Twenty years of building simulation trends: Text mining and topic modeling of the *Bldg-sim* email list archive

Clayton Miller¹, Matias Quintana¹, Jason Glazer²

¹ Building and Urban Data Science (BUDS) Lab, Department of Building, School of Design and Environment (SDE), National University of Singapore (NUS), Singapore

² GARD Analytics, Inc., Arlington Heights, IL, USA

Abstract

The recent growth of the building performance research community has been in parallel with the *onebuilding.org Bldg-sim* email list. This list was formed in 1999 and steadily grew into a major venue for building simulation community announcements, discussions, and questions and answers (Q&A). This paper presents an analysis of the *Bldg-sim* email archive to determine how traffic has grown over the years and what general trends have come and gone in this particular user community. We use a text mining approach to find the most prominent topics over the years and create trend metrics from the frequency of the most common words from those topics. The results illustrate the relative rise and fall of various software tools, organizations, and simulation topics from the portion of the simulation community who used the email list. Visualizations showing the trends are presented and discussed. The paper also discusses the generalizability of results as it should be noted that the users of this email list are primarily English-speaking simulation practitioners from North America. All data and code used in this analysis are available in a reproducible GitHub repository.

Introduction

Many professionals in the building simulation domain are members of the *Bldg-sim* email list¹. This email discussion list was created in 1999 as a means of posting questions, announcements, and discussion-starters for various topics in the building simulation community. The volume of communication on the email list grew and then receded steadily, following the general trends of email list used for such purposes. The reduction in traffic is due to other lists that have gradually branched off *Bldg-sim* that focus on more specific topics such as eQuest, EnergyPlus, and TRNSYS. Also, other channels have taken over some of the load, especially with the growth of the now popular Q&A for building simulation platform called *unmethours.com*².

¹<http://lists.onebuilding.org/listinfo.cgi/bldg-sim-onebuilding.org>

²<https://unmethours.com>

The archive of *Bldg-sim* is publicly available as a source of information about the topics of discussion in the community. The emails from this archive are a source for understanding the trends of subjects, software, and stakeholders. Text mining is a maturing field of data mining that combines various pattern recognition techniques to extract high-quality insight from large amounts of text. One field of text mining is trend analysis; this sector focuses on the extraction of various keywords over a time range to show the growth or recession of various topics. Google trends is a popular website that leverages text sources from across the internet to show the popularity of various terms and concepts³. In various research communities, trend analysis has gained traction as a means of determining where a field is headed or where gaps exist to help target future focus. Some examples are in business analytics (Moro et al., 2015) and psychiatry (Abbe et al., 2016).

This paper provides an analysis of building simulation community trends using data from the *Bldg-sim* email list. We treat the list archive as a *sample of communication* from the domain represented by users that have posted to this specific email list. The methodology of data collection, processing, and text mining is showcased, and the trend analysis results are given in the format of time-series visualizations. A discussion is included that provides context to the analysis and discusses the limitations and generalizability of the results.

Methodology

This section outlines the process of data extraction, text analysis, and visualization. First, the email archives from 1999 to 2019 were downloaded and compiled into text files to be processed. These files were scraped from the public email list web archives^{4,5}. Various attributes were extracted from these files and targeted for analysis such as author, date/time, and

³<https://trends.google.com/trends/>

⁴<http://onebuilding.org/archive/bldg-sim-onebuilding.org/index.html>

⁵<http://lists.onebuilding.org/pipermail/bldg-sim-onebuilding.org/>

content. The topic modeling process utilized the Scikit-learn package in Python using Non-Negative Matrix Factorization (NMF) and TF-IDF weighting (Pedregosa et al., 2011).

The first information extracted was the total volume of traffic on the email list. Figure 1 illustrates the monthly email volume on the list in the last 20 years. For the first five years, email volume hovered around one email a day (~ 30 /month) - a manageable amount for even casual users. Sometime in mid-2007, the volume increased quickly to a peak of 360 emails per month by July 2007. The monthly volume then stabilized at around ~ 150 emails and has been steadily decreasing due to a split in the email list that occurred in 2008 and the introduction of other channels to the community.

Figure 1: Monthly volume of emails from the Bldg-sim email list archive.

Figure 2: Frequency of top twenty words found in the topic modeling process on each annual email content. These and other selected words have been classified in various categories and the frequency of their occurrence in the email archives are illustrated in Figures 3-8. Table 1 illustrates an example of six topics per year with five words per topic.

To conduct the topic modeling, the text was transformed to features in a way that their importance within an email, and across all emails, is captured. A common representation for this is a Bag-of-Words. With this technique, text files are represented by their

word frequencies, ignoring the words' relative position in each document. However, in any language, some words are presumed to be non-informative despite being rather frequent. These words are tagged as *stop words*. Stop words should be removed to avoid using them as a signal for modeling (Silva and Ribeiro, 2003). Examples of these words in the English language are words like *and*, *the*, *her*; in an email scenario, these words can be extended to words such as *from*, *by*, *subject*. A matrix representation of all the documents and their terms can be visualized as follows:

$$\begin{bmatrix} o_0^1 & o_1^1 & \dots & o_t^1 \\ o_0^2 & o_1^2 & \dots & o_t^2 \\ \vdots & \vdots & \ddots & \vdots \\ o_0^n & o_1^n & \dots & o_t^n \end{bmatrix}$$

Where o_m^n is the term frequency count (o) of term m in text n .

In a large collection of texts (also referred to as a *corpus*), even after removing stop words, some words will still be overly-represented relative to how interesting they are for the analysis. Relying purely on the frequency of occurrence will make these terms stand out more than rarer though perhaps more interesting terms. To minimize this effect, a weight coefficient is added to the count features, known as *Term Frequency-Inverse Document Frequency (TF-IDF)*, which will in turn capture more valuable information about the documents for our queries (Ramos, 2003). The weight coefficient in this new representation comes from the IDF portion, which is calculated in Scikit-learn as follows:

$$idf(t) = \log \frac{n_d}{df(d, t)} + 1 \quad (1)$$

where n_d is the total number of texts, and $df(d, t)$ is the number of texts that contain the term t . Finally, the resulting values are normalized by the Euclidean norm:

$$v_{norm} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \dots + v_n^2}} \quad (2)$$

Resulting in the final formula for every term t :

$$\frac{tf(t) * idf(t)}{v_{norm}} \quad (3)$$

With the different text of all documents being represented numerically, the unsupervised learning techniques known as Non-negative Matrix Factorization (NMF) is used to cluster documents (Li et al., 2014) into *topics* and then find the top n highest-rank terms in each topic as the most representative words for such a cluster.

Table 1 shows an example of such an analysis where the top six items for each year of the *Bldg-sim* data

Table 1: The main analysis of this paper clustered ten topics per year with ten words recorded. This table shows topics detected using the methodology using a smaller number of topics to fit in a single page. This table uses six topics clustered per year with the top five words illustrated. The full list of topic detected in this analysis is found in the project Github repository.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1999	doe, 1e, user, version, software	research, ashrae, information, looking, bill	jason, administrator, consultant, analytics	number, jun, bill, see, large	systems, software, better, source, available	thermal, program, hourly, interested, web
2000	doe, doe2, energyplus, powerdoe, program	weather, hourly, john, website, find	heat, water, air, temperature	software, development, effort, university, available	ashrae, research, systems, projects, engineer	public, private, open, source, software
2001	jason, problem, today, gpc18, analytics	weather, doe, looking, street, group	trnsys, engineering, university, software	doe2, heat, 1e, temperature	research, people, ashrae, corp, vice	load, loads, doe, office, equipment
2002	air, systems, control, floor, cooling	jason, il, analytics, busse, highway	powerdoe, dmi-inc, ma, dmi, savings	energyplus, available, software, information	doe, 1e, control, modeling, case	weather, sources, location, architecture, doe2
2003	doe, equest, chiller, doe2, 1e	mass, thermal, air, cooling, heat	tamu, esl, jeff, haberl, texas	yahoo, weather, actual, equest	software, architects, tools, getting, architect	ashrae, leed, modeling, method, standard
2004	air, cooling, equest, vav, water	radiant, heating, systems, temperature, research	yahoo, equest, gerald, daylight, pde	weather, tmy2, energyplus, program	wall, window, lighting, glass, fenestration	vikram, jason, forum, sami, web
2005	equest, glumac, window, construction, vikram	bb, jeff, haberl, tamu, texas	yahoo, spam, gerald, tired, protection	energyplus, engineering, green, software, information	heat, water, recovery, exhaust, pump	floor, thermal, temperature, south, air
2006	air, cooling, heat, water, temperature	energyplus, doe, weather, software, doe2	yahoo, spam, tired, protection, best	natural, ventilation, ramana, koti, appendix	equest, kendra, tupper, vikram, sami	pv, bb, jeff, ibpsa, bipv
2007	loop, heat, pump, water, cooling	software, developing, modeling, studioma, rosiers	baseline, leed, orientation, existing	elevator, vikram, elevators, sami, direct	bb, jeff, haberl, tamu, texas	nexant, unmet, john, equest, hours
2008	equest, weather, floor, wizard, program	air, heat, cooling, water, load	baseline, leed, existing, appendix, ashrae	pae, engineers, jpeg, image, bytes	bnim, ramana, consultant, koti, elements	uk, paul, zed, cfd, carey
2009	yahoo, loads, cooling, roof, load	unmet, hours, rosenberg, jim, michael	fan, power, supply, exhaust, kw	air, leed, baseline, ashrae, leakage	hay, cost, leed, paul	ies, energyplus, cfd, software, equest
2010	baseline, cooling, proposed, air, heating	modeling, leed, ibpsa, conference, ashrae	john, equest, eurek, nwo, voodoo	whiteboxtechnologies, weather, doe, source, joe	perkinswill, vikram, sami, addressee, chi01exmb	window, nick, frame, smith-boucher, glass
2011	baseline, air, fan, proposed, ventilation	sheffer, leed, cost, marcus, hours	3d, font, span, style, div	doe, haberl, weather, comfort, energyplus	nick, smith-boucher, leed, jpeg, caton	epsteinglobal, ce, et, epstein, martin
2012	weather, energyplus_support, yahoo, groups, energyplus	heat, cooling, heating, baseline, water	alex, wichert, robert, mark, jon	modeling, ibpsa, leed, conference, meeting	bre, uk, wales, registered, england	cleantech, analytics, cleantechanalytics, sheelam, khare
2013	ibpsa, energyplus, iesve, modeling, performance	weather, whiteboxtechnologies, joe, binned, huang	fan, wichert, seer, robert, power	residential, floor, dormitory, floors, roof	thumb, loads, christian, cooling, rules	baseline, proposed, air, ventilation, outdoor
2014	openstudio, performance, modeling, ibpsa, training	plug, loads, aeieng, chris, eac1	whiteboxtechnologies, weather, joe, radiation, low	jacobs, ventilation, jim, casework, baseline	fan, curve, vav, reset, archenergy	genopt, daysim, command, batch, bat
2015	actual, app, models, jacob, appendix	designbuilder, training, days, energyplus, openstudio	julien, fan, lighting, gouv, dutel	weather, whiteboxtechnologies, joe, utility, bills	certification, experience, technical, leed, usgbc	ibpsa, usa, news, newsletter, subscribe
2016	performance, experience, webinar, research, leed	openstudio, mc_cid, mc_eid, unqid, psdconsulting	ibpsa, usa, conference, news	weather, whiteboxtechnologies, joe, climate, epw	energyplus, workshop, openstudio, designbuilder, training	jones, jim, water, thermal, intended
2017	cgi, 40gmail, successfully, tresidder, entitled	network, gotraining, attendee, performance, register	jim, buildingperformance, dirkes, leed, linkedin	ibpsa, usa, buildingsimulation2017, news, conference	energyplus, openstudio, workshop, bigladdersoftware, training	weather, joe, whiteboxtechnologies, doe, solar
2018	ibpsa, usa, news, conference, unqid	energyplus, openstudio, bigladdersoftware, workshop, training	weather, wind, joe, rain, whiteboxtechnologies	webinar, performance, tools, research, modeling	tickets, eventbrite, tool, cost, optimization	lighting, linkedin, leed, value, cgi

Figure 3: General simulation terms that are commonly found in the topic discussions and have not fluctuated significantly over time. The top area chart is the number of occurrences of the word in every six months and the lower area chart is the percentage of the word's occurrence relative to the other words in this particular analysis. This visualization method is the same in Figures 4-8. These graphics are best viewed in color.

Figure 4: Occurrence of concepts that are more specific to applications in which more volatility can be seen over time. LEED starts slowly, but eventually becomes the most frequent word in the topic extraction process.

are extracted. In the table, each topic is represented by five words from the process. These words give a sense of the focus of that topic. The analysis for this paper actually uses ten issues each year with ten descriptor words. Figure 2 illustrates the highest frequency words found in those topics throughout the years. Each of the subsequent sections of this paper is focused on grouping the words found in the topics from year to year. These words have been tagged with the categories: *general concepts*, *software and systems*, and *organizations and companies*. The further analysis highlights the *most frequent users*. The overall trends of these categories are the main focus of the following sections. These sections outline the summary of changes of the most frequent words in the total *Bldg-sim* corpus.

General Concepts and Terms

The next few sections analyze the temporal trends of each category to show the frequency of use of various prominent words in the email list. The first set of words interpreted are the most common and somewhat generic terms. Figure 3 illustrates nine of the top words considered generalized terms used in the simulation process. They include words such as *performance*, *engineering*, and other notations that are common in the discussion of simulation of buildings. These terms seem to be reasonably consistent over time, illustrating that general terminology hasn't changed significantly over time. This result is unsurprising as the way we describe the building simulation process has likely not changed over these two decades.

Figure 5: Software tools and engines that most commonly found in the topic models - major shifts have occurred over the years - the most obvious being from doe2 to openstudio. The growth and recession of the discussions of various tools is more volatile than the systems they simulate.

Figure 6: Occurrence of building systems that are being discussed in the topics - these trends are more consistent than the tools. A slow but steady increase in the prominence of lighting and day-lighting can be seen in the trends.

Figure 4 illustrates words that are more specific to certain concepts in simulation as applied to the built environment. This visualization shows significantly more volatility and, therefore, more interesting empirical information about the popularity increase and decrease of these topics. For example, *LEED* slowly grew through the 2000s and has maintained at the same level since 2009. This behavior matches the growth of the LEED model creation process over that period. Even though LEED itself has maintained popularity, it seems that the concept of calculating a *baseline* had its peak between 2008 and 2014, but has since become negligible. This behavior is unusual as perhaps the *baseline* creation process has become mainstream, and the collective learning curve of the community has reached maturity.

Systems and Software

Figure 5 illustrates the modulation of various simulation software platforms discussed on the email list. This visualization turns out to be one of the most fascinating in that it captures the change of emphasis on various programs over time. In the early 2000s, *doe2*, *equest*, and *trnsys* dominated the discussions; this insight follows the logic of the favorite practitioner tools that were commonly used in that time frame. The first program to make a dent in that dominance was *iesve* at around 2009. The *open studio* platform has slowly been gaining ground since about 2014 to the point where it is now the most popular simulation platform discussed. This situation is intuitive in some ways, but engaging in that the popularity only started at that time, as *open studio* had been avail-

Figure 7: Trend of the various non-profit organizations that are most frequent in the topics over time. IBPSA has recently overtaken ASHRAE over time.

Figure 8: Trend of various companies, or for-profit entities, over the last twenty year. There is much more volatility than the non-profit entities.

able for years before 2014. The demand for *equest* has remained reasonably consistent over the years but is also slowly being squeezed out.

Figure 6 captures the trends of the building system types that are simulated and discussed on the email list. There is much less variability over time as compared to the software and specific concepts. The insight gained from this analysis is that while our tools and specific implementation objectives might be changing over time, the core systems that we simulate have remained fairly constant. One interesting variation is the slow, but steady growth of the *lighting* and *day-lighting* topics. These two topics have been overshadowed by heating and cooling system simulation, but are becoming more mainstream with the growth of tools like *radiance*.

Organizations and Companies

Figure 7 shows the non-profit organizations that are among the most frequent words in the topic modeling process. The *ASHRAE* organization has dominated the discussion for many years, likely not just due to the discussion of the organization, but also because of its various standards used in the simulation process. This dominance has slowly eroded over the last decade due to the progression of *IBPSA*. Organizations such as Texas A&M (*tamu*) University are present for many years, likely due to a large number of emails sent by members labs there focused on buildings. The United States National Laboratories such as the National Renewable Energy Laboratory and Pacific Northwest National Lab (*nrel/pnl*) are present over time, as is the US Green Building Council (*usgbc*).

Figure 8 shows the companies, or for-profit entities, that are among the most frequent words in the topic modeling extraction. Most of these companies show up in the email list due to people from those companies participating in the discussions. There is more volatility over time with companies than with non-profits. The results of this analysis should be taken lightly as these companies are only the ones that show up prominently in the topic modeling process, likely because one of their employees becomes a vital contributor of a particular set of discussions. For each year, there are employees from dozens of other companies that don't show up on this analysis that also have significant contributions to the talks.

Super Users

Since people are not just merely part of a topic, we wanted to quantify the users based on the *To/From* fields of the emails. An analysis using this method resulted in a list of over 3,312 people - 126 of which have posted 30 times or more. This situation proves there is a reasonably large, active group of participants, and many of the most active users are participating heavily in the Q&A portions of the email list. To recognize the top users in terms of the number of emails sent on the list, we compiled a list of the top 20 posters of all time, as seen in Table 2. One interesting finding is that the topic modeling process failed to capture many of the top users of the platform. This situation is likely due to the TF-IDF algorithm filtering names that are found in multiple topics, thus relegating the names as words as they are not specific to a single topic. Users who comment on a diversity of issues would not show up in the topic modeling process it seems.

Table 2: Top users on the platform by number of emails from the last twenty years.

User	Email Count
Nick Caton	309
John Aulbach	289
Jason Glazer	275
Joe Huang	254
Jeff Haberl	251
Vikram Sami	246
Jim Dirkes	205
David Eldridge	192
Drury Crawley	184
Chris Jones	184
Varkie Thomas	183
Karen Walkerman	158
Robert Wichert	138
Brandon Nichols	122
Chris Yates	106
Leen Peeters	99
Ramana Koti	97
Carol Gardner	90
Mike Tillou	85
James Hansen	84

Generalizability of results

This paper is not meant as a comprehensive analysis of the entire building simulation community. It would be impossible to collect all correspondence emails from all people in the simulation community to accurately capture the entire community's trends. Therefore, this analysis is only applicable to the population represented by the users of this particular email list. It is outside the scope of this paper to create a detailed demographic or professional analysis of the users of the platform; however, there are general attributes that allow us to make certain assumptions about the generalizability of the results. The following are a list of points to consider when interpreting the usefulness of the results:

- The email list is primarily in the English language; thus the population captured is likely from North America, Australia, New Zealand, and parts of Europe where a large percentage of practitioners and researchers use English as their working language.
- Further, while *ASHRAE* and *IBPSA* have a large international reach, many of the other non-profit organizations and companies are based in the United States. This suggests that the topics, tools, and users are likely biased towards the simulation community in the USA and Canada. This insight is reinforced when looking at the top 20 users of the platform over the years, as most of them are based in North America.
- The prominence of free, open-source simulation tools in the email list points to the fact that these platforms generally only have community-driven support for users. This situation can be compared to for-profit simulation tools that will often include on-demand support from the vendor. Open-source tool users are likely going to be more active on this email list.
- The analysis shows that the use of software tools has transitioned from tools like *equest* to newer platforms like *openstudio*. This transition more closely resembles what would be expected from simulation practitioners rather than researchers.
- The splitting of this more general topic email list into more niche lists that focus on simulation engines like *equest* and *transys* likely have had an influence in the slowdown of traffic for those topics. The impact of those lists was not accounted for in this analysis.
- Other text mining methodologies could produce slightly different results based on the methodology in which the keywords were chosen. Readers are welcome to reproduce the analysis in this paper and use different combinations of words to understand various issues that are important to them.

Conclusion

The analysis of the email archive of the *BLDG-SIM* list has uncovered numerous interesting building energy simulation trends over time. This analysis is relevant to the population of practitioners who have regularly used the platform over the last twenty years. Some of the most prominent insights include:

- Most of the general simulation terminology such as type of system or application has changed little in twenty years
- There have been major shifts in the types of simulation tools with major phase transitions occurring every 5-10 years. These results show exactly when the adoption of new platforms (such as *open studio*) actually occurs, often years after they are introduced.
- One of the key open-source community shifts has been the transition from *doe2* and *equest* to their modern relatives *energyplus* and *open studio*. Interestingly, these transitions occurred long after the introduction of the tools to the community.
- The simulation of *heating*, *cooling* and *ventilation* systems maintains a strong majority, but *lighting* and *daylighting* systems are slowly gaining ground.
- Non-profit organizations' popularity has had shifts on the email list over time, the most prominent being the focal point shift from *ASHRAE* to *IBPSA* over the last five years
- For-profit companies have shown a lot of volatility, mainly due to the activity levels of a few employees. The applicability of this analysis of these entities was found to be limited by this methodology.
- While the email users numbered at above 3,300, the heaviest users (30 or more emails in twenty years) was 126. The top 20 users include several key industry and academic thought leaders; however, the results are mostly limited to the USA and Canada communities.

While many of these conclusions are intuitive, this publication is the first that has attempted to empirically quantify these shifts to detect the speed or length of these transitions.

Future work

Over the last five years, the volume of the email list has reduced due to the increase in use of the www.unmethours.com website. There are less emails, but they are more focused on the non-Q&A simulation topics such as announcements. *Unmethours* has rapidly grown with over 6,834 questions posted as of May 2019. This knowledge source should be analyzed in a similar way to better understand the trends from 2015 and into the future. A combination analysis of emails and Q&A websites would provide a more comprehensive overview of the field. The results from this

source could likely be considered more generalizable as a publicly-accessible website should have more diversity than the *Bldg-sim* email list. The site also has the benefit of collecting other types of user behavior information like *up* and *down* votes, traffic information, and the number of answers per question.

Reproducibility

The entirety of this analysis from data collection to visualization is reproducible through a set of Jupyter notebooks and guidance hosted on Github⁶. The authors encourage feedback and community-driven further analysis.

Acknowledgment

This paper would not be possible without all the enthusiastic people who have used the *Bldg-sim* email list over the years. Special thanks to This paper has no specific funding source.

References

- Abbe, A., C. Grouin, P. Zweigenbaum, and B. Falissard (2016). Text mining applications in psychiatry: a systematic literature review. *International journal of methods in psychiatric research* 25(2), 86–100.
- Li, L., J. Yang, Y. Xu, Z. Qin, and H. Zhang (2014). Documents clustering based on max-correntropy nonnegative matrix factorization. *Proceedings - International Conference on Machine Learning and Cybernetics* 2, 850–855.
- Moro, S., P. Cortez, and P. Rita (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent dirichlet allocation. *Expert Systems with Applications* 42(3), 1314–1324.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.
- Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries Juan. In *Proceedings of the first instructional conference on machine learning*, Volume 242, pp. 133–142.
- Silva, C. and B. Ribeiro (2003). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 2003.* 3, 1661–1666.

⁶<https://github.com/buds-lab/twenty-years-of-bldgsim-textmining>