

University of Sheffield

William Briggs 6 Month Report



William Briggs

Supervisor: Dr Mark Stevenson

Panel: Dr Andreas Vlachos, Professor Richard Clayton

Department of Computer Science

May 7, 2018

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name:

Signature:

Date:

Abstract

Medical data comes in large volumes, making it a challenge for systematic reviewers to process and find relevant information. Being able to apply automation techniques to this field presents itself as a suitable application for natural language processing. We will look at two areas of this field. Biomedical information retrieval by efficiently indexing and querying a large medical database, can we return an optimum set of documents. Stopping criteria, given an existing set of rankings, what is the most efficient way to look at these documents and determine a cut-off (stopping) point.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Literature Survey | 2 |
| 2.1 | Steps of a Systematic Review | 2 |
| 2.1.1 | Question Definition | 2 |
| 2.1.2 | Relevant literature search | 3 |
| 2.1.3 | Data Filtering | 3 |
| 2.1.4 | Data Extraction | 4 |
| 2.2 | Indexing and Querying Medline and Automated Query Generation | 4 |
| 2.2.1 | Automated Query Generation | 4 |
| 2.3 | Stopping Criteria | 5 |
| 2.3.1 | Evaluation Metrics for Finding Stopping Points | 5 |
| 2.3.2 | Existing Stopping Methods | 6 |
| 3 | Research Questions | 10 |
| 3.1 | Automated Decision Making of Relevant Studies | 10 |
| 3.2 | Information Extraction From Studies | 11 |
| 4 | Novel Work | 12 |
| 4.1 | Baseline Approaches to Stopping | 12 |
| 4.1.1 | Percentage cut-off | 12 |
| 4.1.2 | Similarity score cut-off | 13 |
| 4.2 | Random Sample Method to Stopping | 13 |
| 4.2.1 | Curve Fitting | 14 |
| 4.2.2 | Gaussian Process fitting | 16 |
| 4.2.3 | Conclusion on Curve fitting and GP | 17 |
| 4.2.4 | Gaussian Process Conclusion | 18 |
| 4.3 | Indexing and Querying Medline with Limited Information | 18 |
| 4.3.1 | Acquiring Key Information from A Systematic Review Protocol | 18 |
| 4.3.2 | Indexing medline | 19 |
| 4.3.3 | Results | 20 |
| 4.3.4 | Medline automatic query Conclusion | 20 |

| | |
|--|-----------|
| <i>CONTENTS</i> | 4 |
| 5 Future Work | 22 |
| 6 Doctorial Development Programme | 23 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | Visualisation of target method last relevant document selection. C is number of documents in collection. | 7 |
| 2.2 | Example of using knee method to find a stopping point. Image inspired from [15] | 8 |
| 3.1 | Tree diagram example, all credit for diagram goes here [8] | 11 |
| 4.1 | Visualisation of using a confidence interval for predicting a stopping point. . . | 15 |
| 4.2 | Visualisation of using a confidence interval for predicting a stopping point using a gp. We can see the gp is significantly over fitting to the data. | 17 |

List of Tables

| | | |
|-----|---|----|
| 4.1 | Percentage cut-off rates for Test_Data_Sheffield-run-2 | 12 |
| 4.2 | Similarity cut-off comparison for stopping for Test_Data_Sheffield-run-2. Using cosine similarity scores. | 13 |
| 4.3 | Comparison of different of sample method using curve fitting for different CLEF 2017 runs. Sample size = 3. Results are taken as averages over all topics for search method. No topic drop-out | 15 |
| 4.4 | Comparison of different of sample method using curve fitting for different CLEF 2017 runs. lower = lower-bound confidence interval. Sample size = 3. Results are taken as averages over all topics for search method. with 0.5% cut-off | 16 |
| 4.5 | Lowest effort possible to find 70% of relevant documents. | 16 |
| 4.6 | Comparison of different of sample method using gp for different CLEF 2017 runs. lower = lower-bound confidence interval. Sample size = 3. Results are taken as averages over all topics for search method. with 0.5% cut-off | 17 |
| 4.7 | Comparing target method with curve fitting along with confidence interval. Using Sheffield-run-2 | 18 |
| 4.8 | Results for IR medline system. Comparison for both 5000 and 25000 thresholds | 20 |

Chapter 1

Introduction

Medical literature poses interesting challenges for Natural Language Processing (NLP) researchers. The sheer volume of medical data makes it difficult for humans to process efficiently.

Evidence-based medicine has become an important aspect in health care and policy making. One key task is the creation of systematic reviews. Systematic reviews are transparent reviews that aim to pull together and critically analyse and summarise relevant literature to a topical question. The process of creating a systematic review is rigorous and time consuming with varying degrees of complexity in-between steps. This report will look at the existing work done using NLP as part of the systematic review process as well as the novel work done by myself so far.

We first review the stages involved in creating a systematic review. We break the steps down by looking at the PICO strategy [12]. By breaking the steps down, it becomes easier to examine potential candidates for applying NLP techniques to the process. Areas for research are then identified.

We then move on to look at stopping methods for systematic reviews. Stopping methods are about finding a suitable stopping point given a list of ranked documents. Two stopping existing methods are examined; the target method and the knee method.

The work completed so far is then presented. We look at using curves to make predictions of finding a stopping point, including using a Gaussian process. We then present our work for the automatic query generation process by using systematic review protocols as a basis for inferring the query.

Chapter 2

Literature Survey

Systematic reviews have many different stages that propose themselves as a candidate for automation. This section is going to look at the techniques that have been applied for some of these stages in previous literature.

2.1 Steps of a Systematic Review

It is useful for us to break down the steps involved in creating a systematic review into subtasks. This way we can observe what techniques can be applied during the relevant subtasks to improve the efficiency of the process. The following definitions are derived task simplifications from the cochrane tutorial on systematic reviews: [11].

1. Question definition.
2. Relevant literature search.
3. Data Filtering.
4. Data Extraction.
5. Analysis and Data combination.

2.1.1 Question Definition

One of the best known techniques for formulating a systematic review question is known as the PICO strategy [12]. This technique focuses on exposing 4 pieces of information in the systematic review question: patient population, intervention or exposure, comparison or control and outcome.

Example: (credit goes to [12])

”Is animal-assisted therapy more effective than music therapy in managing aggressive behaviour in elderly people with dementia?”

| | |
|---|--------------------------------|
| P | elderly patients with dementia |
| I | animal-assisted therapy |
| C | music therapy |
| O | aggressive behaviour |

A potential point of interest would be attempting to generate these questions automatically given some literature context.

2.1.2 Relevant literature search

After formulating a question, systematic reviews need to search for the relevant literature that surrounds this question.

Large medical database-such as Pubmed [13] contain relevant studies that can be used to create the review. These databases are typically very large and require concise queries to efficiently retrieve data.

Naturally this can be modelled as an information retrieval problem. We have a large number of documents and we wish to retrieve the most relevant ones. One task for the 2017 CLEF conference was to produce a ranking of the most relevant documents for topics [6]. Many techniques have been proposed for ranking of relevant documents, with varying degrees of performance [1] [4] [7].

An important aspect of the relevant literature search step is the construction of the query. Query creators often apply filters (also known as hedges) to increase the effectiveness or/and the efficiency of the searching. Two key attributes for the query are the precision and the sensitivity (aka recall). By including synonymous phrases e.g: quality adjusted life or quality of well-being or disability adjusted life the sensitivity can be increased, but as expense of the precision. The creation of this query is a task that could potentially have some aspects of NLP applied to it.

2.1.3 Data Filtering

The data filter stage involves reducing the amount of documents returned by the initial query down to a smaller subset of relevant document. This is can also be referred to as the abstract screening phrase [6].

The length of this stage is highly dependant on how many documents were returned by the initial query, often in the excess of 5000 studies for a single query. In response to this, stopping criteria methods have been proposed that aim to optimize two key parameters; the effort and the recall. That is to say we want to get as many relevant documents as possible, whilst

looking at the fewest. Examples of approaches include the knee method [15] and the target method [3]. Other techniques could be applied and evaluated such as curve fitting.

2.1.4 Data Extraction

The data extraction phase involves pulling the relevant information from the filtered subset of studies. Examples of important information includes how many people took part in the study and what the results were.

Being able to extract the relevant information from studies presents itself as an information extraction problem. The task to automate the process of extracting relevant information would reduce time and complexities of manually reviewing studies [5].

2.2 Indexing and Querying Medline and Automated Query Generation

Medline is a large collection of medical literature and data from around the world. Typically each entity will contain a title and an abstract containing some information on the study. Whilst Medline as a whole is very easy to access [9], the large size and complexity of the data makes it difficult to retrieve the relevant information.

Being able to create a reliant index of Medline would help with the effectiveness of the queries. As such existing medline indexes and IR systems have been created [10].

2.2.1 Automated Query Generation

Being able to automate query generation for literature searching would save systematic reviewers a significant amount of time. However, medical literature queries are typically complex and contain multiple levels of logical operators and synonymous term look ups. This makes the task of creating a query manually in-its-self a challenging piece of work.

Rapid Automatic Keyword Extraction Algorithm

Rapid automatic keyword extraction (RAKE) is a keyword extraction algorithm was proposed by Rose, Engel and Cramer in 2010 [14]. This algorithm is used for taking the key pieces of information from text and is useful the domain of information extraction. This algorithm is of interest to us as it has potential usage within the field of query generation.

RAKE heavily relies on stop-words and punctuation separators as an indicator for the importance of a phrases and words. RAKE will iterator over sequences of words until a

stop-word or separator is found, this phrase/word is then split and extracted. Frequency of occurrence (tf) and word co-occurrence matrices can then be used to reduce the key-word set down further.

RAKE can be further optimized by specifying minimum term frequency rates to capture more prominent terms.

2.3 Stopping Criteria

Stopping criteria is about finding the optimum point in a set of documents. This could be useful in a decision making process. Consider having 100 relevant documents, where each document contains a binary value. If we looked at 1/3 of these documents and saw a trend of positive values, we could use this to infer the reliability of the remaining documents.

Another use of stopping criteria is when filtering through potentially relevant documents. Consider a query that returns 10000 documents, of which only a small sub-set of these are relevant. Reviewers would need to filter through each of these 10000 documents to pull out the relevant ones. Or it could be that the reviewers are happy to hit a 90% recall of relevant documents, and are happy to miss the remaining 10% in exchange for time-saved.

Two key methods have been proposed for finding stopping points so far, the target method [15] and the knee method. [3]. Both these methods are discussed below 2.3.2

2.3.1 Evaluation Metrics for Finding Stopping Points

In order to evaluate the suitability of our stopping method, we can use two evaluation metrics. The recall, which is simply the number of documents returned for a topic, and effort which is the number of documents we had to look at for a topic.

$$Recall = \frac{R}{|D|} \quad (2.1)$$

Where R is the number of returned relevant documents and D is the set of all relevant documents.

$$Effort = \frac{L}{|D|} \quad (2.2)$$

Where L is the number of returned documents looked at.

Naturally we could exclusively optimized each of the parameters by either returning everything in the document collection ($R = |D|$) or by just looking at a single document. ($L = 1$)

Therefore it becomes difficult for us to evaluate our stopping criteria as we need to consider both of these parameters adjacently.

In response to this we can make use of two more evaluation metrics that were proposed by Cormack and Grossman [3]:

$$reliability = P[acceptable(S) == 1] \quad (2.3)$$

reliability is computed over all searches and is read as the probability of the acceptability being 1. Where acceptability is calculated as:

$$acceptability(S) = \begin{cases} 1, & recall(S) \geq 0.7. \\ 0, & recall(S) < 0.7. \end{cases} \quad (2.4)$$

A stopping point is deemed to be acceptable if 70% of the relevant documents have been found. As such, the reliability is an average over a search method.

2.3.2 Existing Stopping Methods

There are different approaches can be experimented with in finding an optimum stopping point. Consider a percentage cut-off method, where we use the score similarity score for deciding if its worth continuing to look down the rankings:

$$Difference(D_i, D_i + 1) > C \quad (2.5)$$

Where difference returns a score of how close document D_i and $D_i + 1$ are together and C is a cut-off constant. We can expand this to an example:

$$(1 - (0.73/0.75)) * 100 > 0.015 \quad (2.6)$$

Here we are saying if the two documents' scores are above 1.5% then we should stop looking down the rankings.

We a basis of how stopping methods work, we can move on to more established and defined methods.

Target

The target method is an approach that can guarantee a certain a certain level of reliability 2.3.1. It was first proposed by Cormack and Grossman [3].

The target T denotes how many documents we should randomly select from our initial query. A larger value of T will increase the effort required as we are more likely to select a document towards the end of the query set. Documents are looked at until the target point T has been reached.

We first compute a random target set of relevant documents. We then calculate the last document in the target set and mark that as our target point:

$$d_{last} = \underset{d \in T}{\operatorname{argmax}} \operatorname{relrank}(d) \quad (2.7)$$

It must hold that d is in the target set.



Figure 2.1: Visualisation of target method last relevant document selection. C is number of documents in collection.

Increasing our target set size is likely to increase the probability the last document being towards the end of the document collection.

We can calculate the recall of the point by looking at the relevance rank of the last document:

$$\operatorname{recall} = \frac{\operatorname{relrank}_{last}(d)}{R} \quad (2.8)$$

Where R is the number of relevant documents.

For our method to be deemed reliable we must achieve 70% recall with a 95% average over all topics.

$$P\left(\frac{\operatorname{relrank}_{last}(d)}{R} \geq 0.7\right) \geq 0.95 \quad (2.9)$$

Assuming we have a large number of relevant documents R we need to determine cut-off c

$$P(R - \text{relrank}_{last}(d) > cR) = 0.05 \quad (2.10)$$

Which translates to the probability of the remaining relevant documents being higher than the cut-off point should be 0.05.

While this approach is shown to acquire 95% reliability, the effort needed is often significantly highly, often requiring us to look at huge volume of documents.

Knee Method

A different stopping method proposed by [15] is known as the knee method. This approach uses a curve to generate a 'knee', which is then used for predicting a stopping point. This approach is likely to be highly dependant on the quality of the initial rankings. This is because we need a curve that reaches a peak quickly, before flattening out.

We use a vertical line panning the length of the ranking set and use it to calculate the distance from the ranking at each point. The point with the maximum distance is chosen as a suitable stopping point.

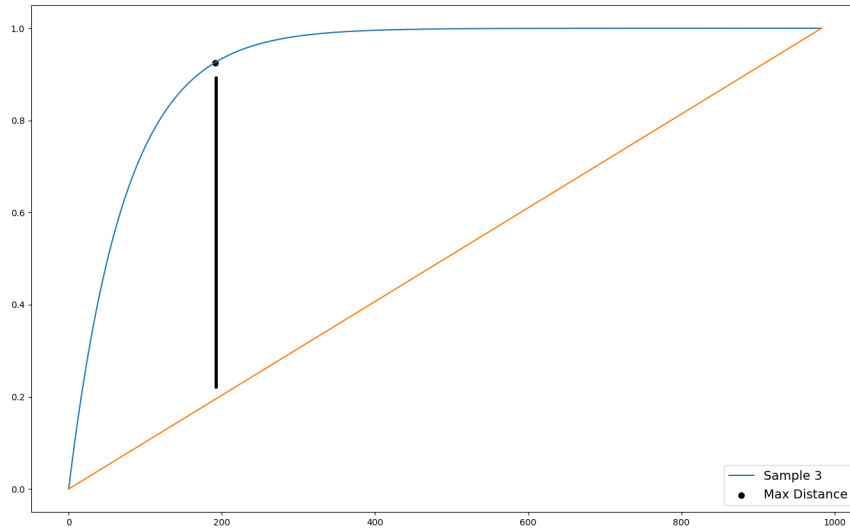


Figure 2.2: Example of using knee method to find a stopping point. Image inspired from [15]

We can see in the above example the method has predicted we look at around 200 of the 1000 documents to achieve a suitable stopping point.

This method also imposes an additional constraint for rankings of a large volume.

It was found that the knee method is a better approach for finding a stopping point than the Target method [3]. The recall was always found to be better and the reliability was found to be the same or higher for 6 out of 8 of the test collections.

Chapter 3

Research Questions

In this chapter we will present some possible research areas for systematic reviews.

3.1 Automated Decision Making of Relevant Studies

Once relevant studies have been identified by their abstracts, reviewers are required to process the studies and extract useful information for the systematic review. Many of these studies will not be useful and can be discarded, but not without the cost of the reviewer having to look at the content.

A useful piece of research would be to determine if a study is relevant to the systematic review question.

There are two potential routes that could be taken for the investigating this task. The first approach would be to use existing information from the systematic review (ie the protocol) and determine if the study contains the information. A second approach would be to a semi-supervised learning method, having the reviewer look at a subset of the studies, so we can then build a classifier for the remaining studies. Both of these approaches can be applied to the abstract screening and the data filtering stage of the systematic review. For the abstract screening we would be trying to predict if a study is relevant, by looking at the abstract. For the data filtering stage we would be looking at the actual text of the study (typically a pdf file)

There are many challenges for this research topic. Not all studies are publicly available and are often protected by publishing licences. This makes it difficult to gather data. Another challenge is having to deal with such a broad range of data, as well as the pdf format studies.

3.2 Information Extraction From Studies

In a similar vein to the first topic 3.1 we could also look at extracting the relevant information from a study. Often empirical data is identified in studies and extracted to generate a tree diagram:

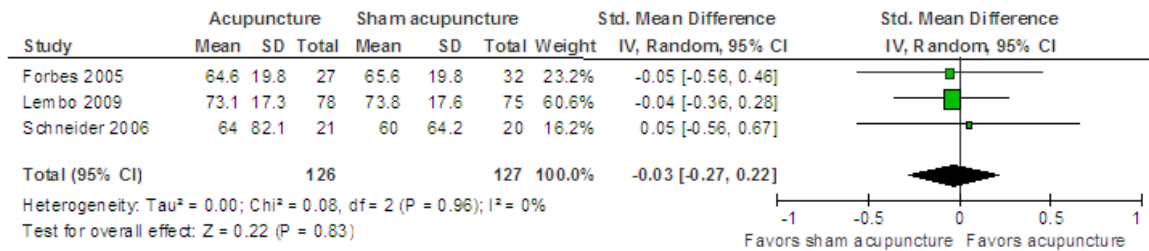


Figure 3.1: Tree diagram example, all credit for diagram goes here [8]

These tree diagrams contain summarized information from the relevant studies. Studies that contain more useful (i.e. higher number of participants) have smaller confidence margins and vice-versa for smaller studies.

An interesting area of research would be identifying empirical data in a study to generate these tree diagrams automatically.

The challenge in this task is gathering training data that is sufficiently marked as to where key information occurs within a study. This would also still involve working with a limited number of studies and pdf documents.

Chapter 4

Novel Work

In this section the work completed so far will be presented. Two main areas of the systematic review process has been focused on. Stopping criteria and indexing/querying pubmed.

4.1 Baseline Approaches to Stopping

As discussed in 2.3.2 there are many baseline approaches we can take for finding a suitable stopping point. Most approaches are heavily dependant on the initial rankings of the document collection, and assume more relevant documents feature towards the start of the collection.

4.1.1 Percentage cut-off

One approach is we can simply take a cut of the document collection and evaluate how many relevant documents we have retrieved. Assuming we are satisfied with 95% reliability, this could be a suitable method.

| % of Documents | recall | reliability | effort |
|----------------|--------|-------------|--------|
| 10% | 0.49 | 0.1 | 0.16 |
| 25% | 0.74 | 0.25 | 0.66 |
| 50% | 0.91 | 0.5 | 0.93 |
| 75% | 0.98 | 0.75 | 1.0 |
| 90% | 0.99 | 0.9 | 1.0 |

Table 4.1: Percentage cut-off rates for Test_Data_Sheffield-run-2

By looking at half of the rankings we achieve a reliability of 0.93.

4.1.2 Similarity score cut-off

Many document rankings will feature a similarity score between the document and the query. As we descend down the rankings, the score decreases in value as documents become less relevant.

We can make use of the similarity score to derive a stopping point. This method works by looking at documents d_i and $d_i + 1$ and determining if the difference between the similarity scores has become too large. This method will work on the basis that documents that are no longer relevant will have a sudden drop in score such that we can identify this as our stopping point.

| $diff(d_i, d_i + 1)$ | recall | reliability | effort |
|----------------------|--------|-------------|--------|
| 0.01% | 0.025 | 0.0023 | 0 |
| 0.05% | 0.12 | 0.1 | 0.1 |
| 1% | 0.359 | 0.333 | 0.333 |
| 2% | 0.88 | 0.86 | 0.86 |
| 5% | 1 | 0.99 | 1.0 |

Table 4.2: Similarity cut-off comparison for stopping for Test_Data_Sheffield-run-2. Using cosine similarity scores.

These results are highly sensitive to the similarity score and show it's difficult to use this score as an affective measure for stopping. We found similarity scores rarely have sudden drops in values, making it difficult to use this method to identify a stopping point.

4.2 Random Sample Method to Stopping

As an approach to determining when to stop looking at document abstracts returned by the query we are proposing a new sampling method. This approach assumes we have sensible ranking algorithm for returning documents for a query.

The first step of this method is to generate a sample set. We will use an interval method for generating our set, i.e select every N th document

$$U = \frac{|D|}{S} \quad (4.1)$$

Where U is the computed randomised subset, D is the document collection and S is the sample size.

We then use this subset U to create a model / baseline for our topic as a way of predicting how many documents one would need to look at to hit a threshold. The intuition behind

this approach is that the rate of which relevant documents occur should be relatively similar when the number of returned documents is the same.

4.2.1 Curve Fitting

Our first approach uses a simple curve fit against a sample set along with a non-linear function.

$$F(x) = n - a \exp^{-kx} \quad (4.2)$$

Where a , k and n are learnt weights and x is an associated return rate for a document. We generate the curve using the non-linear least squares algorithm [16].

The first sample size of 1 is included to show how the effort metric is affected. For us to use sample everything, we would need to look at everything, as such the effort averaged out at 1. In this example we were only concerned in achieving 70% recall, as such even when sampling everything we would really obtain 100% recall at the expense of 100% effort.

Looking at every 3rd document and then producing a prediction curve will reduce our effort. We are still required to look at 1/3 of documents, as such the effort will always be above 0.33.

Relevance Ranking

Our results so far have been based on Test_Data_Sheffield-run-2 [1] of CLEF 2017 Technology Assisted Reviews in Empirical Medicine [2]. Naturally, the reliability of our curve is heavily based on how good the initial rankings are for each topic. We can compare different ranking methods for generating our stopping curve. By looking at the CLEF 2017 technology assisted review task [6] we can determine the best candidates to use. All CLEF 2017 submissions follow a standard format that we use to represent the ranks of the documents:

CD010775 NF 19307324 1 0.27152011529138564 Test-Data-Sheffield-run-2

We also introduce a new column: topics sampled is the number of topics that were evaluated using the curve. This is included as some of the ranking methods do not produce enough relevant documents to generate a suitable curve.

| Submission | recall | reliability | effort | topics sampled |
|---------------------------|--------|-------------|--------|----------------|
| Test_Data_Sheffield-run-2 | 0.68 | 0.53 | 0.50 | 30 |
| Waterloo A-rank-cost | 0.70 | 0.46 | 0.44 | 30 |
| Waterloo B-rank-cost | 0.70 | 0.46 | 0.40 | 30 |
| auth run-1 | 0.74 | 0.53 | 0.43 | 30 |
| auth run-2 | 0.70 | 0.50 | 0.43 | 30 |
| ntu run-1 | 0.71 | 0.43 | 0.68 | 30 |
| ucl full-text | 0.86 | 0.8 | 0.91 | 30 |

Table 4.3: Comparison of different of sample method using curve fitting for different CLEF 2017 runs. Sample size = 3. Results are taken as averages over all topics for search method. No topic drop-out

This second set of results will use a mandatory cut-off parameter for topics with less than 0.5% of relevant documents. The maximum number of topics for this dataset remains at 30.

We also include a confidence interval evaluation for lower bounded range of a 3σ confidence interval. The key advantages of using a curve as method of evaluating stopping criteria is being able to make use of this confidence interval in a real systematic review. In the context of a systematic reviewers at the data filtering stage, we could specify that the system is 95% certain that 70% of relevant documents have been found. At which point the reviewer can decide if its worth continuing to look at documents.

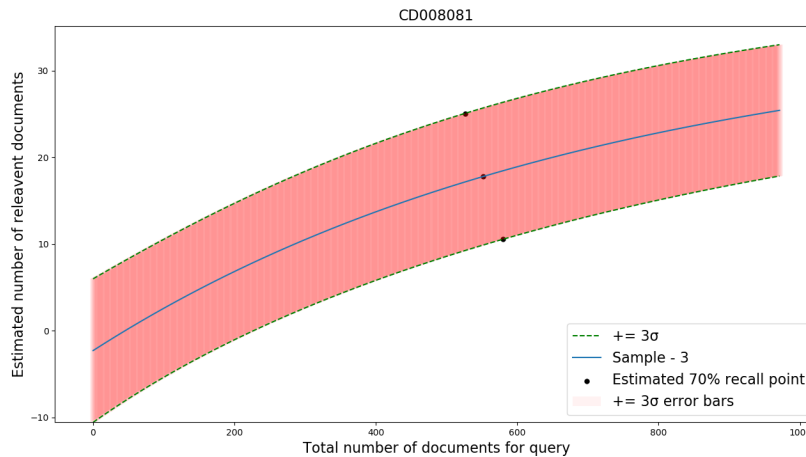


Figure 4.1: Visualisation of using a confidence interval for predicting a stopping point.

| Submission | recall-lower | reliability-lower | effort-lower | topics sampled |
|---------------------------|--------------|-------------------|--------------|----------------|
| Test_Data_Sheffield-run-2 | 0.69 0.74, | 0.52 0.60 | 0.48 0.51 | 23 |
| Waterloo A-rank-cost | 0.71 0.73, | 0.47 0.47 | 0.43 0.44 | 23 |
| Waterloo B-rank-cost | 0.71 0.75, | 0.52 0.82 | 0.41 0.43 | 23 |
| auth run-1 | 0.72 0.74, | 0.52 0.60 | 0.41 0.42 | 23 |
| auth run-2 | 0.70 0.72, | 0.52 0.60 | 0.42 0.43 | 23 |
| ntu run-1 | 0.76 0.74, | 0.56 0.52 | 0.72 0.70 | 23 |
| ucl full-text | 0.86 0.94, | 0.82 0.86 | 0.91 0.95 | 23 |

Table 4.4: Comparison of different of sample method using curve fitting for different CLEF 2017 runs. lower = lower-bound confidence interval. Sample size = 3. Results are taken as averages over all topics for search method. with 0.5% cut-off

We have deliberately compared two of the better participant rankings (Waterloo and auth) and two of the lower performers (ntu and ucl). We can see the quality of the initial rankings significantly influences the performance of our stopping criteria. This suggests there is a important relationship between using a curve to predict a stopping point and how good the initial ranking of documents is.

Some of the ranking methods struggle to produce curves and when combined with a cut-of parameter produce become not worth considering in our evaluation. In this situation we simply returned everything for the given topic, resulting in 100% recall at the expense of 100% effort.

Finally we look at the oracle scores for each run. These are the best possible scores we can achieve for the given rankings. We will assume we are satisfied with 70% recall. Results are taken as averages over all topics.

| Submission | recall | reliability | effort |
|---------------------------|--------|-------------|--------|
| Test_Data_Sheffield-run-2 | 0.7 | 1.0 | 0.11 |
| Waterloo A-rank-cost | 0.7 | 1.0 | 0.07 |
| Waterloo B-rank-cost | 0.7 | 1.0 | 0.06 |
| auth run-1 | 0.7 | 1.0 | 0.08 |
| auth run-2 | 0.7 | 1.0 | 0.09 |
| ntu run-1 | 0.7 | 1.0 | 0.4 |
| ucl full-text | 0.7 | 1.0 | 0.67 |

Table 4.5: Lowest effort possible to find 70% of relevant documents.

4.2.2 Gaussian Process fitting

As an alternate approach to fitting a simple curve, we can apply a Gaussian Process (GP).

We will apply a constant kernel plus a squared-exponential kernel. One of the key challenges when applying a Gaussian Process was the amount of over-fitting.

| Submission | recall-lower | reliability-lower | effort-lower | topics sampled |
|---------------------------|--------------|-------------------|--------------|----------------|
| Test_Data_Sheffield-run-2 | 0.73 0.73, | 0.73 0.73 | 0.50 0.50 | 26 |
| Waterloo A-rank-cost | 0.70 0.70, | 0.40 0.40 | 0.42 0.42 | 25 |
| Waterloo B-rank-cost | 0.73 0.73, | 0.71 0.71 | 0.41 0.41 | 27 |
| auth run-1 | 0.74 0.75, | 0.56 0.60 | 0.42 0.39 | 23 |
| auth run-2 | 0.72 0.73, | 0.52 0.52 | 0.42 0.42 | 23 |
| ntu run-1 | 0.67 0.67, | 0.40 0.40 | 0.64 0.64 | 22 |
| ucl full-text | 0.62 0.62, | 0.52 0.56 | 0.82 0.82 | 25 |

Table 4.6: Comparison of different of sample method using gp for different CLEF 2017 runs. lower = lower-bound confidence interval. Sample size = 3. Results are taken as averages over all topics for search method. with 0.5% cut-off

We can see the gp does a better job at fitting to the topics than the curve method. However, for most of the lower ranking runs the effort is increases significantly, suggesting the gp is not fitting very well to the data.

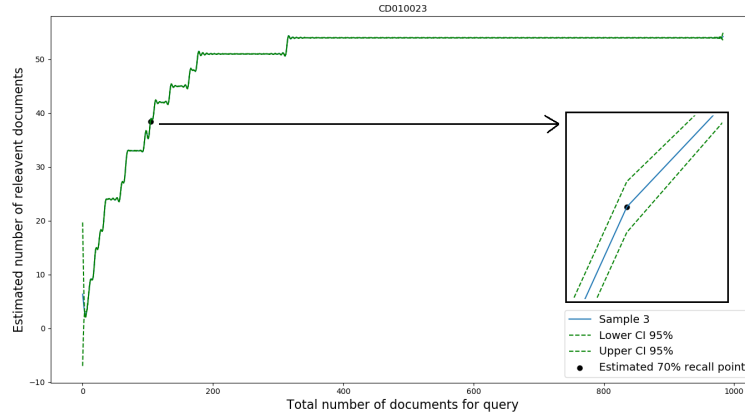


Figure 4.2: Visualisation of using a confidence interval for predicting a stopping point using a gp. We can see the gp is significantly over fitting to the data.

4.2.3 Conclusion on Curve fitting and GP

We implemented two methods for predicting stopping points in ranked medical studies. Our first approach used a general curve to estimate the point in which 70% recall is likely to have been hit. Our second method used a Gaussian Process in the same way. We used a sampling method to generate our curves to makes predictions about the remaining studies.

Comparing with Target Method

| Method | Target | recall | reliability | effort |
|----------------------|--------|--------|-------------|--------|
| Target Method | 10 | 0.952 | 0.96 | 0.652 |
| Sheffield-run2-curve | - | 0.75 | 0.64 | 0.51 |

Table 4.7: Comparing target method with curve fitting along with confidence interval. Using Sheffield-run-2

We can evaluate how well this method does against an existing set of relevance rankings. We will use the Sheffield run data from the CLEF 2017 task [6]. As the target method allows us to specify our level of reliability, we needed a target T of 9 to hit 95% reliability. We can see that the reliability of our method does not come close to the Target method. Our curve, however has a much lower effort.

4.2.4 Gaussian Process Conclusion

So far, we found using a GP was not an appropriate choice due to the level of over-fitting. Further work is needed in determining an efficient kernel for our data.

4.3 Indexing and Querying Medline with Limited Information

CLEF 2018 [2] presents an appropriate sub-task for using a limited amount of information to retrieve relevant documents. Normally, reviewers are required to construct complex Boolean queries to retrieve data from Medline. The objective of CLEF 2018 Sub-Task 1: No Boolean Search [2] is to search effectively and efficiently bypassing the construction of the Boolean query.

4.3.1 Acquiring Key Information from A Systematic Review Protocol

A systematic review protocol is created before the systematic review process is started. A systematic review protocol describes the rationale, hypothesis, and planned methods of the review. The Medline query is created manually with the help of the protocol. Here we are looking to generate a suitable query/relevant information from the protocol to then automatically query Medline.

We used RAKE [14] to extract key-words from a protocol. The minimum word occurrence count is set to 1, as the protocols are typically small. We used a pubmed stop list as the phrase splitting parameter. Example shown below:

Topic: CD008122

Title: Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries

Objective: To assess the diagnostic accuracy of RDTs for detecting clinical *P. falciparum* malaria (symptoms suggestive of malaria plus *P. falciparum* parasitaemia detectable by microscopy) in persons living in malaria endemic areas who present to ambulatory healthcare facilities with symptoms of malaria, and to identify which types and brands of commercial test best detect clinical *P. falciparum* malaria.

endemic countries objective|ambulatory healthcare facilities|rapid diagnostic tests|falciparum parasitaemia detectable|malaria endemic areas|diagnostic accuracy|falciparum malaria

The | symbol represents a separation between a phrase. The protocols are pre-processed as follows: Reference removal, lowercase, words less than N length removed, pubmed stoplist. We decided to not perform any stemming/additional manipulation at this stage, due to uncertainty of query format. The resulting content is stored in a separate file appended with '.kwq' (key-word query).

The key-word-query receives some final pre-processing prior to being loaded into our information retrieval (IR) system. We used a Lancaster stemmer to reduce words down to a base form. The result is as follows:

endem country object amb healthc facil rapid diagnost test falcipar parasitaem detect malar endem area diagnost acc falcipar malar

4.3.2 Indexing medline

Medline was downloaded from the online resource. We processed the xml files and retrieved the information for each study - title, id, abstract. To reduce the size, we store each record into a local database, containing only the relevant information for each study.

We used Apache Lucene to generate an index for the medline local database. The abstract and title were concatenated together. Pre-processing was done using the same format as the query: pub-med stoplist, Lancaster stemmer and lower casing.

We will provide comparisons of various ranking methods as well as the evaluation scores for each.

4.3.3 Results

Results were generated using the eval script from the CLEF 2017/2018 task [6]. We calculated the top N results over the CLEF 2017 training set. We include a random baseline to provide a comparison between results.

| Run | recall | ap | lastrel | wss100 | wss95 | normarea | N |
|--|--------|-------|-----------|--------|-------|----------|-------|
| Random-baseline | 0.05 | 0.02 | 126.7 | 0.0 | -0.0 | 0.024 | - |
| Train-Data-Sheffield-bm25-Run1-objective-only | 0.538 | 0.034 | 3039.051 | 0.101 | 0.108 | 0.431 | 5000 |
| Train-Data-Sheffield-tfidf-Run1-objective-only | 0.313 | 0.007 | 2633.718 | 0.021 | 0.023 | 0.247 | 5000 |
| Train-Data-Sheffield-boolean-Run1-objective-only | 0.538 | 0.034 | 3039.051 | 0.101 | 0.108 | 0.431 | 5000 |
| Train-Data-Sheffield-bm25-Run1-objective-only | 0.68 | 0.034 | 12310.231 | 0.169 | 0.172 | 0.592 | 25000 |
| Train-Data-Sheffield-tfidf-Run1-objective-only | 0.601 | 0.007 | 14883.744 | 0.13 | 0.136 | 0.455 | 25000 |
| Train-Data-Sheffield-boolean-Run1-objective-only | 0.471 | 0.007 | 12974.205 | 0.03 | 0.029 | 0.381 | 25000 |

Table 4.8: Results for IR medline system. Comparison for both 5000 and 25000 thresholds

As we increase the number of documents we return, the recall naturally increases. When we return 25000 documents for each topic, we are able to obtain a total recall rate of over 58%. However, the precision (ap, average precision) is very low, suggesting a significant amount of the documents are not useful. BM25 was found to be the best ranking method, followed by tfidf and boolean.

Improvements could certainly be made to this system:

- MeSH headings would be useful in expanding the range of the query to capture synonymous terms.
- Tokenization could be optimized to capture phrases of different sizes.
- Introducing a cost or stopping point to remove the amount of non-relevant documents. We can see for the 25000 documents set of results the last relevant document was around the 20000 point, meaning we could drop the last 5000 from our result set.

4.3.4 Medline automatic query Conclusion

We built an IR system using Apache Lucerne and compared three separate ranking methods. We found bm25 ranking gave the best results overall.

We found we were able to achieve fair results with a little optimization techniques to the index and query data.

We compared the performance of our system across different return thresholds, naturally finding as we increase the returned number of documents we get a higher recall. This comes at the expense of reduced precision.

We suggested further improvement to our system, such as including a phrase model for more robust features for both index and query.

Chapter 5

Future Work

This chapter will outline a plan for future work. This chapter will relate back to our research questions chapter. 3

We will look at feature extraction for systematic reviews. This will involve looking at content within a study and attempting to identify it automatically. The first step of this process is being able to obtain relevant studies for a systematic review. We will look at techniques for obtaining these in the form of pdf files. Work will also need undertaking in processing a diverse range of pdf documents.

We will investigate classification of studies and determine if we can make a binary decision on whether or not a study is relevant to a research question. To achieve this we will first need to process the content of the studies and decide on a sampling strategy. This might involve using a proportion of the relevant studies as a training examples, and then trying to classify the remainder of the studies. We will also look at using unsupervised methods, by trying to directly use the systematic review as an indicator of relevant studies. We will need to overcome the challenge of condensing large studies (e.g full pdf texts) into relevant chunks.

Finally we will look at alternate approaches to finding stopping points. We found that fitting a GP caused too much over fitting to our sampled rankings. We will look at alternate regression-based machine learning algorithms to finding a stopping point.

Chapter 6

Doctorial Development Programme

- Attended Healtex - UK HEALTHCARE TEXT ANALYTICS CONFERENCE
- Lab demonstrator for module COM4519 Cloud Computing
- Undertook module HAR6169 Study Design and Systematic Review Methods
- Marked assignments for COM3110 Text Processing
- Enrolled on FCE6100 Professional Behaviour and Ethical Conduct
- Completed TRAINING NEEDS ANALYSIS (TNA) form.
- Used Learning Management System (LMS) to attend 6 teacher training courses.
- Gave introduction talk to NLP group.
- Contributed to 2018 CLEF lab.
- Became member of Text Processing for Health Technology Assessment (TePHTA).

Bibliography

- [1] ALHARBI, A., AND STEVENSON, M. Ranking abstracts to identify relevant evidence for systematic reviews: The university of sheffield's approach to clef ehealth 2017 task 2. CLEF 2017 (2017).
- [2] CLEF. Clef. <https://sites.google.com/view/clef-ehealth-2018/task-2-technologically-assisted-reviews-in-en>
- [3] CORMACK, G. V., AND GROSSMAN, M. R. Engineering quality and reliability in technology-assisted review.
- [4] CORMACK, G. V., AND GROSSMAN, M. R. Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. CLEF 2017 (2017).
- [5] JONNALAGADD, S. R., GOYAL, P., AND HUFFMAN, M. D. Automating data extraction in systematic reviews: a systematic review.
- [6] KANOULAS, E., LI, D., AZZOPARDI, L., AND SPIJKER, R. Clef 2017 technologically assisted reviews in empirical medicine overview.
- [7] LEE, G. E. A study of convolutional neural networks for clinical document classification in systematic reviews: Sysreview at clef ehealth 2017. CLEF 2017 (2017).
- [8] MANHEIMER, E., CHENG, K., WIELAND, L. S., MIN, L. S., SHEN, X., AND LAO, L. Acupuncture for treatment of irritable bowel syndrome. John Wiley and Sons, Ltd (2012).
- [9] MEDLINE. medline. <https://www.nlm.nih.gov/bsd/pmresources.html>.
- [10] NLM. nlm. <https://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/finding>
- [11] NUNN, J. cochrane. <http://cccrg.cochrane.org/animated-storyboard-what-are-systematic-reviews>.
- [12] OF TASMANIA, U. pico. <https://utas.libguides.com/SystematicReviews/FormulateQuestion>.
- [13] PUBMED. pubmed. <https://www.ncbi.nlm.nih.gov/pubmed/>.

- [14] ROSE, S., ENGEL, D., AND CRAMER, N. Automatic keyword extraction from individual documents.
- [15] SATOPAA, V., ALBRECHT, J., IRWIN, D., AND RAGHAVAN, B. Finding a “kneedle” in a haystack: Detecting knee points in system behavior.
- [16] SCIPY. leastsquares. https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.least_squares.html#scipy.optimize.least_squares.