Wikitheoria: A Computational Framework for Parsimonious Sociology Theory Construction

Mingzhe Du
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, USA
dum@email.sc.edu

Jose M. Vidal
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, USA
vidal@sc.edu

Barry Markovsky
Department of Sociology
University of South Carolina
Columbia, SC, USA
barry@sc.edu

ABSTRACT

In the social sciences, theory construction refers to the research process of building testable scientific theories to explain and predict observed phenomena in the natural world. Terms represent the theories' concepts or ideas and their meanings are explicated in their definitions. 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 parsimony analysis in theory construction are based on the heuristic approaches. However, it is not always easy for young researchers to understand the theoretical work in a given area because of the problem of "tacit knowledge", which often makes results lack coherence and logical integrity. Therefore, we propose 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. The experiment results showed that the proposed approach achieves the precision of 82%, recall of 82% and accuracy of 81.69%. This study proves the effectiveness of using cloud-based knowledge aggregation system and semantic analysis models for promoting the parsimonious sociology theory construction.

KEYWORDS

cloud computing, semantic similarity, parsimony analysis, embedding, modularized theory construction

1 INTRODUCTION

In the social sciences, theories are used to explain and predict observed phenomena in the natural world[25]. Theory construction is the process of building theories to strict specifications with respect to the clarity of their concepts or ideas through the definitions of terms[24, 25]. 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, 20].

Conventional methods for parsimony analysis in theory construction are based on the heuristic approaches[20, 25]. However, although the young researchers are trained by mentors who are familiar with accepted views in that area, it is not always easy for them to understand the theoretical work in the given areas because of the problem of "tacit knowledge"[1, 14, 24, 25]. It can be difficult to try to acquire a sense of understanding in another theoretical area, and the challenge of how to interpret specific vague or ambiguous terminologies inside the information often makes results lack coherence and logical integrity.

To help with this problem, we propose a computational framework using Google Cloud Platform and semantic textual similarity analysis models to facilitate the parsimonious approach of theory construction[11, 31]. As shown in Figure 1, the proposed approach consists of three components: (1) Sociology theories were modularized and constructed with the cloud-based tools 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 paper are as follows: (1) We propose a computational framework for theory modularization and theory construction. (2) We prove the effectiveness of using embedding models on the semantic similarities of sociological definitions. (3) Additionally, we experiment with textual similarity measurement (cosine similarity) and similarity prediction (KNN) in which the result achieves an accuracy of 81.69%. The results of this study can be further applied to the theory construction of psychology, criminology, and other social sciences.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed cloud-based platform and embedding-based semantic analysis method. Evaluation results and discussions are presented in Section 4. The conclusion is in the last section.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACIT 2019, May 29–31, 2019, Honolulu, HI, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7173-5/19/05…$15.00
https://doi.org/10.1145/3325291.3325355

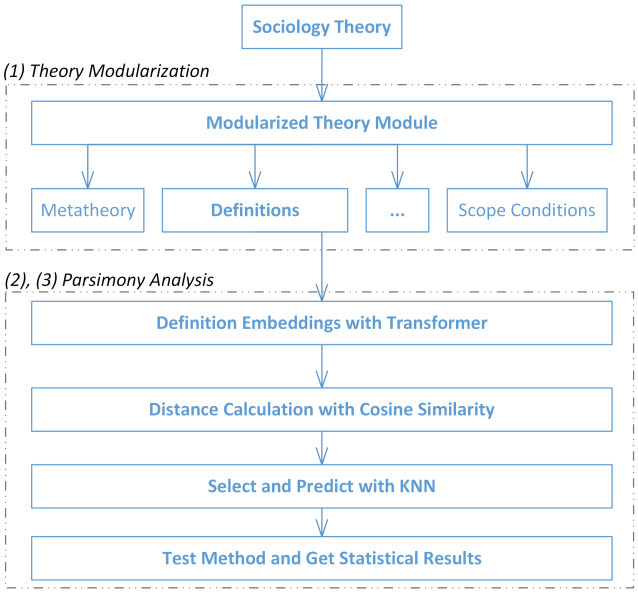

Figure 1: Parsimonious Theory Construction Workflow

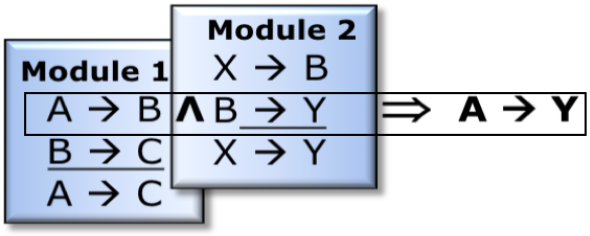

Figure 2: Schematic Illustration

2 RELATED WORK

2.1 Theory Modularization

A theory uses clearly defined terms in propositions that are amenable to the application of logical calculi [16, 19, 24]. 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 [29] 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 [14] described the nature and benefits of modularization, from genetics to social networks to manufacturing. These ideas apply readily to theories [15, 17, 24–26]. Figure 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.

2.2 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. [1, 24, 25]. In a theory, terms are carefully chosen by the theorist to convey ideas or concepts, and their meanings are clearly implicated in the definitions[20, 24, 25]. 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 "denomination" extracted from the Blackwell Encyclopedia of Sociology. 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 the theory already and the sociologist plans to add a new definition D2, he should determine whether D2 is in the 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 wide spectrum of natural language understanding applications[3–6, 11]. 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[10, 12, 27]. 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 [12].

3 PROPOSED APPROACH

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 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 modularized theory construction and parsimony analysis.

Figure 3: Modularizrd Theory Module on Wikitheoria

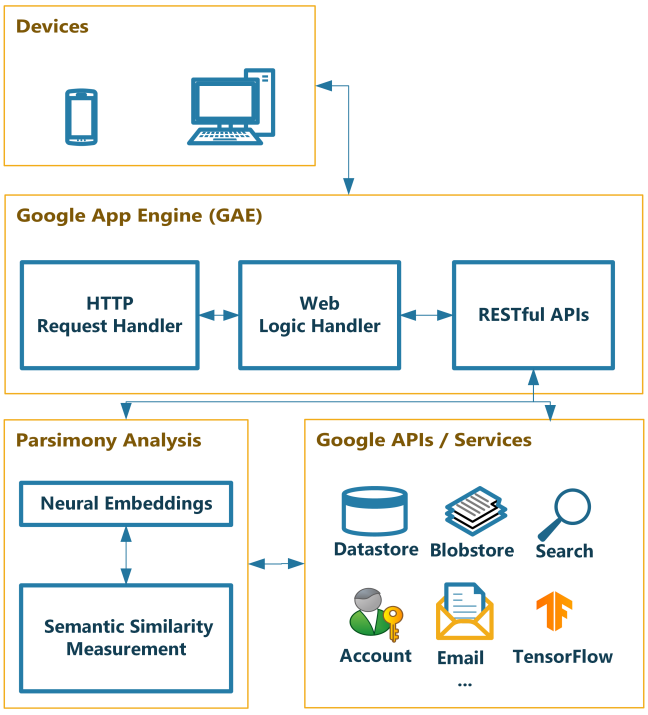

Figure 4: Wikitheoria Platform Infrastructure

3.1 Cloud-based Theory Modularization

Wikitheoria is a cloud-based application which could be accessed with smart phones and computers. With the help of Wikitheoria, a theory is modularized and constructed with several essential components. For example, in Figure 3, the theory "Social Identity Model of Collective Action (SIMA)" is constructed using theory title, key words, metatheory, terms and definitions, propositions, derivations, scope conditions and evidence.

As illustrated in Figure 4, the HTTP requests sent from devices are accepted and processed by the application. The various services such as parsimony analysis, email, account, datastore and blobstore, etc. communicates 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[31] to handle the web related logics. For the parsimony analysis, we encode the sentence definitions with Tensorflow Hub and Keras[2, 21], then train the classifier with Scikit-learn[30].

The proposed application utilizes Google App Engine (GAE), a Software-as-a-Service (SaaS) platform[31], which offers important 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 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.

3.2 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[8, 9, 12, 18, 23, 28, 33].

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[8].

The Universal Sentence Encoder mixed an unsupervised task using a large corpus together showed a significant improvement by leveraging the attention-based Transformer architecture[12]. 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 KNN algorithm.

Cosine Distance[13] between two vectors u and v is defined as

$$sim = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2} \quad (1)$$

To find the potential redundant definition, we employed an approach with KNN model [7]. The model was implemented with Scikit-learn[30], 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.

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(2), recall(3) and accuracy (5) were used to evaluate the semantic similarity measurement.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = 2 * \frac{Precision}{Precision + Recall} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

4 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[22, 32] 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 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, it shows the precision, which indicates the ratio of 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 on both categories, the best prediction accuracy is 81.69%.

As shown in Table 1 and Table 2, the performance of Transformer embedded sociological definitions with 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 the strong performance on recall, precision and accuracy for promoting the parsimonious theory construction.

5 CONCLUSION

In this study, we propose a generic knowledge aggregation framework to facilitate the parsimonious approach of theory construction with a cloud-based theory modularization platform and semantic-based algorithm to reduce the semantic redundancy. 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.

Table 1: The KNN 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: The performance of KNN by category, k = 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		

Our results demonstrated the effectiveness of using 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). 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.

ACKNOWLEDGMENTS

The authors are supported by the SES-1123040 and IIS-1551458 grants from the National Science Foundation. We would like to thank Jake Frederick, Nicolas Harder for annotations and Jing Wang for improving the user interface.

REFERENCES

[1] Kees Aarts. 2007. Parsimonious Methodology. *Methodological Innovations Online* 2, 1 (2007), 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. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*. 265–283.

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

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

[5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. SEM 2013 shared task: Semantic textual similarity. In *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. 32–43.

[6] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *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, 385–393.

[7] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.

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

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

[10] Alexander Budanitsky and Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics* 32, 1 (2006), 13–47.

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

[12] 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. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[13] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics* 8, 1 (2010), 43–48.

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

[15] Olivier Darrigol. 2008. The modular structure of physical theories. *Synthese* 162, 2 (2008), 195–223.

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

[17] Joseph Dippong, Will Kalkhoff, and Eugene C Johnsen. 2017. Status, networks, and opinion change: An experimental investigation. *Social Psychology Quarterly* 80, 2 (2017), 153–173.

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

[19] Lee Freese. 1980. Formal theorizing. *Annual Review of Sociology* 6, 1 (1980), 187–212.

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

[21] Antonio Gulli and Sujit Pal. 2017. *Deep Learning with Keras*. Packt Publishing Ltd.

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

[23] Godfrey N Lance and William Thomas Williams. 1967. A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The computer journal* 9, 4 (1967), 373–380.

[24] Barry Markovsky. 2010. Modularizing Small Group Theories in Sociology. *Small group research* 41, 6 (2010), 664–687.

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

[26] William McCune. 2005. Prover9 and mace4.

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

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

[29] Sabrina Moretti. 2002. Computer simulation in sociology: What contribution? *Social Science Computer Review* 20, 1 (2002), 43–57.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[31] Dan Sanderson. 2015. *Programming Google App Engine with Python: Build and Run Scalable Python Apps on Google's Infrastructure.* " O'Reilly Media, Inc.".

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

[33] Justin Zobel and Alistair Moffat. 1998. Exploring the similarity space. In *Acm Sigir Forum*, Vol. 32. ACM, 18–34.