SEMESTER PROJECT REPORT

Topic Modeling Based Bibliometric Exploration of ICLS and CSCL

Author:
Zhechen Su (zhechen.su@epfl.ch)

Supervisor:
Stian Haklev

Professor:
Pierre Dillenbourg

June 1, 2019

Contents

1 Introduction 3

1.1 Background . . . 3

1.2 Purpose and Scope . . . 4

2 Method 4

2.1 Data Preprocessing . . . 4

2.2 Latent Dirichlet Allocation . . . 5

2.3 Author-Topic Modeling . . . 6

2.4 Dynamic Topic Modeling . . . 6

3 Data Source 7

4 Results and Analysis 8

4.1 Choose Topic Number . . . 8

4.2 Topic Interpretations . . . 9

4.3 Topic Volume . . . 11

4.4 Topic Preference of Conference . . . 12

4.5 Topic Popularity and Trend . . . 13

4.6 Topic Evolution . . . 14

4.6.1 Eye-tracking and Gaze . . . 14

4.6.2 Identity and Equity . . . 14

4.7 Author Topic Preference and Cluster . . . 16

4.8 Author Similarity Exploration . . . 17

4.8.1 Which article is a cross-work product? . . . 18

4.8.2 Who is versatile? . . . 18

4.8.3 Who prefer working with different guys? . . . 20

5 Conclusion and Further Development 22

1 Introduction

1.1 Background

Learning Science plays a critical role in human's daily life. It is an interdisciplinary field of study on teaching and learning that has developed over the past 30 years and covers a wide range of research fields in education, brain science, psychology, cognitive science, information science, and biomedicine. In short, learning science is mainly research on problems like how do people learn, how can they promote learning efficiency. With the advancement of technology and education, learning science is becoming the core concept and a dominant force in reforming education and teaching. At the same time, The number of papers in this field is rising rapidly.

A large amount of information and knowledge enhance the accessibility of researchers, which brings convenience but also confusion. For researchers, it is meaningful and valuable to figure out the trends and overall situation in his field. Thus, there is a need for scientific analysis to conclude and provide insights about the hot spot in fields. Traditional literature bibliometric analysis includes quantitative investigation and visual analysis and is used to determine the development trends of the research, calculate the academic authority of journals, and so on. However, it is not intrinsic to classify articles' topic adequately. Keywords that the author wrote cannot work as labels, because everyone may express similar conceptions in different ways like *Assessment* and *Critique* in Learning Science.

In recent years, topic modeling in the field of machine learning is popular and has also been introduced into bibliometric research. Topic modeling uses statistical algorithms to extract semantic information from a collection of texts and has become an emerging quantitative method for assessing substantial textual data. It takes documents as input and outputs are hidden topical patterns presented across the collection. Also, the model annotates documents according to these topics. This model is not only robust but also a method of non-human supervision.

Initially, topics are seen as the reduced dimensional representation of text. Deerwester et al.[1] suggested using singular-value decomposition(SVD) to reduce dimension, which is known as "latent semantic indexing" (LSI). Then the probabilistic latent semantic indexing method was presented by Hoffmann[2]. It assumes that each document is a mixture of polynomial random variables (topics), and each word in the document is generated by a topic. Different words in the document can be generated by different topics. However, the number of parameters in PLSI model grow linearly as the corpus grows, and there is no good prediction for unobserved text.

Blei[3] claimed a better three-layer Bayesian model in 2003. Latent Dirichlet Allocation(LDA) model takes Dirichlet distribution as the prior distribution, then transforms sparse document-term matrices into fixed low dimensional document-topic matrices. LDA can generate a good distribution of topics, as well as be used to predict topics distribution of the unobserved text. Good generalization ability makes it successful model in the fields of machine learning and information retrieval. Based on the original model, researchers developed numerous extensions like Correlated Topic Models (CTM) and Hierarchical Dirichlet Process (HDP)[4][5]. All of these models have been widely used in bibliometric research to discover hidden structures and semantic topics in an area.

In this study, we used the LDA model to extract the semantic information and generate topic from scientific journals in ISLS. Then, we found the support documents for each topic according to the generated topic space as well as allocate topic distribution to each document. Moreover, in topic level, we analyzed the two conferences' preference and topic percentage change over four years. Next, the Dynamic topic model was applied to explore conceptional word evolution in word level. Lastly, for author level, we calculated their preference similarity and tried to find collaboration patterns among not only persons but also topics.

1.2 Purpose and Scope

With the rapid increment of articles in fields, knowing what papers are talking about and understanding their opinions in a short time are highly valuable. For scientific researchers who are going to start a new work, they may raise questions like what is the development of the main research issues in fields, which are the current research hot spots, which are gradually fading, and so on. Such questions are interesting but usually hard to get standard answers. Also, few articles are written by a single person, so collaboration is another essential aspect a researcher should consider when he starts to write.

In this study, we focus on Learning Science area and take two typical conferences as examples. Topic modeling is employed as a primary method, trying to give more reliable explanations and insights about the questions above. To observe the focus migration inside conferences as well as to understand researchers' preference and cooperation patterns, we conduct this project based on Topic Modeling and give insights into the following aspects:

1. Topic preference of ICLS and CSCL
2. Topic composition of each document
3. Topics' popularity and their tendency over the years
4. Conception evolution inside a topic
5. Authors' preference and their similarity between each other
6. Authors' work pattern

2 Method

2.1 Data Preprocessing

Abassi Nour Ghalia and Guillain Lonore Valentine had parsed the pdf files to text and generated a metadata csv to store author information. Thus, based on their help, our work is to make corpus and dictionary as inputs of the topic model. By using the Python interface, we can easily get a list of contents from the text file in system folder. The contents are raw textual data of each document. There are several steps to manipulate the raw text to the required materials for topic modeling.

1. **Split words based on spaces** This step is the most natural step. English sentences are composed of punctuation, spaces, and words. Just divide the words into arrays based on spaces and punctuation. For example, *Nobody knows how ancient people started using fire* is divided into {*Nobody, knows, how, ancient, people, started, using, fire*}.

2. **Remove stop words** In English, there are a lot of words such as *a*, *the*, *or*, etc., which are often used as articles, prepositions, adverbs or conjunctions. These words are too frequently used in almost every document, so we ignore them all at the time of preprocessing. If there are a large number of such words in our corpus, then it will take more time to train the model and reduce the accuracy of the model. Such as {*Nobody, knows, how, ancient, people, started, using, fire*} remove the stop word and get {*Nobody, ancient, people, started, fire*}

3. **Make bigrams** Sometimes, we cannot simply take the only single word in corpus and ignore the presence of bigrams like *New York*. Independent words cannot well express a particular meaning. So, we search all two words frequently occurring together in documents. If the word pair have a frequency between lower and higher thresholds, we

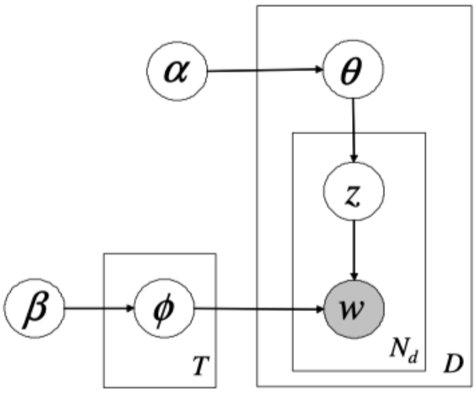

Figure 1: LDA

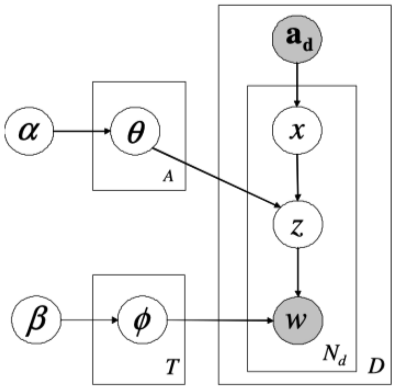

Figure 2: ATM

concatenate these two words and make them as a new word. {*Nobody, ancient, people, started, fire*} becomes {*Nobody, ancient_people, started, fire*} after importing bigrams.

4. **Stemming** Words have singular and plural numbers, as well as -ing and -ed, but when calculating models, different forms of words should be treated as the same. For example, apple and apples, doing and done are the same words, the purpose of extracting stems is to restore to the most basic words. At the same time, we can unify all the capitals. {*Nobody, ancient_people, started, fire*} is processed into {*nobody, ancient_people, start, fire*}

2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation is the most straightforward topic model. LDA intuitively believes that documents are generated from multiple topics. This process is also the document generation process given by LDA. First of all, what we do is limited by the dictionary, that is, the words appearing in the document will not exceed the scope given by the dictionary. For example, the topic "gene" contains a number of words about genes with a high probability, and the topic "evolutionary biology" contains related words of evolutionary biology with high probability.

For each document in the document collection, the LDA model in Figure 1 generates every word in text as follows:

1. For each of T topics, draw words distribution ϕ independently from a symmetric Dirichlet(β) prior;

2. Randomly choose the topic distribution θ in the document from a symmetric Dirichlet(α) prior;

3. Choose a topic z responsible for generating that word, drawn from the θ distribution;

4. Choose word w from the topic distribution θ corresponding to z.

This hierarchical Bayesian model estimate ϕ and θ provides information about the topics that participate in a corpus and the weights of those topics in each document respectively. The core computational problem of topic modeling is the use of observed documents to infer hidden topic structures. Gibbs sampling[6] is used to estimate these parameters. This can also be seen as the inverse of the generative process.

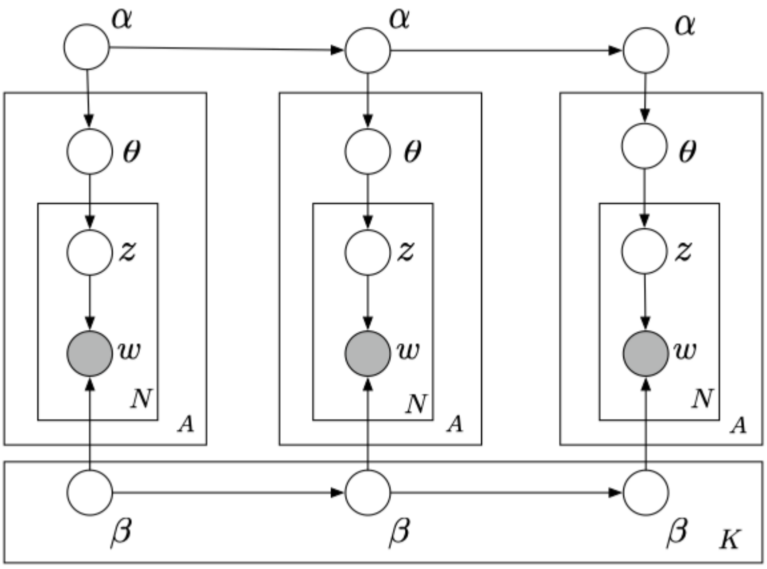

Figure 3: Dynamic topic model

2.3 Author-Topic Modeling

The author-topic model is also an extension of Latent Dirichlet Allocation (LDA), that allows us to learn topic representations of authors in a corpus. LDA describes each document as a mixture of probabilistic topics and each topic as a multinomial distribution over words. While, the Author topic model (ATM) adds an extra author layer over LDA and assumes that the chosen author of the document generates the topic proportion of a given document. As results, every author is associated with multiple documents, and each document can be associated with multiple authors.

The overall process is shown in Figure 2. Different from LDA, the model generates words in each document as follows:

1. From a group of authors a_d, choose an author x who is responsible for generate this word;
2. Associate author x with a distribution θ over topics, chosen from a symmetric Dirichlet (α) prior;
3. Select a topic z according to the author and his topic distribution generated in step 1 and 2;
4. For each of T topics, draw words distribution ϕ independently from a symmetric Dirichlet(β) prior;
5. Generate a word according to the distribution ϕ corresponding to topic z.

The author-topic model subsumes the topic model and author model. Topic models correspond to the case where each document has one unique author, and the author model corresponds to the case where each author has one unique topic. Estimating the parameters ϕ and θ through Gibbs sampling, we obtain information about which topics authors typically write about, as well as a representation of the content of each document in terms of these topics.

2.4 Dynamic Topic Modeling

LDA does not explicitly model temporal relationships. The randomness of the algorithm makes topic different if we train several LDA models for time serials. Dynamic Topic Models (DTM), however, leverages the knowledge of different documents belonging to a different time-slice in an attempt to map how the words in a topic change over time. The Dynamic Topic Model penalizes significant changes from year to year while the beta distributions in Topics over Time are relatively inflexible.

Figure 2 represents the model. The model can be seen as a combination of LDA. The only difference is the change of Dirichlet prior distribution. Specifically, the documents within each time slice are modeled with a topic model of the same dimension, and each topic in time slice t evolves from a corresponding topic in time slice $t-1$. The steps are shown as follows:

1. From the previous time period $t-1$, the Dirichlet distribution $\beta_{t,k}$ of this time period t is generated, indicating the possible text distribution of the topic k:

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 I)\forall k \tag{1}$$

   Where N represents a Gaussian distribution, and the mean is the Dirichlet distribution of the previous time period. This assumption is very intuitive because in each adjacent time period, the distribution of documents and topics does not change significantly.

2. Similarly, the Dirichlet distribution of this time period t is generated from the previous time period $t-1\alpha_t$:

$$\alpha_t|\alpha_{t-1} \sim N(\alpha_{t-1}, \delta^2 I) \tag{2}$$

3. For each of topics, draw words distribution ϕ independently from a symmetric Dirichlet($\beta_{t,k}$) prior;

4. Randomly choose the topic distribution θ in the document from a symmetric Dirichlet(α_t) prior;

5. Choose a topic z responsible for generating that word, drawn from the θ distribution;

6. Choose a word w from the topic distribution θ corresponding to z.

Based on this method, the model ensures topics evolve gradually and can be used to analyze the evolution of (unobserved) topics of a collection of documents over time.

3 Data Source

Data used in this study are derived from two conferences in Learning Science area. The International Society of the Learning Sciences (ISLS) is a professional society which investigates learning itself and to how learning may be facilitated both with and without technology. The society organizes two conferences: CSCL and ICLS. The International Conference on Computer-Supported Collaborative Learning (CSCL), held bi-annually since 1995, focuses on issues related to learning through collaboration and promoting productive collaborative discourse with the help of the computer and other communications technologies. The International Conference of the Learning Sciences (ICLS), held bi-annually since 1996. The contents cover issues and findings across the entire field of the learning sciences. These two conference cover MOOC, codesign, and other well-represented topics under the auspices of the learning sciences field.

There are 1125 pdf papers of CSCL and ICLS from 2014-2018 in our dataset.They consists of year of publication, name of the author with their order, total count of authors, contents of articles, formatted references etc. Data was analyzed to meet the objectives mentioned above. Through the scripts written by Abassi Nour Ghalia (nour.abassi@ep.ch) and Guillain Lonore Valentine (leonore.guillain@ep.ch), we convert these pdf files to txt format, as well as extract author information from metadata. Also, we stored the extracted 2357 authors' data in csv format which can be used in our model later. The data was analysed under Python 3.7 in Jupyter Notebook and we made use of packages including Pandas[7], Gensim[8] and Spacy[9].

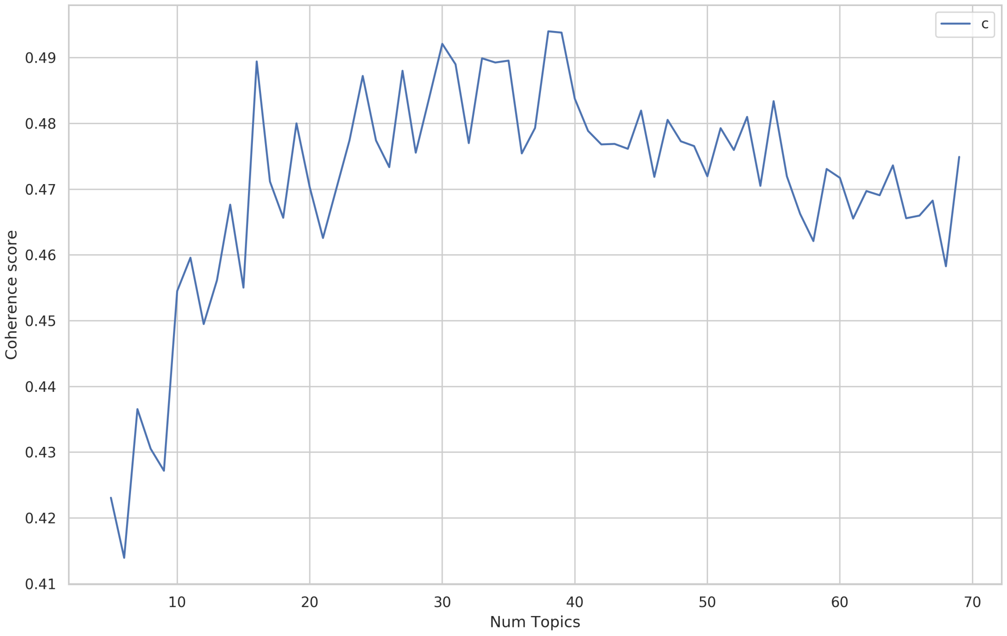

Figure 4: Coherence score of topic model with different numbers of topic

4 Results and Analysis

We generated corpora and dictionaries from 1125 documents and used them to train the LDA, DTM, and ATM models mentioned above. Specifically, each model is individually trained using the optimal number of topics 30. However, the randomness of the algorithms causes the distribution of the keywords they generate are different. Thus, in practice, we train each model multiple times to match most of the topics in the distribution so that the models can share most of the same labels.

In this chapter, we first show how to find the optimal number of topics through grid-search in Section 4.1. Then, in Section 4.2, this parameter is used to train the LDA model and we label each topic based on the word distribution in the topic according to the model's output. We next analyzed the size of each topic in Section 4.3 based on the trained LDA model. In Section 4.4, we focus on the conference level and explore the conference's preferences for the topic and their article receiving pattern. Section 4.5 focuses on analyzing the popularity and future trends of the topic. In Section 4.6, DTM is employed to find the evolution of words within a topic. In the last part, we use ATM to combine the author information with topics and represent the author's topic preferences in the form of a vector. Based on that, the similarity between the authors is quantified by their preference data, and the author's cooperation model can be inferred.

4.1 Choose Topic Number

In the literature corpus, we use LDA, DTM and ATM models to reveal potential knowledge topics for different purposes. The first task of the training model is to determine the number of topics, which is the only parameter that the model needs. Based on the same corpus and dictionary, we assume that the best number of topics for the three models is the same. Hence, we employ the coherence value mentioned in Roder's study[10] to measure the model's performance and reliability. They presented a unifying framework that spans a configuration space of coherence definitions, and the framework is state-of-the-art in terms of topic coherence. The usual approach to finding the optimal number of topics is to enumerate LDA models by the number of topics and pick the ones that give the highest coherence value. Then validate the most suitable one by experts' knowledge to confirm the results, which is called the optimal topic number.

Figure 4 represents the coherence score of the data occurred in these models with different numbers of topics. The result suggests that the data are best accounted for by models incorporating 16, 30, 38, 39 topics. We used manual checks to make sure about the validity and robustness of the model. The model with 16 topics are too general, and its topics mostly appear as a mixture of several real topics. 38 and 39 topics make the model had some meaningless cluster. Whereas, the model with 30 topics has enough fine word distribution in a single topic, although there are still three topics we cannot categorize. Thus, we choose 30 as our optimal topic number and keep it the same in all our models.

4.2 Topic Interpretations

Table 1 displays the top 10 high-frequency words for 27 valid topics in the 30-topic model.The columns following the label are the keywords in each topic and are sorted in descending order according to the likelihood of occurrence in the theme. Referring to Yoon's research[11] and other online documents, topics listed in Table 1 are mostly highly recognizable because many of them have specific words or phrases in the research direction. We select some of these representative topics and explained below.

Topic *Embodied Cognition* contains words like *gesture, object, graph, body, embody, movement.* Embodied cognition emphasis the bodys role in forming cognitive representations[12] which is an interesting topic in learning science research. The body plays a significant role in human cognitive function. Thus researchers are trying to find more insights regarding the role of embodied tools: gesture, action, and analogical mapping. These embodied tools could be leveraged to improve learning in several ways.

Topic *Eye-tracking and Gaze* has words *pair, partner, dyad, dialogue, tutor, gaze, eye_track.* In learning analytics, eye-tracking could prove to be useful to understand the cognitive processes underlying dyadic interaction; in two contexts: pair program comprehension and learning with a Massive Open Online Course (MOOC). The first context is a typical collaborative work scenario, while the second is a particular case of dyadic interaction, namely the teacher-student pair[13]. Both contexts are related to a pair, thus the first several words are explainable in this topic.

Topic *Scaffolding* addresses learning issues with the high frequent terms of *simulation, experiment, virtual, physical, scaffold.* In education, scaffolding refers to a variety of instructional techniques used to move students progressively toward stronger understanding and, ultimately, greater independence in the learning process[14]. People usually mention scaffolding and simulation environments together, so *simulation* is the highest.

Surprisingly, we find a topic *Expertise.* The topic contains words like *expert, novice* which are not meaningful if looking at it individually. However, the topic is about the expert role in learning science research, and it supports many other areas. For instance, topic *Scaffolding* mentioned before, call scaffolder in its context as an expert who might need to manage and control for frustration and loss of interest that could be experienced by the learner[15]. Also, as a more general topic *Problem Solving*, it emphasizes the knowledge and skills possessed by experts but not by novices. Both topics are from other topic but they are related to expertise. Therefore, it is possible that a document is dominated by *scaffolding* and also has a certain proportion of *expertise.*

Other topics in Table 1 are not discussed in detail due to space constraints, however, they also suggest explicit research themes related to learning science. These established topics confirm to our prior knowledge of the subject.

Topic Label	High-frequency Terms									
Sociocultural Context	theoretical	culture	cultural	relation	symposium	mediate	situate	notion	dimension	sociocultural
Embodied Cognition	gesture	object	graph	body	embody	movement	dynamic	physical	embodiment	motion
Epistemological Cognition	argumentation	argument	claim	source	epistemic	reasoning	quality	criterion	scientific_reason	justification
Critique and Revision	critique	guidance	revise	integration	revision	unit	rubric	energy	step	prompt
Feedback and Assessment	feedback	instructor	assignment	quality	section	active	traditional	college	receive	grade
Professional Development	lesson	reflection	plan	professional_development	pedagogical	whole_class	planning	frame	enactment	professional
Second Language Learning	text	read	language	word	literacy	reading	document	sentence	comprehension	english
Problem Solving	phase	solution	problem_solv	transfer	solve	domain	script	interactive	step	guidance
Scaffolding	simulation	experiment	virtual	physical	scaffold	lab	scaffolding	run	investigation	prompt
Eye-tracking and Gaze	pair	partner	dyad	dialogue	tutor	gaze	attention	step	quality	eye_track
Maker	youth	stem	site	mentor	library	maker	local	medium	city	pathway
Coding	network	code	analytic	cluster	qualitative	link	word	visualization	quantitative	behavior
Learning Environments	scientist	disciplinary	phenomenon	investigation	water	authentic	light	uncertainty	frame	unit
Museum	representation	exhibit	museum	map	visitor	physical	visual	location	informal	image
Knowledge Building	knowledge_build	collective	epistemic	contribution	reflection	object	productive	advance	thread	creation
Complex Systems	unit	progression	component	complex_system	biology	population	modeling	map	ecosystem	scale
Multiple Representations	condition	score	item	comparison	receive	control	prompt	average	session	experimental
Online Course	communication	post	mooc	quality	contribution	comment	chat	message	forum	social_media
Conceptual Learning	physics	energy	misconception	element	force	interview	particle	representation	phenomenon	earth
Metacognition	behavior	regulation	awareness	positive	motivation	affect	emotion	self_efficacy	agent	metacognitive
Leadership	team	network	implementation	innovation	leadership	meeting	leader	policy	district	organizational
expertise	engineering	expert	novice	expertise	engineer	artifact	designer	solution	failure	studio
Identity and Equity	identity	position	power	woman	cultural	equity	history	historical	gender	agency
Mathematic	mathematic	mathematical	math	reasoning	function	algebra	productive	attend	event	middle
Software	user	digital	access	app	session	software	application	display	device	tablet
Workshop and Robot	stem	programming	workshop	computational	robot	code	compute	high_school	computing	scratch
Games and Gaming	game	parent	family	player	medium	digital	gaming	facilitator	card	character

Table 1: Top 10 high-frequency terms for each topic in the 30-topic model

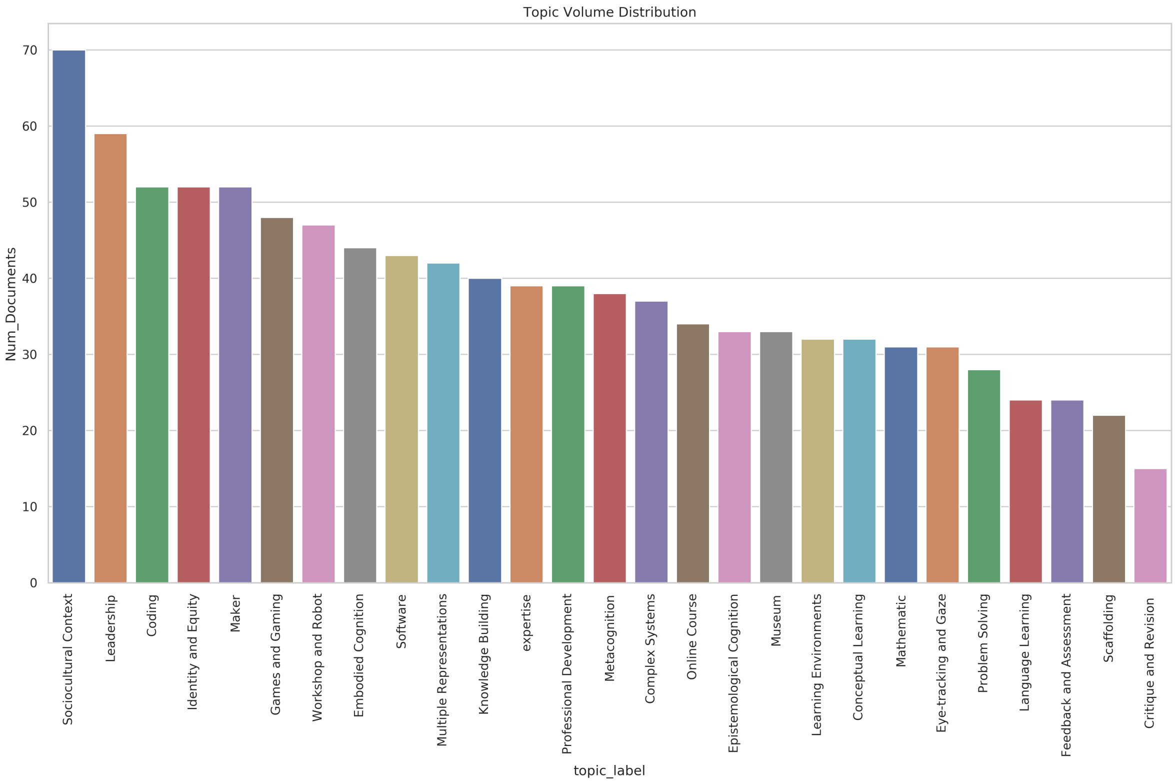

Figure 5: Documents Distribution over last 10 years

4.3 Topic Volume

Figure 5 presents an overview of volumes distribution among topics. y-axis is the number of documents having dominant topic on x value. For a document, dominant topic refers to the topic assigned highest proportion by 30-topic model. Integrating topic proportions for all the text in the corpus, we obtained a topic distribution for the whole corpus, as shown in the Figure 5. The results indicates that the three highest-frequency research topics among ICLS and CSCL are *Sociocultural Context* (70), *Leadership* (59), *Coding* (52), while the three lowest-frequency research topics are *Feedback and Assessment* (24), *Scaffolding* (22), *Critique and Revision* (15).

These Results is explained as the higher volume a topic has, the more widely it is discussed, or the more attention it is paid to in latest four years. For example, the sociocultural is a popular direction and was widely discussed in the earlier (before 2010) literature. Thus, experts with relevant background are willing to conduct research based on it. In other words, researchers' large amount and topic's own widespread make this topic frequently appearing in documents. On the contrary, small-volume topics like *scaffolding* have two possible explanations for their volume. One is that the topic is too specific as a result few people are working on it. The other reason is that this topic is a new area without accumulation, and its difficulty prevents volume increasing.

Interestingly, topics associated with computer and network such as *Coding, Maker, Workshop and Robot, Software* hold document amount more than average. It is reasonable because the technological revolution has changed the world as well as people's learning approaches. Therefore, more and more people are willing to learn computer network related technologies, which has led to the rapid development of relevant education. On the other hand, with the popularity of technology, researchers are more willing to change the way learners learn through the effectiveness of digital media. And the usage of computer and other technology has become

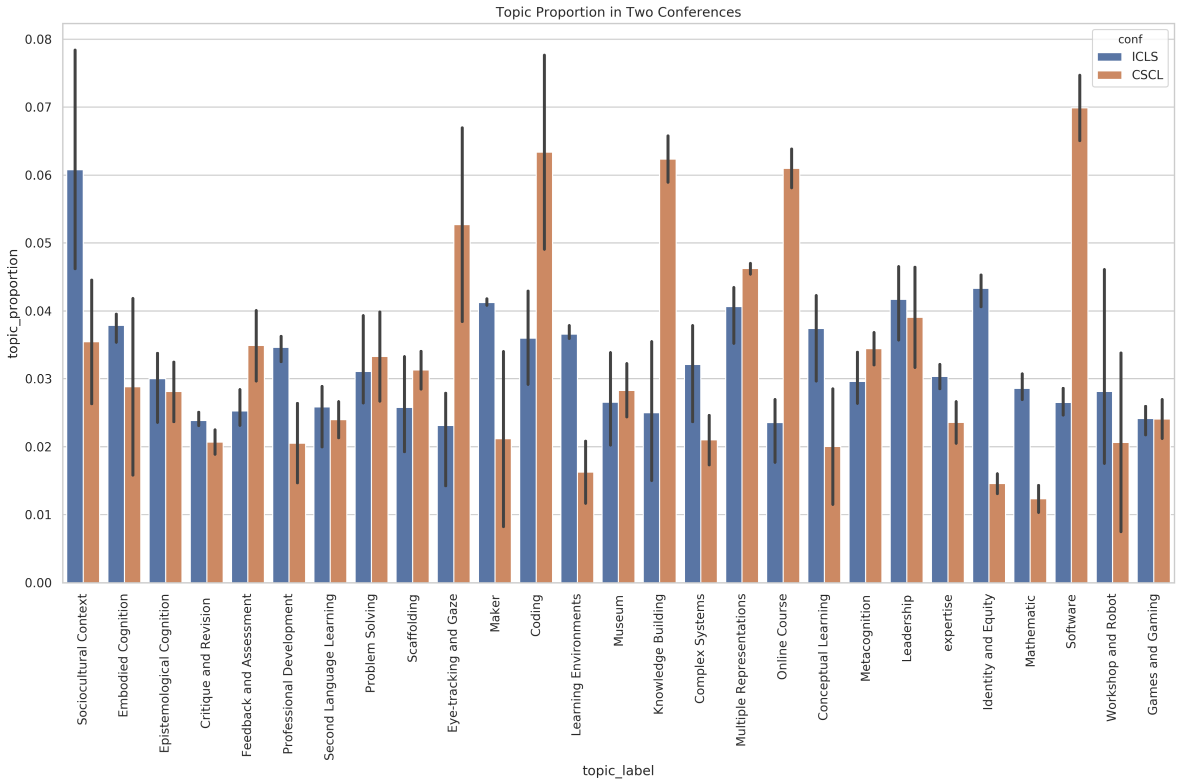

Figure 6: Topic proportion in ICLS and CSCL

an inevitable part of the current education system.

4.4 Topic Preference of Conference

In the conference level, we are interested in the conference's preference. Therefore, instead of using the entire corpus as an input to the model, we separate the two meetings and make our own corpus. The theme model uses the annual corpus of each meeting as input, then calculates the corresponding topic distribution, and then integrates the four years of data. This approach allows us to estimate results more properly with bootstrapping. Figure 6 is the bar plot with error bar. The error bar is the confidence interval where the 95 % of the proportion lies in.

The figure compares the topic distribution of the two conferences of ICLS and CSCL by integrating four years of data. The results show that both of them have obvious topic preference. According to Figure 6, topics in CSCL like *Coding, Online Course, Software* have a much larger proportion than that in CSLS. Whereas, CSLS prefers topics *Sociocultural Context, Identity and Equity* more than the other. As we mentioned in the introduction, CSCL, Computer-Supported Collaborative Learning, is more about technology related topics so that this conference focus on learning and computer. On the contrary, ICLS is about general learning science, and coverage is what matters. Moreover, the standard deviation of CSCL is 0.016, and std is 0.008 in ICLS. This means that ICLS has a more uniform preference for different topics, while in CSLS, significant differences in proportions between topics can be observed.

Based on the topic distribution, we can infer the taste of the conference and predict the probability that a paper will be accepted by the conference. Precisely, the topic distribution of a paper that needs to be predicted is first calculated by the same topic model, and the distribution is compared with that of the conference. If their topic distribution is similar or the most prominent topic of the paper matches the topic that the conference likes, then this paper

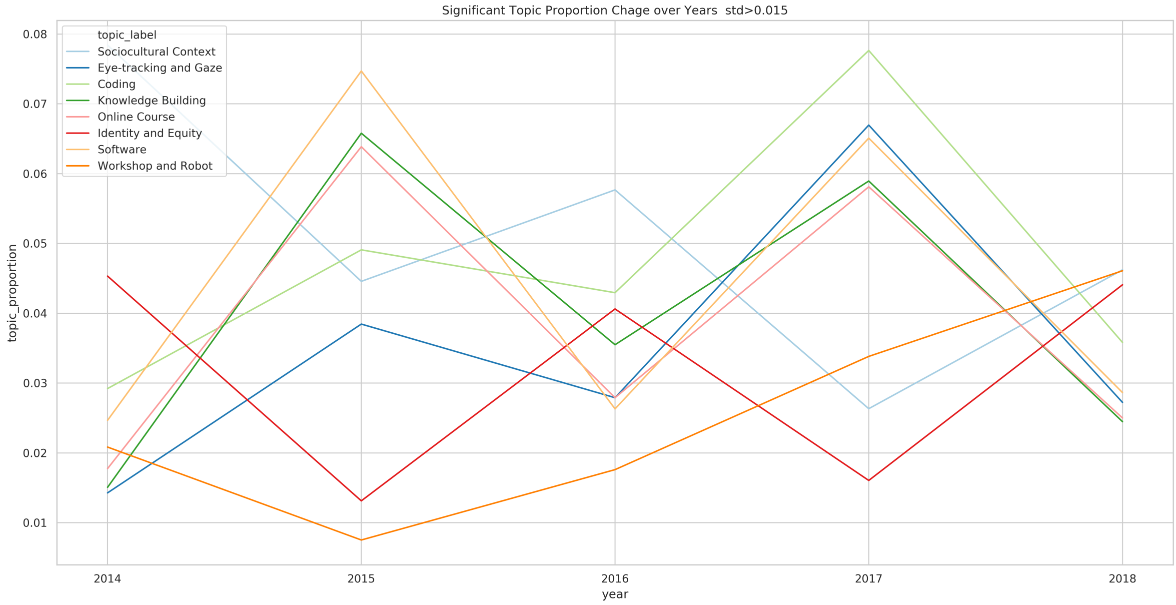

Figure 7: Topic proportion change over years with std>0.015

is likely to be accepted.

4.5 Topic Popularity and Trend

In order to explore the trend of the topic, we looked at the changes in the distribution of topics over the years. We treat all the articles in the two conferences as a whole and input them into the trained LDA model. Then get the topic distribution for each year as an output. By analyzing the changes in the topic over the four-year period, we have drawn a line plot (Figure 8).

Figure 7 shows 4 years' topic percentage changes. In order to increase the readability of the graph, we filter the topic by the degree of change. If a topic's data of 4 years has a standard deviation larger than a threshold, the change of this topic can be seen as significant. Useful and meaningful information is usually carried in such significant changes. In our case, we set the threshold as 0.015 to control line amount less than 10. The plot reveals that, in most case, the lines are in M and W shape, which means every other year's data are not continuous. Consider the fact we have two conferences: ICLS is holded in even year, CSCL in the odd year. Due to two conferences has different preference and starting point, the percentage is up and down frequently.

To eliminating diverse between conference, we set the first two years (2014, 2015 in our case) as original points corresponded to conferences and set their increments to zero. And for years after, calculate proportion increment by subtracting data of the last session. For example, the topic proportion of 2016 should subtract that of 2014. The increment can be positive and negative, respecting to the topic's popularity increase or decrease of this year. Then we accumulated sum each topic's increment, which is shown in Figure 8. The y-axis reflects the accumulated change from original point. In other words, if this line rises, it proves that this topic is being paid more and more attention, and vice versa.

As we can see from Figure 8, the proportion of topic *Workshop and Robot* shoots up after 2016, peaking at 0.044 at 2018. The years between 2015 and 2017 witnesses a moderate growth in the topic *Eye-tracking and Gaze, Maker, Leadership*. Otherwise, percentage of topic *Sociocultural Context, Conceptual Learning, Embodied Cognition* descends drastically in four years,

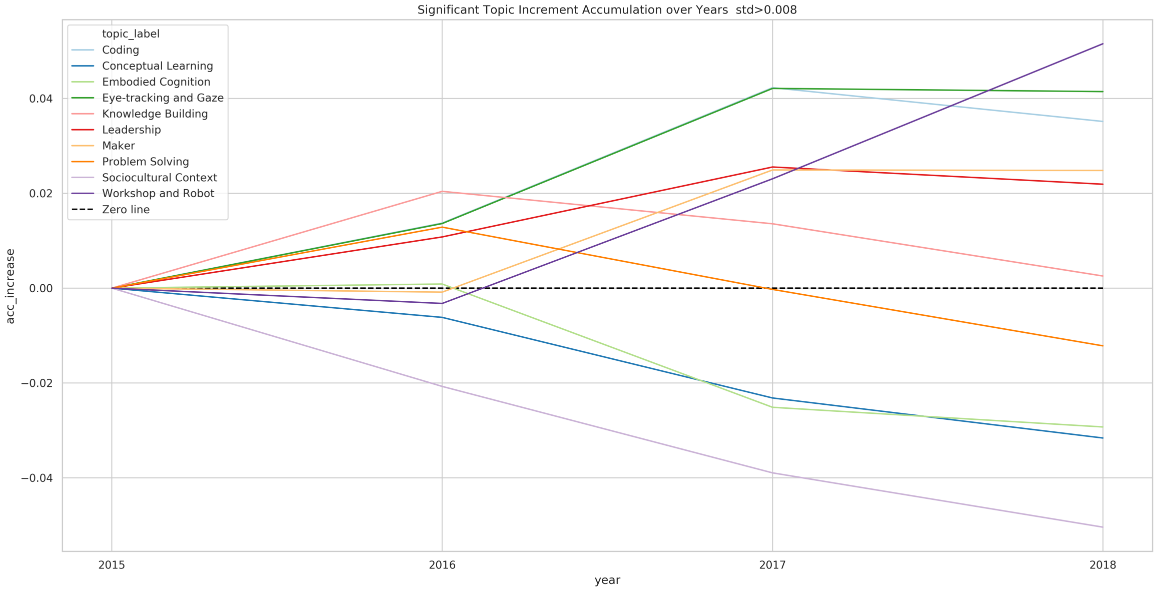

Figure 8: Accumulation of topic proportion change over years

although two of them have high volume in Figure 5. This shows that in the field of learning science, the focus of most scholars' research is from the traditional cognitive-related direction to the technology-related direction. Correspondingly, more and more research related to computer and networks.

4.6 Topic Evolution

In the DTM model, we still use 30 as the number of topics. However, due to the randomness of algorithm, the topic generated by the model is somewhat different from that mentioned in the previous table. So we chose a similar topic model so that most of the tags can be shared to make the analysis concise. We illustrate two typical topics with explanation below. Noted that the lines in the plot are filtered by standard deviation larger than 0.0003 to detect significant change of words. If the word not in plot, it means the the line does not change much during four years.

4.6.1 Eye-tracking and Gaze

In Figure 9, word *dyad*, as the most common word in this topic, first keep raising over two years and peaking in 2016. Then, the number drop until now. And word *tutor* increase slightly. Their unstable values could be interpreted as changes are happening inside the topic since keywords hardly disappear. It is not surprising that the word *misconception* starting from low-value area proliferates. This is not a new conception in the topic, but, in fact, it draws people's attention in this area since 2015, and it is likely more and more people research it.

Furthermore, there is a word *fluency* whose percentage sinks to 0.002 in 2015 and is removed after that. In terms of eye-tracking in learning science, fluency often mentioned with reading fluency together. Thus, the disappearance of the word suggests that researchers do not may too much attention as before.

4.6.2 Identity and Equity

Identity an equity are essential issues in learning science. Note the rising use of word *stem* coupled with the decline of the use of word *culture*. With the growing importance and popularity

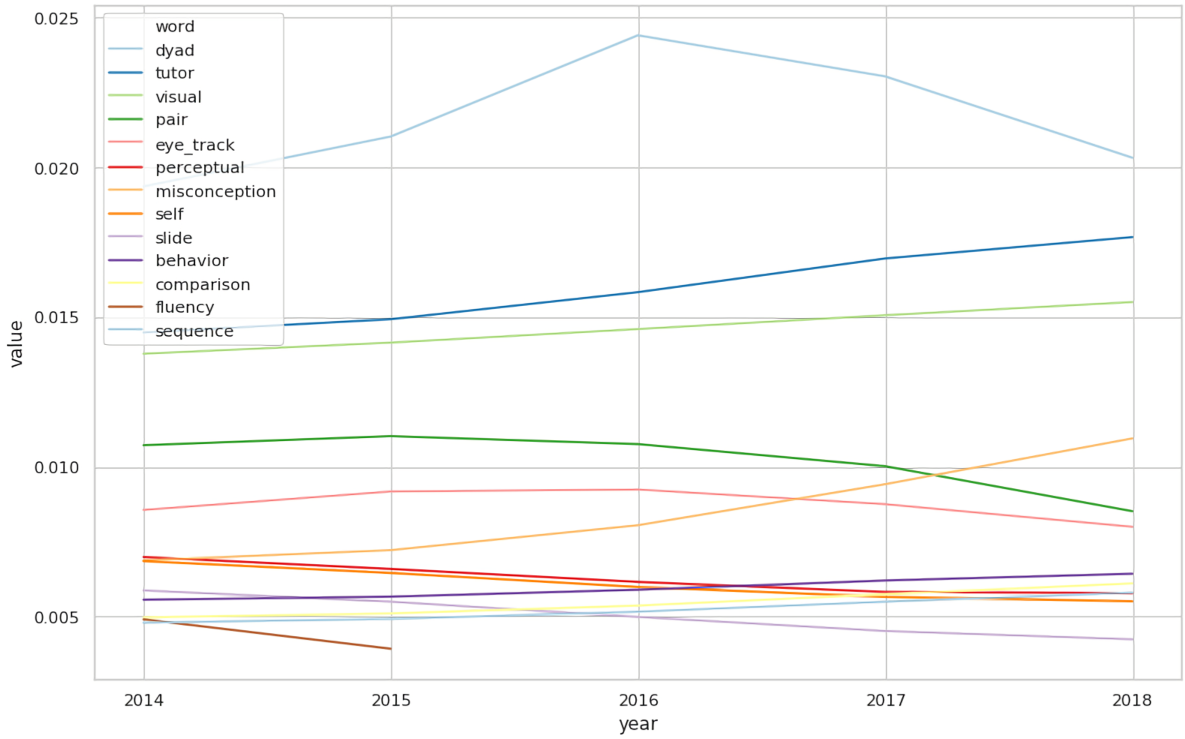

Figure 9: Topic evolution of 'eye-tracking and gaze'

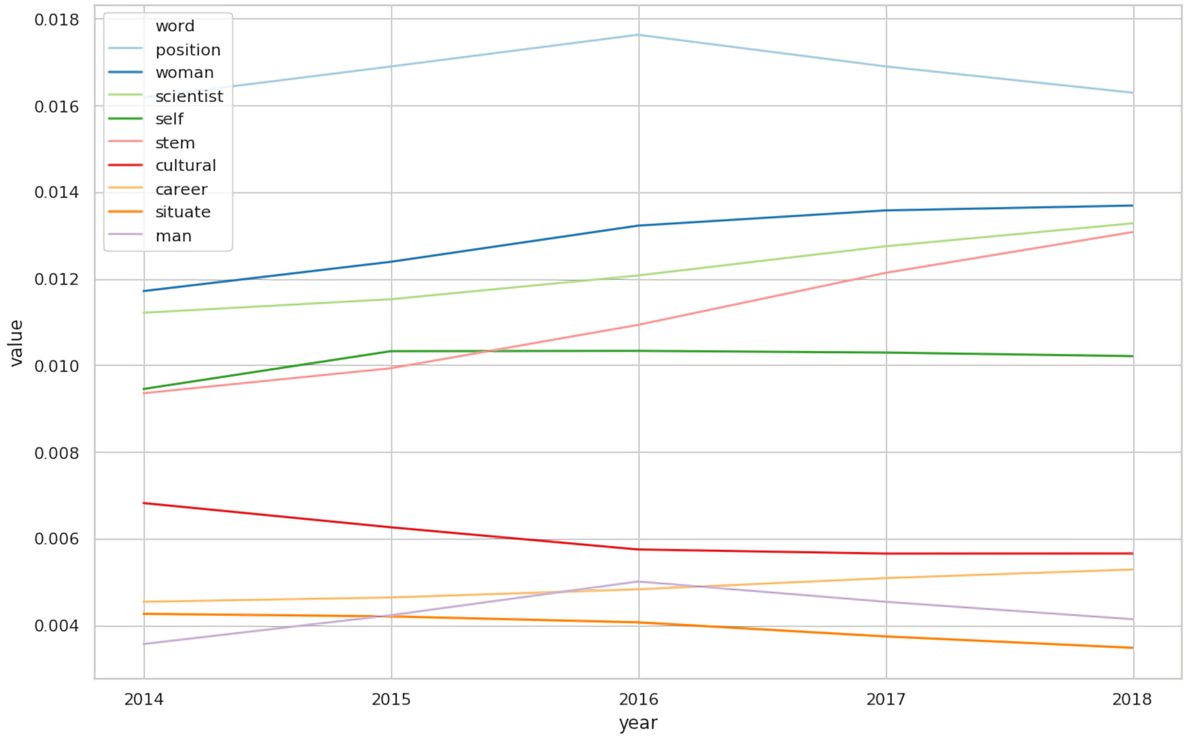

Figure 10: Topic evolution of 'Identity and Equity'

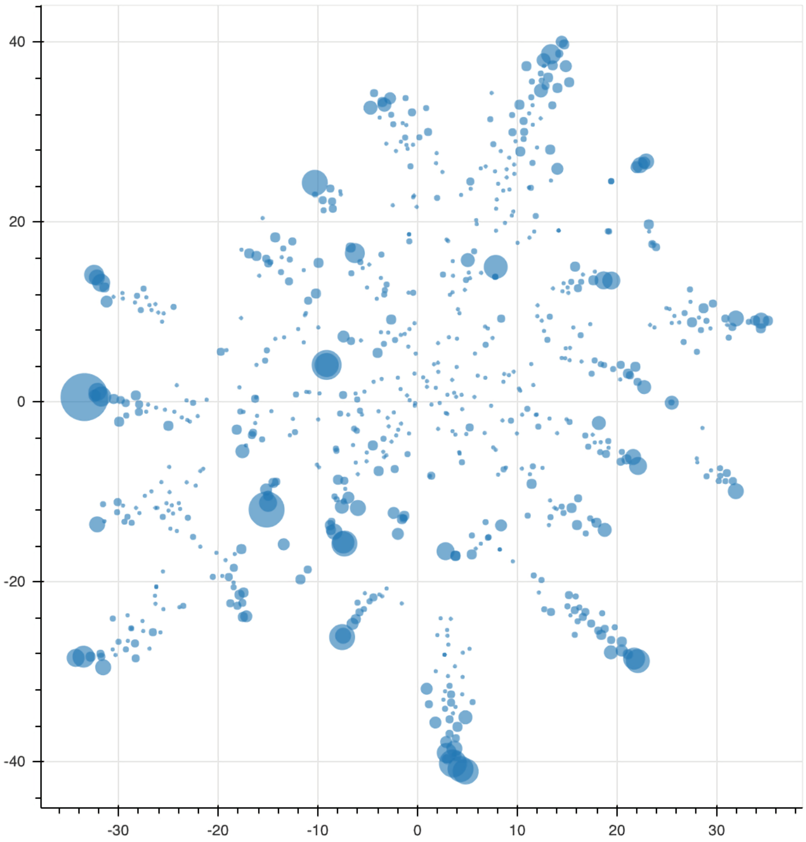

Figure 11: Author topic preference in 2D

of STEM major, people nowadays face equity problem related to STEM more than before, such as providing equity with STEM academic units to ensure student and research success.

Although keyword *man* and *woman* change almost synchronously *woman* have much higher proportion in this topic than that of word *man*. The difference reveals that woman faces more issues corresponded to equity and identity than man and researchers focus more on them.

4.7 Author Topic Preference and Cluster

This part is mainly about the author-topic model. Giving corpus and author of documents, we can acquire topic preference vector of each author. The size of the vector is the same as the topic number and values are between 0 to 1, which denote as the possibility the author would choose the topic. Based on the vectorized author preference, we produce Figure 11. The intuitive author-topic representation takes all the author-topic distributions and embeds them in a 2D space. To do this, we reduce the dimensionality of this data using t-SNE. t-SNE is a method that attempts to reduce the dimensionality of a dataset while maintaining the distances between the points. That means that if two authors are close together in the plot below, then their topic distributions are similar. The circles in the plot are individual authors, and their sizes represent the number of documents attributed to the corresponding author. Large clusters of authors tend to reflect some overlap in interest.

At about (31, 0) on Figure 11, we have a cluster of *problem solving* researchers like Rummel Nikol, Aleven Vincent and Kollar Ingo. They all prefer working on problem solving related area. Topic *Eye-tracking and gaze* is located around (-30, 12) which has authors like Dillenbourg Pierre, Schneider Bertrand and Sharma Kshitij. At the bottom of the plot, (5, -40), Many Asian authors aggregate together and enjoy the same interest of *knowledge building*.

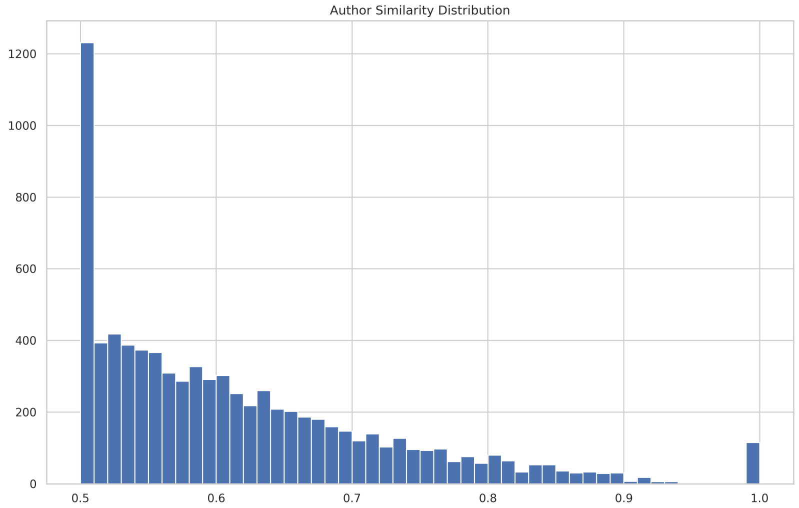

Figure 12: Author similarity distribution among 2357 authors

Authors who belong to the same cluster in the picture have the same topic preferences, so they possibly know each other and even work together. On the other hand, because the distance in the graph represents the degree of difference, if the gap between the two clusters is close, the authors' topic preference difference between them is not large as well, so that there is the possibility of cooperation between two close clusters.

4.8 Author Similarity Exploration

We calculate similarity between authors based on Hellinger distance. The Hellinger distance is a natural way of measuring the distance (i.e. dis-similarity) between two probability distributions. Its discrete version is defined as

$$H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K}(\sqrt{p_i} - \sqrt{q_i})^2} \tag{3}$$

where p and q are both topic distributions for two different authors. The similarity is defined as

$$S(p,q) = \frac{1}{1 + H(p,q)} \tag{4}$$

Since p and q are less than 1, the Hellinger distance must less than 1 as well. So, the similarity value is between 0.5 and 1. If the similarity of two authors is 1, we say they have perfect same preference. If the value is around 0.5, the author pair has nothing in common.

We calculate the similarity between every pair of authors, and Figure 12 shows the similarity value distribution. The graph is highly right-skewed, and there is a large amount of author pair with similarity below 0.51. It is known that every person is unique, and he is not likely to be similar to most of the others. Thus, we assume if a pair of authors has similarity value less than 0.51, the two are distinct, and their preference, as well as research areas, are far from each other. Also, if an author pair's similarity larger than 0.7, these two authors are similar. If multiple authors have completed an article together, then we call there is a collaboration between any two of these authors. For each cooperation, if the similarity between the two authors is less than 0.51, that is, the two authors have distinct research preference, then this cooperation is a cooperation between research directions and it is a cross-work.

Based on this assumption, we would answer three questions below:

1. Which article is a cross-work product?
2. Who is versatile on topics?
3. Who prefer working with different guys?

4.8.1 Which article is a cross-work product?

This question is raised because we want to analyze the typical characteristics of articles written by authors with different topic preferences. If two author write an article together, we say this is a cooperation. For example, if A, B, C are three authors of an article, then there are three times of cooperation: (A, B), (B, C), (A, C). Therefore, for each article, we count cooperation between two authors having similarity lower than 0.51 and divided by the total collaboration number. The ratio that we get represents the diversity of the article. The results get rid of articles with less than 4 authors and are sorted by diversity ratio.

In the Table 2, we can find that most of the top-ranked articles involve multiple directions in learning science. For example, "Examining the Role of Unpacking 3-Dimensional Teaching and Learning in Museum-Based Professional Development", this article covers the knowledge of multiple directions such as *museum* and *professional development*, so the authors almost have different topic preferences making them dissimilar from each other. Such interdisciplinary collaborations generally produce surprising results.

It is worth mentioning that if the number of authors of an article is too large, it will disturb our analysis. For example, "Real-Time Visualization of Student Activities to Support Classroom Orchestration" in the Table 2 is a discussion record with 15 authors. Together, authors projects from various contexts with similar goals. Obviously, the greater the number of authors, the greater the likelihood that they will come from different fields, leading to unreliable statistical results.

4.8.2 Who is versatile?

It is common for researchers to work in multiple areas. But whose researches cover the broadest range of topics? Firstly, we label articles based on topic distribution of each document from the LDA model, then get the dominant topic and count the number of distinct dominant topics with respect to one author's articles. The result is the number of topics that the author used to research on. Lastly, the author's versatile score is calculated as dividing the number of different topics by his total number of articles. The higher versatile score is, the higher tendency the author has to research on different areas. This score not only reflects the breadth of the researcher's work, but also somehow indicates the role of the researcher. For example, to analyze their experiment results, researchers often collaborate with statistical experts, even the experts are not familiar with their areas. It means if some people often work with people in different fields, they are probably data scientists.

From Table 3, We can see that most people in the table are involved in more than five different areas. Surprisingly, Dornfeld Catherine wrote a total of six articles, and each of which belongs to a different topic. Also, Danish Joshua A. wrote 15 articles that involved 10 topics. Due to the limitations of the information, we cannot fully analyze the conclusions contained in the table. But what we can know is that if an authors score is higher, his research direction will be more general and easy to apply to various areas.

Title	Authors	# of authors	# of cooperation	# of cross-work	cross_ratio
Between the Lines: The Role of Curriculum Materials and Teacher ...	[’Sherwood, Carrie-Anne’, ’Allen, Carrie D.’, ’Moorthy, Savitha’, ...]	6	15	13	0.87
Professional Development of Science Teachers in Underserved Communities ...	[’Fuhrmann, Tamar’, ’Fernandez, Cassia’, ...]	4	6	5	0.83
Examining the Role of Unpacking 3-Dimensional Teaching and Learning ...	[’Vaishampayan, Gauri A.’, ’Price, Aaron’, ...]	6	15	12	0.8
Secondary Students Evaluation of Inappropriate Strategies of Reasoning ...	[’Ma, Guanzhong’, ’van Aalst, Jan’, ’Chan, Carol’, ’Wang, Jing’]	4	6	4	0.67
How to Enjoy Writing Papers: Supporting Literature-Based Inquiry ...	[’Eberle, Julia’, ’Schnfeld, Tim’, ’Arukovic, Selma’, ’Rummel, Nikol’]	4	6	4	0.67
Using Machine Learning Techniques to Capture Engineering Design ...	[’Bywater, Jim P’, ’Floryan, Mark’, ’Chiu, Jennifer L.’, ’Chao, Jie’, ...’]	9	36	22	0.61
Beyond Just Getting Our Word Out: Creating Pipelines From Learning ...	[’Jacobson, Michael J.’, ’Lund, Kristine’, ...]	6	15	9	0.6
A Qualitative Exploration of Self- and Socially Shared Regulation ...	[’Hensley, Lauren’, ’Cutshall, Jessica’, ’Law, Victor’, ’Xie, Kui’, ’Lu, Lin’]	5	10	6	0.6
Data Moves: Restructuring Data for Inquiry in a Simulation and ...	[’Wilkerson, Michelle’, ’Lanouette, Kathryn’, ’Shareff, Rebecca’, ...]	9	36	20	0.56
Real-Time Visualization of Student Activities to Support Classroom ...	[’Tissenbaum, Mike’, ’Matuk, Camillia’, ’Berland, Matthew’, ...]	15	105	54	0.51
Enhancing Online Structured Dialogue During Teaching ...	[’Mochizuki, Toshio’, ’Kitazawa, Takeshi’, ’Oshima, Jun’, ...]	5	10	5	0.5
Using a video-based approach to develop pre-service science teachers...	[’Chan, Kennedy’, ’Cheng, Ka Lok’, ’Chan, Carol’, ’Yung, Benny’]	4	6	3	0.5
Criss Crossing Science Domains in Knowledge Building Communities ...	[’Khanlari, Ahmad’, ’Zhu, Gaoxia’, ’Costa, Stacy’, ’Scardamalia, Marlene’]	4	6	3	0.5
Capturing Qualities of Mathematical Talk via Coding And Counting	[’Heyd-Metzuyanim, Einat’, ’Tabach, Michal’, ’Nachlieli, Talli’, ...]	4	6	3	0.5
Networks in Small-Group and Whole-class Structures in Large Knowledge ...	[’FENG, Xueqi’, ’van Aalst, Jan’, ’Chan, Carol’, ’Yang, Yuqin’]	4	6	3	0.5
From Computational Thinking to Computational Action: Understanding ...	[’Tissenbaum, Mike’, ’Sherman, Mark A’, ’Sheldon, Joshua’, ’Abelson, Hal’]	4	6	3	0.5
Uncovering Teachers Pedagogical Reasoning in Science Discussions	[’Clarke, Sherice’, ’Gerritsen, David’, ’Grainger, Rebecca’, ’Ogan, Amy’]	4	6	3	0.5
Design-Activity-Sequence: A Case Study and Polyphonic Analysis of ...	[’Wheeler, Penny’, ’Truan-Matu, tefan’, ...]	4	6	3	0.5
Comprehension SEEDING: Providing real-time formative assessment ...	[’Wylie, Ruth’, ’Chi, Michelene T. H.’, ’Talbot, Robert’, ’Dutilly, Erik’, ...]	7	21	10	0.47
Examining Parent-Child Communication and Affect During Tabletop ...	[’Missall, Kristen’, ’Nanda, Salloni’, ’Courshon, Caitlin’, ...]	7	21	10	0.47

Table 2: Diversity in articles (sorted by cross_ratio)

Name	# of topics	# of docs	versatile_score
Dornfeld, Catherine	6	6	1.00
Polman, Joseph L.	9	10	0.90
Shapiro, R. Benjamin	6	7	0.86
Levy, Sharona	5	6	0.83
Hakkarainen, Kai	5	6	0.83
Rose, Carolyn Penstein	9	11	0.82
Lee, Victor R.	7	9	0.78
Borge, Marcela	6	8	0.75
Acosta, Alisa	5	7	0.71
Wilensky, Uri	5	7	0.71
Hovey, Christopher M	5	7	0.71
Bang, Megan	5	7	0.71
Hmelo-Silver, Cindy E.	5	7	0.71
Chen, Bodong	5	7	0.71
Fischer, Frank	9	13	0.69
Basu, Satabdi	4	6	0.67
Danish, Joshua A.	10	15	0.67
Olsen, Jennifer K.	4	6	0.67
Sayre, Eleanor C.	4	6	0.67
Svihla, Vanessa	4	6	0.67

Table 3: Author versatile score descending order

4.8.3 Who prefer working with different guys?

Cooperation is a very common thing in an academic community. Some scholars like to work with people who have the same research direction, and some are more inclined to reach different people for new inspiration. This part is aiming at find the group of people who prefer working with different guys.

In the first part of this section, we know authors' cooperation and its similarity of author pair. So, by iterating an author's all cooperation, we check whether the cooperation is a cross-work or not. If the similarity is lower than 0.51, we say it is a cross work. Otherwise, it is a similar work. After calculate cross work amount and similar work amount, the results are shown in Figure 13. In order to eliminate interference factor, the plot only shows author having cooperation times larger than 10. If the total cooperation times of a author are too small, he would be easy to reach high ratio and disturb the results.

According to the results in Figure 13, Many authors, including Hurwich Talia and McNamara Danielle, have high cross-work ratio above 0.7 and their total work time is less than 40. And for authors having more than 40, their ratio is generally not so high and raking behind at least 15. This can be interpret as large authors are often professors or professional researchers and they have their own specific direction. Although they might work with his students as supervisors which increases cross work ratio, they mainly focus on his own field. PhDs like Talia raking 1, have more chance to work with different guys in related fields and are more likely to be exposed to new fields.

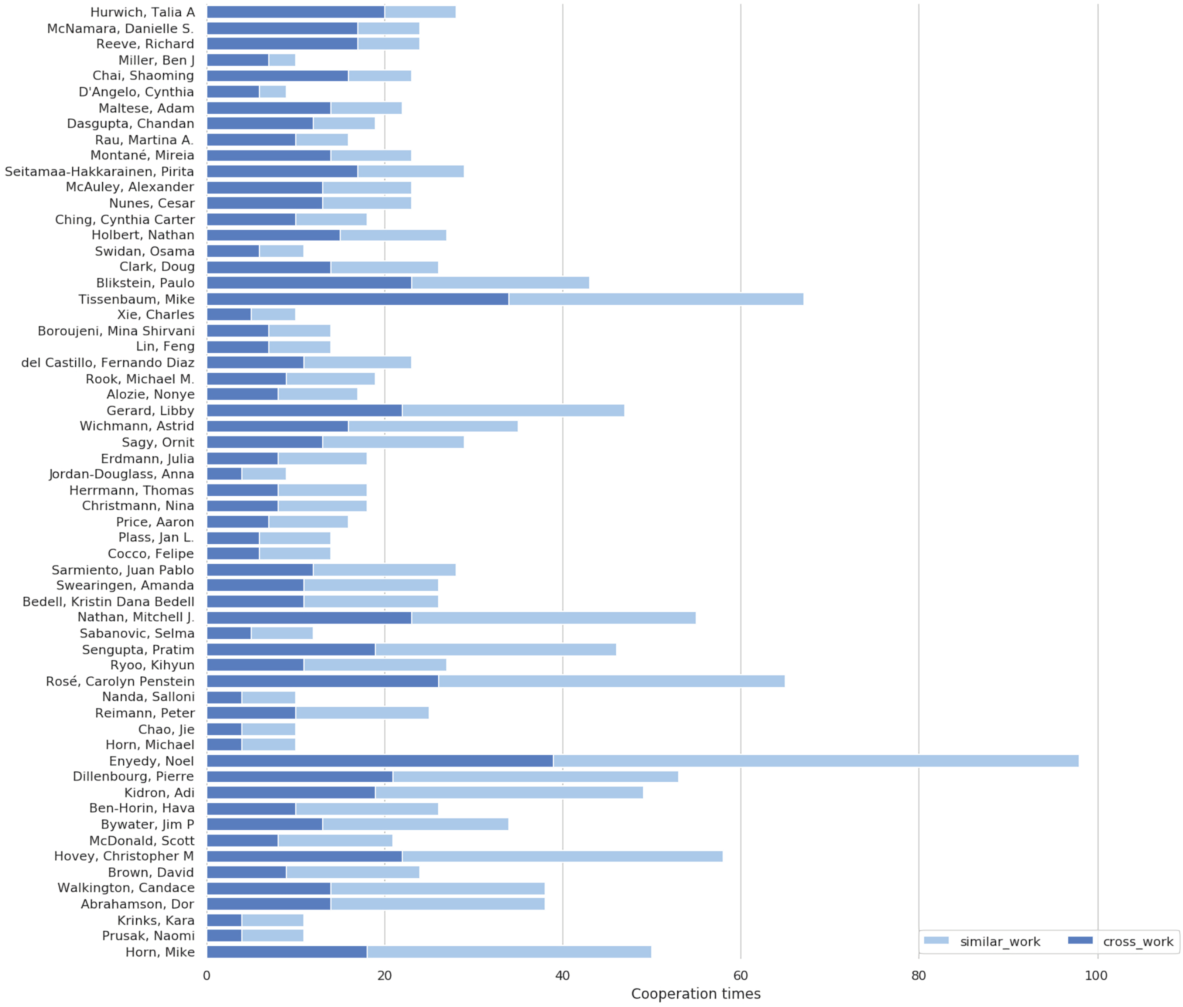

Figure 13: Cooperation and cross-work of Authors (Sorted by cross-work ratio)

5 Conclusion and Further Development

This report analyzes a total of four years of academic articles in ICLS and CSCL using three topic models. We first used grid-search to find the optimal number of topics. Based on this parameter, we trained three models of LDA, DTM, and ATM. First, we discuss the research status of the field by the size and distribution of the topic. After that, we compared the topic preferences of the two conferences and summarized what kind of paper is more likely to be accepted. Then the temporal feature was introduced by DTM to analyze the evolution of the topic in 4 years. Finally, the authors' topic preference is calculated by the ATM model, and we cluster authors according to their similarity between each other. Based on that, we find out some special authors.

Overall, the reliability of the results and the stability of the model need to be improved. Using a larger data set can significantly improve the results of the model, especially for DTM. The short duration of four years is not enough to judge the trend of a keyword in one topic, which means our analysis is only a vague guess about the overall development. More than ten years' data is preferred for temporal analysis. On the other hand, the number of articles in these four years is uneven. 2014 has a much larger number of articles than that in 2015 and 2016. This fact makes time-related analysis not so reliable. Another problem is that topics in models are inconsistent because of the randomness of the training. This may be solved by selecting a larger number of topics like 100 and filtering out certain meaningful topics manually.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” tech. rep.

[2] T. Hofmann, “Probabilistic latent semantic indexing,” in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’99, (New York, NY, USA), pp. 50–57, ACM, 1999.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] D. M. Blei and J. D. Lafferty, “Correlated Topic Models,” tech. rep.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical Dirichlet Processes,” tech. rep., 2005.

[6] “Markov Chain Monte Carlo and Gibbs Sampling,” tech. rep.

[7] W. McKinney, “Data structures for statistical computing in python,” in *Proceedings of the 9th Python in Science Conference* (S. van der Walt and J. Millman, eds.), pp. 51 – 56, 2010.

[8] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. `http://is.muni.cz/publication/884893/en`.

[9] M. Honnibal and M. Johnson, “An improved non-monotonic transition system for dependency parsing,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 1373–1378, Association for Computational Linguistics, September 2015.

[10] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,”

[11] S. A. Yoon and C. E. Hmelo-Silver, “What Do Learning Scientists Do? A Survey of the ISLS Membership,” *Journal of the Learning Sciences*, vol. 26, no. 2, pp. 167–183, 2017.

[12] S. M. Weisberg, N. S. Newcombe, and J. J. Gibson, “Embodied cognition and STEM learning: overview of a topical collection in CR:PI A brief primer on embodied cognition,”

[13] E. Éducative, “Gaze Analysis methods for Learning Analytics THÈSE N O 6696 (2015) ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE PRÉSENTÉE LE 6 NOVEMBRE 2015 À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS,” tech. rep.

[14] J. Van De Pol, M. Volman, F. Oort, J. Beishuizen, and J. B. Nl, “The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support,” *Instructional Science*, vol. 43, pp. 615–641, 2015.

[15] D. Wood, J. S. Bruner, and G. Ross, “THE ROLE OF TUTORING IN PROBLEM SOLVING*,” tech. rep., 1976.