A machine teaching approach to literature surveys*

Thomas Krichel

November 2022

0 Status

This is the “detailed research proposal” for my Lindberg fellowship application. Its latest version is online at http://openlib.org/home/krichel/proposals/tiumen.pdf.

1 Introduction

This is an unusual proposal. You could also call it an outsider or diversity proposal. Here is why.

First, the proposal is supported by an organization I founded and run. The organization is the Open Library Society, Inc., a US 501(c) (3) charity registered in New York. Thus it I did not add an institutional support statement.

Second, the proposal comes from a person who considers himself a librarian but has no actual qualification in the subject other than having been a library school professor.[1]

Third, the tone of this proposal is conversational and a bit blunt. I prefer to say how things are rather then obfuscate my reasoning by bombastic and stiff language. And I try to give my readers a respite from the rather abstract nature of my material by using informal language.[2]

Finally, the proposal is neither based on existing research nor on established practice. It is exploratory.

The proposal’s saving grace is that lies right at the heart of librarianship. It focuses on users working with sets of documents.[3]

The rest of the proposal is structured as follows. In Section 2 I set out what I want to do. In Section 3 I introduce the key concept of machine teaching. This is the most abstract section. Section 4 motivates machine teaching as an approach to information retrieval. It could be skipped if you are already motivated. Section 5 accounts of how the proposed system would work. This is the most complicated section. Section 6 deals with the research expected to come out of the proposal. Section 7 discusses the implementation stages. Section 8 discusses prior work I did. It could be skipped if you are already convinced that I have the skill and experience to conduct this work. Section 9 concludes. I have relegated financial aspects to Appendix A.

2 Task statement

I propose to build a system called Biomed Reviews. It will be online at http://biomed.reviews. It will the first ever application for surveying the literature in PubMed using machine teaching. Let’s look at this statement more closely.

First, let me explain what I mean by “literature in PubMed”. Users will not be interacting with the PubMed website through Biomed Reviews. Instead they will be using the data that the National

[*] I am grateful to Brandon Henry Patterson and Kerri Shaffer (University of Utah) and to Antonio P. DeRosa (Memorial Sloan Kettering) for sending me their winning proposals. Sophie C. Rigny (Library of Congress) provided valuable comments. Any errors are mine only.

[1] The bibliographic sketch belabors this point further.

[2] I hope that reading this proposal will be more enjoyable than getting root canal surgery.

[3] I am a staunchly conservative believer that librarianship is a service profession. It serves to bring documents to users. When looking at recent Lindberg awards it seems to me that they have been given for more marginal concerns. I am not denying that these marginal concerns are important. But they do not appear to be *central* to librarianship.

Library of Medicine (henceforth NLM) provides at [1] and [2].[4] It will be available in Biomed Reviews as soon as possible, by the next day at the latest.[5] Biomed Reviews is limited to PubMed data because PubMed data is freely available. It is large enough to allow for research.

Second, let me explain what I mean by "survey the literature". I mean that users of Biomed Reviews will be looking at the whole or a subset of PubMed by publication span, say the last three or ten years. They will want to know about *all* the papers that have been published on a certain research topic. They will be willing to spend a few hours on this effort. These hours do not have to be crammed into one continuous session. They may be spread over several days. Typical users are those who prepare PhD dissertations, or librarians contributing to systematic reviews. Biomed Reviews will not be suitable for casual users. Neither will Biomed Reviews be suitable for expert users looking for current awareness. For these users I built "bims: Biomed News" [3]. I run it since 2017. Since 2020 researchers at the NLM have maintained LitSuggest, see [4], and [5]. It is a system that is similar to "bims: Biomed News". In some ways, it imitates Bims. So there are already at least two smart systems for current awareness to PubMed.[6] But there is no system for literature review based on machine teaching.

Let me use the next section to explain what I mean by "machine teaching".

3 Machine teaching

I confess: "machine teaching" is a term I invented.

Most people have come across the term "artificial intelligence". Artificial intelligence has been a lot in the headlines of late. For example, machines beat the best humans in chess. Teaching a machine to play chess is essentially easy. The aim is known. We want to win the game. But a literature survey is very different. Every survey has a different topic. Every user has a different approach to the topic.

Most people also have come across the terms "machine learning". Basically, it means that a machine is learning something. Usually it is learning from a vast amount of data. As a librarian, I leave that problem domain to computer scientists. These folks have come up with the idea of "supervised machine learning". Here, the machine is learning through supervision by a human. The supervision produces positive and negative examples. As a librarian, I am interested in the users producing the examples, thereby teaching the machine. I can use computer science methods to make the machine learn. But users teaching rather than machines learning is the focus of my work.

Recall that Biomed Reviews is for literature reviews using PubMed. Therefore we work with PubMed records. Let's just call them records. The positive examples are the records that users classify as relevant. We call them positive records. Conversely, the records that user classify as non-relevant are called the negative records. The union of the positive and the negative records are the training records. Biomed Reviews will use standard machine learning algorithms on the training records. The result of a machine algorithm's work is a machine learning model, henceforth a model. Once we have a model, we can use it to rank records of unknown status by the probability of them being positive. Then users can look at these records. They select further positive and negative training records. The model can then be updated. The remaining unknown records can be reranked.[7]

Surveying the literature using machine teaching uses no search queries whatsoever. Instead the user need is expressed by the training records. At the start, the user is supposed to know about at least

[4] I suspect that "Medline citation data" is the actual correct term.

[5] expect, perhaps on the day of the annual update. But that is just once in December.

[6] In general, as evidenced by the recent survey of [6], current awareness is as huge research area.

[7] What I call "rank" is what the standard computer science literature calls "classify". This is not my usage. Users classify. Computers rank. This distinction is important for the theoretical evaluation of such system, see [7].

one positive record. Since our service addresses the need of non-casual users, we can be optimistic that each of them will arrive with positive PubMed record in hand. If a user knows none, there still are a lot of searching tools.

4 Motivation

You may wonder: “Aren’t conventional searches good enough? Why bother with machine teaching?” If you don’t, just skip this section.

Searches use queries. They are the standard way to retrieve bibliographic data. At this time, they are only way to do it. This proposal challenges that status quo. I understand that it is hard to take that challenge seriously.

Nevertheless I trust you understand what the limitations of searches are. Basically the terms in the query have to be in a record for the record to be found. Artificial intelligence, ontologies and similar natural language processing tools have been used to try to overcome that limitation.[8] But the problem remains that queries are short. A query for “current weather in Jackson Heights, NY” is easily understood. But in biomedical information the situation is much fuzzier. This is where more data about the user need just means better retrieval.

Rare diseases give us a poignant example. Yes, you can search for the name of the disease. And you will find the papers written about the disease ... by authors who were aware of the disease *and* who cared to mention it. But what about the other papers? Papers that are relevant to the disease, but whose authors did not mention the disease? Or papers by authors who did not know about the disease? It may be possible to find such papers by symptoms of the disease. Most diseases have many symptoms. Many symptoms are common across many diseases. And any symptom has various names. Therefore searching via symptom names is a truly Herculean task.

This is a typical scenario where machine teaching can help. When we teach by examples, the examples are full records. Even if we just take a few positive and a few negative examples, there is just much more information in these examples than in a search query. It’s just the mass of data that makes for better information retrieval through examples than through queries.

Let me just give you one example. Consider Jun Maruta’s bims-madeba Biomed News report on “Mal de débarquement syndrome”. In the issue of 2018–09–23, he found a single positive record, see [9]. Note that the displayed record does not mention “Mal de débarquement syndrome”, with or without accent. Jun Maruta found this record using a model trained on 13 positive records and several hundred negative ones.

Now I am not for a moment claiming that machine teaching is always better than searches. I suspect that it often is. But unless we have more evidence, we can’t tell. A lot depends on the information retrieval scenario. You can think of searches as a series of quick shots. Most of the time, searches are isolated from each other. Aggregating the information from various searches is burdensome. Machine teaching is an interactive process of data collection. The results of the previous ranking are used as a basis for the next ranking. Aggregating the records is handled by the software.

5 The teaching process

When we want to find out how a human can teach a machine efficiently, we would like to try out a few different approaches to see which one works best. Unfortunately, each approach would require building a specific user interface. Sadly, building interfaces that is easy to use is a huge job. Here I can only build one. Therefore I need to set out a plan of how the interface will actually work. I have been thinking long and hard about how to do this. Here what I have come up with.

5.1 The initial model

I suggest I should allow the user to put in a single PubMed record at the start. Let’s call that record

[8] See [8], for an example.

the seed record. The user may know many potential seed records. I prefer users to see them found by the system. This will strengthen users' beliefs that Biomed Reviews actually delivers. Conversely, if it does not propose known positive records, that will be useful feedback.

We need negative records in order to create a machine learning model. We can't ask users to enter negative records. I know of two approaches to get negative records. One is the approach of LitSuggest. They use random records. The other one is the one I use for Biomed News. It is described in Subsection 8.2. Biomed Reviews will use exactly the same approach. Thus is it will come with a very carefully selected set of 1000 "scoping records". I remove the scoping record than is closest to the seed record. Thus I start by constructing a model with one positive and 999 negative records. This approach has a great track record with Biomed News.

5.2 The tub

Once we have a model, we can rank all 35 millions PubMed records. Unfortunately, doing that—on the type of computer I can afford—takes a very long time.[9] I have to assume that no user will be willing to wait for this long every time they update their model. Thus I need to rank a subset of records at a time. This is where hard choices have to be made. I made them.

I suggest using a "backward dynamic" approach. Thus I propose to work from the present to the past. I plan to do this in tandem with Biomed News. Thus, to start with, Biomed Reviews users will look at the papers that are in the current Biomed News issue. That is roughly 30000 records. These records came out last week. They are placed in the "tub". Here the tub is just an expression I use to talk about the records that are to be ranked by the model. I say these records are the ones in the tub. So at the start, we have the initial model, and we place the most recent Biomed News records into the tub. There, the records are ranked by the initial model. They are presented one-by-one to the user, by ranked order. The user is asked to classify the first record as relevant or non-relevant. This is done using the arrow keys. Then the next record appears. Any records that have been classified by the user are removed from the tub. At any time the user can interrupt the examination of records. At that point the user can ask for the tub to be updated in place. That means, given the additional classified records, compute a new model and re-rank the records in the tub. Or a user can add older Biomed News data to the tub, and then re-rank all records in the tub. Both operations will will take time. Users will be informed that a calculation is in progress. The browser will auto-reload the page say every minute. Users will start some other task on their computers while this is going on.

As the user adds more records to the tub, the time to rank them increases. There are just more records to rank. Once the tub goes over a certain target size, items that have a low rank will be cleared from the tub. The most likely approach is to set a floor probability.[10] If a record has a lower probability of being relevant than the floor probability, and provided the tub is over a the target size, the record is removed from the tub. It will not reappear.

This may all sound a bit complicated. But it is the best way I can figure out right now to build a system that users may actually want to use. Of course a backward dynamic approach carries some small risk of dynamic bias.[11] Newer papers may talk about the same topic with different words. I feel that the dynamic bias is a risk worth taking to get users onboard with interesting records at the start. Models adapt anyway.

[9] I don't know how long. This will be part of the research on this project.

[10] In later stages of the project this floor probability *may* become user configurable.

[11] Let me give you a crude illustration. Assume your subject is pandemics. As you work your way back in time, reaching 2019, well suddenly COVID-19 disappears from the literature and the pandemic literature looks very different. So the model trained an 2022 to 2020 will be biased to look at vintage 2019 papers.

5.3 The finish

When the users are finished, they can close the survey. When a survey is closed, no changes can be made to it. A future survey could be created that is based on the closed survey. However, given limits of this project, I can not guarantee that I will undertake coding for this facility. In any case surveys will be published in plain text JSON files. All that data will be freely available in bulk as a contribution to open science.

6 The research

As I wrote in Section 1, this proposal aims at exploratory research. It's the best term I could find. The definition given in [10] says that exploratory research is "the preliminary research to clarify the exact nature of the problem to be solved". This is not fully satisfactory for this proposal. I built the first-ever machine teaching system for bibliographic data in 2004, see [11]. Thus I have 16 years of experience with the problem domain of machine teaching for current awareness. Literature surveys are quite different. In [10] I read "Exploratory research is used when the topic or issue is new and when data is difficult to collect." This statement is very fitting.

Here I am going over the main research problems.

6.1 The revelation problem

When users teach a machine, they will be happy to tell us what records are positive. These are the records they find of interest. The users may go on reading the papers that the records describe.

But for machine teaching, we need to know about the negative records. Negative records vastly outnumber positive records. If we don't want to burden users, we will only get a few negative records from them. It is unclear to me at this stage whether Biomed Reviews users will reveal enough negative records as to make for viable models. Or, in other words, it is not clear if I can build an interface that will get users to reveal enough negative records. Therefore, the general research problem is to uncover methods to reveal machine teachers preferences about negative records and/or make reasonable assumptions about other records to assume to be negative. I am not sure at this point, if I need assumed negative records other than the records in the scoping records set. This is an issue that has to be clarified through experimentation.

6.2 The descriptive problem

To evaluate a system, we need user data.[12] We can not get user data until we have Biomed Reviews operating. Biomed Reviews needs to be built with a focus on data collection. Therefore usage should be precisely recorded. How this is done is a major conceptual challenge. Ideally we want each interaction of the user with the system to be recorded. For this exercise, user demographics are not important. It is sufficient for the user to give a single PubMed record, and to use the system. I will set up this "minimal demographic" mode of operation first. Let's call this dynamic data. How to precisely represent dynamic data is a major research question.

Dynamic data is not the only data the system should produce. Ideally we want completed reviews to be publishable and citable, with things like

- a title of the survey
- some user demographics: name, homepage, affiliation
- contact email for the survey (may not be made public)
- the handles of positive PubMed records
- the handles of negative PubMed records
- the handles of unknown PubMed records[13]

[12] I believe in data from real users, not the ones who labor for a gift certificate.

[13] You may wonder why would we bother with the set of unknown handles. Well, if the survey was well-done, then the remaining unknown handles will be unlikely to contain positive records. At a later stage, if we want to renew the survey, presumably we want to first search the records that were not in the set of unknown records.

Users who will be happy to make their review public may not be happy with the dynamic data being public. I will need to figure out how to make both datasets live side-by-side.

6.3 The evaluative problem

The final, more general, question is how to evaluate the system. In [7] I pointed out that the usual suspects like precision and recall don't apply. For a system that has yet to be built no ready-made evaluation method exists. Even if it existed, I believe that I—as the builder and proponent of the system—should not evaluate it. There is an obvious conflict of interest.

7 Project stages

7.1 Starting stage

I want to set up something that is usable as early as possible. That minimal implementation may not have all the features. Users only enter the seed PubMed handle, and drive the system by classifying papers, reranking, and adding data into the tub. Basically, we get some sort of anonymous sandbox system that anybody can just casually test drive. The data saved contains the positive and negative records. That is, the documented outcome will not contain the dynamic data.

It will take up to seven months from me to complete this stage. It is only when this stage is finished that we look for users.

7.2 User involvement

Of course I could promise that Biomed Reviews will take the world by storm. But I am a librarian and researcher, not a politician. Generally finding users for a new system is hard. However, literature surveys are a common library user need. I hope that the Medical Library Association will do some reaching out for me. I will do all I reasonably can to get users.

7.3 Documented phase

At about mid-time, I will work on a paper describing the working of the system other than the actual machine learning.[14] In the second phase of the project I will make sure that I introduce more features. What these features are will depend on the feedback I get from users. But the main job is to document users' actions. I aim at documenting the process, rather than just the results.

7.4 After the end of funding

I expect that after the project is finished, I will publish one research paper about it.

I have an excellent track record of keeping projects running without external funding. For example, I have been running NEP: New Economics Papers, see [12], since 1998 without any subsidies. If there is no further external funding for Biomed Reviews, I will *at least* keep it running as a sandbox system for Biomed News. This will allow candidate Biomed News users to test out my machine teaching technology.

8 Prior work

In this section I am trying to convince you that I have the technical chops to get the project done. If you are already convinced, you can skip this section. The important point is that I already have built parts of the infrastructure needed for Biomed Reviews because I built and maintain Biomed News.

8.1 Pumex

I wrote pumex [13]. It is an indexing software for PubMed records. It runs daily for Biomed News. It downloads the new PubMed file from the NLM ftp site. It finds out what records are new. It puts them in files of records by daily input, so-called "dain",

[14] The Biomed Reviews software will fire up learning by a system command. Thus the software will be machine-learning agnostic. It will be possible to plug into systems that use different machine-learning methods.

see [14]. This has been running since 2015–12–15. Earlier records are classified by a PubMed date field that is no longer available in current records. Partitioning by dain allows to easily document the unknown records at survey completion time.

8.2 Scoping dataset

I have taken great care to build a special subset of PubMed records known here as the scoping set. For Biomed News, I used 1000 such records. These are selected to be as different as possible between themselves.[15] Finding the records for this set is a computationally expensive task. And I am still working to refine the methodology. But the results of my research on this matter so far can be used.[16] I firmly believe that using the scoping papers is better than using random records as LitSuggest does.

8.3 Feature extractions

The machine learning technology I use is based on LibSVM, see [15]. To get this to run, features have to be extracted from the PubMed records. I have honed this craft since 2017. I take care of all the information in the PubMed record, including, for example, author affiliations. I have also worked on the extraction of phrase features that are likely to be significant.

9 Conclusions

I realize that for the committee to actually choose this outsider proposal will be rather courageous. The work I do is truly pioneering. It is the opportunity to create a system for literature review that actually comes from the library community, rather than being dropped onto it by the Googles and the Elseviers of this world.

The comittee could justify its decision by referring to the name of the award. On Donald A.E. Lindberg's Wikipedia page [16], I read "He was known for his work in medical computing,... especially the development of PubMed." This proposal will apply serious computing to PubMed. I use PubMed it in ways he did not think of. For the life of me, I can not imagine a finer proposal to pay tribute to him.

References

[1] National Library of Medicine. Annuxgal baseline. `https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/`, . Accessed: 2022–11–03.

[2] National Library of Medicine. Daily update files. `https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/`, . Accessed: 2022–11–03.

[3] bims: Biomed News web site. `http://biomed.news`. Accessed: 2022–11–11.

[4] National Library of Medicine. Litsuggest. `https://www.ncbi.nlm.nih.gov/research/litsuggest/`, . Accessed: 2022–11–08.

[5] Alexis Allot, Kyubum Lee, Qingyu Chen, Ling Luo, and Zhiyong Lu. LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res*, 49(W1):W352–W358, 07 2021.

[15] Sorry for being vague. To explain how this works is too complicated for this proposal.

[16] Here is how I use these records as a base set of negative records. Assume a user enters the single positive PubMed record. I take a closeness metric to find the record in the scoping set that is closest to that record. I remove this closest record from the scoping papers. I get a model with one positive record, and 999 negative records. Then I rank the records in the tub. I show them to the user. Assume the user classifies five records and classifies ten as negative. She asks Biomed Reviews to rerank the records in the tub. What happens? Well, I build an intermediary model of the five positive and ten negative records. I use that model to rank the remaining 999 scoping papers. I remove the 15 top-ranking papers from scoping record set. So the models still has 1000 papers but the number of scoping papers decreases from 999 to 984. As the user classifies more records I crowd out the scoping papers.

[6] Christin Katharina Kreutz and Ralf Schenkel. Scientific paper recommendation systems: a literature review of recent publications. *International Journal on Digitial Libraries*, 2022. doi: https://doi.org/10.1007/s00799-022-00339-w.

[7] Thomas Krichel. Information retrieval performance measures for a current awareness report composition aid. *Information Processing and Management*, 43:1030–1043, 2007. available at http://openlib.org/home/krichel/papers/sendai.pdf.

[8] Cochrane Collaboration. Project transform final report. https://community.cochrane.org/sites/default/files/uploads/inline-files/Transform/201910_ProjectTransformReport_FINAL_WEB.pdf. Accessed: 2022–11–05.

[9] Jun Maruta. bims-madeba issue of 2018–09–23. http://biomed.news/bims-madeba/2018-09-23. Accessed: 2022–11–10.

[10] Wikipedia. Exploratory research. https://en.wikipedia.org/wiki/Exploratory_research, . Accessed: 2022–11–07.

[11] Michael E. D. Koenig. Nomination for Tony Kent Strix Award. http://openlib.org//home/krichel/tony_kent_strix.pdf, 2022. Accessed: 2022–11–11.

[12] NEP: New Economics Papers web site. http://nep.repec.org. Accessed: 2022–11–09.

[13] Thomas Krichel. pumex. https://pumex.openlib.org, . Accessed: 2022–11–04.

[14] Thomas Krichel. dain files. https://pumex.openlib.org/dain, . Accessed: 2022–11–09.

[15] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[16] Wikipedia. Donald A.B. Lindberg. https://en.wikipedia.org/wiki/Donald_A._B._Lindberg, . Accessed: 2022–11–04.

A Financial details

The total amount sought is $10000 US dollars. This can be sent to the bank account of the Open Library Society. There will be no expense for software because I use open-source software and write all purpose-built software myself. Server space and CPU time will be covered by existing sponsors. I expect Biomed Reviews to stay online after the funding has expired.

The project funding will be used to pay me. I will work on it at about a 70% time commitment for about 15 months. Given the technically demanding nature of this work, I would expect the normal expense on such a person about $4k a month. Roughly, the $10k here buy labour of a market value of over $50k.