# A Study: From Electronic Laboratory Notebooks to Generated Queries for Literature Recommendation

**Oldooz Dianat, Cécile Paris, and Stephen Wan**
CSIRO Computational Informatics
Sydney, Australia
firstname.lastname@csiro.au

## Abstract

Relating one's research to the vast body of scientific knowledge is a difficult task; the sheer volume of literature makes it difficult to keep up-to-date with scientific developments. Particularly when research is on-going, keeping track of related work is especially important to avoid an unintended duplication of effort. We outline a novel approach to this problem that uses the text in an Electronic Laboratory Notebook (ELN) as a representation of an experimental context in the field of Chemistry. The contribution of this work is to situate the literature recommendation task within the context of the user's experimental information needs. We find that our approach to transform the ELN text into queries for use with PubMed is able to recover a subset of user bibliographies. We find that alternative methods for query generation that capture both scientific terminology and salient terms in the ELN complement each other.

## 1 Introduction

Identifying the relationship between one's research and the ever growing body of scientific knowledge is a time-consuming and laborious task. The sheer volume of existing literature makes it difficult to stay up-to-date with new scientific developments. Furthermore, this task is continual: this relationship must be revisited periodically so that one can avoid unintended overlaps with work published in parallel.

To help keep up with advances by our scientific peers, we can use a number of tools to provide continual exposure to newly published work. These can range from collaborative bibliography tools with a social network component, for example, the Mendeley application[1]. However, such tools do not have a mechanism to capture the information needs of a researcher that might change on a daily or weekly basis due to the outcomes of experiments.

In this work, we recognise the increasing use of Electronic Laboratory Notebooks (ELNs) in our research environments to capture a representation of research as it progresses. As part of the project described in this paper, we introduce the novel use of the text in the ELNs as a representation of the user's context—specifically, their current experimental context—that provides insights on their information needs. Our aim is to devise a system to transform this context into queries for a scientific literature search engine, and then suggest references that may be relevant.

This paper describes the initial exploration in generating queries from the on-going experimental context as represented in ELNs. In this work in progress, we investigate the effectiveness of different types of information extracted from ELN content for the purposes of suggesting relevant literature.

The ability to suggest references in the context of reading an ELN entry is potentially useful in many contexts. Indeed, in our user study, we noted that users often mentioned the need to identify relevant literature based on content from the ELN entry. For example, a doctoral student may have identified closely related work to scope the thesis, but may nevertheless want to monitor the literature to ensure that the scoping remains novel. Using our approach, as she writes up her daily work in the ELN, our system would look for and suggest related work to read, reducing the risk that recently published work that is closely related goes by unnoticed.

We conducted our studies with the LabTrove

[1] http://www.mendeley.com/

Oldooz Dianat, Cecile Paris and Stephen Wan. 2013. A Study: From Electronic Laboratory Notebooks to Generated Queries for Literature Recommendation. In *Proceedings of Australasian Language Technology Association Workshop*, pages 62−70.

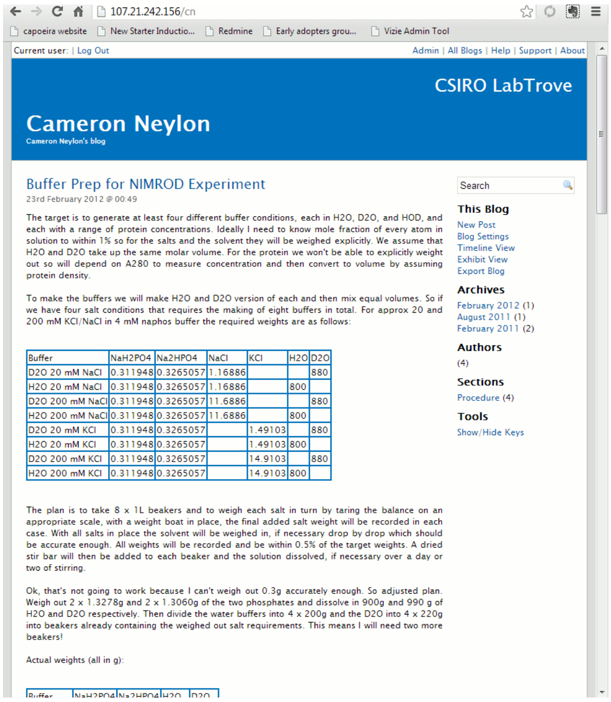


Figure 1: A LabTrove blog entry by Cameron Neylon during his affiliation with the University of Southampton. (Reproduced with permission from Cameron Neylon.)

tool[2] (Milsted et al., 2013), an ELN based on open source webblog software. LabTrove, designed by the University of Southampton, has been designed with Chemistry researchers in mind, allowing them to post daily updates about their research outcomes. Although we focus on chemistry ELN entries in this work, LabTrove is potentially more widely usable by other researchers in the experimental sciences. A screenshot of the LabTrove interface is presented in Figure 1.

Ideally, we would conduct user studies to evaluate the effectiveness of the suggested references for a knowledge discovery task; however, the time required for such a study makes this approach prohibitive for exploratory research. In lieu of such studies, we describe the extent to which generated queries can reconstruct the bibliographies of users, a slightly different scenario to knowledge discovery. An instance of the LabTrove ELN was in use at the chemistry department in the University of New South Wales. Users of LabTrove at the university that were interested in collaborating with us were identified by the university library which helps to host the ELN. The users provided access to their LabTrove entries and their research bibliographies.

[2]http://www.labtrove.org/

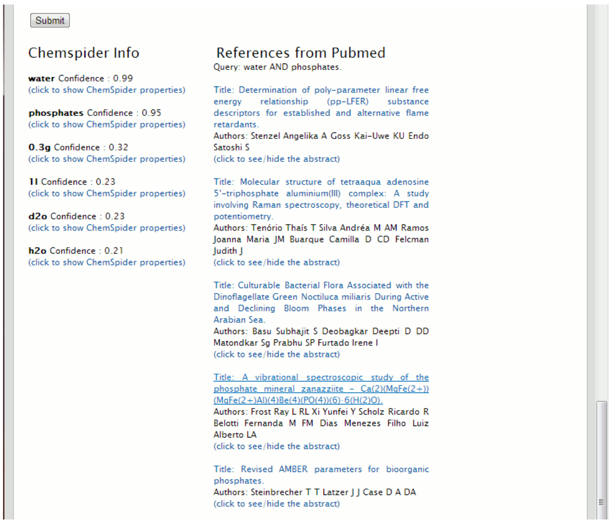


Figure 2: Automatically detected chemical entities and suggested PubMed references are shown after the main blog entry.

This study is based on the blogs and bibliographies of three users. Finding additional data was difficult given our recruiting constraints. Nevertheless, we are able to report on preliminary findings that indicate the extent to which the different query generation methods are able to reconstruct the gold standard bibliographic information. This provides insights as to the strengths and weaknesses of the different approaches to query generation when used for this scenario. We find that alternative methods that capture both scientific terminology and salient terms in the ELN complement each other.

In the remainder of this report, we present an overview of the system in Section 2. We describe the data used in this study in Section 3. The algorithms for generating queries from ELN content are described in Section 4. We present our evaluation of different query generation methods in Section 5. We discuss the results obtained and outline future work in Section 6. Section 7 describes related work in suggesting scientific literature and evaluating these query generation methods. We finish with concluding remarks in Section 8.

## 2 A System Description

We have deployed a version of LabTrove with our code to provide extra linked data at the university for the participants who have volunteered to trial. To provide links to relevant scientific literature from the ELN entries, we instrumented Lab-

Trove such that, as an ELN user reads a blog entry (that he or she is entitled to read), a list of automatically detected chemical entities are presented following the main text entry. These entities are detected using the OSCAR tool for Named Entity Recognition in chemistry literature (version 4 (Jessop et al., 2011)). For a description of earlier OSCAR versions, see Corbett and Murray-Rust (2006) and Corbett et al. (2007)).[3]

In this deployed version, to automatically suggest scientific literature, we use the chemical named entities as queries which are sent to the PubMed Entrez Application Programming Interface (API). This API provides references from the PubMed repository of scientific literature, bibliographic details and abstracts for references matching the query.

We modified the blog display page to provide extra linked data. A screenshot of the CSIRO plugin is presented in Figure 2. Within the interface, the user can decide whether or not to view extra linked data that we have associated with the blog text (clearly indicating, for legal reasons, that this is added data, kept separate to the author's original entry).

The linked data includes relevant chemical properties which are obtained by sending the chemical named entities as queries to the ChemSpider[4] web services maintained by the Royal Chemistry Society. Our plugin for LabTrove also suggests scientific publications retrieved from the PubMed API to help show what existing literature may be relevant to blog content. The user can request suggested references triggering the on-demand retrieval of search results from PubMed. These are presented alongside the list of detected chemical entities. Any user clickthrough data is stored in a log to allow for automatic tuning of the algorithms.[5]

## 3 The LabTrove Users and their Data

We used the blog posts of 3 users who provided matching bibliographies for their blogs, refered to hereafter as L, R and D.[6] An overview of the descriptive statistics of the users' ELN blogs is presented in Table 1. The users belong to the same research group and share the same research supervisor. The supervisor is known to be a strong advocate for the use of ELNs, and the group uses the ELN on a regular basis within their research meetings.

| user | num of posts |
|---|---|
| L | 571 |
| R | 148 |
| D | 1078 |

Table 1: Number of posts for our three users.

In our user study, we found that the main use of the ELN was to record and archive daily experimental data. The ELN is also used, however, for a number of other research tasks, such as:

1. Experimentation in using the ELN itself;
2. Archiving supporting research documents like reference files;
3. Archiving draft publication files; and
4. Record iterations of thesis structure and argument.

As such, the text collection are a heterogeneous collection. In this preliminary investigation, we assume that each blog (containing a series of entries) is about a single research goal and that the user has a single bibliographic file against which we can compare suggested references. However, in reality, not all of the blog entries are related to an overarching research goal that might subsume a series of experiments. Indeed, a blog may span multiple research goals, each deserving a separate set of bibliographic recommendations.

## 4 Query Generation

In our system design, the suggestion of references from ELN blog entries would ideally perform the following broad steps:

1. Represent the user's experimental context as a query;
2. Retrieve scientific publications to suggest (for this user context);
3. Filter candidate suggestions; and
4. Present the suggestions to the user.

[3] The OSCAR tools is run every night to process new ELN entries.

[4] www.chemspider.com

[5] This is a feature to be explored in future work.

[6] A fourth user, W, also provided a bibliography. However, the bibliography was relatively small and did not have a substantial overlap with PubMed references.

To simplify our investigation of suggesting references, in this work, we consider steps 1, 2 and 4 of the problem. We do not include any filters (step 3) to vet the suggestions against a list of references representing the user's prior reading history. Although such a filter would undoubtedly be useful (we return to this point in Section 6), our focus here is in characterising the transformation process from ELN content to query formulation.

For this investigation, we used four approaches for creating queries from the ELN content, specifically:

1. Chemical entities in a single ELN post;
2. The title of a single ELN post;
3. Salient terms from a single post; and
4. Overlapping terms from adjacent posts.

The first method has been deployed for the participants to trial. In this paper, however, we investigate the pros and cons of all methods.

Each method provides an ordered list of candidate query terms. However, the complete set of candidate query terms may be too restrictive to retrieve results. To determine the final set of query terms resulting from each of the four approaches, we use a filtering method for query terms which we refer to here as *iterative back-off*. This filter identifies the largest query set that retrieves results from PubMed. Essentially, the approach, outlined in Algorithm 1, continually drops the least ranked candidate until a non-null set is returned by the PubMed API. In this way, results are as specific as possible.

```
Data: Set of unique words, W
Result: Set of query words, Q where Q ⊂ W
Initialisation;
Q ← W;
Results ← pubmed(Q);
while Results is empty do
    WeakItem ← min_q(score(q) : for q in Q);
    Q ← Q \ WeakItem;
    Results ← pubmed(Q);
end
return Q;
```

**Algorithm 1:** The algorithm for iteratively trying queries until a non-null result is obtained from PubMed.

As a parameter, this filter requires a scoring function, $score(q)$, defined for each set of candidate terms. This function is used for sorting purposes. In the remainder of this section, we describe the four methods and the relevant scoring functions.

### 4.1 Chemical Entities in a Single ELN Post

Intuitively, chemical knowledge related to the user's current work may be useful in the query generation process. A starting point for this is to identify which words and phrases are in fact part of chemistry terminology and then to use these as queries. For each post in an ELN blog, the OSCAR tool provides a list of chemical entities referenced in the text.

Each of these entities has an associated confidence score from OSCAR. For the iterative back-off, we use this confidence score to sort the list of query candidates (based on chemical entities) in reverse order.

### 4.2 The Title of a Single ELN Post

As an alternative to using chemical terms as indicators of the experimental focus of a blog post, we can also use the words from the title. Title words are generally chosen to reflect the focus of the blog. Indeed this heuristic is used in text summarisation approaches to suggest keywords (Edmundson, 1969).

For each post, we retrieve the title, identify the words, and remove stopwords.[7] We use the relative placement within the title as a scoring mechanism for the iterative back-off method.

### 4.3 Salient Terms from a Single Post

To rank unique words (except for stopwords) based on their salience in the text we use one of two standard weighting methods: (1) Term Frequency (TF), or (2) Term Frequency with an Inverse Document Frequency factor (TF.IDF) (for an overview of Information Retrieval methods including TF and TF.IDF, see Manning et al. (2008).)

A priori, it is unclear as to which weighting method will be best, and so we test both variants in this work. The words with a high TF can be interpreted as an indicator of the content of the document. However, some words like "water" may occur often in the user's ELN blog. This could

[7] In the remaining methods, we define words as space delimited tokens with all non-alphanumeric characters replaced by space.

signify that it is a less important reactant in the experiment since it is a common substance used in all the user's experiments. This may be captured by the TF.IDF weighting.

Given a particular weighting scheme, to find the candidate list of query terms, we obtain a reverse sort of the unique words in the text (after removing stopwords) and then apply the iterative back-off approach to obtain a query set.

### 4.4 Overlapping Terms from Adjacent Posts

In this method, we try to make use of more context to find suggested literature. The intuition is that additional contextual information, for example the wider research goal of the user, will help provide better query terms. For example, in some ELN blogs, results for control conditions might be written up in a separate entry to the results for the test conditions for the independent variable. Using content from more than one LabTrove blog entry may thus provide additional experimental context.

We start by considering the preceding post to the post in question, using a Markov assumption that this captures the relevant experimental context. We compile unique words for both $\text{post}_i$ and $\text{post}_{i-1}$. We then take the intersection of these two sets. In this particular study, the list is assumed to be unranked (or tied). However, one could also employ a weighting scheme like TF or TF.IDF to rank the words. To help make the query more specific, we also only keep queries that are longer than 2 words.

We hypothesise that any experimental context that is useful in generating a query will be repeated in the adjacent posts. The advantage to this approach is its simplicity, we do not need to employ computationally expensive methods to identify in advance the set of posts in a blog that corresponds to a single research goal. We borrow from work in multi-document summarisation (for example, see Barzilay et al. (1999)) which treats words mentioned in multiple texts (in this case, both posts) as being particularly important in capturing background information.

## 5 Evaluation

In this investigation, we are interested in testing different query generation methods that are based on the experimental context found in the ELN blog. Although we intend for the suggestion of new literature to be presented during a knowledge discovery task, for simplicity, we examine the effects of the query generation methods on a bibliography reconstruction task for each of the three participant's blogs.

We do note, however, this ground truth version of "relevance" is limited for two reasons. Firstly, the bibliography is not exhaustive: that is, it does not evaluate the ability to count related articles outside the bibliography as useful suggestions and so it may miss relevant work (which is, in a way, the point of suggesting references). Secondly, the bibliography may also be too broad, containing not only work related to the central focus of the blog (or the user's core research), but any literature that the user deemed worth curating. While the evaluation of suggested literature based on bibliographies is not a perfect fit with the knowledge discovery application, it does allow us to study the query generation methods using intrinsic methods.

As an additional constraint in this work, we limit our investigations to PubMed which only contains a subset of research in the Analytical Chemistry, namely those to do with the Life Sciences. Research documented in the ELN that lies outside of this domain cannot be evaluated in this work.

Because of these limitations, the absolute value of the recall and precision metrics is not the focus of the study. Our aim is not to reconstruct the bibliographies. We use the metrics simply to rank the different query generation methods under review in this work.

### 5.1 Preparing the Bibliography Gold Standards

We used the three bibliographies volunteered by the users: L, R and D. The bibliographic files required preprocessing to convert them into sets of PubMed references, against which we compare our suggested references. The bibliographies were originally provided in EndNote format. Each EndNote file was converted into plain text, where each bibligraphic entry was transformed into a reference, one reference per line.[8]

We wrote a Python script to use the article title and date from the reference as search parameters in PubMed. Those entries that retrieved a corresponding PubMed identifier were kept and stored in a gold standard set for evaluation.

[8]We used a free evaluation copy of EndNote X6.0.1 (Bld 6599) for this conversion.

### 5.2 Procedure

We now describe the procedure for computing the suggested references that we wish to evaluate. For this study, we computed a set of references for each approach described above.

For each user blog, we compiled a suggested bibliography by using the following procedure:

1. For each blog entry in the blog, find suggested references (max 100) for the blog entry, using one of the above query generation procedures;

2. Take the union of all suggested references (excluding duplicates) and compare these to the user bibliography.

We repeated this procedure with each method for query generation outlined above. For each application of this procedure, we obtain a set of suggested PubMed unique identifiers. We compare these to the gold standard bibliographic sets (one for each user) of PubMed identifiers, and measure performance using the standard Information Retrieval (IR) metrics of recall and precision (for an overview of IR evaluation, see Salton and McGill (1983)).

### 5.3 Experiment Results

In this section, we provide the raw results from our evaluations against user bibliographies. As highlighted above, given the limitations of this evaluation framework, we are primarily interested in using the relative values to rank our query generation methods and to understand how they may be improved. The recall results are presented in Table 2 and the precision scores are presented in Table 3.

Note that the precision scores are very low because the suggested references are the union of the suggested references for each blog. We note however that Parra and Brusilovsky (2009) also report precision scores in similar ranges, indicating that other researchers have found the problem of literature recommendation to be a difficult problem with regard to precision. We list the precision results here for completeness but base our rankings on recall results, since this indicates the ability to find any relevant results. Due to the small sample size, we are unable to report significance. However, the rankings are still useful in determining which query generation methods show the most promise for further development.

| Method | L | R | D | Ave. |
|---|---|---|---|---|
| OSCAR4 | 3.4% | 2.3% | 2.4% | 2.7% |
| Expt. | 6.9% | 0.2% | 3.2% | 3.4% |
| Title | 3.2% | 3.2% | 4.0% | 3.7% |
| TF.IDF | 5.1% | 1.6% | 4.4% | 3.7% |
| TF | 8.6% | 1.3% | 7.3% | 5.7% |

Table 2: Recall scores (expressed as a percentage) for each method used independently. Legend: Columns show the recall scores for the three blogs and the average recall. "Expt." stands for experimental context.

| Method | L | R | D | Ave. |
|---|---|---|---|---|
| Title | 0.2% | 0.2% | 0.1% | 0.2% |
| TF.IDF | 0.1% | 0.3% | 0.2% | 0.2% |
| Expt. | 0.4% | 0.1% | 0.2% | 0.2% |
| OSCAR4 | 0.2% | 0.7% | 0.1% | 0.3% |
| TF | 0.4% | 0.2% | 0.2% | 0.3% |

Table 3: Precision scores (expressed as a percentage) for each method used independently. Legend: Columns show the precision scores for the three blogs and the average recall. "Expt." stands for experimental context.

We find that, with regard to recall, the best method for suggesting references is based on the Salience (TF) method using term frequencies for choosing keywords.

To determine if the approaches are complementary in nature, we combine them to see the effect on recall. If the margin of improvement is large enough, this suggests that relevant references being retrieved are not overlapping, and that the approaches can usefully be combined. We present the recall results in Table 4 (with precision results in Table 5 presented for completeness).

We find that the best result overall is indeed to use all approaches, for which we see an average recall of 9.3%. This represents almost 60% increase in recall over the best performing single method (Salience TF) which achieved a recall of 5.7% on average. Note however that this combined result is only marginally better than the slightly less complex combination which uses the Title, OSCAR4 and Salience (TF), which obtains a recall of 9.2%.

## 6 Discussion and Future Work

There are a two research avenues we would like to pursue: (1) improving the methods for query generation, (2) conducting further experimentation on

| Method | L | R | D | Ave. |
|---|---|---|---|---|
| M1 | 5.9% | 5.3% | 5.7% | 5.6% |
| M2 | 11.3% | 6.0% | 10.2% | 9.2% |
| M3 | 11.5% | 6.0% | 10.5% | 9.3% |

Table 4: Recall scores (expressed as a percentage) for method used in combination. Legend: Columns show the recall scores for the three blogs and the average recall. M1: Title, OSCAR4 methods; M2: M1 with TF; M3: M2 with Experimental Context.

| Method | L | R | D | Ave. |
|---|---|---|---|---|
| M1 | 0.2% | 0.3% | 0.0% | 0.2% |
| M2 | 0.2% | 0.2% | 0.1% | 0.2% |
| M3 | 0.2% | 0.2% | 0.1% | 0.2% |

Table 5: Precision scores (expressed as a percentage) for method used in combination. Legend: Columns show the precision scores for the three blogs and the average recall. M1: Title, OSCAR4 methods; M2: M1 with TF; M3: M2 with Experimental Context.

performance.

In this study, we found that the Salience (TF) method is the best approach, which accords well with textbook approaches to generic query generation. However, it is interesting to note that using chemical entities retrieves a complementary set of references to the Salience (TF) method, as evidenced by the gain in recall performance as we combine these approaches.

Better methods for incorporating chemistry domain information might still be possible, perhaps by using the IDF approach to model which chemical entities are salient across the entire blog and thus across the experimental context. In addition, we can experiment with the use of the chemical named entities detected by the OSCAR tool that describe chemical processes.

Implementation of a larger experimental context method was not overly successful. Recall that our hypothesis was that what was common between two adjacent posts would be important. Even when using the experimental context with other methods (M3), we only observed a slight benefit.

There are a number alternative approaches to using a larger experimental context. Perhaps it might be the differences and not the similarities between the posts that are more useful as query terms for retrieving literature.

We could also take a different approach to capturing the research goals of the student as captured by the blog. It may be the case that more than one post is required for this purpose, or that the simple adjacency of posts is not sufficient for capturing the context of the overaching research goals in general. If the latter, we could first segment the blog into portions, where each portion represents a linguistically coherent set of text describing laboratory tasks that correlates to some larger research goal. We could then generate a query for each segment. For this task, we might employ text segmentation approaches which use dramatic changes in vocabulary to signify a new topical segment (for example, see Hearst (1994) as the seminal work in such text segmentation approaches). This might also hopefully improve recall since retrieval would be based on segments and not blog posts.

Interestingly, the evaluation results suggest that the blogs might themselves be different. For example, the suggested references for ELN Blog R consistently under-performs compared to Blog L and D. This could be because there are fewer entries in Blog R. As we are using only three blogs (limited by the number of bibliographies we were provided), our results might be heavily affected by the individual variations in the blogs. Ideally, we would repeat this experiment with a larger number of blogs to gain a more stable impression of the strengths and weaknesses of the various query generation methods.

We can also employ post-processing methods on both the query generation and literature retrieval processes. Query expansion methods (for example, see Jones et al. (2006)) could help select additional search terms for the set of query terms selected after the iterative back-off process. In addition, for a real knowledge discovery scenario, we could filter the retrieved references that the user is already aware of.

The evaluation task presented here simply looked at strict comparisons against user bibliographies. As described in Section 5, this approach does not have the ability to reward relevant articles that do not belong to the gold standard. One avenue for future research is to explore methods like those described in (Büttcher et al., 2007) to handle unknown documents for which we have no relevance judgements.[9]

[9]We thank the anonymous reviewers for this suggestion.

We could also consider a looser evaluation which examines articles commonly cited by the suggested references, as is done in (Jha et al., 2013). This would allow the ability to detect older seminal articles that we may not be able to recover using generated queries if that seminal work uses vocabulary that is different to contemporary research. Appropriately handling these by counting them as matched if one or more suggested references cite them may help provide a better understanding of the performance of the system.

Finally, we are now collecting user interface data with which to conduct user studies. By analysing cases where the user clicked the PubMed links based on the abstract of the suggested reference, we may be able to learn if the system is able to present useful recommendation in a real research context.

## 7 Related Work

Representing the user's context as part of an information need is an open research question. In related work by Wan and Paris (2008), the user's reading context was used to summarise Wikipedia text[10]. Similar methods have been used for summarising scientific literature to capture the user's context (Mei and Zhai, 2008).

There are a number of related works sharing the same motivation of helping researchers keep in touch with current scientific developments. Research in automatically generating literature surveys focuses on generating the text of the survey using summarisation methods (for example, see Mohammad et al. (2009)). However, that work does not tackle the problem of suggesting the references themselves. In work by Jha et al. (2013), articles are retrieved from a query provided by the user and a survey is generated from these. The authors used a results expansion method that adds certain cited references from the retrieved articles. Although they also retrieve references, our problem is different in that we have to automate the query generation from some textual documents representing the user's experimental context.

The work described here is more akin to link creation, where we postulate a link from an ELN entry to an article. In most link creation work, there is a pre-existing list of potential candidates to link to. For example, in work on linking Wikipedia pages, the candidate pages are the existing wikipedia pages whose title occurs in the potential linking page (for example, see Milne and Witten (2008)). In our case, such a correspondence between the linking page and the potential link target does not exist.

[10]www.wikipedia.org

Previous work has examined the problem of recommending articles to users, but this has usually been performed using topic modelling approaches to identify similarities amongst articles (Wang and Blei, 2011), or else capitalising on social and collaborative networks for sharing publications like CiteULike[11], Mendeley[12] and Bibsonomy[13] where suggestions are based on collaborative filtering methods (for example, see Bogers and Van den Bosch (2008) and Parra and Brusilovsky (2009)). In these works, the evaluations have opted for task-based user studies (Parra and Brusilovsky, 2009).

## 8 Conclusions

In this work, we explore the problem of using electronic laboratory notebooks to suggest literature to a researcher. The aim is to help researchers keep abreast of scientific developments whilst their work is continuing. We use the notebook entries to generate queries which are sent to PubMed to retrieve scientific literature. In this paper, we presented recall and precision results when comparing against lists of references known to be relevant, which we source from the bibliography files of the ELN users. We find that our combined method for query generation, using both traditional information retrieval methods and chemistry NER achieves 60% improvement over the best performing single method, using term frequency methods. This suggests that the methods presented in this paper are the first steps towards utilising the user's experimental context to suggest literature for a knowledge discovery task.

## 9 Acknowledgements

We thank the Digital Libraries Division of the UNSW for supporting this work. We appreciate the help of Brynn Hibbert and his students L, R, W and D for kindly sharing their blog data and bibliographies. Finally, we thank Cameron Neylon for allowing us to reproduce content from his LabTrove blog for demonstration purposes.

[11]www.citeulike.org
[12]www.mendeley.com
[13]www.bibsonomy.org

## References


Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 550–557, Morristown, NJ, USA. Association for Computational Linguistics.

Toine Bogers and Antal Van den Bosch. 2008. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290. ACM.

Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung, and Ian Soboroff. 2007. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 63–70, New York, NY, USA. ACM.

Peter Corbett and Peter Murray-Rust. 2006. High-throughput identification of chemistry in life science texts. In *Proceedings of the Second international conference on Computational Life Sciences*, CompLife'06, pages 107–118, Berlin, Heidelberg. Springer-Verlag.

Peter Corbett, Colin Batchelor, and Simone Teufel. 2007. Annotation of chemical named entities. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, pages 57–64, Stroudsburg, PA, USA. Association for Computational Linguistics.

H Edmundson. 1969. New methods in automatic abstracting. *Journal of ACM*, 16(2):265–284.

Marti Hearst. 1994. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9 – 16, New Mexico State University, Las Cruces, New Mexico.

David Jessop, Sam Adams, Egon Willighagen, Lezan Hawizy, and Peter M. Rust. 2011. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3(1):41+.

Rahul Jha, Amjad Abu-Jbara, and Dragomir Radev. 2013. A system for summarizing scientific topics starting from keywords. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 572–577, Sofia, Bulgaria, August. Association for Computational Linguistics.

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 387–396, New York, NY, USA. ACM.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824, Columbus, Ohio, June. Association for Computational Linguistics.

David Milne and Ian H. Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

Andrew J. Milsted, Jennifer R. Hale, Jeremy G. Frey, and Cameron Neylon. 2013. Labtrove: A lightweight, web based, laboratory blog as a route towards a marked up record of work in a bioscience research laboratory. *PLoS ONE*, 8(7):e67460, 07.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 584–592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Denis Parra and Peter Brusilovsky. 2009. Collaborative filtering for social tagging systems: an experiment with citeulike. In *Proceedings of the third ACM conference on Recommender systems*, pages 237–240. ACM.

G. Salton and M. J. McGill. 1983. *Introduction to modern information retrieval*. McGraw-Hill, New York.

Stephen Wan and Cécile Paris. 2008. In-browser summarisation: Generating elaborative summaries biased towards the reading context. In *Proceedings of ACL-08: HLT, Short Papers*, pages 129–132, Columbus, Ohio, June. Association for Computational Linguistics.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 448–456, New York, NY, USA. ACM.