.8304
	XGB	**0.85**	**0.85**	**0.8317**	**0.8482**
BERT	KNN	0.54	0.54	0.4516	0.5446
	GNB	**0.55**	**0.56**	**0.4590**	**0.5580**
	LR	0.47	0.54	0.0917	0.5580
	XGB	0.52	0.54	0.2816	.5446

84.82%, precision of 85% and recall of 85%. In contrast, the USE-DAN achieves a slightly lower accuracy of 84.82%.

To evaluate word embeddings versus sentence encodings, we compared Tables 3.3 and 3.4, and observed that the best values for our task are generally obtained using sentence-level embeddings. This result is intuitive, as one would expect the sentence-level encodings to capture the semantic similarities of definitions better. Both Transformer and DAN archives good results with 84.82%, and 83.48% in prediction accuracy. The best accuracy of sentence-level embedding is 3.13% better than the word-level embedding. Transformer in Table 3.3, the ratio of correct predictions on "same concept" in the total number of correct and wrong predictions on "same concept" is 85%. With the recall on "same concept", 85% of the "same concept" definitions are detected from all the "same concept" definitions in the dataset. Considering the overall performance on both categories, the best prediction accuracy is 84.82%.

3.3.3 The Effect of Classifiers with Embeddings

From Table 3.3 and Table 3.4, we could see that the XGB model outperforms other three models in the six of eight embedding methods. The results of the LR are very close to the results of XGB on word-level embeddings. When using sentence encodings, both GNB and LR classifiers work well for the dataset. However, we believe that hyper-parameter tuning might improve the results of XGB, making this classifier more competitive when using sentence encodings. Comparing with other classifiers, KNN doesn't perform well with the word embedding and sentence encodings on the dataset. Given that we are using various sociological books to train the model, even though the word embeddings or sentence encodings are meant to evaluate the semantic similarity between sociological definitions, KNN is still sensitive to noise as it is making its classifications based on the nearest neighbors selected. If the nearest neighbors are noisy, the classification can be wrong. Thus, our study suggests that XGB or LR are good choices as traditional supervised semantic classifiers, but hyper-parameter tuning may be needed to achieve the best performance with XGB.

3.4 Conclusion

In this study, we propose a semantic embedding-based approach to check the semantic consistency of sociological definitions for sociology theory construction. Towards this goal, we performed an extrinsic evaluation, where eight embeddings were used with four supervised models to learn classifiers for semantic analysis. To the best of our knowledge, our work is the first to apply recent state-of-art pre-trained word-level embeddings and sentence-level embeddings models to measure the semantic similarities of sociological definitions, and the first one that evaluates four different types of classifiers on promoting the parsimony for sociology theory construction.

Among the eight types of embeddings, the Transformer pre-trained by Google performed the best on the data used in our study. In particular, the XGBoost classifier was shown to make the best use of the semantic embeddings. Our results demonstrated the effectiveness of using word-level and sentence-level embeddings with semantic analysis models for promoting the parsimonious sociology theory construction. The proposed approach achieves the precision of 85%, recall of 85%, and accuracy of 84.82%.

Chapter 4

Semantic Content-based Recommendation with SOREC

4.1 Related Works

4.1.1 Parsimonious Theory Construction

Across the sciences, theories are used to explain and predict observed phenomena in the natural world. Theory construction is the process of building theories to strict specifications with respect to the clarity of their terms and the logical integrity of their arguments[43, 42, 1]. In mature sciences, terms and their associated definitions are essential components and are carefully chosen to convey ideas or concepts in theory[43, 42, 32]. Our focus here, however, is social science theorizing where quite frequently key terms are not defined explicitly. As a result, to varying degrees, theoretical writing is overly verbose in such fields as political science, economics, sociology, management, anthropology, and others. This verbosity can make it difficult or impossible for readers to glean authors' intended meanings, leading them instead to infer their own meanings.

The purpose of the work we report here was to develop a method that facilitates more parsimonious theorizing in the social sciences. The parsimony principle—as exemplified by the familiar notion of Occam's Razor—can be expressed this way: Given a choice between competing assertions, then all else being equal, the simpler version is preferred. This applies equally well to choosing between competing theories of empirical phenomena, or between alternative definitions for a theoretical term[43, 42, 32]. Conventional social science theory construction is generally informal and abductive. That is, it seeks plausible accounts for empirical observations, but without applying rigorous standards to the semantics and logic through which the theoretical ideas are expressed. The result frequently lacks coherence, consistency, and logical integrity.

Statistical machine learning tools offer a more rigorous scientific approach to guide informal theorizing in the social sciences. In this paper, we propose a recommendation system based on a machine learning model that fosters parsimony in the development of terminological systems for informal theories. The targeted user is the social scientist wishing to develop a new formal theory, or a more rigorous version of an existing theory. This entails not only listing theoretical propositions but also attending to the terms used in those propositions and to the definitions of those terms. Our system intervenes in the process of formulating or selecting the defined terms used to express theoretical propositions[43, 42, 32].

Wikitheoria is a fully functional knowledge aggregation web framework hosted on the Google Cloud Platform[29]. It is built for modularized theory construction in the social sciences. SOREC is one of Wikitheoria's sub-systems constructed for the purpose of facilitating the construction of parsimonious theories[28]. Parsimony favors the use of relatively few definitions (terms), rather than creating new ones when the user goes to add the new definitions.

For example, consider the following two definitions for the term "ambivalence".

- D1: the presence in one person at the same time of two competing or conflicting emotions or attitudes
- D2: simultaneous conflicting feels toward a person or an object

If D1 were in the theory already and the user entered D2 as a new definition, the semantic recommendation system would determine whether D2 is in the theory or is similar to D1. If either is the case, the system recommends using D1. As such, the system helps safeguard against redundancy and fosters the more parsimonious theories.

4.1.2 Recommendation System

A recommender system is defined as “A system that has as its main task choosing certain objects that meet the requirements of users, where each of these objects are stored in a computer system and characterized by a set of attributes.”[35] It helps users to quickly discover the information they need in a specific context through information filtering. Most of these recommendations are implemented in three filtering methods: collaborative filtering, content-based filtering, and hybrid filtering[34, 58, 47, 49, 15, 24].

The collaborative filtering method learns from the users’ past activities and uses their common behavior patterns to make recommendations that the user may be interested in [34, 58]. Content-Based filtering focuses on the characteristics of the recommended item. For example, when searching for a similar pre-defined definition from the lexicon, the recommendation output is based on its syntactic and semantic relatedness[47, 49, 24, 15]. Hybrid filtering is a combination of CF and CBF[15]. According to CBF’s prior knowledge, the primary source of information used in content-based filtering systems is text fragments[47]. A set of encoding methods, typically TF-IDF are used to present the definitions[57]. However, in our study, a number of semantically equivalent sociological terminologies are used to construct definitions. The traditional IR methods work fine on the general domain but fail to capture the semantic similarity in the sociological domain. Therefore, natural language processing and machine learning-based models are currently used to analyze, classify, or measure the latent semantic similarity to support the CBF.

4.1.3 Semantic Analysis

In recent years, word embedding and sentence embedding techniques have gained substantial improvement in natural language understanding[16, 46]. Mikolov et al. have illustrated the effectiveness of neural word representations for similarities and other neural language processing algorithms. For word similarity measurements, 63.7% and 65.6% accuracy were achieved by using Continuous Bag of Words and Skip-gram, respectively, with the corpus of 1.6 billion words[46]. In 2018, Cer et al. demonstrated the excellent performance of Transformer embedding on semantic similarity tasks[16].

Many efforts have been made to develop semantic textual similarity datasets and related models[16, 40, 56]. For example, the SemEval Semantic Textual Similarity (STS) challenges have been organized for over six years. These challenges greatly accelerated semantic textual research. Manually annotated datasets given by STS empowered the improvement and examination of various methods for semantic similarity estimation. Many supervised learning models were shown to be well performed for semantic recommendation[7, 6, 5]. Diverse features such as WordNet-based sentence distance calculation and distributed word embedding representations were shown to perform well for comparable semantic text similarity computations on the general domain[4, 45, 14, 11].

4.2 Method

In this section, we specify the user interface development, details of our data and provide the details of the methodology used to perform our analysis. The general architecture of our proposed recommendation process is illustrated in Figure 4.1

4.2.1 User Interface Development

The structure of Wikitheoria's database provides a foundation for creating the recommendation system that offers suggestions to users based on the content consistency

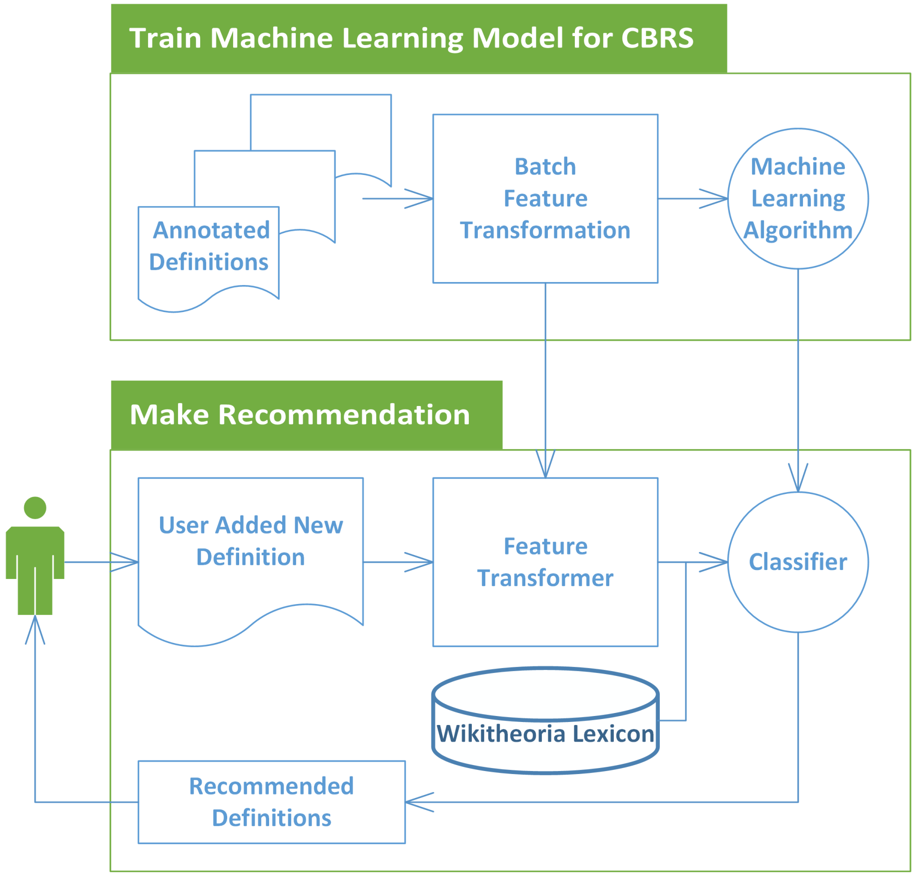

Train Machine Learning Model for CBRS
Annotated Definitions
Batch Feature Transformation
Machine Learning Algorithm
Make Recommendation
User Added New Definition
Feature Transformer
Classifier
Wikitheoria Lexicon
Recommended Definitions

Figure 4.1: An illustration of SOREC recomendation workflow with offline training and online prediction process.

checking from syntactic level and semantic level. As illustrated in Figure 4.2, the overall recommendation system includes three stages of recommendation.

Beginning with a large, pre-assembled lexicon of terms and definitions, our recommendation system analyzes newly offered provisional terms and definitions with respect to their syntactic similarities that previously defined in the lexicon with the first two stages. The syntactic level analysis is a relatively trivial task, which performs the calculation by using Trie-based autocomplete jQuery libraries and traditional information retrieval algorithms, such as Term Frequency Inverse Document Frequency (TF-IDF) and cosine similarity. However, the semantic level analysis for definitions on stage 3 requires extensive machine learning offline training and online prediction process. The experiment details are explained in the following sections.

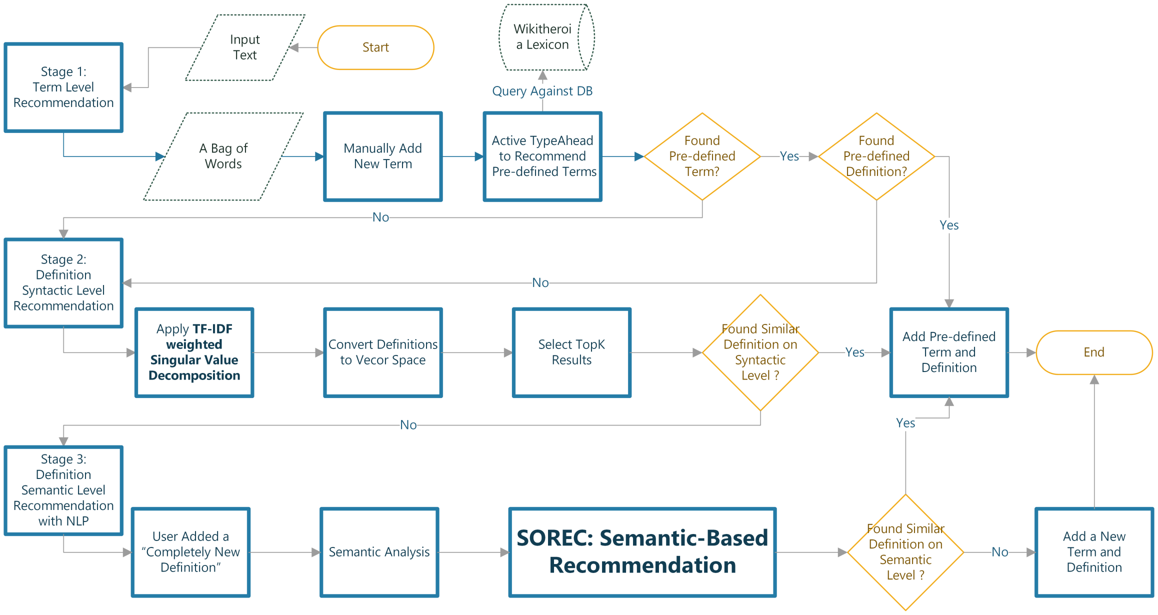


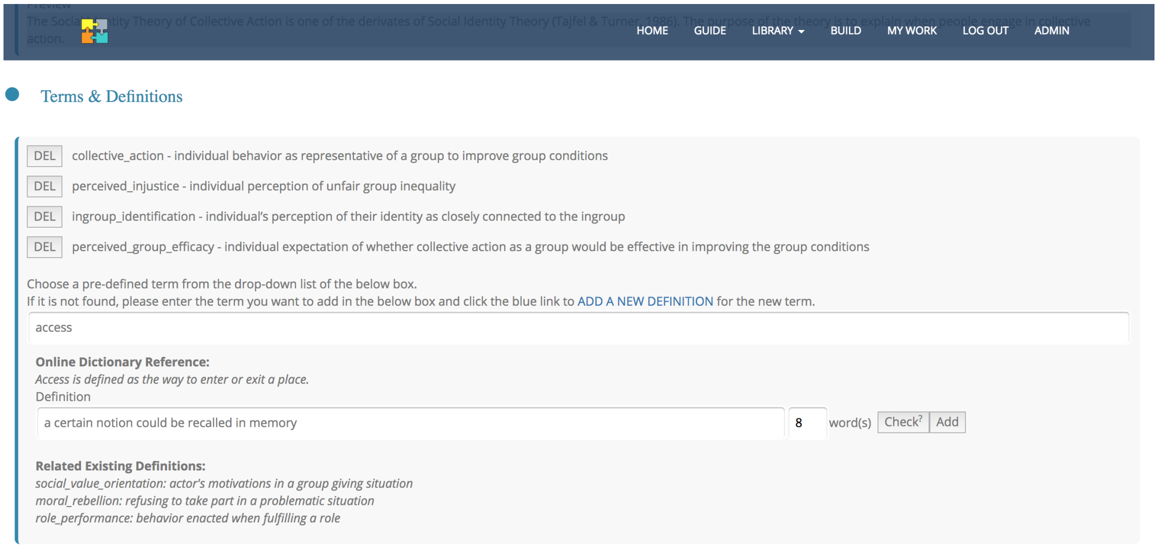
Figure 4.2: An illustration of three stage recommendation with sociological terms and definitions on Wikitheroia platform.

As shown in Figure 4.3, the recommendation system performs a recommendation on both terms and definitions. Our recommendation starts from term recommendation; once the user enters term "access", the text field for a term which associates with Ajax scripts starts querying the lexicon and finds whether this term has been previously defined for the first stage. The user could append the pre-defined terms and definitions to the current theory module, or create their term and definition pair.

The semantic level SOREC recommendation gets triggered when the user clicks "check" for the newly added definition for stage 3. This semantic level recommendation is based on a novel approach that is supported by semantic content-based filtering using the optimized gradient boosting method and features extracted from unsupervised machine learning methods.

4.2.2 Data Preprocessing

Pre-processing is an essential step to improve the accuracy of the model prediction. It can both reduce the time complexity for model training and accelerate the system



Terms & Definitions

DEL collective_action - individual behavior as representative of a group to improve group conditions

DEL perceived_injustice - individual perception of unfair group inequality

DEL ingroup_identification - individual's perception of their identity as closely connected to the ingroup

DEL perceived_group_efficacy - individual expectation of whether collective action as a group would be effective in improving the group conditions

Choose a pre-defined term from the drop-down list of the below box.
If it is not found, please enter the term you want to add in the below box and click the blue link to ADD A NEW DEFINITION for the new term.

access

Online Dictionary Reference:
Access is defined as the way to enter or exit a place.
Definition

a certain notion could be recalled in memory | 8 | word(s) | Check? | Add

Related Existing Definitions:
social_value_orientation: actor's motivations in a group giving situation
moral_rebellion: refusing to take part in a problematic situation
role_performance: behavior enacted when fulfilling a role

Figure 4.3: An illustration of sociological terms and definitions recommendation user interface on Wikitheoria.

response for online model predictions. In our study, pre-processing methods include tokenizing the definitions, removing the stop word, and converting all words to the lower case were applied to the definitions before applying the analysis of the features. We evaluated definition similarities based on the character level and term level briefly described in the following subsections on the basis of an annotated dataset. The TreeBank tokenizer implemented in the NLTK toolkit was used to convert a definition sentence to a list of tokens[9].

4.2.3 Feature Extraction with Definitions

Textual similarity metrics detect similarities between the two definitions. These are then used as features for our machine learning models. Zobel and Moffat analyzed a range of similarity measures in information retrieval. They found there to be no one-size-fits-all metric, i.e., no metric that consistently worked better than others[59]. To optimize our model, we spent considerable effort on feature engineering, experimented with the different combinations of feature sets, and obtained the best result with the features described below.

Descriptive Feature Set

The descriptive feature set includes basic properties of definition sentences. It presents observations about the characteristics of definitions. In this feature set, we calculated their lengths, length difference, character counts (excluding spaces), word counts, and words in common.

Tokenized Feature Set

In general, replaced words, inserted words, and missed words frequently occur in similar definitions. The tokenized feature set calculates the edit distance between one definition and another, i.e., the minimum number of operations that it would take to transform one definition into the other. We first processed definitions as two sets of sorted/unsorted token lists, then evaluated the similarity of pairwise token sets by calculating the minimum number of primitive operations, including insertion, deletion, substitution, or copying of a character required to convert one string into the exact match of the other. Specifically, we calculated the normalized Levenshtein distances [39] of pairwise tokens to generate the features based on the overlap ratio of unsorted token sets, the overlap ratio of sorted token sets, the overlap ratio of an unsorted partial token set and the overlap ratio of a sorted partial token set. The output ratio is on a 0 to 100 scale.

Embedding Feature Set

Tokenized features measure similarity based on exact matches between isolated words, but not their semantic meanings in context. The Universal Sentence Encoder[17] mixed an unsupervised task using a large corpus together showed significant improvement by leveraging the Transformer architecture, which is based on the attention mechanism. We trained our definitions with Tensorflow Hub[1] transformer encoders. Each definition was transformed into a 512-dimensional sentence vector. With the Transformer encoded embedding output, we computed the distance of two definition vectors (u and v) with the kernel functions shown below.

Cosine Distance[19] between two vectors u and v is defined as

$$sim_1 = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2} \tag{4.1}$$

Manhattan Distance[19] computes the distance between two vectors u and v by summing the differences of their corresponding components, which is defined as

$$sim_2 = \sum_i |u_i - v_i| \tag{4.2}$$

Jaccard Distance[23] proposed by Jaccard and Needham measures the dissimilarity between two vectors u and v, is defined as

$$sim_3 = \frac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v} \tag{4.3}$$

Canberra Distance[38] between two vectors is defined as

$$sim_4 = \sum_i \frac{|u_i - v_i|}{|u_i| + |v_i|} \tag{4.4}$$

Euclidean Distance[13] between 1-D arrays u and v is defined as

$$sim_5 = \|u - v\|_2 \tag{4.5}$$

Minkowski Distance[19] between 1-D arrays u and v is defined as

$$sim_6 = \|u - v\|_p = (\sum |u_i - v_i|^p))^{1-p} \tag{4.6}$$

Bray-Curtis Distance[12] is defined as

$$sim_7 = \sum |u_i - v_i| \, / \sum |u_i + v_i| \tag{4.7}$$

[1] http://www.tensorflow.org/hub/modules/google/universal-sentence-encoder-large/3

Skewness and Kurtosis[41] are the parameters used to measure the symmetry of the dataset and the weight of the tail compared to the normal distribution. Skewness is a measure of the symmetry. If the distribution or dataset appears to be the same as the left and right sides of the center point, it is symmetrical. Kurtosis is a measure of tail thickness, i.e., distribution with high kurtosis often has a heavy tail or outliers. Datasets with low kurtosis tend to have a light tail or outliers.

4.2.4 Model

With the features we created in the previous section, our goal was to develop a relevance model that would accurately predict if a user added new definition is semantically similar to an existing definition in theory. Our model is built with XGBoost, proposed by Chen and Gestrin in 2016[18], an optimized distributed gradient boosting library. Gradient boosting is a popular technique that can solve complex regression or classification task by producing and combining a number of weaker and smaller prediction models in the form of decision trees. The model is built in stages and generalized by optimizing a differential loss function.

As a result, gradient boosting combines a number of weak learners into a single, strong learner on an interactive basis. In contrast to linear classifiers (such as logistic regression), decision tree models also can capture non-linear relationships in data. We estimate the best hyperparameter settings for each model using a grid search with 10-fold cross-validation on the training set[37]. Carefully tuning the tree-related hyperparameters (such as the maximum depth of a tree) results in the most significant increase of cross-validation F1 score and accuracy. Tuning the learning rate is effective in preventing overfitting on the training data. Using a large number of estimators results in the best performance overall, with training time increases proportionally.

In our experiment, we chose tree booster in XGBoost as described in this section for all the feature representations' evaluation, in which max tree-depth was 15, step size shrinkage was 0.1, n estimators were 800, and minimum loss reduction was 1.0. As shown in Figure 4.4, firstly, the descriptive features and tokenized features were extracted from the definition pairs. These two feature sets were used to directly calculate the similarity of two definitions with respect to basic properties and edit distance. Then, these features adopted kernel-based Transformer sentence encoding to calculate the similarity of two definitions. All these similarity scores were concatenated as features and evaluated in the machine learning XGBoost model.

4.2.5 Evaluation

For this study, the XGBoost model evaluation was performed using 10-fold cross-validation over 2235 sociological definition pairs, including 959 positive samples and 1276 negative samples. Each fold contained 1811 definitions pairs for training, 200 pairs for validation, and 224 pairs for testing.

The final result for the supervised semantic content-based filtering was calculated by averaging the results of each fold. The quality and correctness of the proposed method is evaluated as 1) True positive (TP), the number of correct predictions on "same concept"; 2) True negative (TN), the number of correct predictions on "different concept"; 3) False positive (FP), the number of wrong predictions on "same concept"; 4) False negative (FN), the number of wrong predictions on "different concept". The precision (4.8), recall (4.9), F-measure (4.10) and accuracy (4.11) were used to evaluate the recommendation system.

$$Precision = \frac{TP}{TP + FP} \tag{4.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.9}$$

$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.11}$$

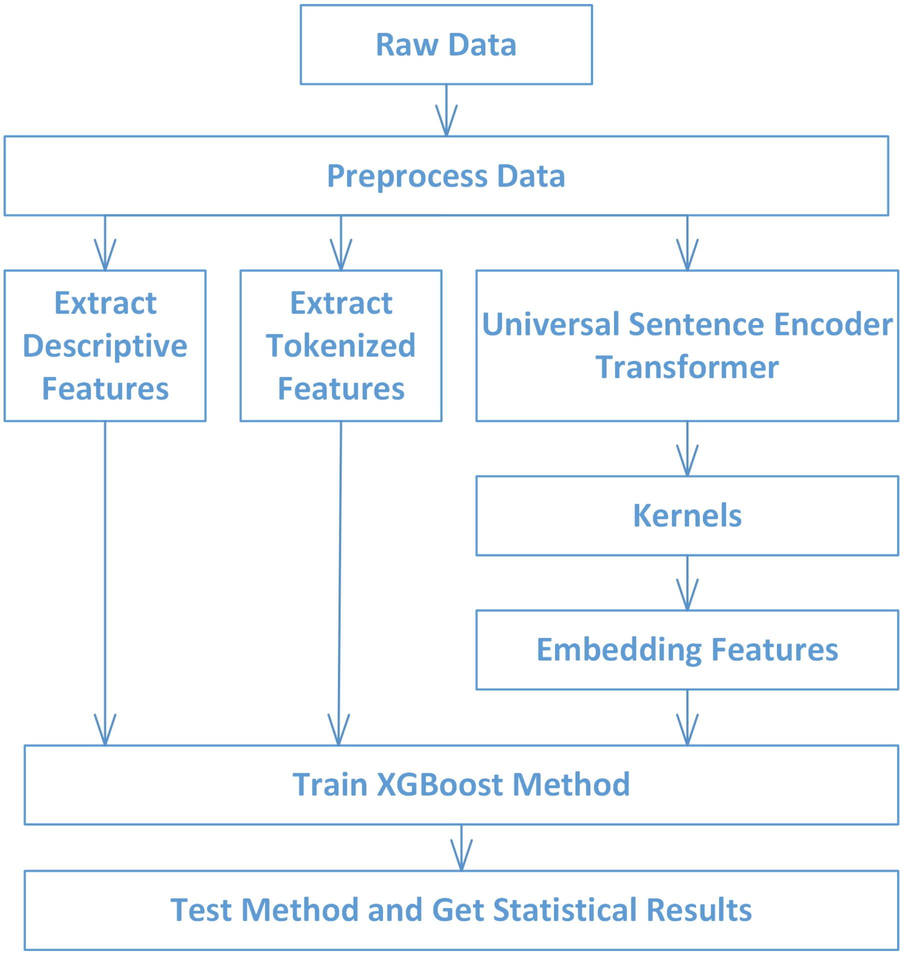

Raw Data
Preprocess Data
Extract Descriptive Features
Extract Tokenized Features
Universal Sentence Encoder Transformer
Kernels
Embedding Features
Train XGBoost Method
Test Method and Get Statistical Results

Figure 4.4: An illustration of SOREC prediction workflow with three feature sets and definition transformation.

4.3 Results and Discussion

In this section, we validate the effectiveness of our proposed method from two experiments. First, we evaluate the proposed method in terms of classical metrics of precision, recall, F-measure, and accuracy to justify the usefulness of each feature set in our method. Second, we break down the results and compare the improvement with the normalized confusion matrix.

4.3.1 Precision, Recall, F-measure and Accuracy

In this study, our feature representations are extracted from three different categories: descriptive feature set (DF), token feature set (TF), and embedding feature set (EF). Table 4.1 shows results for corresponding feature categories that were used as model input.

To evaluate the effectiveness of different categories, we performed several experiments on different combinations of feature categories. From the results, both DF and TF obtained moderate precision, recall, and accuracy. Comparing with EF or DF, the range of increase in precision, recall, and accuracy is 5% to 12%. Between three categories, our experiments show that the EF contributed more to the performance compared with other feature sets. When concatenating EF with DF or TF, the increase in precision varies from 5% to 15%. This is an expected result because embedding-based features enclose the transfer learning from the billion words corpus and the measurement from multiple dimensional spaces.

To capture the semantics, EF shows that it is a promising method for representing definitions as vectors while capturing semantics. Table 4.1 shows a significant improvement in precision, recall, F-measure, and accuracy when DF, TF, and EF are concatenated.

The experimental results indicate that the model with a combination of three feature categories outperforms the individual performance of each category and the combination of any two categories, which means that these feature sets complement each other. Although the embedding-based feature set obtained the most promising performance on precision and accuracy, the combination of three categories by a supervised algorithm had the best performance on all metrics. The proposed supervised semantic analysis model achieves the best precision of 86%, the best F-measure of 84.42%, and the best accuracy of 86.16%. We also compared Transformer-embedded definitions with average pooling Google News pre-trained embeddings on the same

XGBoost model described in the previous section. As shown in Table 4.2, Transformer outperforms Word2Vec by 2% on all metrics.

4.3.2 Usefulness of Each Feature Category

In the previous section, we saw a steady improvement when three categories were gradually added to the prediction model. For the rest of this subsection, we conducted a three-step experiment to justify the usefulness of each feature set and to inspect the effectiveness of each category.

DF considers the length related descriptions of definitions. As shown in Figure 4.5 and Table 4.1, the TP is 49% and TN is 85%. The result is expected, as in the general domain, similar definitions tend to be of similar length. When definitions substantially differ in length, the model tends to predict dissimilarity. But in sociology, the "same concept" could be defined with one short sentence or multiple sentences if terminologies are densely used to support definitions. In our dataset, the average length of a definition is 17.87, and the average length difference between pair definitions is 9.67. The definition pairs with high relative length differences tend to be harder to predict correctly for the "same concept." With DF representation, the model achieved an average accuracy of 69.19% and F-measure of 57.8%.

Next, we validate the usefulness of TF. For the tokenized feature set, we calculated the edit distance between one definition and another. As shown in Figure 4.6 and Table 4.1, with DF and TF, the recommendation quality for correctly recommending the "same concept" definitions have been improved by 16%. The F-measure is 11.97% higher than the previous step, and the prediction accuracy achieves 75.89%.

With the added EF, from Figure 4.7 and Table 4.1, we see a 19% improvement on TP and 4% improvement on TN. 84% of "same concept" definitions are correctly predicted, and 88% of "different concept" definition pairs are also correctly predicted. This indicates the sociological semantic relatedness could be well represented by calcu-

Table 4.1: The comparison results of prediction on sociological definition test dataset with 10-fold cross validation.

Model	Precision	Recall	F-measure	Accuracy
TF-IDF + W2V + SVD	0.68	0.67	0.5780	0.6741
DF	0.70	0.69	0.5868	0.6919
TF	0.77	0.76	0.6936	0.7633
EF	0.84	0.83	0.8140	0.8348
DF-TF	0.76	0.76	0.7066	0.7589
DF-EF	0.85	0.85	0.8265	0.8482
TF-EF	0.84	0.84	0.8275	0.8437
DF-TF-EF	0.86	0.86	0.8442	0.8616

Table 4.2: The comparison results of prediction with Word2Vec and Transformer.

Embedding	Precision	Recall	F-measure	Accuracy
Word2Vec	0.84	0.84	0.8258	0.8437
Transformer	0.86	0.86	0.8442	0.8616

lating the embedding distance from multiple dimensions. The F-measure is improved by 1.67%, and the best prediction accuracy achieves 86.16%.

As shown in Figure 4.8, the baseline recommendation system is based on the TF-IDF, Word2Vec, and Singular Value Decomposition (SVD) model. To evaluate the performance of the baseline model, we encoded definitions with TF-IDF Word2Vec, applied SVD, then fed them to the same XGBoost model we constructed in Section III. From Figure 4.8 and Table 4.1, this model shows a weak ability to predict “same concept” definitions with TP of 50%. Comparing with the baseline model, the proposed semantic content-based recommender system improves the lexicon semantic similarity recommendation and achieves the best performance.

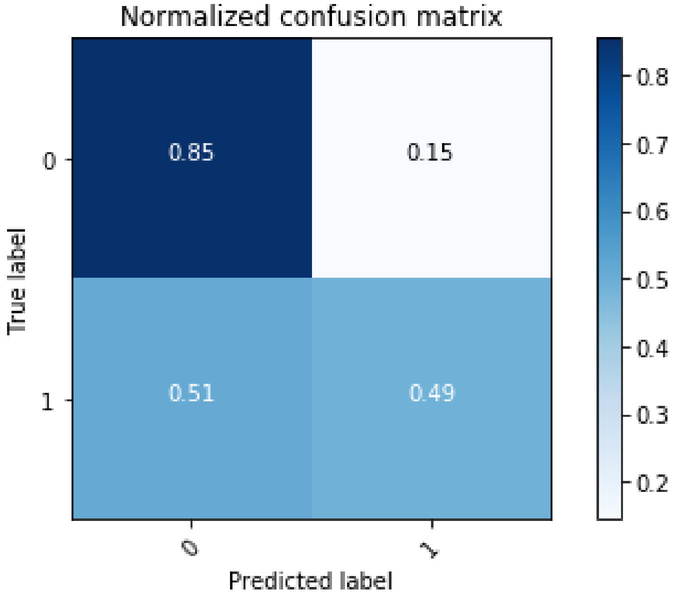


Figure 4.5: The results with descriptive features in normalized confusion matrix.

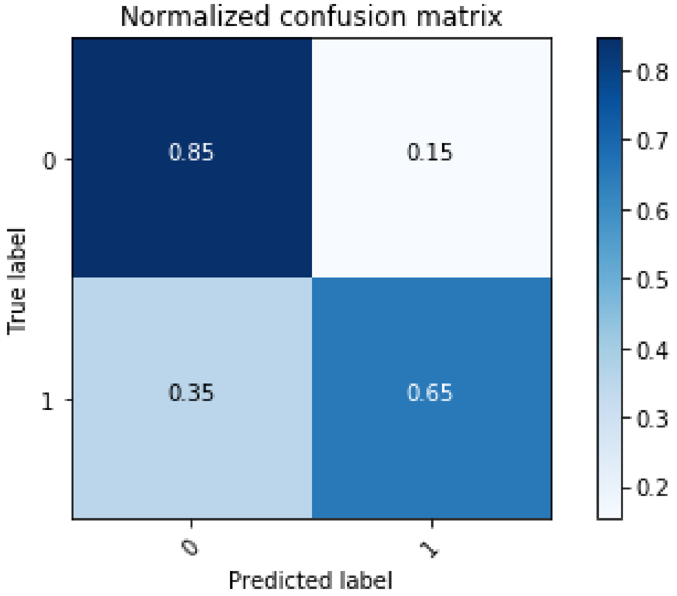


Figure 4.6: The results with descriptive features and tokenized features.

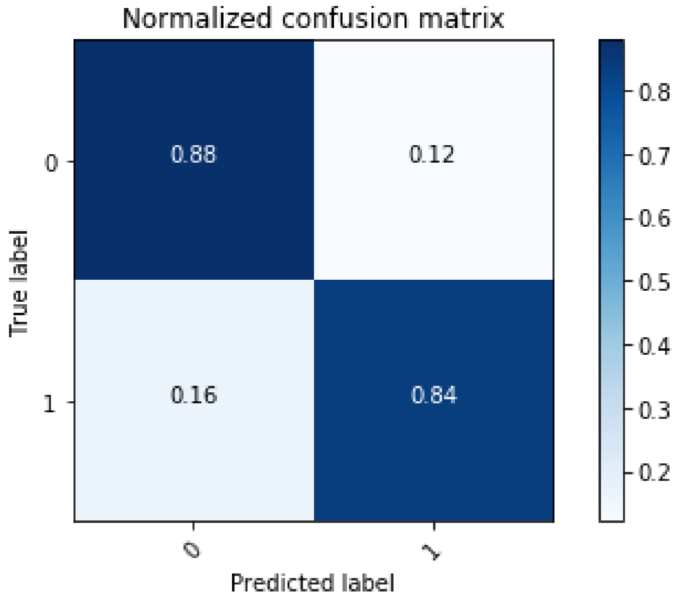


Figure 4.7: The results of prediction when descriptive features, tokenized features and embedding features are concatenated together.

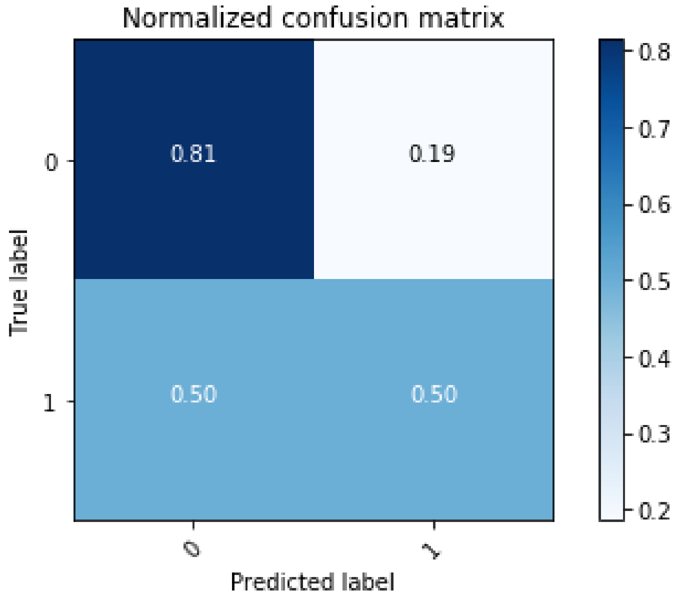


Figure 4.8: The results of conventional content-based recommendation system prediction accuracy with TF-IDF, Word2Vec and SVD.

4.4 Conclusion

In this study, we proposed a novel semantic content-based recommender system for sociological theory construction. To the best of our knowledge, there is neither a similar recommender system nor published research on the semantic evaluation of sociological definitions. We demonstrated the need for a semantic recommender for semantic level analysis and the effectiveness of our proposed approach to understanding the semantic similarity of terminologies and definitions in the sociological domain. Another important contribution of this study is to provide a solid baseline as well as a sociologists-annotated benchmark dataset for future studies in this research area.

Our results revealed that the descriptive features, the edit distance based tokenized features, and the kernel function based embedding features complement each other. Notably, the high-level features consist of the embedding vector distances calculated from unsupervised kernel functions, with the help of an XGBoost model increased the overall performance of the recommender system.

The proposed CBRS is deployed and serving as a part of the Wikitheoria platform. Theory construction is a typical research process in a lot of human science-related disciplines, such as Psychology, Criminology, etc. Our sociology-domain specific semantic definition recommendation can also be applied to various content-based recommendation applications for parsimonious theory construction in these disciplines.

Chapter 5

Conclusion

5.1 Summary

In summary, this research solves the problem of “how to facilitate sociology researchers to build parsimonious theories” in three parts.

In Chapter 2, we show that using cloud-based theory modularization with semantic-based parsimony analysis with the machine learning model is a viable approach. The proposed work mainly describes what theory modularization is, what parsimony analysis is in theory construction, how we formalized these concepts and implemented with Google Cloud, an implementation of a machine learning method with neural embedding, and the initial results we could achieve.

In Chapter 3, we mainly focus on the parsimony analysis. Since there are many neural embeddings and machine learning models, we wish to gain an insight of which embedding method is the best for our domain of knowledge, and which machine learning model could better capture the semantic similarity. So we experiment with eight different neural embedding methods on four representative machine learning models. Furthermore, we develop 11 semantic features to enhance prediction accuracy. Through the experiments, we find that the tree-boosted XGBoost classifier with attention machenism-based Transformer representation performs the best on our dataset.

In Chapter 4, we know that neural embedding with the machine learning method performs well on parsimony anaysis. But it is still a theoretical work on the algorithm level. Our framework needs an actual tangible product to serve the researchers. So, we propose and implement SOREC, a semantic content-based recommendation system(CBRS). CBRS has 26 features and validate the effectiveness of each feature set. We also compare our proposed CBRS with the conventional CBRS on our dataset, and it is substantially better (+19% accuracy). In this study, we prove the effectiveness of the proposed CBRS, and we establish a baseline for studies in this research area.

5.2 Future Works

This work is an initial step towards a promising new direction. In future work, we plan to incorporate other types of deep learning architectures such as convolutional neural network, deep belief network, and recurrent neural network. Further performance boost may be possible when using such deep learning models since these models can explicitly take the context and ordering of words into account. Moreover, we plan to explore deeper architectures by applying data normalization techniques to help model stability and a better local optimum. In this study, we establish a solid baseline for the semantic content-based recommendation on the sociology lexicon, however we believe the quality of the recommendation could be further improved with transfer learning with massive domain knowledge encapsulated literature and knowledge graph infused machine learning algorithms.

Theory construction is a common research process in a lot of human science-related disciplines. We wish the applications and research methodologies described in this study could be further extended to support the theory construction and parsimony analysis in psychology, criminology, and other social sciences.

Bibliography

[1] Kees Aarts, *Parsimonious methodology*, Methodological Innovations Online **2** (2007), no. 1, 2–10.

[2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., *Tensorflow: A system for large-scale machine learning*, 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

[3] Titipat Achakulvisut, Daniel E Acuna, Tulakan Ruangrong, and Konrad Kording, *Science concierge: A fast content-based recommendation system for scientific publications*, PloS one **11** (2016), no. 7, e0158423.

[4] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa, *A study on similarity and relatedness using distributional and wordnet-based approaches*, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2009, pp. 19–27.

[5] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe, *Semeval-2014 task 10: Multilingual semantic textual similarity*, Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), 2014, pp. 81–91.

[6] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo, *Sem 2013 shared task: Semantic textual similarity*, Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity, vol. 1, 2013, pp. 32–43.

[7] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre, *Semeval-2012 task 6: A pilot on semantic textual similarity*, Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth

International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2012, pp. 385–393.

[8] Naomi S Altman, *An introduction to kernel and nearest-neighbor nonparametric regression*, The American Statistician **46** (1992), no. 3, 175–185.

[9] Steven Bird, Ewan Klein, and Edward Loper, *Natural language processing with python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.", 2009.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, *Enriching word vectors with subword information*, Transactions of the Association for Computational Linguistics **5** (2017), 135–146.

[11] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning, *A large annotated corpus for learning natural language inference*, arXiv preprint arXiv:1508.05326 (2015).

[12] J Roger Bray and John T Curtis, *An ordination of the upland forest communities of southern wisconsin*, Ecological monographs **27** (1957), no. 4, 325–349.

[13] Heinz Breu, Joseph Gil, David Kirkpatrick, and Michael Werman, *Linear time euclidean distance transform algorithms*, IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995), no. 5, 529–533.

[14] Alexander Budanitsky and Graeme Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness*, Computational Linguistics **32** (2006), no. 1, 13–47.

[15] Robin Burke, *Hybrid web recommender systems*, The adaptive web, Springer, 2007, pp. 377–408.

[16] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia, *Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation*, arXiv preprint arXiv:1708.00055 (2017).

[17] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al., *Universal sentence encoder*, arXiv preprint arXiv:1803.11175 (2018).

[18] Tianqi Chen and Carlos Guestrin, *Xgboost: A scalable tree boosting system*, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016, pp. 785–794.

[19] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert, *A survey of binary similarity and distance measures*, Journal of Systemics, Cybernetics and Informatics **8** (2010), no. 1, 43–48.

[20] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes, *Supervised learning of universal sentence representations from natural language inference data*, arXiv preprint arXiv:1705.02364 (2017).

[21] David Cornforth and David G Green, *Modularity and complex adaptive systems*, Intelligent Complex Adaptive Systems, IGI Global, 2008, pp. 75–104.

[22] Olivier Darrigol, *The modular structure of physical theories*, Synthese **162** (2008), no. 2, 195–223.

[23] John Davey and Elizabeth Burd, *Evaluating the suitability of data clustering for software remodularisation*, Proceedings Seventh Working Conference on Reverse Engineering, IEEE, 2000, pp. 268–276.

[24] Marco De Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro, *Semantics-aware content-based recommender systems*, Recommender Systems Handbook, Springer, 2015, pp. 119–159.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805 (2018).

[26] Joseph Dippong, Will Kalkhoff, and Eugene C Johnsen, *Status, networks, and opinion change: An experimental investigation*, Social Psychology Quarterly **80** (2017), no. 2, 153–173.

[27] Mingzhe Du, Zaid Alibadi, Jose Vidal, and Barry Markovsky, *Towards parsimonious sociology theory construction with neural embeddings and semantic analysis.*

[28] Mingzhe Du, Jose Vidal, and Barry Markovsky, *Sorec: A semantic content-based recommendation system for parsimonious sociology theory construction.*

[29] Mingzhe Du, Jose M. Vidal, and Barry Markovsky, *Wikitheoria: A computational framework for parsimonious sociology theory construction.*

[30] Jinbo Feng and Shengli Wu, *Detecting near-duplicate documents using sentence level features*, Database and Expert Systems Applications, Springer, 2015, pp. 195–204.

[31] Lee Freese, *Formal theorizing*, Annual Review of Sociology **6** (1980), no. 1, 187–212.

[32] Hugh G Gauch and Hugh G Gauch Jr, *Scientific method in practice*, Cambridge University Press, 2003.

[33] Antonio Gulli and Sujit Pal, *Deep learning with keras*, Packt Publishing Ltd, 2017.

[34] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl, *Evaluating collaborative filtering recommender systems*, ACM Transactions on Information Systems (TOIS) **22** (2004), no. 1, 5–53.

[35] Hosein Jafarkarimi, Alex Tze Hiang Sim, and Robab Saadatdoost, *A naive recommendation model for large databases*, International Journal of Information and Education Technology **2** (2012), no. 3, 216.

[36] Ammar Ismael Kadhim, Yu-N Cheah, Inaam Abbas Hieder, and Rawaa Ahmed Ali, *Improving tf-idf with singular value decomposition (svd) for feature extraction on twitter*.

[37] Ron Kohavi et al., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, Ijcai, vol. 14, Montreal, Canada, 1995, pp. 1137–1145.

[38] Godfrey N Lance and William Thomas Williams, *A general theory of classificatory sorting strategies: 1. hierarchical systems*, The computer journal **9** (1967), no. 4, 373–380.

[39] Vladimir I Levenshtein, *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet physics doklady, vol. 10, 1966, pp. 707–710.

[40] Yuqing Mao, Kimberly Van Auken, Donghui Li, Cecilia N Arighi, Peter McQuilton, G Thomas Hayman, Susan Tweedie, Mary L Schaeffer, Stanley JF Laulederkind, Shur-Jen Wang, et al., *Overview of the gene ontology task at biocreative iv*, Database **2014** (2014).

[41] Kanti V Mardia, *Measures of multivariate skewness and kurtosis with applications*, Biometrika **57** (1970), no. 3, 519–530.

[42] Barry Markovsky, *Modularizing small group theories in sociology*, Small group research **41** (2010), no. 6, 664–687.

[43] Barry Markovsky and Murray Webster Jr., *Theory construction*, The Blackwell Encyclopedia of Sociology (George S. Ritzer, ed.), Malden, MA: Blackwell, 2nd ed., (in press).

[44] William McCune, *Prover9 and mace4*, 2005.

[45] Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al., *Corpus-based and knowledge-based measures of text semantic similarity*, AAAI, vol. 6, 2006, pp. 775–780.

[46] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 2013, pp. 3111–3119.

[47] Raymond J Mooney and Loriene Roy, *Content-based book recommending using learning for text categorization*, Proceedings of the fifth ACM conference on Digital libraries, ACM, 2000, pp. 195–204.

[48] Sabrina Moretti, *Computer simulation in sociology: What contribution?*, Social Science Computer Review **20** (2002), no. 1, 43–57.

[49] Michael J Pazzani and Daniel Billsus, *Content-based recommendation systems*, The adaptive web, Springer, 2007, pp. 325–341.

[50] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., *Scikit-learn: Machine learning in python*, Journal of machine learning research **12** (2011), no. Oct, 2825–2830.

[51] Jeffrey Pennington, Richard Socher, and Christopher Manning, *Glove: Global vectors for word representation*, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[52] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, *Deep contextualized word representations*, arXiv preprint arXiv:1802.05365 (2018).

[53] Dan Sanderson, *Programming google app engine with python: Build and run scalable python apps on google's infrastructure*, " O'Reilly Media, Inc.", 2015.

[54] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha, *Singular value decomposition and principal component analysis*, A practical approach to microarray data analysis, Springer, 2003, pp. 91–109.

[55] Donghui Wang, Yanchun Liang, Dong Xu, Xiaoyue Feng, and Renchu Guan, *A content-based recommender system for computer science publications*, Knowledge-Based Systems **157** (2018), 1–9.

[56] Yanshan Wang, Naveed Afzal, Sijia Liu, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Sunyang Fu, and Hongfang Liu, *Overview of the biocreative/ohnlp challenge 2018 task 2: Clinical semantic textual similarity*, Proceedings of the BioCreative/OHNLP Challenge **2018** (2018).

[57] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal, *Data mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2016.

[58] Feng Zhang, Ti Gong, Victor E Lee, Gansen Zhao, Chunming Rong, and Guangzhi Qu, *Fast algorithms to evaluate collaborative filtering recommender systems*, Knowledge-Based Systems **96** (2016), 96–103.

[59] Justin Zobel and Alistair Moffat, *Exploring the similarity space*, Acm Sigir Forum, vol. 32, ACM, 1998, pp. 18–34.